Use of a Likelihood-Based Mixed-Effects Model Repeated Measures Analysis for a Clinical Trial in Major Depressive Disorder

Baldeo K. Taneja[1], Thomas D. Kinghorn[2], John F. Reinhard, Jr.[3], Andrew C.G. Uprichard[4]
[1]Supernus Pharmaceuticals, Inc., 1550 East Gude Drive, Rockville, MD 20850
[2]MDS Biostatistics, 2200 Renaissance Blvd. Suite 400, King of Prussia, PA 19406
[3]Repligen Corporation, 41 Seyon St, Bldg. #1, Waltham, MA 02453
[4]EPIX Pharmaceuticals, Inc., 4 Maguire Road, Lexington, MA 02421

Abstract

Statistically valid analyses of longitudinal trials can be problematic in the presence of missing data. In a recent randomized, double-blind, placebo-controlled, 8-week clinical trial to assess the efficacy and safety of PRX-00023 in subjects with major depressive disorder, about one-quarter of the subjects withdrew prematurely. In anticipation of such a dropout, a likelihood-based mixed-effects model repeated measures analysis had been prospectively chosen as the primary analysis. In this paper, the appropriateness of the design and the related statistical analysis in the presence of missing values is highlighted with results of the study itself. Emphasis is placed on how the data-based robust efficacy results strengthened the sponsor's position to take a bold decision. In addition, we propose a three-pronged solution to the problem of missing data in the highly-regulated environment.

Key Words: Major Depressive Disorder, Longitudinal Data, Dropouts, Missing Data, MMRM Analysis, Sensitivity Analysis.

1. Introduction

A characteristic of virtually all studies of psychiatric conditions is a high percentage of missing data. In a review of studies by Khan, A. et al. (2001a, 2001b), dropout rates in antidepressant clinical trials averaged 37%, similar between the active drug and placebo, and were in 50%-60% range for trials of antipsychotics, somewhat greater on active drug than on placebo, and intermediate among trials with active controls. Treatment effects are often evaluated by comparing change over time in the outcome measures in these trials (longitudinal clinical trials). Repeated measures analyses are routinely performed for this purpose. In the literature, "mixed-effect model repeated measures" is used interchangeably with "mixed model repeated measures" and is abbreviated as MMRM.

Missed visits and early study termination are common occurrences in Central Nervous System (CNS) clinical trials. Missing values can be problematic for statistically valid analyses of longitudinal data. The problem is how to best deal with the missing values in clinical trials such as the antidepressant trial highlighted here.

2. An Antidepressant Clinical Trial

EPIX Pharmaceuticals, Inc. conducted an antidepressant clinical trial (EPX-CP-020) to assess the efficacy and safety of the partial 5-HT1A receptor agonist PRX-00023 in subjects with major depressive disorder. The primary objective of this study was to evaluate the efficacy of PRX-00023, administered twice daily (BID) over 8 weeks, in subjects with major depressive disorder (MDD) as determined by change from baseline in the Montgomery-Asberg Depression Rating Scale (MADRS) score. The study was listed at www.clinicaltrials.gov as NCT00448292.

2.1 Methodology

This was a randomized, 8-week, double-blind, placebo-controlled, multi-center, flexible-dose, study to evaluate PRX-00023 BID in adults with MDD as defined by DSM-IV (Diagnostic and Statistical Manual of Mental

