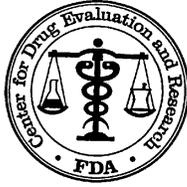


**CENTER FOR DRUG EVALUATION AND
RESEARCH**

APPLICATION NUMBER:

021879Orig1s000

STATISTICAL REVIEW(S)



U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Translational Science
Office of Biostatistics

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/Serial Number: 21,879

Drug Name: Zenvia (Dextromethorphan/Quinidine combination)

Indication(s): Pseudobulbar Affect

Applicant: Avanir

Date(s): Submission: April 23, 2010

Review Priority: Priority

Biometrics Division: Division of Biometrics I

Statistical Reviewer: Tristan Massie, Ph.D.

Concurring Reviewers: Kun Jin, Ph.D., Team Leader
Jim (Hsien Ming) Hung, Ph.D., Director, Division of Biometrics 1

Medical Division: Division of Neurology (HFD-120)

Clinical Team: Devanand Jillapalli, M.D.
Ron Farkas, M.D. Team Leader
Russell Katz, M.D., Division Director

Project Manager: Susan Daugherty

Keywords: Count Data; Longitudinal Data; Negative Binomial Distribution

Table of Contents

LIST OF TABLES.....	3
LIST OF FIGURES.....	4
1 EXECUTIVE SUMMARY	5
1.1 CONCLUSIONS AND RECOMMENDATIONS	5
1.2 BRIEF OVERVIEW OF CLINICAL STUDIES	5
1.3 STATISTICAL ISSUES AND FINDINGS	6
2 INTRODUCTION	7
2.1 OVERVIEW.....	7
2.2 DATA SOURCES	8
3 STATISTICAL EVALUATION	8
3.1 EVALUATION OF EFFICACY	8
3.1.1 <i>Study 123</i>	8
3.1.1.1 Study Design and Analysis Plan	8
3.1.1.2 Disposition of Subjects	13
3.1.1.3 Demographic Characteristics	14
3.1.1.4 Sponsor’s Results.....	16
3.1.1.5 Reviewer’s Results.....	20
3.1.1.5.1 Primary Endpoint	20
3.1.1.5.2 Effect of Protocol Amendment to Increase Sample Size in MS subgroup	25
3.1.1.5.3 Assessment of the Impact of Missing Data	25
3.1.1.5.4 Primary Longitudinal Model Issues	27
3.1.1.5.5 Exploratory Analysis of Episodes by Episode Type	29
3.2 EVALUATION OF SAFETY	31
4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS	32
4.1 GENDER, RACE AND AGE	32
4.1.1 <i>Gender</i>	32
4.1.2 <i>Race</i>	32
4.1.3 <i>Age</i>	32
4.2 OTHER SPECIAL/SUBGROUP POPULATIONS	33
4.2.1 <i>Underlying Primary Disease Diagnosis</i>	33
4.2.2 <i>Individual Sites</i>	36
5 SUMMARY AND CONCLUSIONS	38
5.1 STATISTICAL ISSUES AND COLLECTIVE EVIDENCE	38
5.1.1 <i>Statistical Issues</i>	38
5.1.2 <i>Collective Evidence</i>	39
5.2 CONCLUSIONS AND RECOMMENDATIONS	42

LIST OF TABLES

Table 1 Patient Disposition.....	13
Table 2 Time from Diagnosis (Months)a in ALS Patients	14
Table 3 Demographics of Randomized Patients.....	15
Table 4 Primary Longitudinal Negative Binomial Model- Laughing/Crying Episode Rates (ITT Population).....	16
Table 5 Non-longitudinal Negative Binomial Model –Total Episode Rates (ITT Population)	17
Table 6 GEE Model for Number of Laughing/Crying Episodes (ITT Population)	18
Table 7 Change from Baseline in CNS-LS (ITT Population).....	20
Table 8 Patients’ Average Daily Rate of Laughing and/or Crying Episodes.....	21
Table 9 Total Sum of Laughing and Crying Episodes by Period(ITT Population)	22
Table 10 Patient’s Min/Max Daily Episode Counts (ITT Population)	23
Table 11 Change from Baseline in Average Daily Laughing+Crying Episode Counts.....	24
Table 12 Summary Statistics for Daily Episode Rate by Completion Status	25
Table 13 Laughing + Crying Episodes for those that died	26
Table 14 Average Weekly Episode Count by Episode Type.....	30
Table 15 Summary of Average Weekly Laughing Episodes by Underlying Disease.....	31
Table 16 Incident (Episode) rate ratios by Age Group	33
Table 17 Average Daily Laughing Crying Counts in MS Subgroup.....	34
Table 18 All Post Baseline laughing+crying daily episode counts in MS patients.....	35
Table 19 Incident Rate Ratios of AVP20/Placebo in MS patients based on Various Models	36
Table 20 Number of Episodes per Week During Treatment by Study	40
Table 21 Weekly Average Episode Rate by Study and Episode Type	41

LIST OF FIGURES

Figure 1 CNS-LS scores over Time in ITT Population	19
Figure 2 Biweekly Estimated Incident Rate Ratios	28
Figure 3 Site Specific Treatment Effect Estimates for AVP30 vs. Placebo	37
Figure 4 Site Specific Treatment Effect Estimates for AVP20 vs. Placebo	38

1 EXECUTIVE SUMMARY

1.1 Conclusions and Recommendations

The efficacy data from trial 123 suggests that both the 30 mg Dextromethorphan (DM)/ 10 mg Quinidine (Q) combination as well as the 20 mg DM/ 10 mg Q combination were superior to placebo in controlling the number of inappropriate laughing plus crying episodes associated with pseudobulbar affect in the mixed study population of amyotrophic lateral sclerosis (ALS) and multiple sclerosis (MS) patients. It was previously concluded from the original NDA submission that the 30 mg Dextromethorphan / 30 mg Quinidine combination was superior to placebo in study 106 conducted in MS patients with pseudobulbar affect and superior to the two components in study 102 conducted in ALS patients with pseudobulbar affect. The primary endpoint for the earlier trials was the change from baseline in the CNS-LS score averaged over the treatment period. The differences from placebo in terms of the CNS-LS were also nominally significant in trial 123. The primary model of episode counts suggests that there may be no additional benefit of 30/10 over that of 20/10 compared to placebo. Although a prespecified secondary analysis suggests a possible additional benefit of 30/10 this is not judged very persuasive by this reviewer as it seems to be sensitive to outliers and also is not supported by the simple median changes from baseline in episode rates (medians are more robust to outliers).

1.2 Brief Overview of Clinical Studies

Following his observation of a palliative effect of the Dextromethorphan/Quinidine (DM/Q) combination on pseudobulbar affect (PBA) in ALS patients, Dr. Smith conducted a placebo controlled crossover study systematically to evaluate the efficacy and tolerability of DM 30 mg/Q 75 mg in a population of patients with neurological disorders experiencing PBA. Results of the study showed significantly greater relief of PBA during treatment with DM/Q than with placebo (CNS-93), and supported the decision to develop DM/Q as a treatment for PBA associated with a variety of neurological disorders.

Although DM 30 mg had been tested in several clinical investigations, the associated Q doses ranged from 50 to 200 mg. Based on the initial PK studies in healthy volunteers, Q 30 mg was selected for evaluation in clinical efficacy studies since it caused near maximal inhibition of CYP2D6-mediated metabolism of DM.

Clinical efficacy and safety studies were completed in ALS patients with PBA (Study 99-AVR-102), MS patients with PBA (Study 02-AVR-106), and patients with diabetic peripheral neuropathic (DPN) pain (Studies 01-AVR-105, and 04-AVR-109).

A long-term, open-label safety study was also performed (02-AVR-107). Following submission of the NDA for use of DM 30 mg/Q 30 mg for the treatment of PBA, the FDA suggested that a combination containing a lower dose of Q should be investigated to potentially reduce risks of the higher dose such as QT prolongation.

The pivotal phase 3 study of 2 Zenvia formulations containing Q 10 mg (DM 30 mg/Q 10 mg, and DM 20 mg/Q 10 mg) has now been completed in PBA patients with either ALS or MS as the underlying neurologic disease. Thorough QT studies in healthy volunteers have demonstrated that QT interval prolongation is dependent on plasma concentration of Q, and that the predicted changes in QT interval with Zenvia will be limited because plasma Q concentrations are at the

low end of the concentration-response curve (Studies 05-AVR-119, 08-AVR-126, and 09-AVR-128).

Only study 123 is reviewed here since the other studies were reviewed, previously, at the time of the original application.

1.3 Statistical Issues and Findings

The primary longitudinal negative binomial model found the ratio of laughing plus crying episode rates for AVP20 over placebo to be slightly better numerically than the episode rate ratio of AVP30 over placebo (both statistically significant compared to placebo). The supportive non-longitudinal negative model analysis of the total post-baseline sums of laughing plus crying episode counts suggests the opposite ordering of AVP20 and AVP30. However, the suggested ordering of AVP30 and AVP20 obtained from this analysis seems to be sensitive to some extreme outlier counts in the AVP20 group, many of which came from one particular site. The two groups' results from this model are very similar if data from this site is excluded (see section 3.1.1.5.1). In addition, several other models as well as the simple group medians of the changes from baseline in episode rate suggest that there is little difference between AVP20 and AVP30 (but both are nominally significant compared to placebo).

The standard errors of treatment effect estimates are smaller for the primary longitudinal random effects negative binomial model than for the non-longitudinal negative binomial model that was used to model episode counts in the prior two studies. For the longitudinal model each daily count for a subject is an observation of the dependent variable, whereas for the non-longitudinal model the subject's sum of the counts over all post-baseline days is the sole observation of the dependent variable. Methods to estimate the standard errors of the parameter estimates based on re-sampling the data and re-running the model on the resulting data over and over suggest that the longitudinal model underestimates the standard errors by as much as a factor of 2. This underestimation of the standard error suggests that actual p-values should be larger than reported. It may be related to the primary longitudinal model's potential oversimplification of the within patient correlation (among the patients' set of 84 postbaseline daily episode counts). The model incorporates a single random effect parameter to address this correlation, but there are $84 \times 83 / 2 = 3,486$ different pairs of daily counts per subject. It seems unlikely if only due to the sheer magnitude of distinct pairs that all of the corresponding correlations are equal. At any rate, the underestimation of standard errors appears to not be so great as to alter the statistical significance of the comparisons of AVP20 and AVP30 with placebo.

There is some evidence that there may be less of a treatment effect on laughing than crying or possibly even no effect on laughing but it should be acknowledged that the study was only powered for the combination of laughing and crying. This analysis was motivated by the observed trend in study 102 of a numerically smaller effect on Laughing than on Crying in terms of both episode counts and items of the CNS-LS. However, it must be noted that these studies were not powered to differentiate laughing episode specific treatment effects from crying episode specific treatment effects. Nevertheless, this potentially smaller effect on laughing is supported by independent analyses of episode counts and the sum of the 4 laughing items of the 7 item CNSLS endpoint in two of the three studies.

There was a slight imbalance between the placebo and the drug groups in deaths in study 123. There were 7 Deaths in study 123 all of which occurred in ALS patients (1/64 in placebo, 3/68 in AVP20 mg and 3/65 in AVP 30 mg). A Fisher's exact test comparing the combined drug groups to placebo concludes there is not enough evidence, one-sided $p=0.275$, to reject the null hypothesis that the probability of death is the same among these two groups. This test was conducted post-hoc as a quick and simple way to assess this unexpected death imbalance, given the relatively low overall death rate in the trial.

2 INTRODUCTION

2.1 Overview

Of the two prior trials on which the original application was based, only study 102 in ALS patients compared the combination to each of the individual components of this combination drug product.

Study 102 was a multicenter, randomized, double-blind, controlled, parallel, three-group study of the treatment of pseudobulbar affect in ALS patients. It compared AVP-923 administered orally, two times a day (every 12 hours) for 28 days (the first dose will be taken in the P.M. of Day 1, and the final dose will be taken in the A.M. on Day 29). The last day (Day 29) was to be the last day the patient was on study and could have occurred anywhere between Day 26 and Day 32. Patients were to be randomized to one of three groups to receive either AVP-923 (a capsule containing dextromethorphan hydrobromide [30 mg] and quinidine sulfate [30 mg]), dextromethorphan hydrobromide (30 mg), or quinidine sulfate (30 mg).

The primary efficacy endpoint in study 102 was the CNSLS score. The number of episodes as recorded in the patient diary was one of the secondary endpoints.

The other study, numbered 106, involved the treatment of pseudobulbar affect in MS patients. It compared AVP-923 to placebo.

Following the receipt of the approvable letter which suggested that lower dose formulations should be developed, and based on a series of meetings and discussions with the FDA it was suggested by the Agency that Avanir could perform an additional clinical study (07-AVR-123) assessing the safety and efficacy of a new lower dose formulation of DM/Q. As discussed and agreed upon with the Agency under a special protocol assessment (SPA), Study 07-AVR-123 entitled "*A Double-Blind, Randomized, Placebo- Controlled, Multicenter Study to Assess the Safety and Efficacy and to Determine the Pharmacokinetics of Two Doses of AVP-923 (Dextromethorphan/Quinidine) in the Treatment of Pseudobulbar Affect (PBA) in Patients with Amyotrophic Lateral Sclerosis and Multiple Sclerosis*" could serve as the final confirmatory phase 3 study, depending on the results.

Two new lower dosage strengths of DM/Q, DM 30 mg/Q 10 mg or DM 20 mg/Q 10 mg, were developed as a solid, oral-dosage capsule containing the same excipients as the original DM 30 mg/Q 30 mg formulation. These lower dosage strengths were studied versus placebo in the 07-AVR-123 trial.

The new dose formulations of Zenvia are further characterized as follows:

- Zenvia 30/10: each capsule contains DM 30 mg and Q 10 mg,
- Zenvia 20/10: each capsule contains DM 20 mg and Q 10 mg

Note that when the focus is only on study 123 the drug groups are referred to as AVP30 and AVP20, omitting the specific dose of Q for convenience since it is the same for both drug groups.

The DB phase of this study (#123) was conducted at 52 sites, 36 in the United States and 16 in Latin America (11 in Argentina and 5 in Brazil). Overall, 326 subjects were randomized, 110 were assigned to AVP-923-30, 107 to AVP-923-20, and 109 to placebo. Of the subjects randomized, 224 (68.7%) were at investigative sites in the U.S. and 102 (31.3%) were at sites in Latin America.

2.2 Data Sources

At the time of review the sponsor's study data for trial 123 was contained in the following directories.

<\\cdsesub1\EVSPROD\NDA021879\0035\m5\datasets\study-07-avr-123\tabulations>

<\\cdsesub1\EVSPROD\NDA021879\0035\m5\datasets\study-07-avr-123\analysis>

The data for the primary analysis was contained in the ADAEF data set in the Analysis directory.

At the time of review the sponsor's study report was contained in the following directory.

<\\cdsesub1\EVSPROD\NDA021879\0035\m5\53-clin-stud-rep\535-rep-effic-safety-stud\pseudobulbar-affect\5351-stud-rep-contr\study-07-avr-123>

3 STATISTICAL EVALUATION

3.1 Evaluation of Efficacy

3.1.1 Study 123

The first subject's first visit took place on 07 December 2007 and the last subject's completion date was 23 June 2009. The original protocol was dated October 5, 2007. The protocol was amended once on June 9, 2008. The statistical analysis plan is dated June 19, 2009. The amendment included the provision for increasing the number of MS patients per group from 30 to 42 (the number of ALS patients per group stayed at 60).

3.1.1.1 Study Design and Analysis Plan

Objective

The objectives of this study were to evaluate the safety, tolerability and efficacy of two different doses of AVP-923 (capsules containing either 30 mg of dextromethorphan hydrobromide and 10

mg of quinidine sulfate [AVP-923-30] or 20 mg of dextromethorphan hydrobromide and 10 mg of quinidine sulfate [AVP-923-20]) when compared to placebo, for the treatment of Pseudobulbar Affect (PBA) in a population of patients with amyotrophic lateral sclerosis (ALS) or multiple sclerosis (MS) over a 12-week period.

Study Design

This was a multicenter, randomized, double-blind, placebo-controlled, three-arm parallel study for the treatment of PBA in patients with ALS or MS, with AVP-923 capsules administered orally, two times a day (every 12 hours) during a 12-week period. Patients were to be recruited from a population of patients with ALS or MS who had been clinically diagnosed as suffering from PBA. The operational definition for PBA is “a syndrome characterized by outbursts of crying and/or laughing that are incongruous with, or out of proportion to, the underlying emotion.”

Patients were to be randomized in a 1:1:1 ratio to receive one of the two dose levels of AVP-923 (AVP-923-30 [DM30/Q10] or AVP-923-20 [DM20/Q10]) or placebo for 84 days (the last day of treatment was to be the last day the patient was on study and was to occur anywhere between Day 81 a.m. and Day 87 a.m.). Three hundred and twenty six patients (197 patients with ALS and 129 patients with MS) were enrolled at approximately 60 centers (40 US sites and 20 international sites). Approximately 65 subjects with ALS and 43 subjects with MS were randomly assigned to each of the three treatment groups (AVP-923-30, AVP-923-20 or placebo).

This study was to be randomized by center and by patient underlying neurological disorder (ALS and MS). Eligible patients undergoing the screening period were to be provided with a diary card and instructed to record all laughing and/or crying episodes over a 7-day period prior to entering into the study (randomization). Patients were to be randomly assigned into one of the three treatment groups to receive AVP-923-30 or AVP-923-20 or placebo in a double-blind manner. Patients had to return to the study site for the Baseline visit (Day 1) within 2 days after completion of the 7-day baseline recording period in the diary card. The patient must have had episode counts in the diary for at least a four-day period to determine the baseline episode count. Patients were to take one capsule of study medication in the morning during the first week of the study, and then they were to start taking the study medication twice daily (every twelve hours) for the remaining 11 weeks of the study to complete a 12-week treatment period. The study was to consist of the following visits: Screening (Day -28 to -1), Baseline (Day 1), Visit 2 (Day 15), Visit 3 (Day 29), Visit 4 (Day 57) and Visit 5 (Day 84). For analysis purposes, an additional End of Study (EOS) visit was to be determined, to include Day 84 and Early Termination.

STUDY SAMPLE SIZE

A sample size of approximately 306 total patients, 102 patients in each randomized treatment group with 60 ALS and 42 MS in each of the three arms, was planned for this study. Based on sample size calculations using experience from previous studies, it was estimated that this sample size would be sufficient to detect a 36% reduction in mean episode rates relative to placebo with at least 90% power. It was expected that the longitudinal analysis that was to be used for this study would have somewhat higher power, due to the increased precision that would result from taking within-subject variability into account in the analysis.

Centers that enrolled less than one subject in each of the three treatment groups were to be treated as a single center for analysis purposes and the data from such centers was to be pooled.

Analysis Populations

INTENTION-TO-TREAT (ITT) POPULATION

The Intention-To-Treat (ITT) population refers to all patients randomized. Analysis for the ITT population was to be based on the randomized treatment assigned (regardless of the actual treatment received).

EFFICACY EVALUABLE (EE) POPULATION

The Efficacy evaluable population refers to patients who are protocol adherent. Patients were to be considered protocol adherent if they completed the Day 84 Visit or completed the End-of-Study Visit within 48-hours of discontinuation and if they had taken 80% of their scheduled doses prior to discontinuation of the study medication.

SAFETY POPULATION

The Safety population was to consist of all patients who received at least one dose of study drug.

ENDPOINTS

The primary efficacy endpoint is the number of laughing and/or crying episodes as recorded in the patient diary.

The secondary efficacy endpoints include:

1. Patient score on the Center for Neurologic Study-Lability Scale (CNS-LS) for the assessment of PBA status
2. Patient score on the SF-36 Health Survey (SF-36)
3. Patient score on the Neuropsychiatric Inventory (NPI-Q)
4. Patient score on the Beck Depression Inventory (BDI-II)
5. Patient score on the Pain Rating Scale (PRS) - MS patients only.

All efficacy analyses were to be conducted using two-sided hypothesis tests at the 0.05 significance level. All analyses were to be performed using SAS 9.1 (or higher) and/or Stata 10.1 (or higher). The longitudinal random effects negative binomial model was to be used for the analysis of the primary efficacy endpoint.

PRIMARY EFFICACY ENDPOINT ANALYSIS

The primary efficacy analysis was to be based on the changes from baseline in laughing/crying episode rates recorded in the patient diary and estimated using longitudinal negative binomial regression on the daily episode counts. Daily laughing/crying episode counts were recorded in patient daily diaries. A baseline “daily” episode count was to be calculated based on a patient’s pre-treatment entries recorded at the Baseline visit. The number of pre-treatment days (between 4 and 7) with non-missing episode counts was to be determined as well as the total number of reported episodes over those days. The baseline daily episode rate was then to be calculated as: Baseline episode rate = (number of pre-treatment episodes)/ (number of pre-treatment days with non-missing counts). Daily episode rates at each visit and at the end of study (EOS) were to be determined similarly, using all available non-missing counts for the previous 7 days. The primary outcome is the additional reduction in episode rates experienced with AVP-923-30 compared to placebo. The primary analysis was to adjust for baseline episode rate and study-site differences. Mean changes

in each group were to be assessed using the intention-to-treat population. Primary efficacy analysis was to compare trends in episode rates for the AVP-923-30 (DM30/Q10) dose group and the placebo group. A secondary analysis of the primary efficacy endpoint was to compare episode rates for the AVP-923-20 (20DM/10Q) dose group and the placebo group.

The primary endpoint is the daily laughing and/or crying episode counts. Previous studies have shown that the between-patient variability is likely greater than a simple Poisson model would predict, but may be well-described by a negative binomial model with constant dispersion, which is actually a continuous mixture of Poisson distributions with Poisson rate having a gamma distribution. Thus the longitudinal random effects negative binomial model was to be used for the analysis. The models setting will be described below.