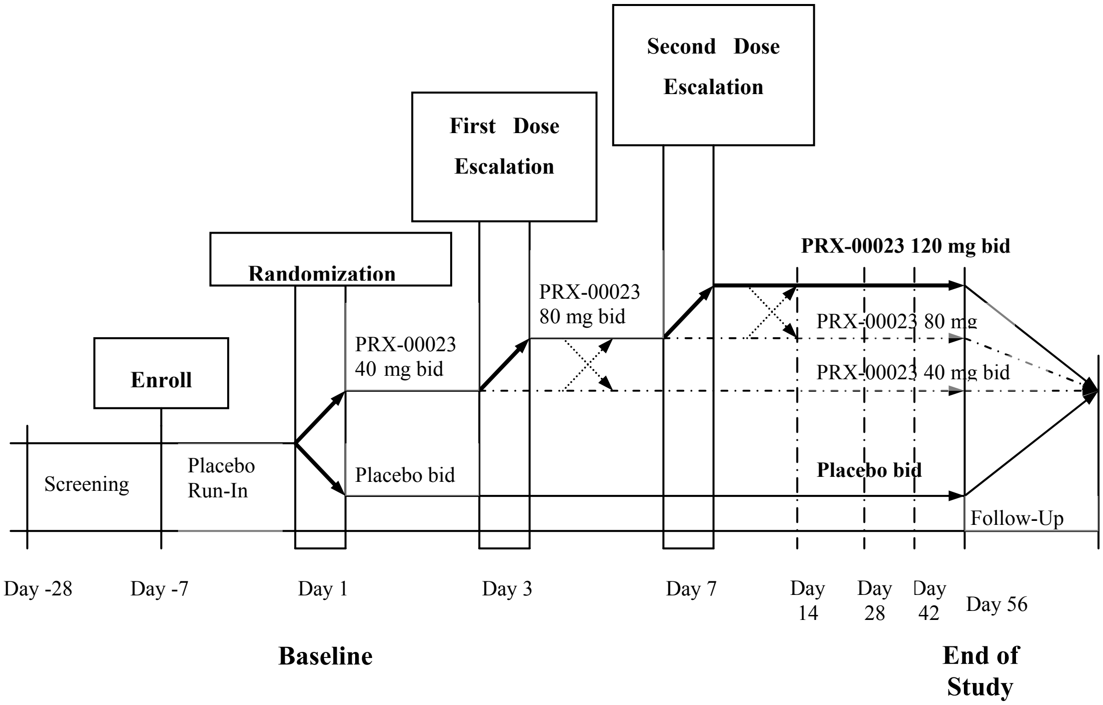

Figure 1: The Design Schematic for EPX-CP-020

Disorders, 4th Edition) criteria (single episode [296.21-296.23] or recurrent [296.31-296.33]). Adult men and women (18-65 years) with a diagnosis of MDD who fulfilled the eligibility criteria were enrolled into a one-week single-blind placebo run-in period followed by random assignment in a 1:1 ratio to a double-blind, dose-escalating, twice daily (BID) regimen of either PRX-00023 (40-80-120 mg) or placebo. Figure 1 depicts the design schematic.

During the 3-week screening period, informed consent was first obtained from each subject prior to conducting any study-related procedures. Medical histories and concomitant medications were reviewed and a physical examination was performed. Baseline depressive symptoms were assessed using the Hamilton Depression 17-item (HAM-D17) rating scale.

Upon receipt and confirmation of qualifying laboratory results, subjects returned to undergo a battery of pre-baseline psychiatric surveys and initiation of the 1-week single-blind placebo run-in period. Subjects that remained eligible were assigned a placebo run-in kit and instructed to take study drug BID with meals for 7 days. Subjects were re-evaluated at the end of the placebo run-in period for dosing compliance. Subjects found to be non-compliant with the BID dosing regimen (<75% of issued doses taken) were excluded from the study.

Eligible subjects were assigned a randomized double-blind drug kit and study drug was dispensed with instructions for taking the initial dose (i.e., PRX-00023 40 mg or placebo BID). Throughout the 8-week dosing period, subjects had periodic site visits for the purpose of interim safety, tolerability, and efficacy assessments.

The double-blind dosing period consisted of 2 scheduled dose escalations on Day 3 and Day 7. No dose-escalation occurred after Week 4. In addition, following either dose-escalation, the study drug dose could have been down-titrated (to the previous dose) at the discretion of the Investigator based on subject tolerability. Only one dose reduction was permitted after Week 4. Randomization was facilitated by an Interactive Voice Response System (IVRS). All subjects were monitored for adverse events (AEs) and changes in concomitant medications throughout the study.

A post-study evaluation occurred via telephone between one and two weeks after the last dose of study medication (either after study completion or early discontinuation from the study) to review safety and new or resolved AEs. Subjects were asked to return to the clinic in some circumstances after Week 8 in order to follow up on adverse events or abnormal lab findings. Subjects who discontinued prior to the Week 8 visit were asked to complete all of the Week 8 efficacy and safety evaluations prior to study exit.

2.2 Statistical Considerations

2.2.1 Sample Size
A target sample size of 140 completed subjects per treatment arm was intended to provide approximately 90% power (α=0.05) to detect a change in the MADRS score by 3.0 units, given a standard deviation of 7.7 units. Enrollment of 180 subjects per arm was intended to provide 140 evaluable subjects after accounting for discontinuances during the 8-week trial.

2.2.2 Safety Analysis
The Safety Population was defined as all subjects who received at least one dose of study drug. Safety variables were summarized by treatment group (PRX-00023 or placebo) in the Safety Population.

2.2.3 Efficacy Analysis
The intent-to-treat (ITT) population was defined as all subjects who had a valid baseline assessment, had MADRS total scores of 20 or greater at the end of the placebo run-in, were randomized, received at least one dose of study medication, and had at least one post-baseline efficacy assessment while on study drug. The primary statistical analysis involved the ITT population, and used change from baseline in MADRS score as the primary efficacy variable.

Secondary endpoints included the responder rate (the % of subjects with ≥ 50% decrease in their baseline MADRS scores) and the remission rate (the % of subjects with MADRS values ≤ 10) at the end of the treatment phase. Additionally, we recorded the Clinical Global Impressions of both Severity and improvement. The Quick Inventory of Depression Symptomatology, self-rated version (QIDS-SR), was completed by the subjects, independent of the other measures.

A mixed-effect model repeated measures (MMRM) was used for the primary efficacy analysis. The dependent variable was the change from baseline in MADRS score. The model included treatment, protocol-specified visit, treatment-by-visit interaction, and center as fixed effects; the baseline MADRS score as a covariate; and visit as a repeated measure. The model included an unstructured covariance matrix.

The primary hypothesis test was the test of the difference between least squares means at Week 8 change from baseline. All baseline measures were the results of assessments conducted at the end of the placebo run-in period, but prior to the subject's first dose during the double-blind dosing period.

2.2.4 Sensitivity Analysis
In the presence of a high drop-out rate, performance of a sensitivity analysis is crucial. The purpose of sensitivity analysis is to see whether different analyses under different set of missingness assumptions provide robust efficacy results. The extent to which efficacy results are stable across such analyses provides confidence in the statistical conclusions. The following taxonomy of missing data assumptions (Rubin (1976)) is now common in the statistical literature: MCAR (Missing Completely At Random), MAR (Missing At Random), and MNAR (Missing Not At Random). Accordingly, the following sensitivity analyses were conducted:

- MCAR-based LOCF (last observation carried forward) analysis because of its historical use
- MAR-based MMRM analyses with various covariance structures
- MNAR-based analyses.

2.3 Results

2.3.1 Trial Profile
Table 1 provides a summary of subject disposition for all subjects. A total of 419 subjects were screened; 46 of these were screen failures. A total of 373 subjects went on to placebo run-in period; 13 of these were excluded from

the study for non-compliance with the BID dosing. A total of 360 subjects were randomized. Out of 360 subjects randomized in 1:1 fashion (180 in each group), 177 (98.3%) subjects actually received PRX-00023, and 175 (97.2%) received placebo. The number of subjects completing the study was similar between the 2 treatment groups: 133 (73.9%) of PRX-00023-treated subjects and 132 (73.3%) of those who received placebo.

Table 1: Subject Disposition – All Subjects

	PRX-00023 **N (%)**	**PLACEBO** **N (%)**	**TOTAL** **N (%)**
Subjects Screened	-	-	419
Number of Screen Failures	-	-	46
Subjects in Placebo Run-In Period	-	-	373
Subjects Not Eligible	-	-	13
Subjects Randomized	180 (100%)	180 (100%)	360 (100%)
Subjects Received Double-Blind Medication	177 (98.3%)	175 (97.2%)	352 (97.8%)
Subjects Completed Study	133 (73.9%)	132 (73.3%)	265 (73.6%)
Subjects Discontinued	44 **(24.4%)**	43 **(23.8%)**	87 **(24.2%)**
Reasons for Discontinuation:			
Adverse Event	6 (3.3%)	3 (1.7%)	9 (2.5%)
Death	0 (0.0%)	0 (0.0%)	0 (0.0%)
Consent Withdrawal	10 (5.6%)	14 (7.8%)	24 (6.7%)
Non-Compliance	3 (1.7%)	8 (4.4%)	11 (3.1%)
Lost to Follow-Up	14 (7.8%)	11 (6.1%)	25 (6.9%)
Other	11 (6.1%)	7 (3.9%)	18 (5.0%)

Note: Denominator for percentages is the number of randomized subjects in a particular group.