Let Y_{it} denote the total number of episodes recorded in the diary of patient i at time t . For purposes of statistical analysis, “time -1” ($t = -1$) is the last day before the patient receives study drug (Day 1). Thus, the pre-randomization diary entries will have negative times, while on-study dates will have positive times. The first dose is to be taken at the site in the morning of the randomization day ($t = 1$).

Let G_{i1} and G_{i2} indicate the treatment groups to which patient i is randomized, $G_{i1}=1$ if patient i is randomized to AVP 20 and $=0$ otherwise. Similarly, $G_{i2}=1$ if patient i is randomized to AVP-923-30 and $=0$ otherwise. In addition, let $C_{i1}, C_{i2}, \dots, C_{i,k-1}$ indicate the study site for patients in the sites 2, 3, \dots, k respectively. Let P_t be a pre-randomization period indicator that is 1 prior to randomization ($t < 0$) and is 0 after randomization ($t > 0$). Similarly, R_t is used to denote the post-randomization period. That is, $R_t = 1 - P_t$. Thus R_t and P_t are the same for all patients at a specified time point.

In addition, let D_i denote the patient's diagnosis, coded as 0 for amyotrophic lateral sclerosis (ALS) and 1 for multiple sclerosis (MS). The conditional mean episode rate for the i_{th} patient at time t can be denoted by λ_{it} and it can be assumed that the dispersion parameter is constant.

The longitudinal NB1 model with

$$\log(\lambda_{it}) = \mu + \pi R_t + \beta_1 G_{i1} R_t + \beta_2 G_{i2} R_t + \sum \gamma_j C_{ij} + \delta D_i$$

can be used. The comparison for episode rates experienced with AVP-923-30 or placebo will be based on the estimate of $\exp(\beta_2)$ in the longitudinal model, and the comparison for episode rates experienced with AVP-923-20 or placebo will be based on the estimate of $\exp(\beta_1)$ in the model. The estimate of δ indicates the difference in patients with ALS and the patients with MS. The following Stata example code estimates this random effect longitudinal NB1 negative binomial regression model. In the example, s2-s20 are indicator variables for study sites.

```
generate g1 = (rx == "AVP-923-20")
generate g2 = (rx == "AVP-923-30")
generate t = day-1
generate P = (t<=0)
generate R = (t>0)
generate rx1 = R*g1
generate rx2 = R*g2
xtnbreg count R rx1 rx2 D s2-s20, i(patientid)
```

PRIMARY EFFICACY ENDPOINT SENSITIVITY ANALYSIS

A non-longitudinal negative binomial (constant dispersion) model analysis of the sums of the episode counts over the double blind phase (with an offset based on the number of non-missing diary days), as was used for the previous trials, was to be carried out as a sensitivity analysis. The baseline episode rate was to be used as a covariate in this sensitivity analysis. In addition, the generalized estimating equation (GEE) method may be used. This can be modeled as:
$$\log(\lambda_{it}) = \mu + \pi R_t + \beta_1 G_{i1} R_t + \beta_2 G_{i1} R_t + \sum \gamma_j C_{ij} + \delta D_i$$
with $\text{Var}(Y_{it}) = \psi \lambda_{it}$, where ψ is the dispersion parameter. A negative binomial distribution function and compound symmetry correlation variance structure can be used to fit the model.

SECONDARY EFFICACY ENDPOINT ANALYSIS

The secondary efficacy endpoints were to be examined in the following order:

- (1) mean change in the Center for Neurologic Study-Lability Scale (CNS-LS) score
- (2) mean change in Neuropsychiatric Inventory (NPI-Q)
- (3) mean change in the SF-36 Health Survey (SF-36)
- (4) mean change in the Beck Depression Inventory (BDI-II)
- (5) mean change in Pain Rating Scale (PRS) score in MS patients.

All secondary efficacy variables were to be analyzed as differences between Day 84 (not to include Early Termination) and baseline values. Except for the SF-36 Reported Health Transition item, the analyses of these endpoints were to be performed by multiple regression models that included treatment as the fixed effect and baseline value, study site, and diagnosis (ALS or MS) as covariates; the changes from baseline value to other applicable visits, such as for CNS-LS, were to be analyzed similarly. In addition, for pain scores, a responder analysis for pain improvement was to be presented in a figure, showing percent of MS patients improved versus percent improvement in pain from baseline. Where applicable, baseline values were to be chosen as the latest non-missing value prior to start of treatment. For PRS scores, baseline values were to be calculated as the average of scores recorded in the pre-treatment diary.

For the SF-36 Reported Health Transition item, change from baseline to Day 84 was to be based on three overall categories: Improved, No Change or Worsened. Treatment values were to be compared to placebo using a chi-square test for row mean score differences.

ADDITIONAL ANALYSES

The analysis plan stated that additional analyses to clarify clinical understanding of the treatments and/or generalizability of the findings may also be performed and would include:

- Time to onset of action (a 30% decrease from baseline in Laughing + Crying episode count)
- Number of episode-free days
- Percentage of patients showing remission (no episodes during the last 14 days of study participation)
- Percentage of patients showing clinical response (40% decrease in episode rate at the end of the study)
- Analysis of episode rates and CNS-LS by diagnosis (ALS or MS)
- Analysis of episode rates and CNS-LS by SSRI usage status.

3.1.1.2 Disposition of Subjects

The disposition of subjects is summarized in Table 1. Overall, 326 subjects were randomized, 110 were assigned to AVP-923-30, 107 to AVP-923-20, and 109 to placebo.

A total of 283 subjects (86.8%) completed the study, and 43 (13.2%) withdrew from the study. Across the treatment groups, the number and proportion of subjects who completed the study ranged from 88 (82.2%) in the AVP-923-20 group to 101 (91.8%) in the AVP-923-30 treatment group. Of the 43 subjects who withdrew from the study, 19 (17.8%) were in the AVP-923-20 group, 15 (13.8%) were in the placebo group, and 9 (8.2%) were in the AVP-923-30 group. Overall, the most frequent reasons for withdrawal were withdrawal of consent (3.4% of subjects), lost-to-follow-up (1.8% subjects), AE (1.8% of subjects), and SAE (1.8% of subjects). Seven subjects, all with ALS as the primary disease, died during the DB phase of the study.

Table 1 Patient Disposition

Subject Category	Number of Subjects (%)				OLE Phase AVP-923-30 (N = 253)
	DB Phase			Overall (N = 326)	
	AVP-923-30 (n = 110)	AVP-923-20 (n = 107)	Placebo (n = 109)		
Subjects screened				332 (100)	
Subjects with no reported diagnosis				3 (0.9)	0
Subjects randomized	110 (100)	107 (100)	109 (100)	326 (100)	
Subjects with ALS	65 (59.1)	68 (63.6)	64 (58.7)	197 (60.4)	146 (57.7)
Subjects with MS	45 (40.9)	39 (36.4)	45 (41.3)	129 (39.6)	107 (42.3)
Subjects in the United States	77 (70.0)	72 (67.3)	75 (68.8)	224 (68.7)	167 (66.0)
Subjects in Latin America	33 (30.0)	35 (32.7)	34 (31.2)	102 (31.3)	86 (34.0)
Subjects dosed	110 (100)	107 (100)	109 (100)	326 (100)	253 (100)
Subjects completing study	101 (91.8)	88 (82.2)	94 (86.2)	283 (86.8)	235 (92.9)
Subjects who withdrew	9 (8.2)	19 (17.8)	15 (13.8)	43 (13.2)	18 (7.1)
Reason for withdrawal					
Lost to follow-up	1 (0.9)	3 (2.8)	2 (1.8)	6 (1.8)	1 (0.4)
Exacerbation of MS	1 (0.9)	0 (0.0)	1 (0.9)	2 (0.6)	2 (0.8)
Adverse event	1 (0.9)	5 (4.7)	0 (0.0)	6 (1.8)	3 (1.2)
Serious adverse event	2 (1.8)	3 (2.8)	1 (0.9)	6 (1.8)	5 (2.0)
Medication refusal due to AE	2 (1.8)	2 (1.9)	0 (0.0)	4 (1.2)	0 (0)
Withdrew consent	2 (1.8)	2 (1.9)	7 (6.4)	11 (3.4)	3 (1.2)
Protocol violation	0 (0.0)	2 (1.9)	1 (0.9)	3 (0.9)	2 (0.8)
Other	0 (0.0)	2 (1.9)	3 (2.8)	5 (1.5)	2 (0.8)

Source: DB phase, [Section 14.1, Table 2](#); OLE phase, [Section 14.2, Table 2](#).

DB = double-blind; OLE = open-label extension; ALS = amyotrophic lateral sclerosis; MS = multiple sclerosis; AE = adverse event.

Note: For the DB phase, percentages are based on the number of subjects randomized in each treatment group and overall, except for screened subjects. For the OLE phase, percentages are based on the number of subjects in the OLE phase.

Note: This table was copied from page 51 of the sponsor's study report

3.1.1.3 Demographic Characteristics

There were no statistically significant between-group differences for any of the demographic characteristics in either the ITT or EE populations. The mean age ranged from 50.27 years in the placebo group to 53.08 years in the AVP-923-30 group. The majority of subjects were Caucasian, with percentages of Caucasian subjects ranging from 72.7% in the AVP-923-30 group to 76.1% in the placebo group. As expected from the distribution of study sites, Hispanics accounted for the next highest percentage of subjects, ranging from 19.1% in the AVP-923-30 group to 19.6% in the AVP-923-20 group. Subjects from other ethnic groups did not exceed more than 5.5% in any one treatment group.

The sponsor reported the durations of disease in ALS patients as in Table 2.

Time from diagnosis of ALS at the time of randomization in the DB phase was markedly different among the 3 treatment groups as reported by the sponsor. Mean time from diagnosis was 22.68 months in the AVP-923-30 group, 16.33 months in the AVP-923-20 group, and 13.36 months in the placebo group. Three of the most relevant independent prognostic factors for higher risk of death in ALS subjects are longer disease progression, age at onset, and the presence of bulbar symptoms. Table 2 summarizes the differences in time from diagnosis in the ALS population in each DB treatment group.

Table 2 Time from Diagnosis (Months)^a in ALS Patients

Parameter	Time from Diagnosis (Months) ^a		
	AVP-923-30 (n = 65)	AVP-923-20 (n = 68)	Placebo (n = 64)
Mean (SD)	22.68 (29.8)	16.33 (22.87)	13.36 (18.01)
Median	14.4	10.26	7.48

Source: [Section 14.1, Listing 1](#).

ALS = amyotrophic lateral sclerosis; SD = standard deviation.

^aAt time of randomization.

Note: this table copied from page 90 of sponsor's study report

This reviewer was unable to verify the numbers in the table exactly instead finding mean durations of 21.7 for AVP30, 16.4, for AVP20, and 13.1 for Placebo.

Based on these numbers both a t-test and a Wilcoxon rank sum test yield a p-value of 0.055 for comparing AVP30 and placebo, the two groups with the most different mean durations of ALS. Note that one of the 30 mg patients was missing the month so the disease onset which was calculated as 59 months assuming January 1, as the missing month and day may have been as much as 11 months later. Thus, the duration could be as low as 48. A few other patients were missing the day so their durations could be up to a month shorter.

Ten (2 AVP30, 7 AVP20, and 1 Placebo) of these patients had no post-baseline data and two of these 10 (1 AVP20 and 1 AVP30) had no baseline episode diary data either.

Table 3 Demographics of Randomized Patients

Characteristic	DB Phase			OLE Phase
	AVP-923-30 (n = 110)	AVP-923-20 (n = 107)	Placebo (n = 109)	AVP-923-30 (N = 253)
Age^a (years)				
n	110	107	109	253
Mean (SD)	53.08 (11.016)	50.81 (11.114)	50.27 (11.939)	52.02 (11.301)
Median (min, max)	54.5 (29.0, 76.0)	50.0 (28.0, 80.0)	50.0 (25.0, 75.0)	51.00 (26.0, 80.0)
Ethnicity (n)				
Caucasian	80 (72.7%)	80 (74.8%)	83 (76.1%)	198 (78.3%)
Black	6 (5.5%)	2 (1.9%)	4 (3.7%)	6 (2.4%)
Asian	1 (0.9%)	0	1 (0.9%)	2 (0.8%)
Hispanic	21 (19.1%)	21 (19.6%)	21 (19.3%)	44 (17.4%)
Other	2 (1.8%)	4 (3.7%)	0 (0.0)	3 (1.2%) ^c
Sex (n)^b				
Male	46 (41.8%)	53 (49.5%)	50 (45.9%)	116 (45.8%)
Female	64 (58.2%)	54 (50.5%)	59 (54.1%)	137 (54.2%)
Height (cm)^b				
n	110	107	109	253
Mean (SD)	168.35 (9.420)	168.86 (9.546)	169.15 (9.576)	168.89 (9.613)
Median (min, max)	169.5 (139.7, 188.0)	168.0 (147.0, 190.5)	167.6 (149.9, 191.0)	168.00 (139.7, 191.0)
Weight (kg)				
n	110	107	109	246
Mean (SD)	73.25 (14.324)	74.09 (15.921)	76.85 (20.388)	74.11 (17.247)
Median (min, max)	73.05 (44.0, 110.7)	72.80 (44.5, 127.1)	73.5 (44.4, 145.1)	70.95 (43.3, 145.1)
Pulse rate (bpm)				
n	110	107	109	252
Mean (SD)	76.04 (11.031)	76.27 (9.991)	75.04 (9.196)	74.98 (10.920)
Median (min, max)	76.0 (50.0, 110.0)	76.00 (60.0, 107.0)	76.0 (45.0, 100.0)	74.00 (52.0, 113.0)
Body temperature (°C)				
n	110	107	109	252
Mean (SD)	36.44 (0.454)	36.38 (0.457)	36.34 (0.420)	36.37 (0.448)
Median (min, max)	36.45 (35.0, 37.3)	36.5 (34.6, 37.4)	36.40 (35.2, 37.3)	36.40 (34.2, 37.3)
Systolic BP (mm Hg)				
n	110	107	109	252
Mean (SD)	121.98 (12.388)	123.26 (13.564)	122.61 (13.625)	120.72 (13.680)
Median (min, max)	120.0 (97.0, 163.0)	121.0 (97.0, 184.0)	120.0 (100.0, 174.0)	120.00 (77.0, 160.0)
Diastolic BP (mm Hg)				
n	110	107	109	252
Mean (SD)	74.74 (9.682)	76.36 (10.667)	77.04 (9.624)	74.24 (10.138)
Median (min, max)	75.0 (49.0, 102.0)	78.00 (57.0, 110.0)	78.0 (57.0, 102.0)	75.00 (20.0, 100.0)

Source: DB phase, Section 14.1, Table 4.1; OLE phase, Section 14.2, Table 4.1.

ITT = intent to treat; DB = double blind; OLE = open-label extension; SD = standard deviation; min = minimum; max = maximum; bpm = beats per minute; BP = blood pressure.

^aAge was calculated as the number of full years completed from date of birth until screening date in DB phase.

^bSex and height for subjects in the OLE phase were from screening visit of the DB phase of the study.

Note: Copied from page 53 of sponsor's study report

3.1.1.4 Sponsor's Results

Analysis of Primary Efficacy Variable

In the overall ITT population, subjects treated with AVP-923-30 experienced approximately half as many episodes of inappropriate laughing, crying, and laughing and crying as subjects receiving placebo ($\exp(-0.6326) = 53.12\%$); the number of these episodes was significantly lower in the AVP-923-30 group than in the placebo group ($p < 0.0001$). Similarly, subjects in the overall ITT population treated with AVP-923-20 experienced approximately half as many episodes of inappropriate laughing, crying, and laughing and crying as subjects receiving placebo ($\exp(-0.6727) = 51.03\%$), and the number of these episodes was significantly lower in the AVP-923-20 group than in the placebo group ($p < 0.0001$).

Table 4 Primary Longitudinal Negative Binomial Model- Laughing/Crying Episode Rates (ITT Population)

Parameter	Estimate	Standard Error	95% Confidence Interval	P-value
Intercept	0.9414	0.1116	(0.7228, 1.1601)	<0.0001
Difference (Post vs Pre)	-0.7661	0.0263	(-0.8177, -0.7146)	<0.0001
Exp (Difference (Post vs Pre))	0.4648	0.0122	(0.4414, 0.4894)	
Treatment (AVP-923-30 vs Placebo)	-0.6326	0.0372	(-0.7054, -0.5597)	<0.0001
Exp (Treatment (AVP-923-30 vs Placebo))	0.5312	0.0197	(0.4939, 0.5714)	
Treatment (AVP-923-20 vs Placebo)	-0.6727	0.0360	(-0.7433, -0.6021)	<0.0001
Exp (Treatment (AVP-923-20 vs Placebo))	0.5103	0.0184	(0.4755, 0.5477)	
Difference (ALS vs MS)	0.4575	0.0702	(0.3199, 0.5952)	<0.0001
Exp (Difference (ALS vs MS))	1.5802	0.1110	(1.3770, 1.8134)	

Note: Model includes adjustment for sites, with small sites pooled.

Copied from page 406 of sponsor's study report

The non-longitudinal analysis was performed as a sensitivity analysis on the total number of laughing and crying episodes over the entire double-blind phase. This was the same analysis method as prespecified for the earlier trial in MS patients, study number 106. Instead of assuming each day's episode count is negative binomially distributed it assumes the sum of the counts over the entire double blind period is negative binomially distributed. P-values were computed using a negative binomial regression (constant dispersion) model with baseline episode rate, pooled study site and underlying disease diagnosis included as covariates. In the overall ITT population, the effect of AVP-923 was statistically significant when compared with placebo ($p < 0.0001$ for AVP-923-30 and $p < 0.04$ for AVP-923-20). Results for the nonlongitudinal analysis for only crying episodes were consistent with the results for laughing and crying episodes combined ($p < 0.0001$ for AVP-923-30 and $p = 0.003$ for AVP-923-20). From an analysis of laughing episodes only, the effects of AVP-923-30 and AVP-923-20 were

not statistically significant compared with placebo. Results for the EE population were consistent with those for the ITT population.

Table 5 Non-longitudinal Negative Binomial Model –Total Episode Rates (ITT Population)

Population	Combined			ALS Only			MS Only		
	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Laughing+Crying		N=312			N=186			N=126	
Parameter	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Treatment (AVP-923-30 vs Placebo)	-0.5442	0.1156	<0.0001	-0.7255	0.1500	<0.0001	-0.2148	0.1710	0.209
Treatment (AVP-923-20 vs Placebo)	-0.2180	0.1048	0.037	-0.2475	0.1256	0.049	-0.1846	0.1841	0.316
Baseline rate	0.0323	0.0020	<0.0001	0.0335	0.0023	<0.0001	0.1199	0.0287	<0.0001
Diagnosis (ALS vs MS)	0.0761	0.1640	0.643						
Intercept	-0.2258	0.2233	0.312	0.0275	0.1855	0.882	-0.9833	0.2858	0.001
Exp (AVP-923-30 vs Placebo)	0.5803		<0.0001	0.4841		<0.0001	0.8067		0.209
Exp (AVP-923-20 vs Placebo)	0.8041		0.037	0.7808		0.049	0.8314		0.316

Note: P-values are computed using negative binomial regression (constant dispersion) for the total number of episodes over the double-blind period, with baseline episode rate, and pooled study site included as covariates. Diagnosis is included as a covariate in the analysis of the combined population.

Note: This table was copied from page 410 of sponsor’s study report

A GEE model was performed as a second sensitivity analysis of the primary endpoint, with compound symmetry correlation variance structure, assuming a negative binomial probability distribution for the number of daily laughing and crying episodes and with independent variables including treatment (AVP-923-30 vs. placebo and AVP-923-20 vs. placebo), period (before or after treatment), underlying disease diagnosis (ALS vs. MS), and site category (U.S. or non-U.S.). In the overall ITT population, subjects treated with AVP-923-30 had approximately half as many daily laughing, crying, and laughing and crying episodes as subjects in the placebo group ($\exp(-0.7478) = 47.34\%$) and that this treatment effect was statistically significant ($p = 0.0002$). The effect due to AVP-923-20 compared to placebo was not statistically significant ($p = 0.2622$) under this GEE model.

Reviewer’s Comment: It is not clear why the sponsor reported the results from the GEE model adjusted for site with sites categorized according to US vs. non-US, as opposed to adjusting for each individual site as in the primary model. The low dose is nominally significant (IRR=.546, $p < 0.0001$) when the latter model is used as can be seen in section 3.1.1.5.4.

Table 6 GEE Model for Number of Laughing/Crying Episodes (ITT Population)

Parameter	Estimate	Standard Error	95% Confidence Interval	P-value
Intercept	1.1967	0.0901	(1.02008, 1.37325)	<.0001
Exp(Intercept)	3.3091	0.2981	(2.77341, 3.94816)	
Difference (Post vs Pre)	-1.8082	0.2945	(-2.38552, -1.23095)	<.0001
Exp[Difference (Post vs Pre)]	0.1639	0.0483	(0.09204, 0.29202)	
Treatment (AVP-923-30 vs Placebo)	-0.7478	0.1995	(-1.13883, -0.35675)	0.0002
Exp[Treatment (AVP-923-30 vs Placebo)]	0.4734	0.0945	(0.32019, 0.69994)	
Treatment (AVP-923-20 vs Placebo)	-0.2755	0.2457	(-0.75709, 0.20612)	0.2622
Exp[Treatment (AVP-923-20 vs Placebo)]	0.7592	0.1866	(0.46903, 1.22890)	
Difference (ALS vs MS)	0.6593	0.1773	(0.31166, 1.00684)	0.0002
Exp[Difference (ALS vs MS)]	1.9333	0.3429	(1.36570, 2.73694)	

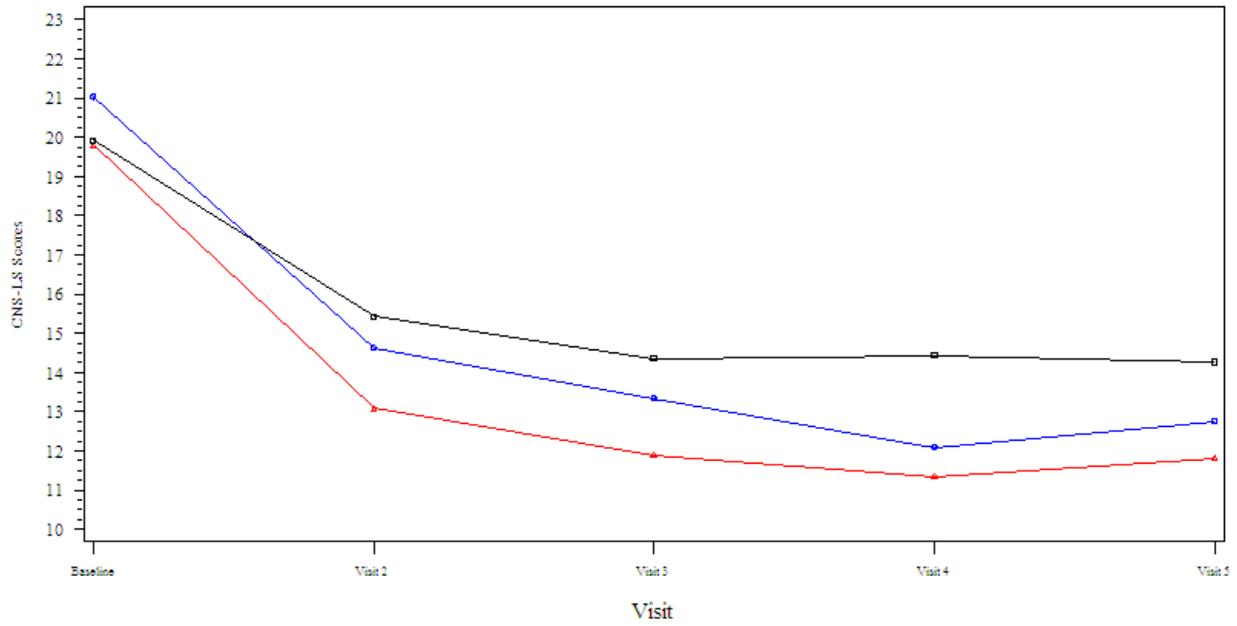
Note: The table is based on the Generalized Estimating Equation with Exchangeable correlation matrix. The daily laughing and crying episodes are assumed to have a negative binomial probability distribution. The independent variables include Treatment (after randomization), Period (prior or after randomization), Diagnosis (ALS vs MS), and Site Category (US vs Non-US).

Note: This table was copied from page 419 of sponsor's study report

Analysis of Secondary Efficacy Variable

The CNS-LS is a 7-item self-report questionnaire that measures the frequency and severity of PBA episodes, including assessments of labile laughter and labile tearfulness, and provides a score for total PBA. For the DB phase of the study, decreases from baseline at all study visits (Days 15, 29, 57, and 84) in CNS-LS total scores were shown in both AVP-923 treatment groups and in the placebo group, using the ITT population. The differences were nominally significant between the AVP-923-30 group and the placebo group at all study visits and were nominally significant between the AVP-923-20 group and the placebo group at Days 57 and 84, but not at Days 15 and 29. Figure 1 shows the CNS-LS scores plotted over time. Both AVP30 and AVP20 were significant compared to placebo and there was little difference between AVP20 and AVP30. Note that although the AVP20 curve (connected means) is consistently above the AVP30 curve it appears to be mainly due to the baseline difference.

Figure 1 CNS-LS scores over Time in ITT Population



Treatment ▲▲▲ AVP-923-30 ●●● AVP-923-20 ■■■ Placebo

Note: This figure copied from page 59 of sponsor's study report

Results using the EE population in the DB phase were consistent with the results using the ITT population.

Table 7 shows the CNS-LS results for the ITT Population.

Table 7 Change from Baseline in CNS-LS (ITT Population)

	AVP-923-30 (N = 110)	AVP-923-20 (N = 107)	Placebo (N = 109)	P-value [1]
Visit 5 (Day 84)				
N	103	96	101	
Mean (Std Dev)	11.82 (4.582)	12.76 (4.996)	14.27 (5.221)	
95% C.I.	(10.92, 12.71)	(11.75, 13.77)	(13.24, 15.30)	
Median	11.00	12.00	14.00	
Min , Max	7.0 , 24.0	7.0 , 24.0	7.0 , 27.0	
Change from Baseline to Day 84				
N	103	96	101	
Mean (Std Dev)	-8.17 (6.104)	-8.24 (6.126)	-5.72 (5.280)	0.0008
95% C.I.	(-9.37, -6.98)	(-9.48, -7.00)	(-6.77, -4.68)	
Median	-7.00	-8.00	-5.00	
Min , Max	-26.0 , 3.0	-23.0 , 6.0	-22.0 , 6.0	
P-value [2]	0.0002	0.0113		

Note: The CNS-LS is a seven-item, self-administered questionnaire that measures the perceived frequency and severity of pseudobulbar affect (PBA) episodes. A CNS-LS score of 13 or higher may suggest PBA.

[1] P-value for treatment effect in multiple regression model with baseline, site and diagnosis (ALS/MS) as covariates.

[2] P-value based on contrast comparing active treatment with placebo.

Copied from page 530 of sponsor's study report

3.1.1.5 Reviewer's Results

3.1.1.5.1 Primary Endpoint

The average numbers of post-baseline period diary entries were 74, 72, and 75 days for placebo, AVP20, and AVP30, respectively. The medians were 82, 83, and 83 and the ranges were 1 to 88, 1 to 87, and 0 to 85.

Table 8 gives summary statistics for the average daily laughing and crying episode counts by period. This simple summary adjusts for the fact that the sum will tend to be lower when a patient terminates early because the patient has fewer days than a completing patient by using the mean instead of the sum. The table also shows summary statistics for the number of days with non-missing episode counts.

Table 8 Patients' Average Daily Rate of Laughing and/or Crying Episodes

Group	Stat	Base Days	Baseline Episode rate	P.B. Days	P.B. Episode rate
AVP30	Mean	6.83	4.6	74.9	.954
	Median	7	2.93	83	.307
	S.D.	1.21	9.48	19.8	1.39
	Min	1	0	0	0
	Max	16	95.9	85	9.83
AVP20	Mean	7.47	6.71	72	2.44
	Median	7	3.07	83	.385
	S.D.	6.24	12.9	24	7.92
	Min	3	.143	1	0
	Max	70	78.9	87	57.9
Placebo	Mean	6.78	4.44	74.5	2.08
	Median	7	2.46	82	.857
	S.D.	1.08	7.61	17.8	3.04
	Min	2	0	1	0
	Max	13	69	88	18.8

Table 9 gives summary statistics for the sum total of all laughing and crying episode counts by period. This simple summary does not adjust for the fact that the sum will tend to be lower when a patient terminates early because the patient has fewer days than a completing patient. However, as seen in the previous table the groups are reasonably balanced with respect to the number of days with non-missing counts per patient.

The mean of the baseline period episode sum is considerably higher for AVP20 than placebo. Based on a negative binomial regression model for the baseline period total sum of laughing+crying episodes the estimated episode rate ratio of AVP20/placebo is 1.226, p=.049 suggesting that the AVP20 baseline episode rate is higher than placebo (note: AVP30 /placebo=.924, p=.463). However, the medians are closer and a nonparametric Wilcoxon rank sum test did not corroborate this nominal significance between AVP20 and placebo during the baseline period, p=0.203.

Table 9 Total Sum of Laughing and Crying Episodes by Period(ITT Population)

statistic	Group					
	Placebo(N=108)		AVP20(N=106)		AVP30(N=108)	
	Baseline total episode sum	Post-Baseline Total Episode sum (N=107)	Baseline total episode sum	Post-Baseline Total Episode sum (N=100)	Baseline total episode sum	Post-Baseline Total Episode sum (N=107)
Mean	30.58	149.73	53.09	156.89	31.71	69.52
Median	18	59	21.5	25	19	24
S.D.	53.12	218.14	111.37	489.82	66.44	109.95
Range	0, 483	0, 1575	1, 714	0, 3413	0, 671	0, 816

Table 10 gives summary statistics for the patient min and max total daily episode count. Thus, for example, the AVP 30 mg group minimum of the maximum patient total count is the smallest of the 106 patients' maximum total daily episode counts. This provides more information about patient's extreme counts which is not well captured by the patient total sum or mean count. Ninety nine percent of daily episode counts were < 22. There were 18 daily laughing+crying counts (7 of these occurred during the baseline period) arising from just 4 patients that were greater than 100.

Table 10 Patient’s Min/Max Daily Episode Counts (ITT Population)

Treat Group		Patient min Count	Patient max count
AVP30(N=106)	Min	0	0
	Mean	0.085	4.849
	Median	0	4
	Max	6	45
AVP20(N=100)	Min	0	0
	Mean	.51	12.48
	Median	0	3
	Max	17	456
Placebo (N=107)	Min	0	0
	Mean	0.299	8.589
	Median	0	5
	Max	10	110
All	Min	0	0
	Mean	0.294	8.565
	Median	0	4
	Max	17	456

In contrast to the previous studies, here the sponsor chose to prespecify an analysis that focuses on the individual daily post-baseline episode counts rather than analyzing the sum over the entire post-baseline period. The assumptions of these two models (i.e., individual day count and post-baseline sum) differ. The daily model treats each daily count as following a negative binomial distribution. For the same patient each post-baseline day is assumed to have the same expected count. For the analysis based on the sum over the period the sum total of all counts over the post-baseline period is assumed to follow a negative binomial distribution. Obviously, because it models a sum over many days the distribution for the sum over the period tends to follow a negative binomial distribution with a greater mean than that for the daily model. Because not all patients completed 84 days this total D.B. period model requires an offset, i.e., an adjustment to reflect the fact that all other things being equal a patient with more diary days completed can be expected to have a higher D.B. total episode count than a patient with fewer days completed. In focusing on the day the sponsor’s model does not require this offset but it does make necessary for an alternative assumption, in particular, that the expected daily count doesn’t change over the

entire 84 day post-baseline period. Despite their different approaches the conclusions from these two models agree on the significance of both AVP dose groups compared to placebo.

The adjusted estimated log of the mean postbaseline daily count from the primary longitudinal model are:

Placebo	AVP20	AVP30
1.5531	.7920	.8259

For the baseline period the corresponding mean over all groups is 3.3434.

The primary longitudinal model results were not sensitive to excluding the very large (>100) daily counts (without: IRR=.509 AVP20/Placebo and .548 AVP30/Placebo vs. with: IRR=.510 AVP20/Placebo and .532 AVP30/Placebo).

The Incident (Episode) Rate Ratio (IRR) is .590 ($p < 0.001$: .470, .739) for AV30/Placebo and .801 for AV20/Placebo ($p = 0.037$: .652, .984) based on NB regression of the total post-baseline laughing plus crying episode sums adjusted for baseline, disease, and sites. If we don't adjust for baseline in the model as in the previous studies which had no baseline period we find .768 (95% CI: .603, .978) for AV30/Placebo and .631 (95% CI: .496, .802) for AV20/Placebo. These results are sensitive to some very large total episode counts. Ninety five percent of the post-baseline total laughing+crying episode counts were ≤ 483 . There were 14 patients with post-baseline total episode counts greater than 500. Seven of these patients (4 AVP20 and 3 placebo) and all 4 patients with counts above 1000 came from site 121 (N=22). A sensitivity analysis excluding all data from site 121 yielded estimated incident rate ratios of .656 ($p < .001$) for AVP30/Placebo and .699 ($p = .004$) for AVP20/Placebo based on the non-longitudinal negative binomial model. Note that the primary longitudinal model was relatively insensitive to the exclusion of data from this site. There is an alternative non-longitudinal negative binomial model for the post-baseline total episode sums that differs only in it's assumption about how the variance of the total post-baseline episode count depends on the mean. Instead of assuming the variance is proportional to the mean, as in the model just reported, it assumes the variance depends on the sum of the mean and the square of the mean. If we use this alternative negative binomial model the estimated IRRs are .466 for AVP30/Placebo ($p < .001$) and .467 ($p < .001$) for AVP20/Placebo based on all of the post-baseline total count data.

Because the negative binomial models may be unfamiliar and have a lot of assumptions this reviewer also performed a simple nonparametric test. As seen in Table 11 the simple (unadjusted) median change from baseline in the average daily count shows little difference between AVP20 and AVP30, but both groups' changes from baseline are judged nominally significant ($p = .0001$ and .0008, respectively) compared to placebo based on the Wilcoxon rank sum test.

Table 11 Change from Baseline in Average Daily Laughing+Crying Episode Counts

Group	N	Bsln Median of Daily Avg	Post Bsln Median of Daily Avg	Median Chg from Bsln Daily Avg	Mean Chg from Bsln Daily Avg	Std Dev of Chg from Bsln Daily Avg
AVP30	105	2.929	0.307	-2.093	-3.632	9.108
AVP20	100	3.071	0.385	-2.115	-4.432	7.143
Placebo	106	2.464	0.857	-1.032	-2.404	5.789

Also, a simple nonparametric Wilcoxon rank sum test of the post-baseline total episode counts (sums) unadjusted for baseline yields a p-value of .0001 for the AVP30 vs. placebo comparison and .0011 for the AVP20 mg vs. placebo comparison.

3.1.1.5.2 Effect of Protocol Amendment to Increase Sample Size in MS subgroup

The protocol was amended after the study was underway and one of the resulting changes was an increase in the sample size of MS patients from 90 to 120. Total episode counts analyses based on only first 270 patients, i.e., the originally planned sample size yielded incident rate ratios for AVP to placebo as follows:

0.688 (p=0.001) AV20 and 0.562 (p<0.001) for AV30. The estimated IRRs for this sample were .448 (p<0.001) and .482 (p<0.001) based on the corresponding longitudinal analysis.

3.1.1.5.3 Assessment of the Impact of Missing Data

The proportions in each group completing the study as reported by the sponsor were 92, 82, and 86 for AVP30, AVP20, and Placebo, respectively. About 75% of randomized patients recorded episode data on 70 days or more in the planned 84 day post-baseline period.

Rerunning the primary analysis in completers (defined for this analysis as those with at least 78 days of past-baseline diary entries) the resulting incident rate ratios of AVP relative to placebo are .531 (p<0.001) for AVP20 and .572 (p<0.001) for AVP30. Based on the non-longitudinal negative binomial model the incident rate ratios in this completers subgroup (72, 76, and 79% of placebo, A20, and A30; Total N=238) were .796 (AVP20/Placebo, p=0.059) and .594 for (AVP30/Placebo, p<0.001). The AVP20/Placebo ratio became nominally significant if instead of adjusting for all sites we just adjusted for whether the site was based in the U.S. or not. Table 12 shows summary statistics for the average daily episode count by completion status.

Table 12 Summary Statistics for Daily Episode Rate by Completion Status

Completion Status	Statistic	AVP30	AVP20	Placebo
Dropout	N	8	18	15
	Mean	.537	5.938	3.057
	Median	0	.75	1
	Min	0	0	0
	Max	12	108	67
Completer	N	101	88	94
	Mean	.984	1.961	1.943
	Median	0	0	0
	Min	0	0	0
	Max	45	456	110

Also, this reviewer found that significance of the primary result remained after an imputation filling in with the patient's last week's data when the planned duration of the treatment period was not completed. In addition, an analysis of just the last week of post-baseline data found nominally significant treatment differences for AVP20 and AVP30 compared to placebo.

Sensitivity Analysis for Patients that Died

As a means of checking for informative censoring of episode counts due to death this reviewer conducted a sensitivity analysis in which the sum of the episode counts for those patients that died was replaced with the highest observed daily rate which was 58 times the patients' number of post-baseline diary entries. Under these conditions the estimated incident rate ratios to placebo become .847 (for AVP20, p=0.214) and .734 (for AVP30, p=0.017). Therefore, the primary result at least for AVP30 mg seems not too sensitive to assuming high episode rates for those that died.

Table 13 Laughing + Crying Episodes for those that died

Subjid	Trt	Bsln Days	Bsln Episodes	Bsln Rate	Post Bsln Days	Post Bsln Episodes	Post Bsln Rate
126501	1	7	4	.5714286	55	5	.0909091
133501	2	7	32	4.571429	56	150	2.678571
135501	2	4	10	2.5	23	48	2.086957
135508	1	7	15	2.142857	9	0	0
135511	3	7	14	2	.	.	.
301501	2	7	15	2.142857	84	21	.25
301504	1	7	20	2.857143	84	38	.452381

3.1.1.5.4 Primary Longitudinal Model Issues

It is unusual to include baseline measurements in the model as measurements of the dependent variable, albeit with a baseline/post-baseline indicator effect in the model. If we omit the baseline data, instead using a baseline average count per day as a covariate, we find incident rate ratios of AVP to placebo of:

.497 ($p < 0.001$) for AVP20 and .721 ($p = 0.001$) for AVP30. The AVP30 ratio is slightly sensitive to the inclusion of the baseline covariate without it the ratio is smaller .586 ($p < 0.001$).

There were 224 cases out of more than 22000 episode count data records where a daily laughing episode count was missing when the corresponding daily crying episode count was available or vice versa. The sponsor treated both daily laughing and crying counts as missing in such cases. However, this reviewer found that the results were not sensitive to various other assumptions for such partially missing counts. For example, if the last count of the same type was carried forward or if the maximum count of the same type up to that time was used the conclusions were still the same.

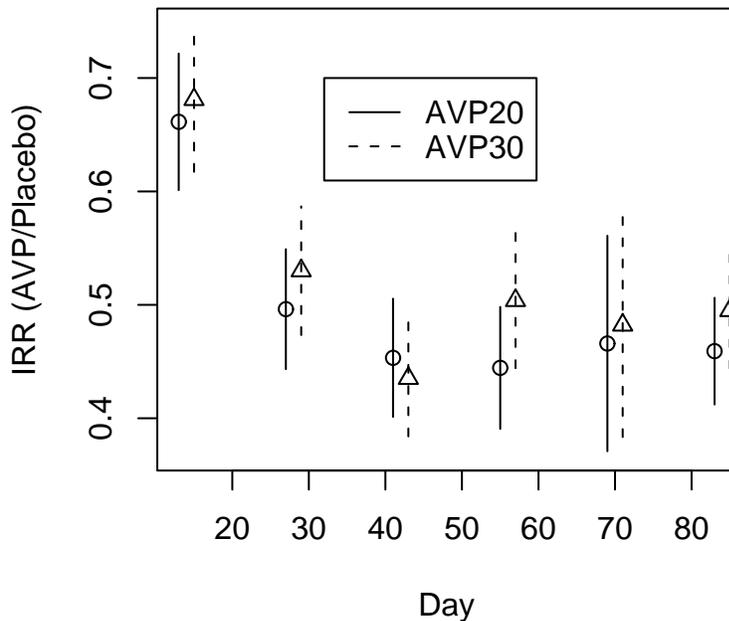
The primary model assumes that each subject's dispersion parameter, d , which controls how much the variance exceeds the mean [$\text{var} = \text{mean} * (1 + d)$] is related to a Beta distribution as follows: $1/(1+d) \sim \text{Beta}$. Like the Normal distribution the Beta distribution depends on two parameters, but unlike the Normal distribution it's underlying random variable only can assume values between 0 and 1. The primary longitudinal model assumes that the dispersion is constant over the pre-treatment and treatment periods. However, if the model is applied to each period separately it appears that the two parameters associated with the Beta distribution random effect may differ for the two periods. The first Beta parameter is estimated to be 5.54 with a 95% C.I. of (4.45, 6.88) for pre-treatment, whereas it is estimated to be 1.84 (1.53, 2.21) for the treatment period. The second Beta parameter is estimated to be 1.84 with a 95% C.I. of (1.55, 2.19) for pre-treatment, but it is estimated as .70 with a 95% C.I. of (.60, .81) for the treatment period.

The sponsor assumed that the expected daily episode count for any particular patient does not change over the 84 day post-baseline period though it differs for patients from different sites, treatment groups, or primary diseases. However, if we add post-baseline day as a covariate in the model the corresponding estimated coefficient is significant ($p < 0.001$). Nevertheless, the estimated incident rate ratios estimated from the model including time as a covariate are nominally significant and almost the same as for the primary model (0.510 for AVP20/Pl and 0.532 for AVP30/Pl).

One may question whether the primary analysis model assumption that the incident rate ratios are constant over the 84 day treatment period is true. One way to get an idea about this is shown in the following graph which displays the estimated incident ratios based on each 2 week period. While, it appears that they may be roughly constant after the first couple weeks they certainly do not appear constant over the whole treatment period. However, a model which allowed the

average daily count to be different for each 2 week period (but the same within the 2 week period see Figure 2) suggests a nominally significant treatment difference at day 84.

Figure 2 Biweekly Estimated Incident Rate Ratios



The considerably larger standard errors (about 2.5 times bigger) of the parameter estimates that one obtains for the primary analysis model when the bootstrap (resampling the data with replacement) is used suggests that the model can not be entirely trusted. The true precision of the estimates seems to be less than indicated by the model which means that the model overstates the significance of some parameter estimates (e.g., confidence intervals should be wider than the model indicates). It could be that the longitudinal model's use of a single random effect for the dispersion isn't rich enough to characterize the within patient correlation for as many as 84 postbaseline timepoints per patient. Although an unstructured covariance matrix would be impractical the number of correlation parameters that would be required for an unstructured covariance matrix would be $84 \cdot 83 / 2 = 3486$, so a single random effect may be a serious oversimplification. Actually, because the baseline counts are simultaneously modeled with the post-baseline, here there would be even more parameters.

A jackknife estimate of the standard error of the parameter estimates also yielded a higher standard error for the parameter estimates. The idea of the jackknife approach is to re-run the analysis N times where N is the number of patients and patient i is excluded from the i th analysis. This seeing how the parameter estimates change without a patient's data can provide insight into whether one patient has a big impact on the result based on the full data set. This suggests that the standard errors based on applying the analysis to the full data set may be a bit

too low which translates into exaggeratedly small p-values. However, even with the bootstrap or jackknife based standard errors for the parameter estimates the Zenvia dose groups were still nominally significant compared to placebo.