2.3.2 Demographics at Baseline (ITT Population)

Demographic characteristics are summarized in Table 2 for the ITT population.

Table 2: Demographic and Other Baseline Characteristics (ITT Population)

Variable		**PRX-00023**	**PLACEBO**	**TOTAL**
Age (yrs)	N	155	168	323
	Mean (SD)	40.8 (12.40)	38.8 (12.11)	39.8 (12.27)
	Min – Max	20 – 65	18 – 63	18 – 65
Sex	N	155	168	323
	Male	68 (43.9%)	63 (37.5%)	131 (40.6%)
	Female	87 (56.1%)	105 (62.5%)	192 (59.4%)
Race	N	154	168	322
	White	114 (74.0%)	129 (76.8%)	243 (75.5%)
	Black or African American	26 (16.9%)	27 (16.1%)	53 (16.5%)
	Asian	4 (2.6%)	3 (1.8%)	7 (2.2%)
	Native American or Alaska Native	1 (0.6%)	0 (0.0%)	1 (0.3%)
	Native Hawaiian or Pacific Islander	0 (0.0%)	1 (0.6%)	1 (0.3%)
	Other	9 (5.8%)	8 (4.8%)	17 (5.3%)
Ethnicity	N	154	168	322
	Hispanic or Latino	20 (13.0%)	28 (16.7%)	48 (14.9%)
	Non-Hispanic or Non-Latino	134 (87.0%)	140 (83.3%)	274 (85.1%)
Height (cm)	N	155	167	322
	Mean (SD)	169.2 (9.28)	168.9 (9.24)	169.0 (9.26)
	Min – Max	142.2 – 193.0	144.3 – 198.1	142.2 – 198.1
Weight (kg)	N	155	168	323
	Mean (SD)	85.4 (24.66)	86.3 (24.71)	85.9 (24.68)
	Min – Max	49.0 – 291.0	46.3 – 181.4	46.3 – 291.0

Note: SD=Standard Deviation

Characteristics were similar between treatment groups, except that there were more women in the placebo group, 105 (62.5%) compared to the PRX-00023 group, 87 (56.1%). Age ranged from 18 to 65 years, with a mean age of 39.8. Most subjects were white (74%) and the number of black subjects was similar between groups (about 16%). There were 20 (13.0%) Hispanic subjects in the PRX-00023 group and 28 (16.7%) in the placebo group. Mean height was about 169 cm, overall and in both treatment groups and the mean weight was similar between groups, 85.4kg in the PRX-00023 group and 86.3 kg in the placebo group.

2.3.3 Efficacy Results (ITT Population)

MADRS scores are summarized for Baseline, Week 8, and Week 8 change from Baseline for the ITT population in Table 3. The active drug (PRX-00023) brought the mean MADRS score down from 30.8 to 21.4 in 8 weeks (a reduction of 9.4 points) whereas the placebo performed in a similar fashion and brought the mean MADRS score down from 30.2 to 20.3 in 8 weeks (a reduction of 9.8 points). The results, while comparable, appeared to favor the placebo group.

Table 3: Summary of MADRS Scores (ITT Population)

		PRX-00023	PLACEBO	TOTAL
Baseline	N	155	168	323
	Mean (SD)	30.8 (4.31)	30.2 (4.32)	30.5 (4.32)
Week 8	N	141	154	295
	Mean (SD)	21.4 (11.25)	20.3 (10.21)	20.8 (10.72)
Week 8 Change from Baseline	N	141	154	295
	Mean (SD)	-9.4 (10.95)	-9.8 (10.02)	-9.6 (10.46)

Figure 2 below shows the mean MADRS change from baseline over time with 95% confidence intervals (vertical bars). There was virtually no separation between the two groups at any study week.