The GEE model, designated by the sponsor as a sensitivity analysis, has some important different assumptions than the primary longitudinal random effects negative binomial model. In particular, while the primary model assumes a random dispersion parameter for each patient the GEE model assumes a common dispersion parameter for patients with the same baseline values of covariates in the model. The GEE model also assumes that any two daily counts from the same patient have the same correlation. The primary model made no such direct assumption on the correlation but the random dispersion translates into a similar assumption. The GEE model also only assumes the underlying distribution is negative binomial up to the first two distribution moments (mean and variance) whereas the primary model assumes the distribution is exactly negative binomial (all moments not just the mean and variance).

The sponsor's reported GEE analysis grouped treated sites as either U.S. or non-U.S. in contrast to the primary analysis where there were effects for individual sites in the model. This reviewer found that if the latter is done for the GEE method then the low dose also (note: AVP30/placebo IRR=.419, $p<.0001$) is nominally significant compared to placebo IRR=.546, $p<0.0001$.

3.1.1.5.5 Exploratory Analysis of Episodes by Episode Type

The primary analysis treats laughing and crying items the same by just adding them together but they may not be interchangeable with respect to the drug effect. In fact, while the sum of all laughing and crying post-baseline episodes was significant in favor of AVP30/30 in study 102 an exploratory analysis of only laughing episodes was not (but crying only was nominally significant). The same pattern was true for the sum of CNSLS items of a specific type (e.g., laughing). Thus, there is a lingering question of whether there is an effect on laughing episodes only, at least in ALS patients. While there may be less power to detect an effect on laughing only episodes since there are fewer episodes, it still seems like an important question to investigate, especially given the observed pattern of laughing specific results in study 102.

Incident rate ratios based on the primary Longitudinal model restricted to only laughing episodes in study 123 were .720 ($p<0.001$) for AVP20/Placebo and .745 ($p<0.001$) for AVP30/Placebo. However, these estimates were based on the model assuming no treatment group differences during baseline and there appeared to be differences between the treatment groups in baseline laughing rates. Furthermore, the primary analysis called for checking this and including baseline adjustments if they were significant. When this is done IRRs are .784 ($p=0.002$) for AVP20/Placebo and 1.002 ($p=.984$) for AVP30/Placebo. Note that this AVP30/Placebo IRR numerically favors placebo. Longitudinal analysis was again in the wrong direction for AVP30 when the model adjusted for baseline differences by way of a covariate and only analyzed post-baseline daily laughing episodes. IRRs under these conditions are: .960 ($p=.692$) for AVP20/Placebo and 1.36 ($p=0.021$) for AVP30/Placebo. When we only analyze post baseline episodes with no adjustment for baseline, IRRs become: .648 ($p<0.001$) for AVP20/Placebo and .914 ($p=0.456$) for AVP30/Placebo. Therefore, from these various models, it seems far from clear from the longitudinal analysis that there is a significant treatment effect on laughing only

episodes for AVP30 compared to placebo. Since AVP30 vs. Placebo was the first comparison in the testing hierarchy used as an adjustment for multiple testing, significance for AVP20 vs. Placebo shouldn't be formally claimed without being preceded by significance of AVP30 vs. Placebo. Since results based on the longitudinal model are inconsistent it may be worthwhile to examine the results for the non-longitudinal negative binomial model for post-baseline total laughing episode counts.

Table 14 shows the analyses of laughing episodes only, as well as crying episodes only where the analysis is based on the non-longitudinal negative binomial model for post-baseline total episode counts of the particular type.

These results are consistent with a smaller effect on laughing than crying.

Table 14 Average Weekly Episode Count by Episode Type

Episode Type	Group	N	Mean	Median	Min	Max	Incident Rate Ratio(SE)* AV/Placebo
Laugh	A20	107	12.92447	.6285141	0	390.25	.860(.115) p=.258
	A30	110	3.310825	.5833334	0	30.27711	.800(.117) p=.129
	P	109	7.494465	1.46737	0	130.25	N/A
Cry	A20	98	4.805814	1.257028	0	127.5	.67 p=0.002
	A30	106	3.434867	.7916666	0	38.54217	.60 p<0.001
	P	105	7.656668	3.278481	0	60.16666	N/A

*Incident rate ratio based on negative binomial regression model adjusted for baseline als/ms, and site

Table 15 shows a summary and analysis of laughing only episodes by underlying disease. While the incident rate ratio of AVP30/Placebo was estimated as slightly smaller in ALS patients than in MS patients a test for interaction between disease and treatment was not significant, thus suggesting that any observed differences between the disease specific IRRs may be due to chance alone. The treatment difference estimates from the analysis of change from baseline in the laughing items of the CNSLS (not shown here) showed a very similar pattern to that for the laughing episode counts but again a test for interaction between underlying primary disease and treatment group was not significant (p=0.45).

Table 15 Summary of Average Weekly Laughing Episodes by Underlying Disease

Underlying Disease	Statistic	Group		
		AVP20	AVP30	Placebo
MS	N	39	45	45
	Mean	2.78	2.24	3.82
	Median	0.40	0.21	1.20
	Min	0.00	0.00	0.00
	Max	26.83	30.28	38.80
	IRR*	.853 (p=.529)	.859 (p=.530)	N/A
ALS	N	68	65	64
	Mean	19.57	4.06	10.09
	Median	0.96	0.67	2.27
	Min	0.00	0.00	0.00
	Max	390.25	24.67	130.25
	IRR*	.858 (p=.332)	.763(p=.145)	N/A

*Incident rate ratio based on negative binomial regression model adjusted for baseline, treatment, and site

The secondary efficacy measure CNS-LS, which was primary in the previous studies, is a sum of 7 items each scored from 1 to 5. Three of the CNS-LS items relate to crying and four relate to laughing. If we conduct an exploratory analysis of the sum of the laughing items of the CNS-LS the treatment group difference is not nominally significant for AVP30 compared to placebo. The overall baseline mean for the sum of these items was 10.44. The least squares mean laughing total is 6.62 (S.E.=.336) for AVP30 as compared to 7.38 (S.E.=.337) for Placebo. The p-value for the comparison is p=0.0609. For AVP20 the LS Mean is 7.25 (S.E.=.350) which has a p-value of .7486 for the comparison to placebo. For the total CNS-LS the sponsor prespecified analyzing the area under the curve over time (AUC) of the change from baseline rather than change from baseline to the last visit. The least squares mean of the area under the curve for the double blind period of the laughing total is 7.07 (S.E.=.266) for AVP30 as compared to 7.60 (S.E.=.266) for Placebo. The p-value for this comparison is p=0.096. For AVP20 the LS Mean is 7.61 (S.E.=.277) with p=0.9813 for the comparison to placebo. For crying items only both AVP groups' differences in AUC from placebo were nominally significant (-1.68 for AVP20, p<0.0001 and -1.58 AVP30, p<0.0001). Analysis of change from baseline to the last visit for crying items of the CNS-LS was similar.

These results are consistent with a smaller effect of the drug on laughing than crying.

3.2 Evaluation of Safety

General Safety was not reviewed here; please see the medical review.

Of special note there were 7 Deaths in study 123 all of which occurred in ALS patients (1/64 in placebo, 3/68 in AVP20 mg and 3/65 in AVP 30 mg). A Fisher's exact test comparing the combined drug groups to placebo concludes there is not enough evidence, one-sided p=0.275, to reject the null hypothesis that the probability of death is the same among these two groups. This

test was conducted post-hoc as a quick and simple way to assess this unexpected death imbalance, given the relatively low overall death rate in the trial.

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

4.1 Gender, Race and Age

Subjects were 25 to 80 years of age with a mean age of approximately 51 years. About 45.7% of subjects were Male. Approximately 74% were Caucasian, 4% were Black, and 1% were Asian. Nineteen percent (19%) of all subjects were of Hispanic origin. Sixty percent (60%) of subjects had underlying ALS and 40% of subjects had underlying MS.

4.1.1 Gender

About 45.7% of patients were Male. Estimated longitudinal negative binomial model based incidence rate ratios (drug/placebo) were smaller in the male subgroup (.43, .42 for 20, 30 over placebo) than in the female subgroup (.62, .64 for 20, 30 over placebo) suggesting greater reductions of events compared to placebo in males, but reductions in the female subgroup were still numerically favoring Avanir.

4.1.2 Race

Estimated Incident Rate Ratios of AVP to placebo based on the primary longitudinal negative binomial model supplemented with treatment by race interaction effects were .453 for AVP20/Placebo and .493 for AVP30/Placebo in Caucasians (N=241); .825 for AVP20/Placebo and .663 for AVP30/Placebo in Hispanics (N=63); and .595 for AVP20/Placebo and .543 for AVP30/Placebo in Others (N=20). Thus, there was no compelling evidence of a differential effect of the treatment by race.

4.1.3 Age

Incident rate ratios of AVP to placebo are shown in Table 16 by age group. Although they are noticeably variable all were nominally significant according to the longitudinal negative binomial model.

Table 16 Incident (Episode) rate ratios by Age Group

Age Group	Comparison Groups	Incident Rate Ratio
<45	AVP20/Pla	.595
<45	AVP30/Pla	.856
45-51	AVP20/Pla	.555
45-51	AVP30/Pla	.387
52-59	AVP20/Pla	.740
52-59	AVP30/Pla	.373
>=60	AVP20/Pla	.263
>=60	AVP30/Pla	.626

Estimated IRRs of AVP to placebo were .166 for AVP20 and .782 for AVP30 in those 65 and up; they were .556 for AVP20 and .495 for AVP30 in those < 65 years of age. All of these IRRs were nominally significant according to the longitudinal negative binomial model. Thus, there was no compelling evidence of a differential effect of the treatment by age.

4.2 Other Special/Subgroup Populations

4.2.1 Underlying Primary Disease Diagnosis

Sixty percent (60%) of subjects had underlying ALS and 40% of subjects had underlying MS. Overall, primary longitudinal analysis based incident rate ratios of AVP to placebo were .510 for AVP20/Placebo and .531 for AVP30/Placebo ($p<.0001$ for both). The primary model allows for different episode rates by underlying primary disease diagnosis but assumes the treatment difference is the same regardless of underlying primary disease diagnosis. If we modify the model to permit the treatment difference to vary with underlying primary disease diagnosis we find the following.

Estimated longitudinal model based incident rate ratios (drug/placebo) were smaller in the ALS subgroup (.419, .461 for 20, 30 both $p<.001$) than in the MS subgroup (.741, .655 for 20, 30 both $p<.001$) suggesting greater reductions of events compared to placebo in ALS, but reductions in the MS subgroup were still numerically favoring Avanir. These were obtained by adding interactions between ALS and postbaseline treatment group to the primary model.

Based on the non-longitudinal model adjusted for baseline sum and sites estimated IRRs in the MS subgroup were: .828 (AVP20/placebo, $p=.279$) and .829 (AVP30/Placebo, $p=.267$).

In the ALS subgroup these were .768 ($p=0.035$) and .485 ($p<0.001$) for AVP20 and AVP30 over placebo, respectively.

When the longitudinal model was run in ALS patients only (i.e., excluding data from MS patients) the AVP20/placebo estimate was .354 ($p<0.001$) and AVP30/Placebo was 0.375 ($p<0.001$). When the longitudinal model was run in MS patients only the AVP30/placebo estimate (.888, $p=.041$) was reasonably similar to that reported above for the model of all data incorporating interactions for treatment by disease. However, the estimate of the incidence rate ratio of AVP20/Placebo favored placebo numerically (1.010, $p=.868$). Therefore, the subgroup result for AVP20 vs. Placebo in MS patients bears further investigation. Table 17 shows the

average daily laughing+crying episode counts in the MS subgroup. These summary statistics do not suggest that AVP20 is numerically worse than placebo in the MS subgroup.

Table 17 Average Daily Laughing Crying Counts in MS Subgroup

Period	Group	N	Mean	Median	Min	Max
Baseline	AVP30	45	3.93	3.43	0.00	12.86
	AVP20	39	3.16	2.83	0.14	8.29
	Placebo	45	3.29	2.36	0.00	13.71
Post - Baseline	AVP30	45	1.05	0.21	0.01	9.83
	AVP20	39	0.88	0.21	0.00	6.35
	Placebo	45	1.41	0.67	0.00	7.44

It is striking to note that among the AVP 20 mg and placebo group patients 61% of post baseline laughing+crying episode counts were 0. It is possible that the longitudinal negative binomial model is not fitting well in MS patients because there are more zero counts in the data than a negative binomial model can accommodate. In fact in the statistical literature there are zero inflated negative binomial models which may be more appropriate for this situation. In such a situation the counts are assumed to be a mixture of two distributions: a point mass at 0 and a negative binomial. A zero inflated negative binomial model with adjustment for within patient correlation suggested nominal significance for AVP20 relative to placebo (Rate Ratio AVP20/Placebo=0.65). The Poisson distribution is another commonly used distribution for count data. The random effect Poisson analogue to the primary longitudinal negative binomial random effect model suggests that the IRR of A20 to placebo is 0.624. Also, a nonparametric Wilcoxon ranksum test on the post-baseline sums comparing AVP20 to placebo in MS patients gives a p-value of 0.0547 with AVP20 ranksums smaller than expected, thus suggesting a non-significant but numerically lower event rate for AVP20.

Aside from modeling we can look at simple summaries of the data as in Table 17. The mean over patients of the patients' mean daily post-baseline episode counts was 1.05 for AVP30, 0.88 for AVP20 and 1.41 for placebo. The mean over patients of the patient's maximum daily post-baseline episode count was 4.7 for AVP30, 3.38 for AVP20 and 8.73 for placebo.

The mean over patients of the median daily patient post-baseline episode count was 0.74 for AVP30, 0.67 for AVP20 and 0.97 for placebo. These numbers don't seem to agree with the longitudinal negative binomial model estimate of a numerically higher incident rate for AVP20 compared to placebo in MS patients.

Table 18 shows the frequency of episode counts based on all post-baseline daily records (pooling over subjects) in MS subjects. The high proportion of zero counts is notable.

Table 18 All Post Baseline laughing+crying daily episode counts in MS patients

Episodes	A30		A20		Placebo		Total
	N	%	N	%	N	%	
0	2,244	.6682549	1,914	.658864	1,930	.5732106	6,088
1	386	.1149494	358	.1232358	491	.1458271	1,235
2	266	.0792138	210	.0722892	338	.1003861	814
3	157	.046754	167	.0574871	198	.0588061	522
4	83	.0247171	101	.0347676	146	.043362	330
5	54	.016081	56	.0192771	90	.02673	200
6	40	.0119119	35	.0120482	52	.015444	127
7	13	.0038714	37	.0127367	28	.008316	78
8	18	.0053603	19	.0065404	23	.006831	60
9	4	.0011912	5	.0017212	14	.004158	23
10	83	.0247171	3	.0010327	23	.006831	109
11	4	.0011912	0	0	9	.002673	13
12	3	.0008934	0	0	8	.002376	11
13	1	.0002978	0	0	4	.001188	5
14	1	.0002978	0	0	0	0	1
15	1	.0002978	0	0	1	.000297	2
16	0	0	0	0	1	.000297	1
20	0	0	0	0	4	.001188	4
21	0	0	0	0	1	.000297	1
45	0	0	0	0	1	.000297	1
100	0	0	0	0	3	.000891	3
110	0	0	0	0	2	.000594	2
Total	3,358		2,905		3,367		9,630

Also, as shown in Table 19 other reasonable models suggest that the AVP20/Placebo Incident Rate Ratio at least numerically favors AVP20. Therefore, the Longitudinal Negative Binomial Model seems to be alone in suggesting that the AVP20/Placebo incident rate ratio favored placebo numerically in MS patients and so we can downplay that result on this basis, as well as the fact that the study was not powered to detect an effect in the MS subgroup.

Table 19 Incident Rate Ratios of AVP20/Placebo in MS patients based on Various Models

Model	Estimated Rate Ratio AVP20/Placebo	p-value
Longitudinal Negative Binomial	1.01	.868
Longitudinal Zero Inflated Negative Binomial	.624	.039
Longitudinal Poisson	.624	.001
GEE Negative Binomial	.559	.034
Non-Longitudinal Negative Binomial	.828	.301

4.2.2 Individual Sites

In data from US sites only IRRs were .508 ($p < 0.001$) for 20/placebo and .447 ($p < 0.001$) for 30/placebo based on the primary model. These were .801 ($p = 0.077$) and .529 ($p < 0.001$) based on the non-longitudinal model. In non-US sites only they were .467 ($p < 0.001$) and .614 ($p < 0.001$), respectively, based on the longitudinal model. According to the non-longitudinal model these were .668 ($p = .025$) and .819 ($p = 0.219$). In summary, AVP30 was slightly less impressive compared to placebo in non-US sites than in US sites, but it was still nominally significant in both.

Figure 3 shows site specific model based log incident rate ratios (AVP30/Placebo). Negative values favor AVP30. The size of the circle is proportional to the size of the corresponding site.

Figure 3 Site Specific Treatment Effect Estimates for AVP30 vs. Placebo

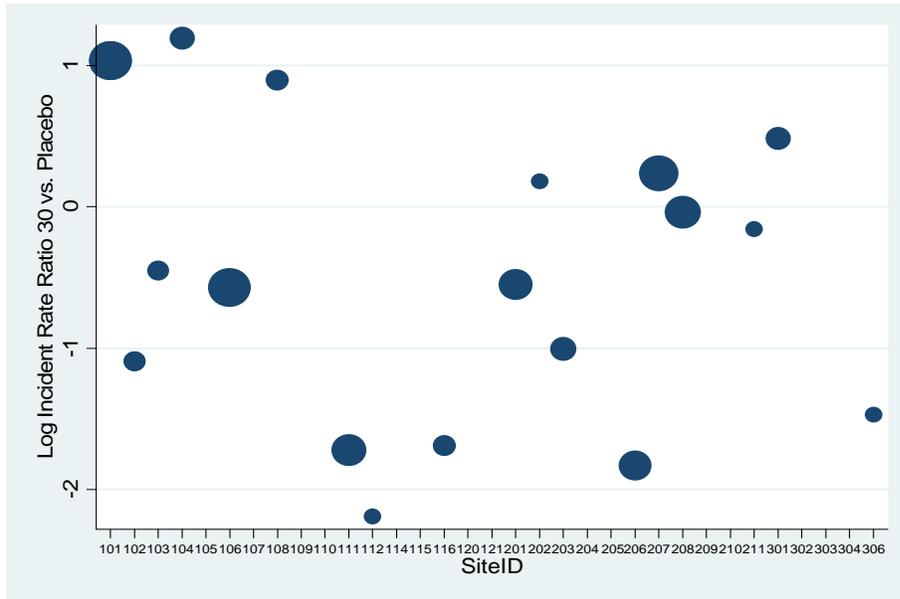
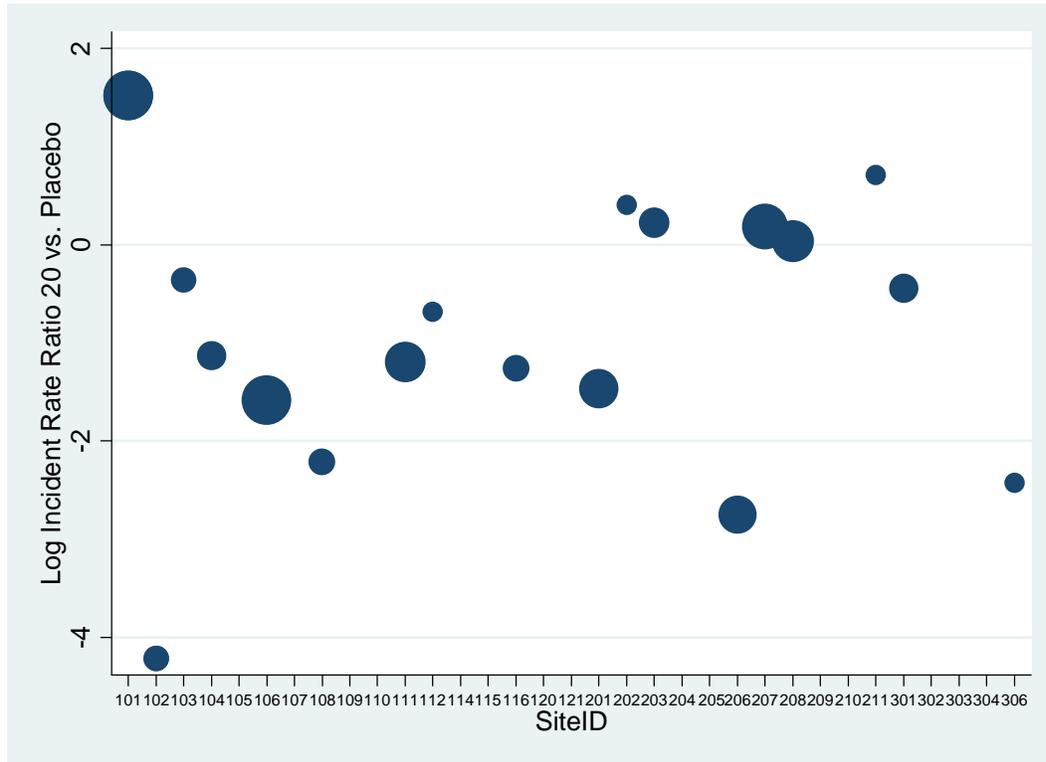


Figure 4 shows site specific model based log incident rate ratios (AVP20/Placebo). Negative values favor AVP20.

Figure 4 Site Specific Treatment Effect Estimates for AVP20 vs. Placebo



Most of the site specific estimated incident rate ratios of drug over placebo favor the drug. In addition, the overall results for AVP20 and AVP30 compared to placebo were not sensitive to the exclusion of data from any one site.

5 SUMMARY AND CONCLUSIONS

5.1 Statistical Issues and Collective Evidence

5.1.1 Statistical Issues

The primary longitudinal negative binomial model found the ratio of laughing plus crying episode rates for AVP20 over placebo to be slightly better numerically than the episode rate ratio of AVP30 over placebo (both statistically significant compared to placebo). The supportive non-longitudinal negative model analysis of the total post-baseline sums of laughing plus crying episode counts suggests the opposite ordering of AVP20 and AVP30. However, the suggested ordering of AVP30 and AVP20 obtained from this analysis seems to be sensitive to some extreme outlier counts in the AVP20 group, many of which came from one particular site. The two groups' results from this model are very similar if data from this site is excluded (see section

3.1.1.5.1). In addition, several other models as well as the simple group medians of the changes from baseline in episode rate suggest that there is little difference between AVP20 and AVP30 (but both are nominally significant compared to placebo).

The standard errors of treatment effect estimates are smaller for the primary longitudinal random effects negative binomial model than for the non-longitudinal negative binomial model that was used to model episode counts in the prior two studies. For the longitudinal model each daily count for a subject is an observation of the dependent variable, whereas for the non-longitudinal model the subject's sum of the counts over all post-baseline days is the sole observation of the dependent variable. Methods to estimate the standard errors of the parameter estimates based on re-sampling the data and re-running the model on the resulting data over and over suggest that the longitudinal model underestimates the standard errors by as much as a factor of 2. This underestimation of the standard error suggests that actual p-values should be larger than reported. It may be related to the primary longitudinal model's potential oversimplification of the within patient correlation (among the patients' set of 84 postbaseline daily episode counts). The model incorporates a single random effect parameter to address this correlation, whereas, for example, a typical mixed model for repeated measures analysis with an unstructured correlation matrix would require $84 \times 85 / 2 = 3,570$ parameters, though this extreme number would probably not be practical. At any rate, the underestimation of standard errors appears to not be so great as to alter the statistical significance of the comparisons of AVP20 and AVP30 with placebo.

There is some evidence that there may be less of a treatment effect on laughing than crying or possibly even no effect on laughing but it should be acknowledged that the study was only powered for the combination of laughing and crying. Nevertheless, this observation is supported by independent analyses of episode counts and the sum of the 4 laughing items of the 7 item CNSLS endpoint in two of the three studies.

There was a slight imbalance between the placebo and the drug groups in deaths in study 123. There were 7 Deaths in study 123 all of which occurred in ALS patients (1/64 in placebo, 3/68 in AVP20 mg and 3/65 in AVP 30 mg). A Fisher's exact test comparing the combined drug groups to placebo concludes there is not enough evidence, one-sided $p=0.275$, to reject the null hypothesis that the probability of death is the same among these two groups. This test was conducted post-hoc as a quick and simple way to assess this unexpected death imbalance, given the relatively low overall death rate in the trial.

5.1.2 Collective Evidence

Table 20 shows the average number of Laughing plus Crying Episodes per Week for the various clinical efficacy trials of the Dextromethorpan/Quinidine combination in its various forms. The estimated incident ratios of AVP to non-AVP group (DM, Q, or Placebo where applicable) based on a non-longitudinal negative binomial model (with constant dispersion) for the post-baseline total laughing plus crying episode counts are shown. For the sake of comparison with the earlier studies which were in one specific underlying disease, study 123 is shown by underlying disease subgroup, as well as overall. There was no placebo group in study 102 but rather the AVP30/30 combination was compared to each of its two components administered alone. Although in study 123 based on the non-longitudinal model AVP30/10 appears to have a numerically lower

event rate than AVP20/10 overall, the opposite was true for the longitudinal (daily count) model (e.g., when baseline data is excluded because the other studies had no baseline period IRRs are: .425 AVP20/Placebo, $p < .001$; .586 AVP30/Placebo, $p < .001$). Therefore, overall, it appears likely that the combination of Dextromethorphan and Quinidine has activity in patients with pseudobulbar affect. Based on study 123 AVP 30/10 may not provide any additional efficacy benefit beyond that of AVP 20/10.

Table 20 Number of Episodes per Week During Treatment by Study

	Group	N	Mean	Median	Range	Incident Rate Ratio* (AVP/DM or Q)
Study 102 (ALS)	AVP (30/30)	67	9.4	2.5	0-116	N/A
	DM 30 mg	33	34.4	4.8	0-727	.650 ($p = .050$)
	Q 30 mg	37	13.0	6.3	0-49	.549 ($p = .003$)
Study 106 (MS)	AVP-923(30/30)	75	4.7	1.3	0-80.0	.536 ($p < 0.001$)
	Placebo	73	11.5	19.43	0-129.8	N/A
Study 123 Overall	AVP(30/10)	109	6.68	2.15	0.00-68.82	.639 ($p < .001$) [Basel. Adj.: 0.591, $p < .001$]
	AVP(20/10)	106	17.07	2.69	0.00-405.25	.772 ($p = .036$) [Basel. Adj.: 0.802 $p = .035$]
	Placebo	109	14.55	6.00	0.00-131.25	N/A
Study 123 (MS subset)	AVP(30/10)	45	7.32	1.50	0.08-68.82	.831*
	AVP(20/10)	39	6.14	1.50	0.00-44.42	.755*
	Placebo	45	9.89	4.68	0.00-52.08	N/A
Study 123 (ALS subset)	AVP(30/10)	64	6.25	2.33	0.00-25.16	.536*
	AVP(20/10)	67	24.05	3.71	0.00-405.25	.765*
	Placebo	64	17.80	8.47	0.00-131.25	N/A

*based on Negative binomial regression model with constant dispersion adjusted for sites and treatment. The IRRs are unadjusted for baseline to be consistent with analysis of earlier studies

Table 21 summarizes weekly average episode rates by study and specific episode type. At least in ALS patients effects of AVP seemed to be numerically bigger for crying type episodes than for laughing type episodes. This analysis was motivated by the observed trend in study 102 of a numerically smaller effect on Laughing than on Crying. However, it must be noted that the study was not powered to differentiate laughing episode specific treatment effects from crying episode specific treatment effects.

Table 21 Weekly Average Episode Rate by Study and Episode Type

		Group	N	Mean	Median	Min	Max	IRR(AVP/DM or Q or Placebo)*
102	Laughing	AVP30/30	67	6.75	1.21	0	116.67	N/A
		DM 30	33	30.44	1.5	0	726.55	.757(p=.276)
		Q 30	37	6.37	1.69	0	45	.751 (p=.244)
	Crying	AVP30/30	67	2.64	.25	0	66	N/A
		DM 30	33	3.96	0.70	0	21	.532 (p=.011)
		Q 30	37	6.63	4.10	0	30.75	.277(p<.001)
106	Laughing	AVP30/30	76	2.52	0.09	0	64.94	.511 (p=0.002)
		Placebo	74	4.78	0.75	0	105.75	N/A
	Crying	AVP30/30	76	2.20	0.57	0	34.00	.521(p=0.001)
		Placebo	74	6.70	2.83	0	51.57	N/A
123 MS	Laughing	AVP20/10	39	2.78	0.40	0	26.83	.866 (p=.582)
		AVP30/10	45	2.24	0.21	0	30.28	.887 (p=.630)
		Placebo	45	3.82	1.2	0	38.8	N/A
	Crying	AVP20/10	39	3.42	0.83	0	31.42	.808 (p=.290)
		AVP30/10	43	5.02	1.18	0	38.54	.950 (p=.789)
		Placebo	43	6.43	2.92	0	44.28	N/A
123 ALS	Laughing	AVP20/10	68	19.57	0.96	0	390.25	.809 (p=.296)
		AVP30/10	65	4.06	0.67	0	24.67	.627(p=.018) #
		Placebo	64	10.09	2.27	0	130.25	N/A
	Crying	AVP20/10	59	5.72	1.42	0	127.5	.544 (p=.001)
		AVP30/10	63	2.35	0.6	0	22.4	.467(p<.001)
		Placebo	62	8.51	3.6	0	60.17	N/A

*Study 123 estimated IRRs are based on non-longitudinal negative binomial model and not adjusted for baseline in order to be consistent with earlier studies

baseline adjusted estimate of IRR is .764 (p=.148)

5.2 Conclusions and Recommendations

The efficacy data from trial 123 suggests that both the 30 mg Dextromethorphan / Quinidine 10 mg combination as well as the DM 20 mg / Q 10 mg combination were superior to placebo in controlling the number of inappropriate laughing plus crying episodes associated with pseudobulbar affect in the mixed study population of amyotrophic lateral sclerosis (ALS) and multiple sclerosis (MS) patients. It was previously concluded from the original NDA submission that the 30 mg Dextromethorphan / 30 mg Quinidine combination was superior to placebo in study 106 conducted in MS patients with pseudobulbar affect and superior to the two components in study 102 conducted in ALS patients with pseudobulbar affect. The primary endpoint for the earlier trials was the change from baseline in the CNS-LS score averaged over the treatment period. The differences from placebo in terms of the CNS-LS were also nominally significant in trial 123. The primary model suggests that there may be no additional benefit of 30/10 over that of 20/10 compared to placebo. Although a prespecified secondary analysis suggests a possible additional benefit of 30/10 this is not judged very persuasive by this reviewer as it seems to be sensitive to outliers and also is not supported by the simple median changes from baseline in episode rates (medians are more robust to outliers).

This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature.

/s/

TRISTAN S MASSIE
10/14/2010

KUN JIN
10/14/2010
I concur with the review.

HSIEN MING J J HUNG
10/15/2010



U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Pharmacoepidemiology and Statistical Science
Office of Biostatistics

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/Serial Number: 21, 879

Drug Name: Neurodex (Dextromethorphan plus Quinidine)

Indication(s): Pseudobulbar Affect

Applicant: Avanir

Date(s): Submission Date: Jan 30, 2006
3 month extension for QT study

Review Priority: Priority

Biometrics Division: Division of Biometrics I

Statistical Reviewer: Tristan Massie, Ph.D.

Concurring Reviewers: Kun Jin, Ph.D., Team Leader
Jim Hung, Ph.D., Director, Division of Biometrics 1

Medical Division: Neurologic Drugs

Clinical Team: Ron Farkas, M.D., Ph.D.
Wilson Bryan, M.D., Team Leader

Project Manager: Melina Griffis

Keywords: Dropouts, Combination Drug, Poisson, Negative binomial

Table of Contents

LIST OF TABLES.....	3
LIST OF FIGURES.....	4
1 EXECUTIVE SUMMARY	5
1.1 CONCLUSIONS AND RECOMMENDATIONS	5
1.2 BRIEF OVERVIEW OF CLINICAL STUDIES	5
1.3 STATISTICAL ISSUES AND FINDINGS	6
2 INTRODUCTION	9
2.1 OVERVIEW.....	9
2.2 DATA SOURCES	11
3 STATISTICAL EVALUATION	12
3.1 EVALUATION OF EFFICACY	12
3.1.1 <i>Study AVR-102</i>	12
3.1.1.1 Study Design.....	12
3.1.1.2 Efficacy Measures.....	13
3.1.1.3 Statistical Methods and Sample size	13
3.1.1.4 Disposition of Patients	15
3.1.1.5 Patient Demographics	15
3.1.1.6 Sponsor’s Results.....	17
3.1.1.7 Reviewer’s Results.....	19
3.1.1.7.1 Primary Analysis.....	19
3.1.1.7.2 Sensitivity of Primary Analysis Result to Missing Data	21
3.1.2 <i>Study AVR-106</i>	36
3.1.2.1 Study Design.....	36
3.1.2.2 Patient Disposition	38
3.1.2.3 Patient Demographics	38
3.1.2.4 Sponsor’s Results.....	40
3.1.2.4.1 Primary Analysis.....	40
3.1.2.4.2 Secondary Analyses	41
3.1.2.5 Reviewer’s Results.....	42
3.1.2.5.1 Primary Analysis.....	42
3.1.2.5.2 Assessment of Sensitivity to Dropouts.....	43
3.1.2.5.3 Secondary Analyses	44
3.2 EVALUATION OF SAFETY	47
4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS	48
4.1 GENDER, RACE AND AGE	48
4.2 OTHER SPECIAL/SUBGROUP POPULATIONS	50
5 SUMMARY AND CONCLUSIONS	51
5.1 STATISTICAL ISSUES AND COLLECTIVE EVIDENCE	51
5.2 CONCLUSIONS AND RECOMMENDATIONS	54
I. APPENDIX	55

LIST OF TABLES

Table 1 Study 102: Baseline Demographics and Disease Characteristics-ITT population (reviewer’s results).....	16
Table 2 Study 102: Mean Change in CNS-LS Scores ^a – ITT Population (N=129, Excluding Poor Metabolizers).....	17
Table 3 Study 102: Differences in Mean CNS-LS Scores ^a - ITT Population (N=129, Excluding Poor Metabolizers)	18
Table 4 Study 102: Analysis of Change from baseline in CNSLS score by Population and Timepoint	20
Table 5 Study AVR-102: Sensitivity Analyses for Change from baseline in CNSLS at day 29	21
Table 6 AVR-102: Mean CNSLS Change at Day 29 or Last Assessment by Termination Reason	23
Table 7 Study 102: Analyses of Laughing and Crying items separately Least Square Means	25
Table 8 Study 102: Summary Statistics for Number of Days with Non-missing Episode Diary Data	26
Table 9 Study 102: Historically Reported Episode Rate before Study and Episode Rate during Study.....	26
Table 10 Study 102: Estimated Treatment Effects and Fits from Various Models for Episode Counts	30
Table 11 Study 102: Negative Binomial Model Analyses of Sum of Laughing and Crying Episode Counts with imputation for patients with no post-baseline data	31
Table 12 Study 102: Treatment Effects and Model Fits for only Laughing episode counts.....	32
Table 13 Study 102: Treatment Effects and Model Fits for Various Models of only Crying Episode Counts.....	33
Table 14 Study 102: Analyses of Other Secondary Endpoints	34
Table 15 Study 106: Patient Demographics	39
Table 16 Study 106: Disease Characteristics.....	39
Table 17 Study 106: Baseline scores on Efficacy Measures	40
Table 18 Study 106: Change in Center for Neurological Study-Lability Scale (CNS-LS) Score— ITT Population ..	41
Table 19 Study 106: Number of Episodes of Inappropriate Laughing and/or Crying —ITT Population.....	41
Table 20 Study 106: Primary Analysis and Additional Analyses for Assessing Sensitivity to Dropouts	44
Table 21 Study 106: Mean CNSLS Change at Day 85 or Last Assessment by Termination Reason.....	44
Table 22 Study 106: Analysis of the Number of Episodes of the Laughing or Crying Type	45
Table 23 Study 106: Analyses of Other Secondary Endpoints	46
Table 24 Study 102: Analysis of Last Change in CNSLS by Gender	48
Table 25 Study 106: Analysis of Last Change in CNSLS by Gender (LSMean)	48
Table 26 Study 102: Analysis of Last Change in CNSLS by Age Group (<65 vs. > 65).....	49
Table 27 Study 106: Analysis of Last Change in CNSLS by Age Group (LSMeans).....	49
Table 28 Study 102: Analysis of Last Change in CNSLS by Race	50
Table 29 Study 106: Analysis of Last Change in CNSLS by Race	50

LIST OF FIGURES

Figure 1 Study 102: Patient Disposition.....	15
Figure 2 Study 102: Change in CNS-LS over Time for AVP-923 Non-Completers (excluding 4 AVP-923 patients with no post-baseline data).....	24
Figure 3 Study 102: Comparison of Poisson and Negative Binomial Model Fits to Episode Count Data.....	28
Figure 4 Study 102: Treatment Group Differences on Change from Baseline in CNSLS (Averaged over Time) within Sites.....	35
Figure 5 Study 106: Patient Disposition.....	38
Figure 6 Study 106: Change from Baseline in CNSLS scores by Visit Week.....	43
Figure 7 Study 106: Treatment Group Differences in Change in CNSLS (Averaged over Time) by Site.....	47

1 EXECUTIVE SUMMARY

1.1 Conclusions and Recommendations

Efficacy data on pseudobulbar affect from a study in Multiple Sclerosis (MS) patients showed that the AVP-923 combination of 30 mg Dextromethorphan and 30 mg Quinidine was significantly better than placebo in treating pseudobulbar affect in the study. An earlier study conducted in Amyotrophic Lateral Sclerosis (ALS) patients with pseudobulbar affect compared the same combination of Dextromethorphan and Quinidine to the individual components of the combination. By design this study had a shorter follow-up (1 month) than what is normally expected in ALS patients and the company did not follow the division's advice to lengthen the follow-up. In addition, while the combination was significantly better than the components on the primary efficacy measure, change from baseline in Center for Neurologic Study-Lability Scale (CNS-LS) score, it was not clearly significantly better in terms of the analysis of the laughing and crying episode counts which the agency had encouraged the company to use as the primary efficacy measure. The sponsor's statistician correctly reported that an assumption underlying the sponsor's prespecified method for the analysis of the episode counts (sponsor designated secondary endpoint) was not supported by the study data and that it is well known that ignoring this fact would lead to p-values that are misleadingly small. No back-up analysis method was specified in the protocol. Several reasonable alternatives to the prespecified method failed to find a significant difference while one other method advocated by the sponsor did. There are no precedents for primary endpoints in pseudobulbar affect because it is a new indication. If one deems the sponsor's pre-specified primary endpoint as a valid endpoint for the indication then the ALS study suggests that the combination is superior, in terms of efficacy, to each of its individual components for pseudobulbar affect in ALS patients after up to one month of treatment. However, the p-value of 0.001 for the primary analysis seems to be optimistic since it excludes 4 patients with no post-baseline efficacy measures all of whom were in the combination group and some sensitivity analyses including these patients result in p-values greater than 0.05 (see section 1.3 for details). Therefore, while the study is considered positive it may not have the strength and robustness one would expect in the case where there is only one study comparing the combination to each of its components. The placebo controlled study in MS patients with pseudobulbar affect lends some support to the efficacy of the drug combination but only relative to placebo, i.e., not relative to the individual components of the combination because they were not included in the design.

1.2 Brief Overview of Clinical Studies

Avanir performed two pivotal randomized, double-blind, controlled, multicenter Phase 3 efficacy studies of the effect of AVP-923 on Pseudobulbar affect (PBA) in two different patient populations. A one-month study (Study 99-AVR-102) that compared AVP-923 (N=70) to each of its components (Dextromethorphan (DM) (N=33) and Quinidine (Q) (N=37)) was completed in 140 patients with Amyotrophic Lateral Sclerosis (ALS), and a 3-month study (Study 02-AVR-106) that compared AVP-923 (N=73) to placebo (N=74) was completed in 147 patients with Multiple Sclerosis (MS).

There were 17 investigators in the ALS study, 99-AVR-102, and all of them were located in the U.S. Patient's ages ranged between 33 and 72 and the mean age was about 55 years. Nearly 90 percent of the patient population was white and about 61 % was male.

In the MS study, AVR-106, there were 18 U.S. investigators and 4 Israeli investigators. Ages ranged between 21 and 71 and the mean age was 45 years. Nearly 91 percent of the patient population was white and about 17 % was male.

Early in development there was a small crossover study, CNS-93, in 12 subjects. This study used 75 mg Q in the combination instead of the 30 mg used in the later trials. The meeting minutes for the pre-NDA meeting held on 5/17/2004 addressed this study as follows. "The non-IND study CNS-93 may be of limited relevance since it used a quinidine dose of 75 mg instead of the proposed clinical dose of 30 mg".

1.3 Statistical Issues and Findings

In study 102 the AVP-923 group had 8 (11%) patients drop out due to toxicity before the week 2 assessment, whereas the DM and Q groups had no dropouts before week 2. Four of these 8 AVP-923 patients had an early post-baseline assessment and 4 did not. The latter 4 patients were the only patients who did not have any post-baseline CNSLS measures but all of them were members of the AVP-923 group (4/70=5.7%). Six other AVP-923 patients dropped out due to toxicity before the week 4 assessment and one died due to ALS complications, according to the sponsor. Note that 4 other AVP-923 patients and 1 Quinidine patient were not considered to be completers by the sponsor, despite having CNSLS assessments at or near days 15 and 29, because they refused to take the medication due to toxicity.

Average baseline scores on the primary efficacy measure, the CNSLS, were 20 for the AVP-923 group, 21 for DM, and 22 for Q (possible range is 0 to 28). Both single component groups had slightly worse scores at baseline and the AVP-923 vs. Q comparison of the baseline CNSLS scores approached nominal significance ($p=0.065$). A similar trend was observed for the Visual Analog Scale quality of life (VAS QOL) and quality of relationships (VAS QOR) ratings at baseline. The global test for any differences among the three VAS QOL means was nominally significant ($p=0.024$). The Q group was 12.2 points higher than the AVP-923 group on the quality of life VAS ($p=0.011$) and 11 points higher on the quality of relationships VAS ($p=0.039$). The Q group also had a higher percentage of patients with the bulbar (as opposed to limbic) type of ALS than the AVP-923 group (62% vs. 43% $p=0.057$). In the presence of baseline differences on variables associated with an efficacy measure the reported treatment group differences on that measure may not be due to the treatment alone.

Based on the primary analysis which was a site, treatment group, and baseline adjusted ANCOVA of the difference between the baseline and the average of the day 15 and 29 CNSLS scores the comparison between the AVP-923 group and the DM group is significant ($p=0.001$) as is the AVP-923 vs. Q group comparison ($p<0.0001$). The primary analysis utilized the last observation carried forward for those patients with only one post-baseline efficacy assessment.

A mixed model analysis of repeated measures using all observed post-baseline CNSLS data and an analysis restricted to the completers population supported the primary analysis results.