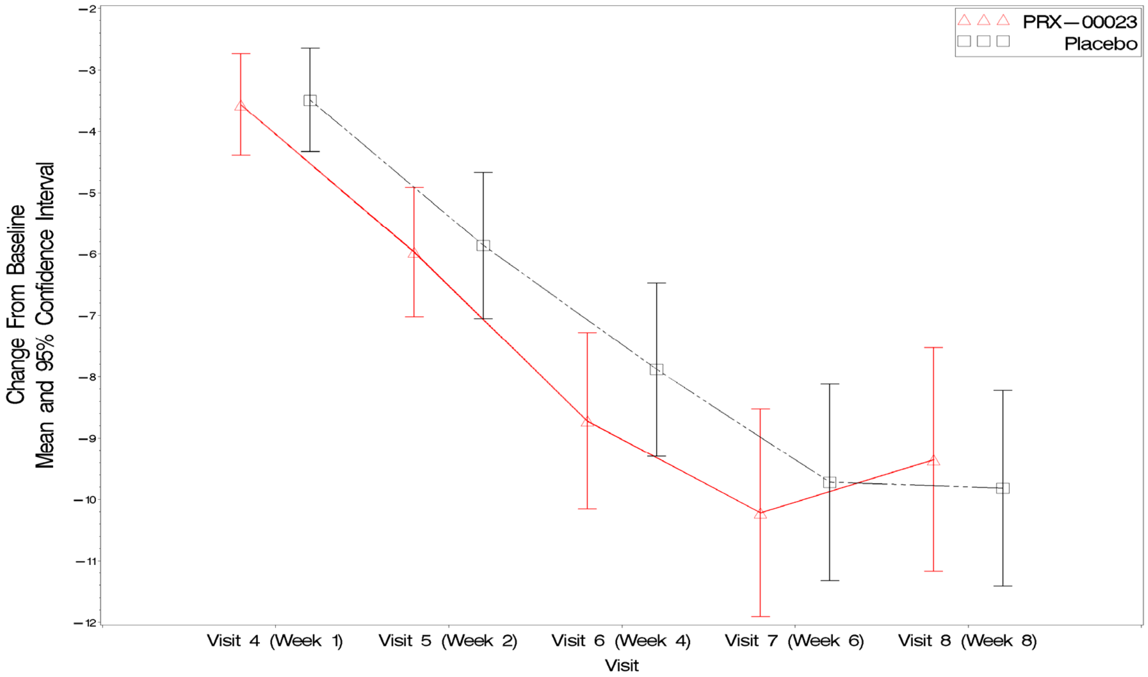

Figure 2: MADRS Change from Baseline over time (ITT Population)

Mixed-effect model for repeated measures (MMRM) analysis estimated drug effect on the basis of observed data. MADRS results for change from Baseline to Week 8 are presented in Table 4 for ITT population. The between-groups difference (0.7) was not statistically significant (p=0.585). Additionally, there were no statistically significant results for any of the secondary efficacy parameters.

Table 4: MMRM Analysis of MADRS Score: Week 8 Change from Baseline (ITT)

	Statistic	**Result**
Week 8 Change From Baseline	N	295
	p-Value	0.585
	Treatment Difference	0.7
	SE	1.22
	95% CI	(-1.73, 3.06)

Note: Treatment difference=least-squares mean differences, which are adjusted for inequalities in number of subjects and baseline values, SE=adjusted standard error of the mean, 95% CI=95% adjusted confidence interval.

2.3.4 Sensitivity analysis

The sensitivity analysis results based on MCAR, MAR and MNAR assumptions were similar (not shown here) yielding the same conclusion for the treatment effect at Week 8.

3. Discussion

In order to understand the potential impact of missing data on the results of a longitudinal clinical trial, the mechanisms leading to the missingness must be considered. In planning for this trial, it was anticipated that the dropout rate would be 25%: as it turned out the dropout rate for the active drug and placebo were 24.4% and 23.8% respectively. These numbers are consistent with other CNS trial results in MDD. The MAR assumption of missingness is typically more plausible in antidepressant clinical trials when extensive efforts are made to observe all the outcomes and the factors that influence them while subjects are following the protocol-defined procedures. Use of MAR is further supported in that the consequences of departures from MAR can be evaluated through sensitivity analyses, and MAR-based statistical methods are often robust to departures from MAR.

On the basis of relevant theory and current practice regarding the analysis of longitudinal clinical trials intended to support regulatory approval of antidepressants, we chose the likelihood-based MMRM analysis as our primary efficacy analysis which is an appropriate choice for the statistical analysis under the MAR assumption. No formal statistical tests for MAR, MCAR, or MNAR were conducted. Instead, we conducted sensitivity analyses under MAR, MCAR, and MNAR assumptions covering all relevant scenarios. These analyses gave us consistent efficacy results providing us with strong confidence in our conclusions.

The present study results were convincing and negative. We believe the basic principles of clinical trial methodology, including statistical design and related analysis, held firm and that this was a case of a failed drug, not a failed study. The field of antidepressants is punctuated with efficacious drugs failing in clinical trials, but in this case the sponsor was sufficiently convinced by the robustness of the data to terminate development of the compound.

4. Proposal for the Future

In the highly-regulated environment in which new medicinal products are developed, sponsors would be well-advised to think carefully about the handling of data in a situation where a high dropout rate is to be anticipated. The current thinking in some statistical circles, as well as, in the FDA, is that missing data in longitudinal clinical trials may not be avoidable, but the best thing to do is to minimize it. Minimizing missing data requires a three-pronged approach with considerations from:

- regulatory
- clinical trial design
- statistical analysis.

4.1 Regulatory Considerations

This was an area of recent discussion (Soon, G. (2008)), but the take-home message must be based on the principle of no surprises. Whenever possible, securing buy-in from the respective regulatory body is a must if (a) it is anticipated that the trial will be used to support registration, and (b) there is a high likelihood of missing data. In the case of the current trial, the sponsor specifically asked the FDA for feedback on the proposed protocol and statistical analysis plan.

Sponsors can also proactively think of requesting a special protocol assessment (SPA) of their protocol by the FDA. In the request for SPA, the sponsor should pose focused questions concerning specific issues regarding the protocol, protocol design (including proposed size), study conduct, study goals, and/or data analysis for the proposed investigation. Although the questions should be specific to the protocol and should not address overall development strategies, the role of the study in the overall development plan should be clear to the Agency for it to answer the protocol-specific questions. And, for that to happen, the sponsor should have had a meeting with the review division so that the division is aware of both the developmental context in which the protocol is being reviewed and the questions that are to be answered.

FDA recommends that a sponsor submit a protocol intended for SPA to the Agency at least 90 days prior to the anticipated start of the study. The Agency's assessment will be based primarily on the questions posed by the sponsor, the underlying data, the assumptions, the information described by the sponsor, and relevant Agency policies and guidance documents. Comments from the FDA review team should be sent to the sponsor within 45 days of receipt of the request for SPA.

4.2 Clinical Trial Design Considerations

The current trial incorporated a number of measures again aimed at minimizing "noise" and maximizing the value of whatever data were collected during the course of the study. Such measures included a placebo run-in that was intended to identify non-compliant subjects, primarily on the basis of dosing. In addition to dosing compliance, the use of excluded drugs was also identified (using antibody based analysis of urine samples) and used to remove subjects that were non-compliant with other aspects of the protocol. In some cases, positive urinary drug screens were subjected to secondary, confirmatory analyses involving more precise methodology (e.g. mass spectrometry). Positive results for hepatitis B and C, using antibody detection, was subjected to confirmatory analysis by nucleic acid testing. These confirmatory tests helped reduce the overall screen failure rate, allowing otherwise evaluable subjects to participate in the trial.

A second consideration in this study was the use of different depression measures for screening/inclusion and the subsequent demonstration of efficacy. HAM-D17 scale was used for screening/inclusion whereas MADRS scale was used for efficacy. It was reasoned that raters, under pressure to identify subjects, might exhibit subtle bias administering the depression rating scale used at the screening visit, inflating the baseline scores. As a further check, a MADRS value of ≥ 20 at randomization was used to define inclusion into the ITT population.

Consistent application of the MADRS was considered crucial for increasing the precision of the estimated effect size and the consequent power of the study. To attain this outcome, raters were initially qualified through a three step process. Potential raters were surveyed as to their educational background (B.S. in psychology or R.N. preferred), their duration and frequency of use of the rating scales. These, experienced raters were shown video, mock patient interviews which they scored and their scores were matched against the desired scores for the 10 items of the MADRS. Discrepancies were discussed as part of the qualification process. Lastly, raters were required to submit a videotape of themselves conducting a mock interview. The video was rated by an independent organization (United BioSource Corporation [UBC], Wayne, PA) against pre-determined criteria of duration and anticipated level of detail of questioning.