Carrying baseline forward is usually discouraged as a method for imputing missing data in the division of Neurologic drugs because it can lead to underestimating the variance of the group difference and thus to a biased test. In study 102 there were 4 patients in the combination group with no post-baseline primary efficacy measures as compared to 0 in the other groups. Usually one focuses on the ITT population modified to exclude these patients as long as they are few in number and not all in one group. Since they are all in one group and the sample sizes are small in this case it is important to assess their potential impact on the results and carrying their baseline scores forward is one way to accomplish this. In study 102 if we impute no change, i.e., carry the baseline forward, for those who were last assessed on the CNSLS before day 23, i.e., more than a week before the intended final assessment time, and for those who had no post-baseline CNSLS assessments (4 patients - all DM/Q) the p-value for the DM/Q vs. DM comparison increases to 0.083 and that for the AVP-923 vs. Q comparison increases to 0.005. Therefore, the significance of the primary analysis result may be affected by changing assumptions regarding the dropouts. For the sake of completeness, if we focus on the usual MITT population where these 4 AVP-923 patients are excluded then the 0.083 p-value for the DM comparison reduces to 0.042. Instead of carrying the baseline forward we could use the more traditional approach of carrying the last observation forward for dropouts with some post-baseline CNSLS scores and examine the effect of a worst case like imputation for the 4 patients with no post-baseline CNSLS scores. In particular, if we impute a change from baseline of +5 for the 4 AVP-923 dropouts with no post-baseline CNSLS scores, which is one point worse than the worst observed change, then the resulting p-values are 0.056 for the AVP-923 vs. DM comparison and <0.05 for the AVP-923 vs. Q comparison. In this reviewer's opinion considering that it is a p-value from a worst-case analysis the AVP-923 vs. DM p-value of 0.056 may be close enough to 0.05 in this case. Therefore, the primary analysis result in study 102 doesn't seem too sensitive to several reasonable assumptions regarding the missing data.

In study 102 the sponsor excluded patients that were randomized but were poor metabolizers from the primary analysis, as stipulated in the statistical analysis plan. There were 5 (7%) AVP-923 patients, 3 (9%) DM, and 3 (8%) Q patients that were determined to be poor metabolizers of cytochrome P450 2D6. The primary analysis result is not sensitive to the inclusion of these patients.

The results for the analysis of the counts of all episodes of the laughing or crying type based on the sponsor's prespecified analysis of the episodes in study 102 are not robust and there is evidence that the assumptions of the model are not satisfied. The observed distribution of the number of episodes does not fit the Poisson distribution proposed by the sponsor for the analysis of episodes in study 102. The sponsor acknowledged this and prespecified a more appropriate negative binomial model instead of the Poisson model for the analysis of episodes in the following study (106). Numerous alternatives to the Poisson model fail to detect a group difference between AVP-923 and DM in the average number of laughing and crying episodes per

week in study 102. This episode count endpoint was designated as secondary by the sponsor but the division indicated a preference for it being primary in several meetings with the sponsor. There is no established precedent for a primary endpoint because this is a new indication. If the division still held its initial preference for the episode count endpoint over the CNSLS then the interpretation of the study outcome could be quite different.

Although the CNSLS, the primary endpoint, contains both laughing and crying items if one considers only the laughing items of the CNSLS (4 of the 7 CNSLS items) in study 102 then the group differences between Avanir and Quinidine and Avanir and Dextromethorphan are not statistically significant. Also, analyses of the episodes of laughing recorded in patients' diaries fail to detect a significant difference between Avanir and Quinidine or Avanir and Dextromethorphan. Results from the analysis of change from baseline in the sum of the CNSLS crying items and the analysis of crying episode counts were also concordant, but in contrast to the laughing results the crying results reached nominal significance. In the placebo controlled study Avanir was significantly better than placebo on the change in the sum of the CNSLS laughing items and the laughing episodes counts, but since this is a combination drug the placebo-controlled trial result doesn't rule out the possibility that one of the components of the combination is enough for laughing episodes.

The sponsor used a non-parametric O'Brien test in an effort to control the type I error for the secondary endpoints. The O'Brien test combines the patient's ranks on each of the endpoints into a single measure (sum of the patient's ranks on each endpoint) and thus requires only one test. The problem with the O'Brien test is that it doesn't indicate which secondary endpoints are significant, only that some combination or composite of them is. There is also a question of whether or not all of the secondary endpoints provide information that is distinct enough from the primary efficacy measure. Secondary endpoints include number of episodes of crying and number of episodes of laughing, Visual Analog scale score for Quality of Life, and Visual Analog scale score for Quality of Relationships. The Pain intensity rating scale was an additional secondary endpoint in study 106 only. The sponsor reported that the O'Brien test was significant for both studies. However, it doesn't indicate which endpoints are significant so it doesn't really avoid the multiplicity problem. In fact, in study 102 this reviewer found a lack of clear significance between the AVP-923 and DM groups in terms of the sums of the episode counts of the laughing or crying type and in study 106 this reviewer found that the AVP-923 vs. placebo comparison on the change from baseline to the end of the study on the pain intensity rating scale was not nominally significant. So the significance of the secondary endpoints depends on the multiplicity adjustment method and the sponsor did not choose an appropriate one.

In study 106, 74 patients were randomized to AVP-923 and 76 were randomized to placebo. There were 21 dropouts in each group (about 28% for each groups). The average baseline CNSLS score was 21 (the possible range is 7 to 35). For the ITT population, excluding those with no post-baseline CNSLS scores, the difference in group least squares mean changes from baseline in CNSLS score averaged over all available post-baseline visits was estimated to be 4.4 points (+/- .74 S.E., $p < 0.0001$). If the comparison was based on the change from baseline at day 85 (or LOCF), instead of averaging over the entire period as prespecified by the sponsor, the

group difference was slightly smaller but still statistically significant: 3.9 points (+/- .86 S.E., $p < 0.0001$). These results seem to be robust to several reasonable assumptions regarding missing data since analysis of the completers population and a mixed model repeated measures analysis still resulted in nominally significant p-values. Therefore, study 106 seems to support the superiority of AVP-923 to placebo for treating pseudobulbar affect in MS patients.

Although this drug is a combination of two drugs the sponsor has conducted only one study comparing the combination to each of the single components. Ideally, a drug combination should be demonstrated statistically significantly superior to each of its components in two studies.

2 INTRODUCTION

2.1 Overview

AVP-923 is being developed by Avanir for the treatment of pseudobulbar affect (PBA), a condition for which no treatments are currently approved. PBA is an affective disinhibition syndrome characterized by loss of emotional control, typically expressed as episodes of involuntary crying and/or laughing and is associated with neurological disease or injury. AVP-923 is a novel combination drug product comprised of two approved drugs, Dextromethorphan Hydrobromide USP (DM) and Quinidine Sulfate USP (Q). The combination contains 30 mg DM and 30 mg Q. DM is thought to have an effect on pseudobulbar affect but may be metabolized too quickly. The addition of Quinidine is designed to slow the metabolism of DM.

Avanir was advised that, since pseudobulbar affect occurs with several diseases, consideration should be given to investigating the product in at least two different disease populations. Data obtained from randomized, controlled, multicenter clinical trials in ALS patients (Study 99-AVR-102; comparison to the individual components alone) and in MS patients exhibiting PBA (Study 02-AVR-106; comparison to placebo) are presented to support the indication of PBA for AVP-923.

Relevant Meetings and Correspondence

Key points conveyed to the sponsor during a July 21, 1999 teleconference:

- **In determining an appropriate measurement of effectiveness for this product, consideration should be given to counting the number of episodes of loss of emotional control.**

- **Consideration should be given to doing a multifactorial study to explore the effects of the combination product versus the individual components to show that a clinical effect is a direct result of the combination product and not one individual component. Additionally, it would be useful to explore the dose response of dexamethorphan.**
- **Since pseudobulbar affect occurs in disease states other than ALS, consideration should be given to investigating this product in at least two different disease populations.**
- **The proposed duration of a one month trial is considered too short. Ordinarily, we would require studies to be of at least three month in duration.**

According to Avanir's meeting minutes for the teleconference of 8/24/2000:

FDA prefers a more straightforward statistical comparison than described in the protocol,

Presumably, the preceding comment refers to the sponsor's plan to analyze the difference between the baseline CNS-LS score and the average of the post-baseline CNS-LS scores instead of the simpler and more commonly used change from baseline to last visit.

On May 23, 2002 comments sent to the sponsor included the following:

We would continue to encourage you to use episode counts as your primary endpoint as opposed to the selected endpoint (CNS-LS).

According to the meeting minutes for End of Phase 2 meeting held on 8/15/2002:

Regarding the primary endpoint CNS-LS:

- **The sponsor was informed that it appears acceptable for the proposed clinical trial in MS.**

In summary, the meeting minutes and correspondence suggest that the agency voiced a preference for the episode counts as a primary endpoint over the CNSLS and stated that one month duration of the ALS study was too short.

2.2 Data Sources

The data for study AVR-102 can be found at the following location:

[\\CDSESUB1\evsprod\n021879\0004\m5\datasets\99-avr-102\listings\](\\CDSESUB1\evsprod\n021879\0004\m5\datasets\99-avr-102\listings)

The ASSESS.xpt dataset contains the efficacy measures.

The study report for AVR-102 is contained in <\\Cdsesub1\evsprod\n021879\0004\m5\53-clin-stud-rep\535-rep-effic-safety-stud\pseudobulbar-affect\5351-stud-rep-contr\study-99-avr-102>.

The datasets for AVR-106, are located in the following directory.

[\\CDSESUB1\evsprod\n021879\0002\m5\datasets\99-avr-106\listings\](\\CDSESUB1\evsprod\n021879\0002\m5\datasets\99-avr-106\listings)

The ASSESS.xpt dataset contains the efficacy measures.

The report for AVR-106 is contained in

<\\Cdsesub1\evsprod\n021879\0002\m5\53-clin-stud-rep\535-rep-effic-safety-stud\pseudobulbar-affect\5351-stud-rep-contr\study-02-avr-106>.

3 STATISTICAL EVALUATION

3.1 Evaluation of Efficacy

3.1.1 Study AVR-102

The date of first enrollment was 11 January 2001 and the date the last patient completed the study was 30 April 2002.

3.1.1.1 Study Design

Objectives

The objectives of the study were to compare and evaluate the safety, efficacy, and tolerance of AVP-923 (dextromethorphan hydrobromide [30 mg] and quinidine sulfate [30 mg]) taken twice daily relative to dextromethorphan hydrobromide [30 mg], and relative to quinidine sulfate [30 mg], in a population of ALS patients who exhibited pseudobulbar affect.

Study Design

This was a multicenter, randomized, double-blind, controlled, parallel, three-group study of the treatment of pseudobulbar affect in ALS patients with AVP-923 administered orally, two times a day (every 12 hours) for 28 days (the first dose will be taken in the P.M. of Day 1, and the final dose will be taken in the A.M. on Day 29). The last day (Day 29) was to be the last day the patient was on study and could have occurred anywhere between Day 26 and Day 32.

Approximately 12 centers were to be identified. Patients were to be randomized to one of three groups to receive either AVP-923 (a capsule containing dextromethorphan hydrobromide [30 mg] and quinidine sulfate [30 mg]), dextromethorphan hydrobromide (30 mg), or quinidine sulfate (30 mg).

The primary efficacy endpoint was the CNSLS score. Secondary efficacy endpoints were: 1) the number of episodes as recorded in the patient diary; 2) the Visual Analog Scale (VAS) response for Overall Quality of Life; and 3) the Visual Analog Scale (VAS) response for Quality of Relationships.

3.1.1.2 Efficacy Measures

Pseudobulbar Affect

Pseudobulbar affect was to be assessed using Center for Neurologic Study-Lability Scale (CNS-LS). The CNS-LS (Appendix I) is a 7-item report measure that provides a score for total pseudobulbar affect including assessments of labile laughter and labile tearfulness. The range of possible scores is 7 to 35. The CNS-LS requires approximately 5 minutes to complete. Patients were required to complete the CNS-LS at the screening visit, at the first day on study prior to taking their first dose of study medication, and at the Day 15 and Day 29 visits. In order to be included in the study, patients must have had clinically diagnosed pseudobulbar affect and have attained a score of 13 or above on the Center for Neurologic Study-Lability Scale at the Day 1 clinic visit.

3.1.1.3 Statistical Methods and Sample size

Statistical Methods

The primary efficacy endpoint was the CNS-LS score. Secondary efficacy endpoints were patient episode counts, the Visual Analog Scale response for Overall Quality of Life (VAS-QOL), and the Visual Analog Scale response for Quality of Relationships (VAS-QOR). Efficacy comparisons, for primary and secondary analyses, were to be tested using a two-sided 5% significance level, and both tests (AVP-923/DM and AVP-923/Q) had to show statistically significant differences in order to permit a claim of superiority of AVP-923.

a. Primary Efficacy Analysis

The primary efficacy analysis was to be based on the improvement in CNS-LS score, where individual improvement was to be measured as the difference between the baseline scores (pre-treatment) and the average of Day 15 and Day 29 scores (post-treatment). Mean improvements in each group were to be assessed using the intention-to treat patients. The statistical analysis plan followed the procedures described by Frison and Pocock (1992). The primary test of efficacy for the AVP-923/DM and AVP-923/Q comparisons was to be based on the ANCOVA method described in Frison and Pocock. This utilizes the average of the Day 15 and Day 29 scores as the dependent variable. The baseline score was to be used as a covariate and the model would also contain adjustments for center and treatment. It was expected that neither DM nor Q would show significant evidence of symptom improvement in the absence of the other.

b. Secondary Efficacy Analyses

Secondary endpoints were to be analyzed using generalized linear models having a model structure that was parallel to the one used for the primary endpoint and they included:

- Episode counts
- Mean improvement in VAS response (Overall Quality of Life)
- Mean improvement in VAS response (Quality of Relationships)

As a supplement to the secondary analyses reported above, adjustment for multiple comparisons due to multiple endpoints in the secondary analyses were to employ the nonparametric method of O'Brien (1984). According to the sponsor this method maintains overall Type I error rates, but has greater power for detecting effects than the Bonferroni method when the treatment affects more than one of the secondary endpoints. The secondary endpoints of laughing and crying were to be expressed as mean counts per week, with greater values representing worse outcomes. The secondary endpoints of VAS-QOL, and VAS-QOR were to be expressed as $[(Y_{15}+Y_{29})/2]-Y_B$, where Y_{15} , Y_{29} , and Y_B are VAS values at day 15, day 29 and baseline respectively. Since for these parameters larger values represent better outcomes, outcomes were to be reverse-coded for implementation of the O'Brien method.

In addendum 1 to the protocol it was stated that episode counts are most readily modeled as Poisson random variables. Let λ be the mean of the Poisson random variable for the i^{th} patient at the j^{th} time. The model is given by $\lambda_{ij} = \mu_i + \gamma Y_{ij0}$ where $\lambda_{ij} = \log[E(Y_{ij1} + Y_{ij2})]$ and Y_{ij1} and Y_{ij2} are the counts at days 15 and 29, respectively.

c. Pre-specified Subset Analysis

As noted in section IIB of the protocol, approximately 7% of the population has reduced activity of P450 2D6; these individuals are referred to as "poor metabolizers" of DM. It was expected that any poor metabolizers who were assigned to the DM arm of the study would have responses that were similar to those receiving AVP-923. The poor metabolizer phenotype was expected to have no effect within either of the other two study arms. If any poor metabolizers were randomized to the DM arm of the study, a subset analysis that excluded those patients from the DM portion of the study was to be carried out. Other patients in the DM only arm identified as taking concomitant medications also inhibiting P450 2D6 metabolism might also be excluded from this analysis.

The statistical analysis plan which is dated November 30, 2001 (prior to the completion of the study) changed this from a pre-specified subset to the primary analysis population. It states that the primary efficacy analysis will be based on the subset of randomized patients that are not determined to be poor metabolizers of cytochrome P450 2D6 on genotyping.

Sample Size Calculations

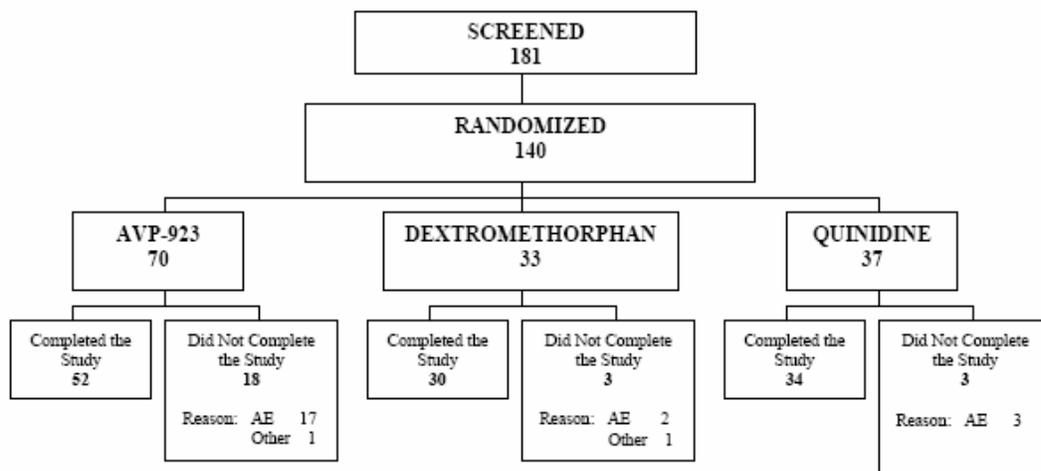
The sample sizes of 48 patients in the AVP-923 group, and 24 patients in each of the DM and Q groups were expected to be sufficient to detect a difference in CNS-LS score of 5.5 between the DM/Q and each component. These calculations were based on standard deviations of 7, 5, and 3 in the DM/Q, DM, and Q groups, respectively. (Power is approximately 85% based on two-sided, 5% test, assuming baseline-Day 15 and baseline-Day 29 correlations of 0.3 and Day 15-Day 29 correlation of 0.7). The assumptions on which sample sizes are based are drawn from a small 14- patient crossover study (Smith et al., 1995), in which DM/Q patients had a mean change from baseline of -6.6 points with standard deviation of 7.5, and placebo treated patients had a mean change of +0.83 with a standard deviation of 3.2.

3.1.1.4 Disposition of Patients

Figure 1 shows the disposition of patients in study 102. A total of 140 subjects were randomized to treatment; 70 were in the AVP-923 group, 33 were in the DM group, and 37 were in the Q group. Note that the sponsor had planned for only 48 subjects in the AVP-923 group and 24 subjects in each of the other treatment groups based on their sample size calculations. Twenty six percent of the AVP-923 group did not complete the study as compared to 9% of the DM group and 8% of the Q group. Seventeen of the 18 AVP-923 dropouts were attributed to adverse events.

One Quinidine patient and 4 AVP-923 patients were not considered to be completers by the sponsor despite having CNSLS assessments at or near both days 15 and 29 because they refused to take the medication due to toxicity.

Figure 1 Study 102: Patient Disposition



3.1.1.5 Patient Demographics

Five (7%) of the 70 AVP-923 patients, 3 (9%) of the 33 DM patients, and 3 (8%) of the 37 Q patients were determined to be poor metabolizers of cytochrome P450 2D6. The Intent-to-Treat Population as defined by the sponsor excludes these patients but includes all other randomized subjects who are not “poor metabolizers” of cytochrome P450 2D6. Table 1 shows the baseline demographics and baseline disease characteristics by group for all randomized patients.

Table 1 Study 102: Baseline Demographics and Disease Characteristics-ITT population (reviewer's results)

Variable	Levels	AVP-923 (N=70)	DM (N=33)	Q (N=37)	All	Any Group differences Pvalue	AVP-923 vs. DM Pvalue	AVP-923 vs. Q Pvalue
RACE	ASIAN	0 (0.0)	1 (3.0)	0 (0.0)	1 (0.7)	0.368	0.252	0.578
RACE	BLACK	2 (2.9)	0 (0.0)	0 (0.0)	2 (1.4)	0.368	0.252	0.578
RACE	CAUCASIAN	63 (90.0)	28 (84.8)	34 (91.9)	125 (89.3)	0.368	0.252	0.578
RACE	HISPANIC	5 (7.1)	3 (9.1)	3 (8.1)	11 (7.9)	0.368	0.252	0.578
RACE	OTHER	0 (0.0)	1 (3.0)	0 (0.0)	1 (0.7)	0.368	0.252	0.578
SEX	FEMALE	25 (35.7)	15 (45.5)	15 (40.5)	55 (39.3)	0.630	0.344	0.624
SEX	MALE	45 (64.3)	18 (54.5)	22 (59.5)	85 (60.7)	0.630	0.344	0.624
AGE	Mean (SD)	55.5 (12.9)	54.3 (12.0)	55.5 (9.9)	55.2 (11.9)	0.905	0.663	0.954
HEIGHT	Mean (SD)	67.6 (4.1)	67.1 (3.7)	68.1 (4.3)	67.6 (4.1)	0.553	0.464	0.607
WEIGHT	Mean (SD)	169.5 (34.8)	172.0 (47.3)	178.9 (40.1)	172.6 (39.4)	0.714	0.852	0.416
ALSTYPE	BULBAR	30 (42.9)	15 (45.5)	23 (62.2)	68 (48.6)	0.151	0.804	0.057
ALSTYPE	LIMBIC	40 (57.1)	18 (54.5)	14 (37.8)	72 (51.4)	0.151	0.804	0.057
EPIISODES PER WEEK (RETROSPECTIVE)	Mean (SD)	23.1 (31.2)	36.0 (63.8)	20.6 (19.1)	25.5 (39.4)	0.254	0.134	0.864
BASE HAMD	Mean (SD)	5.2 (4.3)	4.0 (3.0)	5.8 (4.2)	5.1 (4.0)	0.114	0.104	0.482
BASE VAS-QOL	Mean (SD)	35.1 (26.3)	44.4 (27.9)	47.3 (27.3)	40.5 (27.4)	0.024	0.080	0.011
BASE VAS-QOR	Mean (SD)	31.5 (28.1)	37.8 (28.9)	42.5 (29.4)	35.9 (28.8)	0.107	0.284	0.039
BASE CNSLS	Mean (SD)	20.3 (5.6)	21.1 (6.1)	22.3 (5.8)	21.0 (5.8)	0.181	0.511	0.065
PHENO	EXTENSIVE METABOLIZER	61 (88.4)	30 (90.9)	32 (86.5)	123 (88.5)	0.778	0.461	0.956
PHENO	POOR METABOLIZER	5 (7.2)	3 (9.1)	3 (8.1)	11 (7.9)	0.778	0.461	0.956
PHENO	ULTRA RAPID METABOLIZER	3 (4.3)		2 (5.4)	5 (3.6)	0.778	0.461	0.956
XN	*2XN	7 (10.1)		2 (5.4)	9 (6.5)	0.143	0.058	0.404
XN	WT/WT	62 (89.9)	33 (100.0)	35 (94.6)	130 (93.5)	0.143	0.058	0.404

The Q group had 20% more Bulbar ALS than the AVP-923 group (p=0.06). The baseline Visual Analog Scale (VAS) Quality of Life (abbreviated as VASQOL or VQOL) scores were higher on average in the DM and Q groups than in the AVP-923 group. They were 9.3 (p=0.080) and 12.2 (p=0.011) points higher respectively (higher scores indicate a more negative effect of laughing and crying episodes on quality of life). The baseline scores on the VAS quality of relationships were higher on average in the DM and Q groups than in the AVP-923 group. They were 6.3 (p=0.284) and 11.0 (p=0.039) points higher respectively (higher scores indicate a more negative effect of laughing and crying episodes on quality of relationships). The Q group also had a numerically higher (worse) average baseline CNSLS score than the AVP-923 group (22.3 vs. 20.3, p=0.065). The sponsor only presented a summary of baseline demographics and disease characteristics with poor metabolizers excluded. However, the comparability of the groups at baseline was similar regardless of whether the poor metabolizers were included or excluded.

3.1.1.6 Sponsor's Results

The primary efficacy analysis was the change from baseline in CNS-LS scores, adjusted for center and baseline CNS-LS score. The descriptive statistics for the ITT Population (excluding poor metabolizers) are in the following table.

Table 2 Study 102: Mean Change in CNS-LS Scores^a – ITT Population (N=129, Excluding Poor Metabolizers)

Change in Score	AVP-923 (N=65)	DM (N=30)	Q (N=34)
n	61	30	34
Mean	-7.39	-5.12	-4.91
Std Dev	5.37	5.56	5.56
Median	-6.50	-4.50	-4.25
Min/Max	-24.00/0.0	-25.00/2.0	-21.00/2.0

^a Change in CNS-LS scores was defined as the mean of scores on Day 15 and Day 29 minus the baseline (Day 1) score.

As prospectively specified in the statistical analysis plan, the differences in mean improvement in CNS-LS scores, adjusted for center and baseline CNS-LS scores, were analyzed by using linear regression according to the ANCOVA method of Frison and Pocock, which uses the average over time of the response. The results of this analysis are in Table 3. The results of additional analyses without any adjustments or with an adjustment for baseline CNS-LS score alone are also shown in this table.

Table 3 Study 102: Differences in Mean CNS-LS Scores^a - ITT Population (N=129, Excluding Poor Metabolizers)

Statistics	AVP-923 vs DM	AVP-923 vs Q
Unadjusted difference in mean score	-2.27	-2.47
Std Err	1.22	1.17
p-value	0.0652	0.0366
Difference in mean score adjusted for baseline		
CNS-LS score	-2.97	-3.65
Std Err	1.03	1.00
p-value	0.0046	0.0004
<i>Difference in mean score adjusted for baseline</i>		
<i>CNS-LS score and center^b</i>	-3.29	-3.71
<i>Std Err</i>	1.00	0.97
<i>p-value</i>	0.0013	0.0002

^a Change in CNS-LS scores was defined as the mean of the scores on Day 15 and Day 29 minus the baseline (Day 1) score.

^b Analysis in *italics* was pre-specified in the Statistical Analysis Plan.

Reviewer's Comment: *The discrepancy between the results adjusted for baseline score and the results unadjusted for baseline score may be explained by the fact that the effect of the baseline score is significant. This means that including the baseline score in the model, as planned, reduces the variability attributed to noise (error) which will tend to increase significance of tests (unless the estimated treatment difference is smaller when the baseline score is in the model) since the test statistics are proportional to the reciprocal of the mean square error and a larger test statistic is more significant. Note also that a Wilcoxon rank sum test of change from baseline for the AVP-923 vs. DM comparison yields a p-value of 0.067 while a Wilcoxon rank sum test of the percent change from baseline is nominally significant for the AVP-923 vs. DM comparison (p=0.013). This may add to the argument for baseline adjustment and make the baseline adjusted (ANCOVA) result more believable.*

3.1.1.7 Reviewer's Results

3.1.1.7.1 Primary Analysis

Both the DM and the Q group had numerically higher (worse) average **baseline** scores on the primary efficacy measure, CNSLS, than the AVP-923 group. The difference between the Q and AVP-923 groups **at baseline** approached significance at the nominal level (22.3 vs. 20.3, $p=0.065$).

Of the 140 patients randomized into the study 11 (5 AVR; 3 DM; and 3 Q) were determined to be poor metabolizers of cytochrome P450 2D6 and were thus excluded from the sponsor's primary analysis as specified in the analysis plan. The sponsor calls this reduced population the ITT population although it excludes these 11 randomized patients, all of whom had post-baseline CNS-LS outcome data. The following table shows that including the 11 poor metabolizers in the analysis has little effect on the primary analysis result for the CNSLS. The significance of the results also does not depend on whether we analyze the change from baseline in CNSLS scores at day 29 or the average of the changes from baseline at days 15 and day 29, as prespecified by the sponsor.

For the Modified ITT (MITT) population, which excludes ITT patients without post-baseline CNSLS data, the difference in group least squares mean CNSLS changes from baseline averaged over time was estimated to be 4.2 points ($\pm .93$ S.E., $p<0.0001$) for the AVP-923 vs. Q comparison and 3.1 ($\pm .96$, $p=0.0016$) points for the AVP vs. DM comparison, using the available period for patients that did not complete the study. It is more common in clinical trials submitted to the FDA division of Neurology to base the treatment group comparison on the change from baseline at the end of the study (or last follow-up) instead of averaging the change over the entire double blind treatment period as the sponsor prespecified. In this trial the difference in group least squares mean changes at the end was estimated to be 3.6 points ($\pm .96$ S.E., $p=0.0003$) for the AVP-923 vs. Q comparison and 2.7 ($\pm .98$, $p=0.0066$) points for the AVP vs. DM comparison using the available period for patients that did not complete the study. Thus, although there is no difference in the significance of the group difference between the change averaged over the entire period and the change at the last measurement, the group difference is slightly larger in the averaged changes than the last changes.

Table 4 Study 102: Analysis of Change from baseline in CNSLS score by Population and Timepoint

Population	Timepoint	Statistic	AVP-923	DM	Q
MITT		N	66	33	37
	Baseline CNSLS	Mean	20.3	21.1	22.3
	Average of Days 15 and 29	Mean Change	-7.67	-5.15	-4.68
		Std Dev.	5.54	5.32	5.53
		Median	-6.50	-5.00	-4.00
		Min / Max	-24.0 / 0.00	-25.0 / 2.00	-21.0 / 4.00
		LS Mean	-7.82	-4.72	-3.59
p-value*		0.0016	<0.0001		
MITT Excluding Poor Metabolizers		N	61	30	34
	Baseline CNSLS	Mean	20.1	21.4	22.3
	Average of Days 15 and 29 (Sponsor's primary analysis)	Mean Change	-7.39	-5.12	-4.91
		Std Dev.	5.37	5.56	5.56
		Median	-6.50	-4.50	-4.25
		Min / Max	-24.0 / 0.00	-25.0 / 2.00	-21.0 / 2.00
		LS Mean	-7.39	-4.09	-3.67
p-value*		0.0013	0.0002		
MITT		N	66	33	37
	Baseline CNSLS	Mean	20.3	21.1	22.3
	Day 29 or Last CNSLS	Mean Change	-7.64	-5.48	-5.32
		Std Dev.	5.66	5.59	5.87
		Median	-6.50	-5.00	-5.00
		Min / Max	-24.0 / 1.00	-25.0 / 2.00	-22.0 / 4.00
		LS Mean	-7.85	-5.13	-4.25
P-value*		0.0066	0.0003		
MITT Excluding Poor Metabolizers		N	61	30	34
	Baseline CNSLS		20.1	21.4	22.3
	Day 29 or Last CNSLS	Mean Change	-7.36	-5.47	-5.29
		Std Dev.	5.51	5.84	5.90
		Median	-6.00	-5.00	-5.00
		Min / Max	-24.0 / 1.00	-25.0 / 2.00	-22.0 / 4.00
		LS Mean	-7.45	-4.52	-4.05
p-value*		0.0059	0.0011		

*based on ANCOVA model adjusted for baseline score, treatment, and center (small centers 5,10, and 20 pooled)

3.1.1.7.2 Sensitivity of Primary Analysis Result to Missing Data

The following table gives this reviewer’s results for the change in CNSLS at day 29 or last observation carried forward for the modified ITT (MITT) population and for the change in CNSLS at day 29 for the Completers population. Comparing these results is one simple way to assess the potential impact of dropouts on the primary analysis result. Completers are defined here somewhat arbitrarily as patients whose last assessment was beyond day 22 (not more than a week before the protocol specified last assessment day). Note that according to the statistical analysis plan the day 29 visit was allowed to fall between days 26 and 32. If the definition was changed to reflect this there would be 1 more AVP-923 patient, no more DM patients, and 3 more Q patients that did not “complete” but the results are virtually identical either way.

Table 5 Study AVR-102: Sensitivity Analyses for Change from baseline in CNSLS at day 29

ANALYSIS	GROUP	N	MEAN BASELINE CNSLS	LSMEAN CNSLS CHANGE (S. E.)	LSMEAN DIFFERENCE FROM AVP-923 (S. E.)	P-VALUE
MITT-LOCF	AVP-923	66	20.3	-7.9 (0.6)	.	.
	DM	33	21.1	-5.1 (0.8)	-2.7 (1.0)	0.007
	Q	37	22.3	-4.3 (0.8)	-3.6 (1.0)	<0.001
Completers (i.e., last assessment after day 22 with no imputation)	AVP-923	56	20.2	-8.4 (0.7)	.	.
	DM	30	21.6	-5.3 (0.9)	-3.1 (1.0)	0.003
	Q	35	22.6	-4.1 (0.8)	-4.3 (1.0)	<0.001

10 AVP-923 3 DM and 2 Q patients were last assessed on CNSLS before day 23

The similarity of the results for the MITT and Completers populations lends some support to the primary analysis.

Potential Impact of the 4 AVP-923 patients with no post baseline data on the primary analysis

There were 0 patients with no post-baseline efficacy data in the DM and Q groups but there were 4 such patients in the AVP-923 group. The reason provided for termination was “patient refused to take medication due to toxicity”. The 4 AVP-923 dropouts that did not have any post-baseline CNSLS measures dropped out at days 14, 14, 31, and 8, respectively, and took an unknown number, 16, 13, and 4 doses, respectively. The last of these patients provided 4 days of episode counts each of which was zero, but the other patients provided no episode count or other efficacy data. The worst change from baseline in CNSLS over all groups was +4 points. The worst in the AVP-923 group was +1 point. In order for the four AVP-923 dropouts with no post-baseline efficacy data to impact the results some of them would have to have gotten dramatically worse over the course of the trial. For example, if each of them worsened (increased) by 5 points then the p-value for the comparison between AVP-923 and DM would be 0.056. This is similar to an ANCOVA analysis of the ranks of the changes from baseline in CNSLS where the worst rank is assigned to the AVP-923 dropouts with no post-baseline CNSLS measures. Note that the p-value for the AVP-923 vs. Q comparison would still be significant, p=0.005. If each worsened by 2 points, which is one point worse than the worst observed change among the dropouts in the same

group, then the p-value for the AVP-923 vs. DM comparison would be 0.032 and for the AVP-923 vs. Q comparison it would be 0.002. Considering that the p-value is just increased to 0.056 under a worst case type scenario for the 4 AVP-923 dropouts with no post-baseline efficacy data it seems unlikely that these patients would have worsened enough to impact the primary analysis result. Therefore, the significance of the primary analysis seems valid despite the existence of these 4 AVP-923 patients with no post-baseline efficacy measures.

Carrying baseline forward is usually discouraged as a method for imputing missing data in the division of Neurologic drugs because it can lead to underestimating the variance of the group difference and thus to a biased test. It is used in some disciplines though and it does provide an alternative way to include all ITT patients in the analysis, as well as, another sensitivity analysis. In study 102 if we impute no change, i.e., carry the baseline forward, for those who were last assessed on the CNSLS before day 23, i.e., more than a week before the intended final assessment time, and for those who had no post-baseline CNSLS assessments (4 patients - all DM/Q) the p-value for the DM/Q vs. DM comparison increases to 0.083 and that for the AVP-923 vs. Q comparison increases to 0.005. Therefore, the significance of the primary analysis result may be affected by changing assumptions regarding the dropouts. For the sake of completeness, if we focus on the usual MITT population where these 4 AVP-923 patients are excluded then the 0.083 p-value for the DM comparison reduces to 0.042 under baseline carried forward imputation. Between the primary analysis results and the baseline carried forward results the p-values for the AVP-923 vs. DM comparison range from <0.001 to 0.083 which suggests that the primary analysis result is somewhat sensitive to certain assumptions regarding the missing CNSLS scores. However, in this reviewer's opinion because the results are still reasonable under several worst case type imputation schemes the significance of the primary analysis result seems to be validated.

Mean Changes in CNSLS by Termination Reason

Table 6 shows the mean changes from baseline at the last assessment time for dropouts according to the termination reason. It shows that dropouts from the AVP-923 group averaged 4 points worse on the CNSLS than AVP-923 completers. Dropouts averaged slightly worse for the other groups too but those results are less reliable because of the smaller number of dropouts. Based on the observed means there is no obvious bias resulting from the dropouts.

Table 6 AVR-102: Mean CNSLS Change at Day 29 or Last Assessment by Termination Reason

COMPLETER*	REASON	N	AVR		N	DM		N	Q	
			Baseline Mean (S. D.)	Change Mean (S. D.)		Baseline Mean (S. D.)	Change Mean (S. D.)		Baseline Mean (S. D.)	Change Mean (S. D.)
NO	PATIENT REFUSAL-TOXICITY	12	19.6 (6.2)	-4.7 (5.4)	2	17.0 (0.0)	-2.5 (4.9)	2	18.5 (2.1)	-2.0 (4.2)
NO	PATIENT REFUSAL-NON-TOXICITY	1	16.0 (.)	-4.0 (.)	1	14.0 (.)	-2.0 (.)	.	.	.
NO	DEATH	1	28.0 (.)	0.0 (.)
NO	DISALLOWED TREATMENT	1	17.0 (.)	-8.0 (.)
NO	ALL REASONS	14	19.8 (6.2)	-4.3 (5.1)	3	16.0 (1.7)	-2.3 (3.5)	3	18.0 (1.7)	-4.0 (4.6)
YES	N/A	52	20.4 (5.4)	-8.5 (5.5)	30	21.6 (6.2)	-5.8 (5.7)	34	22.7 (5.9)	-5.4 (6.0)

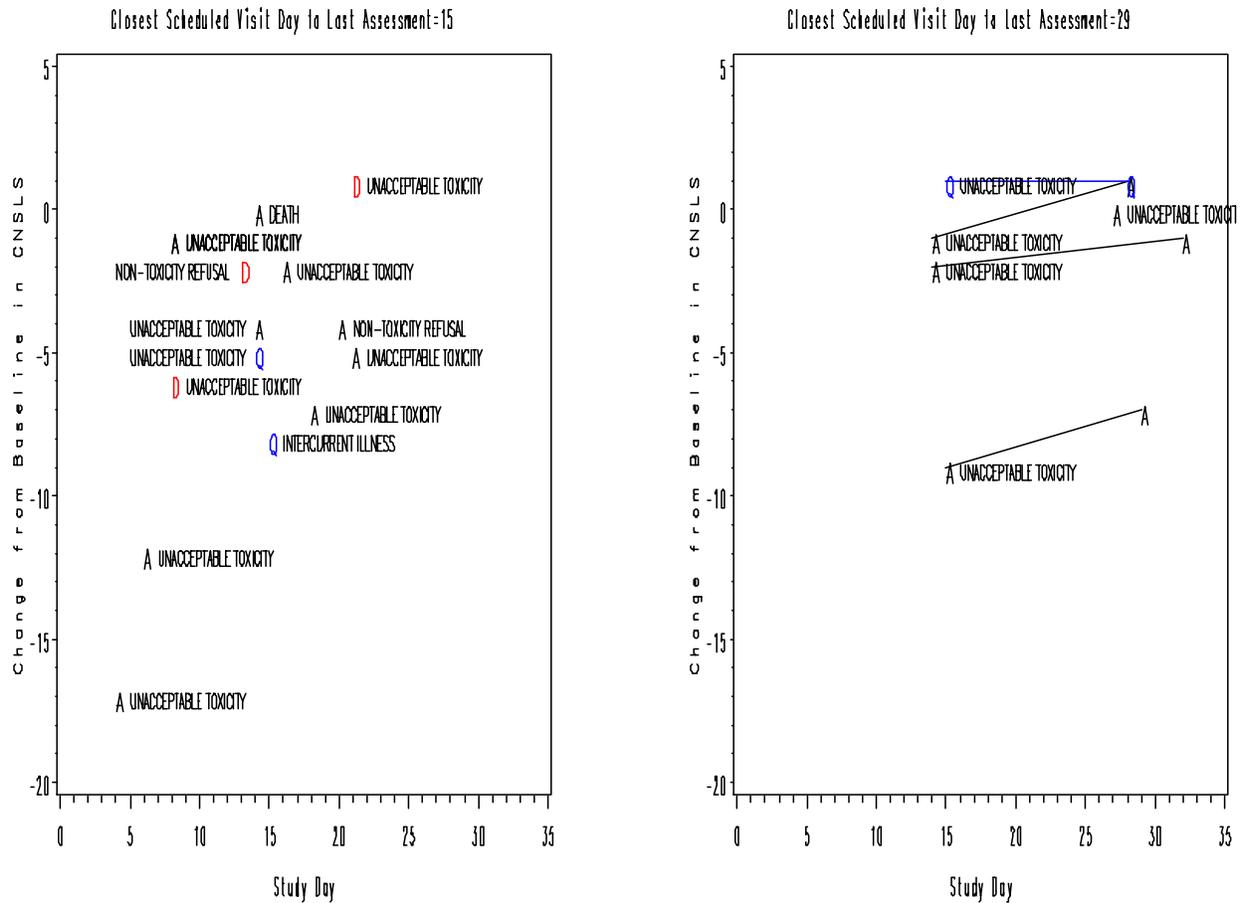
*Completer by Sponsor's definition

Mixed Model for Repeated Measures Analysis

This reviewer also investigated a mixed model for repeated measures analysis as a sensitivity analysis. This approach models all of the observed post-baseline CNSLS scores. The model included baseline score as a covariate, center effects, treatment group effects, visit (as a class variable) and effects for the interaction between visit and treatment group. The within subject covariance structure for repeated measures was specified as unstructured. The repeated measures model (MMRM) estimates the group difference in least squares means at week 4 to be 3.7 points (+/- 1.0 S.E., p=0.0002) for the AVP vs. Q comparison and 2.8 points (+/- 1.0 S.E., p=0.0058) for AVP vs. DM. The same model suggested that the differences decreased from week 2 to week 4. In particular, at week 2 the differences were estimated to be 4.9 (+/- 1.0) for AVP vs. Q and 3.5 (+/- 1.1 S.E.) for AVP vs. DM. Thus, the AVP and Q group mean change in CNSLS scores came an estimated 1.2 (+/- 0.7 S.E.) points closer from week 2 to week 4. The AVP and DM group mean changes came closer numerically also, but to a lesser degree, 0.7 (+/- .8 S.E.) points. This is in line with the previous observation that the group difference based on the average of day 15 and day 29 looked slightly more impressive than that based on day 29 alone. In summary, the group comparisons of CNSLS scores at day 29 based on a repeated measures analysis back up the primary analysis results.

Figure 2 shows the change in CNSLS scores over time for non-completers and the reason for dropping out. The plotting symbol A is used for the AVP-923 group and D and Q are used for the DM and Q groups, respectively. Note that negative changes are better. Completers in the DM and Q groups averaged -4.6 and -3.0 at day 15 and -5.3 and -4.4 at day 29, respectively. Thus, on average patients in the DM and Q groups showed improvement between day 15 and day 29 to a greater, though not significant, degree than that seen in the AVP-923 group.

Figure 2 Study 102: Change in CNS-LS over Time for AVP-923 Non-Completers (excluding 4 AVP-923 patients with no post-baseline data)



Separate Analyses of CNSLS Laughing and Crying Items

It was not planned to analyze the laughing and crying items separately but it seems reasonable to check that the overall effect is not just on laughing or just on crying. The results of these analyses are shown in Table 7. This reviewer noticed that if one ignored the three items in the CNSLS related to crying (items 1, 3, and 6), i.e., focused on the change from baseline in the sum of the four laughing items at day 29 (or LOCF) then the AVP-923 group was not nominally significantly different from either the DM or the Q group. If we use the sponsor’s approach of averaging over the changes from baseline at day 15 and 29 the AVP-923 vs. DM comparison is nominally significant but the AVP-923 vs. Q comparison is not. The same conclusions result whether or not we include the poor metabolizers. This analysis of only laughing items was not planned and, thus, potentially not adequately powered. However, the result is corroborated by a similar lack of significant group differences in the laughing episode counts, as will be described below.

On the other hand, if we focus on the sum of the crying items then differences between AVP-923 and DM as well as AVP-923 and Q are nominally significant in all cases. The analyses of laughing episode counts and crying episode counts (to be described below) showed a similar trend (i.e., nominal significance for crying episodes and a lack of significance for laughing episodes). This suggests the possibility that the combination is not more effective than its components for controlling the inappropriate laughing aspect of the disease.