Ongoing rater performance was subjected to one further check. Subjects were asked to complete a self-rated assessment of their depression symptoms, the QIDS-SR, that was completed (by the subjects) independent of and prior to the observer rated scales. Moreover, the QIDS-SR data were not shared with the MADRS raters but were evaluated by an independent organization (UBC) for concordance with the MADRS. Extreme deviations, e.g. changes in opposing directions, were a basis for rater examination and possible remediation.

A further consideration was the amount of time spent with the subjects performing the assessments. We sought to minimize the time spent, reasoning that such observations might bias the subject's expectation of improvement. Many studies perform both the HAM-D17 and the MADRS scales, in addition to other measures involving other ancillary attributes of the disease. Thus, lengthy (2 hour) evaluations are possible. In such instances, the observations may become interventions that may increase the placebo response.

Another consideration was to use flexible dose instead of a fixed dose, but being careful not to underdose the subjects. Khan, A. et al. (2003) suggested, on the basis of FDA summary basis of approval reports of various antidepressant trials, that the antidepressant dose schedule may influence trial outcome due in part to a significantly lower magnitude of symptom reduction with placebo in flexible dose trials ($p<0.05$) compared to fixed dose trials.

4.3 Statistical Analysis Considerations

Reference has already been made to the fact that "missingness is informative" and should not necessarily result in the "throwing out" of what can be valuable information. In the current antidepressant trial, we chose the likelihood-based MMRM analysis as our primary efficacy analysis (followed by a range of sensitivity analyses) for a number of reasons (Mallinckrodt, C. H. et al. (2008)):

- It is a principled approach to deal with missing values in longitudinal data
- It uses all available data to compensate for missing values
- It accounts for correlation between measures on the same subject
- It has greater flexibility for handling time-effects, thereby handling missing data more appropriately
- It is most appropriate under MAR assumptions
- It provides superior control of Type I and Type II error when compared to LOCF
- It is easier to implement in practice (programming) than LOCF
- It is more efficient and reliable as a method of primary efficacy analysis.

In summary, the use of a likelihood-based mixed-effects model repeated measures (MMRM) analysis for a clinical trial in major depressive disorder resulted in robust data upon which the sponsor was able to make a rapid data-based decision to terminate development of a drug. In conjunction with regulatory input and a strong and appropriate clinical trial design, the statistical methodology allowed this decision to be made with confidence.

Acknowledgements

We would like to thank the EPX-CP-020 team at EPIX Pharmaceuticals, Inc. for their encouragement and support in developing this paper. This paper was developed when Baldeo K. Taneja and John F. Reinhard, Jr. were working with EPIX Pharmaceuticals, Inc.

References

Khan, A., Khan, S.R., Leventhal, R.M. and Brown, W.A. (2001a): Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: A replication analysis of the Food and Drug Administration database. *Int'l Journal of Neuropsychopharmacology*, Vol. 4, pp. 113-118.

Khan, A., Khan, S.R., Leventhal, R.M. and Brown, W.A. (2001b): Symptom reduction and suicide risk among patients treated with placebo in antipsychotic clinical trials: An analysis of the Food and Drug Administration database. *American Journal of Psychiatry*, Vol. 158, pp. 1449-1454.

Khan, A., Khan, S.R., Walens, G., Kolts, R. and Giller, E.L. (2003): Frequency of Positive Studies Among Fixed and Flexible Dose Antidepressant Clinical Trials: An Analysis of the Food and Drug Administration Summary Basis of Approval Reports, *Neuropsychopharmacology.* Vol. 28, pp. 552-557.

Mallinckrodt, C.H., Lane, P.W., Schnell, D., Peng, Y. and Mancuso, J.P. (2008): Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal*, Vol. 42, pp. 303-319.

Rubin, D.B. (1976): Inference and missing data. *Biometrika,* Vol. 63, pp. 581-592.

Soon, G. (2008): Minimizing missing data: Design and regulatory considerations. *2nd Annual FDA/DIA Statistics Forum (April 14-16, 2008)*, Bethesda, MD.