Table 7 Study 102: Analyses of Laughing and Crying items separately Least Square Means

Endpoint/Timepoint	Group	Baseline	LS Mean Change (S. E.)	LS Mean Difference (S. E.)	P-value*
Last change in Laughing items	AVP-923	11.3	-3.5(0.5)	.	.
	DM	11.7	-2.4(0.6)	-1.1(0.7)	0.096
	Q	12.4	-3.1(0.6)	-0.4(0.7)	0.526
Change in Laughing items averaged over time	AVP-923	11.3	-3.6(0.4)	.	.
	DM	11.7	-2.1(0.6)	-1.5(0.6)	0.020
	Q	12.4	-2.6(0.5)	-1.0(0.6)	0.100
Last change in Crying items	AVP-923	9.0	-3.9(0.3)	.	.
	DM	9.4	-2.3(0.4)	-1.7(0.5)	0.001
	Q	9.9	-0.9(0.4)	-3.0(0.5)	0.000
Change in Crying items averaged over time	AVP-923	9.0	-3.8(0.3)	.	.
	DM	9.4	-2.1(0.4)	-1.7(0.5)	< 0.001
	Q	9.9	-0.7(0.4)	-3.1(0.5)	< 0.001

* based on ANCOVA model with adjustments for baseline score, centers, and treatment

Secondary Analyses

Because the FDA Division of Neurology had expressed a preference for having the primary endpoint based on the episode counts in meetings with the sponsor this reviewer did a thorough examination of the episode count data. Furthermore, there is no established precedent for a primary endpoint because this is a new indication. If the division still held its initial preference for the episode count endpoint over the CNSLS then the interpretation of the study outcome could be quite different.

This reviewer found that the significance of the results for the secondary analysis of the number of episodes of laughing and crying (adjusting for the number of days reported) were not robust. In particular, the significance of the difference in the episode counts for the AVP-923 vs. DM comparison is questionable as detailed below.

Nine of the 67 (13%) AVP-923 patients who contributed episode data had no laughing or crying episodes during the double blind period. Four of 33 (12%) DM patients and 1 of 37 (3%) Q patients also had no laughing or crying episodes. Of the patients with no post-baseline laughing or crying episodes some had periods of observation that were considerably shorter than the 29 days planned in the protocol: 1 AVP-923 patient had only 4 days, 1 DM patient had 13 days and the Q patient with no episodes had only 8 days of observation. The following table contains summary statistics for the number of days of observation for laughing or crying episodes for all patients, i.e., including patients with either zero or non-zero episode counts. Although the differences were not statistically significant the average numbers of days with non-missing

episode diary data for the DM and Q groups were both about 3 days higher than the AVP-923 group average.

Table 8 Study 102: Summary Statistics for Number of Days with Non-missing Episode Diary Data

Treatment Group	Number of Episode Diary Days				
	N	Min	Mean	Median	Max
AVP-923	67	2.00	23.97	29.00	32.00
DM	33	3.00	27.00	29.00	34.00
Q	37	8.00	26.84	28.00	31.00

Table 9 shows the rates for episodes of the laughing or crying type before the study (provided retrospectively), and during the study. There was one outlier. This was a DM patient who had a total count of 3010 laughing episodes during the 29 days in the study, an average of more than 100 laughing episodes per day and more than 9 1/2 times the next highest episode count. In contrast, the same patient reported 0 crying episodes during the study. Laughing counts on individual days for this patient ranged from 23 to 170. At baseline the patient estimated her historical episode rate to be just 35 episodes of either the laughing or crying type per week. The summary statistics are provided both with and without the outlier included and they make it clear that this patient has a large effect on the estimates for the DM group.

Table 9 Study 102: Historically Reported Episode Rate before Study and Episode Rate during Study

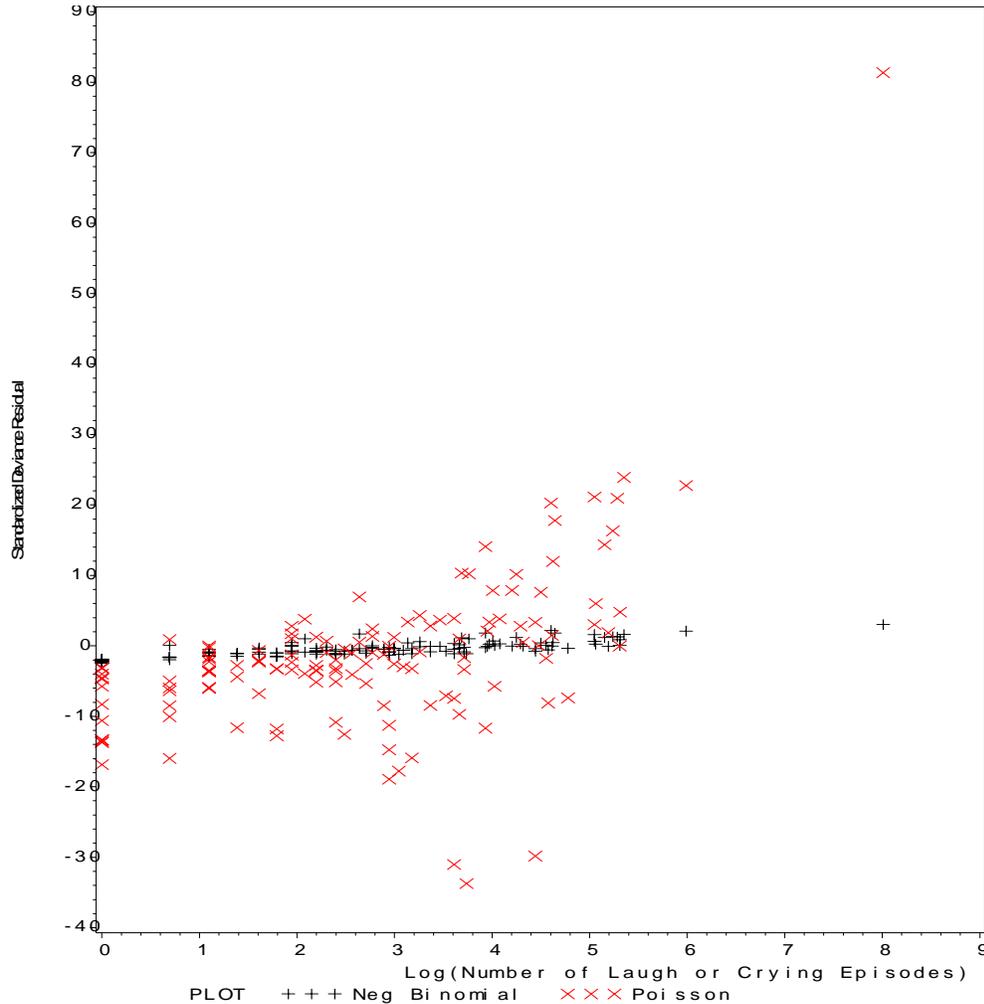
Treatment Group	Historically Reported Number of Episodes per Week				Number of Episodes Per Week During Treatment			
	N	Mean	Median	Std Dev	N	Mean	Median	Std Dev
AVP-923	70	23.1	13.0	31.2	67	9.4	2.5	20.3
Q	37	20.6	14.0	19.1	37	13.0	6.3	14.1
DM	33	36.0	12.0	63.8	33	34.4	4.8	125.8
DM (excluding outlier)	32	36.0	11.0	64.8	32	12.8	4.7	19.7
DM (impute 398=next worst count overall for outlier)	33	36.0	12.0	63.8	33	15.3	4.8	24.2

Since episode counts must be non-negative and integer valued they are not normally distributed and so typical analysis methods like ANCOVA are not usually used for them. The Statistical Analysis Plan specified a Poisson regression model for the analysis of treatment effects on episode counts. The Poisson model assumes that the episodes occur at a constant rate over the follow-up period but allows the episode counts to depend on the site, the treatment, and the number of days with non-missing diary entries. This is a common first model for count data but it requires the strong assumption that the variance of the counts is equal to the expected number

of counts, which is frequently not supported by real data. In “Notes on Episode Count models”, section 7.5.1 of the sponsor’s report for study 102, the statistician for the sponsor acknowledged that “The actual data show strong evidence that there is more variability than the Poisson model would predict (overdispersion), and consequent lack of fit. This particular departure from the Poisson model understates standard errors, so p-values for treatment effects are too small (over-significant).”

The simplest application of the negative binomial distribution is to the distribution of the total number of coin tosses until a certain number, say k , heads are obtained. It is increasingly applied to count data, like we have here as well, because it allows for the variance to increase with the mean, which is typical of overdispersed count data. In the following figure is a plot of the standardized Pearson residuals for the Poisson model as well as for a Negative Binomial model. The residuals are the differences between the observed counts and those predicted for each patient by the respective models. The dramatically wider spread of the residuals based on the Poisson model indicates the relative shortcomings of that model. A very similar result is seen for the episode counts in study 106 (to be described later), thus confirming the inappropriateness of the Poisson model for this episode count data. The deviance (a function of the likelihood) divided by the degrees of freedom is a measure of goodness of model fit. It should be close to 1 for an adequately fitting model. The value is 130.8 for the Poisson model and 1.4 for the negative binomial model which indicates the poor fit of the Poisson model and reasonably good fit of the negative binomial model. As stated by the sponsor’s statistician because the prespecified model does not fit the data it tends to report p-values that are too small (biased towards the alternative) for treatment effects and, therefore, we need to investigate other methods of analyzing the episode count data. The sponsor did not propose any back-up analysis method to be used in the event that the data were overdispersed.

Figure 3 Study 102: Comparison of Poisson and Negative Binomial Model Fits to Episode Count Data



The most common parametric models for count data other than the Poisson are negative binomial models. There are actually two common negative binomial count models. They differ in the dependence of the variance on the mean, μ . For the first, denoted NB1, the variance is $\mu^*(1 + \delta)$ whereas for the second, denoted NB2, the variance is $\mu^*(1 + \alpha*\mu)$. Only NB2 is implemented in the SAS software. However, both are implemented in STATA software. The sponsor's statistician advocates the NB1 model claiming that the quantile-quantile plot (qqplot) of the NB2 residuals suggests they are farther from a normal distribution. This reviewer did not observe any compelling difference between the qqplots of the NB1 and NB2 residuals. In addition, because the assumed distribution of the episode counts is negative binomial rather than normal, the residuals will at best only be approximately normally distributed. Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which are functions of the log likelihood that are penalized for the number of explanatory variables in the model, are commonly used to help choose between models. Models with smaller AIC and BIC values are preferable. Model NB1 has a larger change in AIC and BIC than model NB2 when patient 08-016 who had an

outlying laughing count of 3010 episodes (average > 100 per day) is dropped. Model NB2 seems preferable to this reviewer since it is less sensitive to the exclusion of the outlier patient 08-016. Also, for model NB1 the estimated difference between DM and AVP-923 increases when 08-016 is excluded which is counterintuitive since this DM patient had an enormous incidence rate (3010 episodes; avg. > 100 per day). On the other hand, for model NB2 the estimated difference between DM and AVP-923 decreases when this patient is excluded, as one would expect. Model NB2 has smaller AIC and BIC values and a larger log likelihood than model NB1 whether we include, exclude, or impute the next largest count for the outlier patient which suggests that it is the better of the two negative binomial model choices.

For the NB2 negative binomial model the AVP-923 vs. Q comparison is not statistically significant whether or not the outlying DM patient is included ($p=0.107$ with; $p=0.086$ without). The AVP-923 vs. DM comparison is nominally significant when the outlying DM patient with 3010 episodes is not excluded ($p=0.017$). However, the AVP-923 vs. DM comparison is not nominally significant with the outlying patient excluded ($p=0.343$) or with the next worst episode count (398) observed in the overall population imputed for this patient ($p=0.132$). This suggests that according to the NB2 model the significance of the AVP-923 vs. DM comparison hinges on the outlier patient and, therefore, is suspect. In contrast, for the NB1 negative binomial model both comparisons with AVP-923 are nominally significant whether or not the outlier DM patient is included (AVP-923 vs. DM: $p=0.050$ with, $p=0.013$ without).

One way to resolve the discrepancy between the two negative binomial models would be to compare the groups with respect to the number of episodes per day averaged over the study period using a nonparametric rank sum test. The nonparametric approach also takes care of the outlier patient by replacing the episode count with its rank. Since the reported count is more than 10 times higher than any other count it seems questionable. The patient may have had the highest episode count but may not have really had 10 times more episodes than the person with the next highest count. In terms of ranks the patient still has the highest score but the use of the rank limits the influence of the reported score, which is desirable given the uncertainty surrounding the exact value. The Cochran-Mantel-Haenszel test with modified ridit scores yields a p-value of 0.129 for the AVP-923 vs. DM comparison when the outlier patient is included and 0.190 when the outlier is excluded. The corresponding p-values in the subgroup of non-poor metabolizers are 0.132 and 0.195, respectively. Therefore, the nonparametric test suggests that the AVP-923 and DM groups are not significantly different in terms of the rate of episodes of either the laughing or crying type. On the other hand, the AVP vs. Q comparison was significant in favor of the AVP-923 group according to the Cochran-Mantel-Haenszel test with modified ridit scores ($p=0.0015$ with poor metabolizers and $p=0.0086$ without poor metabolizers).

A few AVP-923 patients provided only a few days of data but had high episode rates. For example, patient 0604 had an average of 117 episodes per week extrapolating from 3 days of data and patient 0308 had an average of 46 episodes per week based on 2 days of data. This reviewer noticed that excluding patients with less than 3 days of data resulted in a nominally significant comparison between the AVP-923 and DM groups ($p=0.032$) in terms of the Cochran-Mantel-Haenszel test. The AVP-923 vs. DM comparison also reached nominal significance ($p=0.009$) in the completers subgroup. We should keep in mind though, that the

AVP-923 group had a lower proportion of completers and excluding dropouts may cause bias, especially considering the high episode rates observed for the 2 AVP-923 dropouts mentioned above.

Table 10 shows the comparisons between AVP-923 and DM and Q for Sum of Laughing and Crying events by model and how the outlier patient is handled. The parameters for DM and Q represent the effects of those groups relative to the AVP-923 group. The parameter Alpha is associated with the NB2 model, for which the variance of the mean, μ , is $\mu + \alpha\mu^2$.

The parameter Delta is associated with the NB1 model, for which the variance of the mean, μ , is $\mu(1 + \delta)$. Alpha=0 or Delta=0 implies that a Poisson model may be appropriate since for a Poisson the variance of the mean is μ . However, one can see in the table that in no case do the confidence intervals for alpha and/or delta include 0.

Table 10 Study 102: Estimated Treatment Effects and Fits from Various Models for Episode Counts

Model	AIC*	Parameter	Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
Poisson w/o 08-016	6881.297	Dm	0.773	0.036	21.520	0.000	0.702	0.843
		Q	0.754	0.035	21.510	0.000	0.685	0.823
Poisson w/ 08-016	16051.560	Dm	1.841	0.030	62.070	0.000	1.783	1.899
		Q	0.845	0.035	24.240	0.000	0.777	0.914
NB1 w/ 08-016	1221.307	Dm	0.431	0.220	1.960	0.050	0.000	0.863
		Q	0.599	0.203	2.950	0.003	0.201	0.998
		Delta [#]	126.172	21.527			90.309	176.277
NB1 w/o 08-016	1304.072	Dm	0.524	0.211	2.480	0.013	0.110	0.937
		Q	0.788	0.193	4.090	0.000	0.410	1.166
		Delta [#]	51.968	8.392			37.869	71.317
NB1 w/ imputation for 08-016	1243.616	Dm	0.572	0.210	2.730	0.006	0.161	0.983
		Q	0.771	0.194	3.980	0.000	0.391	1.152
		Delta [#]	57.705	9.348			42.007	79.270
NB2 w/ 08-016	1218.654	Dm	0.888	0.371	2.390	0.017	0.160	1.616
		Q	0.552	0.342	1.610	0.107	-0.119	1.222
		Alpha [#]	1.928	0.218			1.545	2.406
NB2 w/o 08-016	1256.628	Dm	0.315	0.332	0.950	0.343	-0.336	0.967
		Q	0.540	0.314	1.720	0.086	-0.076	1.156
		Alpha [#]	1.717	0.199			1.367	2.155
NB2 w/ imputation for 08-016	1236.807	Dm	0.497	0.330	1.510	0.132	-0.150	1.143
		Q	0.543	0.319	1.700	0.089	-0.083	1.169
		Alpha [#]	1.742	0.201			1.390	2.183
CMH nonparametric	N/A	Dm	N/A	N/A	N/A	0.145	N/A	N/A
		Q	N/A	N/A	N/A	0.003	N/A	N/A

* Akaike's Information Criterion- a measure of model fit. Smaller values indicate better fit

On the basis of the negative binomial model, NB2, for the episode counts of the laughing or crying type with the outlier included the DM mean was estimated to be 2.4 times higher than the AVP-923 mean, 95% C.I. (1.2, 5.0) p=0.02, and the Q mean was estimated to be 1.7 times higher, 95% C.I. (0.9, 3.4) p=0.11. Note that Table 10 gives the estimated treatment group difference in the natural logarithms of the expected counts. Exponentiating the coefficient for DM gives the estimated ratio of the episode rates (DM to AVP-923) to be $e^{0.888}=2.4$, as described above. The other reported ratios of rates can be obtained from the table similarly.

Without the outlier the DM group mean was estimated to be 1.4 times the AVP-923 mean, 95% C.I. (0.7, 2.6) p=0.34 and the Q mean was estimated to be 1.7 times higher, 95% C.I. (0.9, 3.2), p=0.09.

For the alternate negative binomial model, NB1, with the outlier the DM mean was estimated 1.5 times higher than the AVP-923 mean, 95% C.I. (.9995, 2.4) p=0.050, and the Q mean was an estimated 1.8 times higher than the AVP-923 mean, 95% C.I. (1.2, 2.7) p=0.003.

Without the outlier the DM mean was estimated 1.7 times higher than the AVP-923 mean, 95% C.I. (1.1, 2.6) p=0.013, and the Q mean was an estimated 2.2 times higher than the AVP-923 mean, 95% C.I. (1.5, 3.2) p<0.001. It seems counterintuitive that the DM to AVP-923 ratio of episode rates should increase when the DM patient with the excessively high count is excluded but that is what the NB1 model suggests. This is just one of the reasons why the NB2 model, which doesn't have this property, seems preferable.

There were 3 AVP-923 patients that had no episode diary data and 0 in each of the other groups. Because of this slight imbalance it seems reasonable to investigate their potential effect on the analysis by imputing their rate based on the patient's retrospective report of episodes per week prior to the study and the date of last dose. After these imputations we find that again the NB2 model is preferred over NB1 because it has lower AIC and BIC. Nevertheless, as shown in Table 11 neither negative binomial model has statistically significant comparisons for both AVP-923 vs. DM and AVP-923 vs. Q. Again this suggests that AVP-923 group was not clearly significantly better in terms of the episode counts than both of the other groups.

Table 11 Study 102: Negative Binomial Model Analyses of Sum of Laughing and Crying Episode Counts with imputation for patients with no post-baseline data

Model	Obs	AIC*	Parameter	Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
NB1	140	1340.097	dm	0.354	0.219	1.620	0.106	-0.075	0.784
			q	0.541	0.201	2.690	0.007	0.146	0.935
			delta	127.650	21.579			91.649	177.794
NB2	140	1288.836	dm	0.805	0.378	2.130	0.033	0.065	1.545
			q	0.401	0.343	1.170	0.243	-0.273	1.074
			alpha	1.948	0.217			1.567	2.423

* Akaike's Information Criterion- a measure of model fit. Smaller values indicate better fit

Analysis of Laughing Episodes Only

If we focus on laughing episodes only then the AVP-923 group was not significantly different from either the DM group or the Q group, no matter which analysis method we use and no matter whether we include or exclude the outlier patient. This is shown in the following table. Note that the Poisson model is not considered in the above statement because of the significant lack of fit / overdispersion problem associated with it, which yields p-values that are smaller than they should be. Although this is a secondary analysis and therefore it is potentially underpowered the lack of nominal significance is consistent with a similar finding for the analysis of the change from baseline in the sum of only the laughing items of the CNSLS. In the next section on the analysis of crying episodes only it will be seen that the analysis of crying episodes only and the analysis of change from baseline in only the crying items of the CNSLS were also consistent. However, in contrast to the laughing results, in the crying case nominal significance was reached on both endpoints. Thus, neither the DM vs. AVP-923 comparison nor the Q vs. AVP-923 comparison was nominally significant for the CNSLS laughing items or the laughing episode counts but both comparisons were nominally significant for the CNSLS crying items and the crying episodes counts.

Table 12 Study 102: Treatment Effects and Model Fits for only Laughing episode counts

Model	AIC*	Param	Coef.	Std. Err.	Z	P>z	[95% Conf.	Interval]
Poisson w/o 08-016	6462.662	dm	0.710	0.043	16.540	0.000	0.626	0.795
		q	0.401	0.046	8.780	0.000	0.311	0.491
Poisson w/ 08-016	15919.920	dm	2.069	0.034	60.340	0.000	2.002	2.136
		q	0.520	0.045	11.460	0.000	0.431	0.609
NB1 w/ 08-016	1000.778	dm	0.279	0.256	1.090	0.276	-0.223	0.781
		q	0.287	0.247	1.160	0.244	-0.196	0.771
		delta	197.654	43.033			128.998	302.850
NB1 w/o 08-016	1064.735	dm	0.320	0.257	1.240	0.214	-0.185	0.824
		q	0.390	0.245	1.590	0.112	-0.091	0.871
		delta	72.602	14.693			48.830	107.949
NB2 w/ 08-016	997.411	dm	0.939	0.565	1.660	0.097	-0.169	2.047
		q	0.090	0.557	0.160	0.871	-1.000	1.181
		alpha	3.680	0.456			2.887	4.691
NB2 w/o 08-016	1029.233	dm	0.109	0.509	0.210	0.831	-0.890	1.107
		q	0.077	0.498	0.150	0.877	-0.899	1.053
		alpha	3.384	0.429			2.640	4.337
CMH nonparametric	N/A	dm	N/A	N/A	N/A	0.246	N/A	N/A
		q	N/A	N/A	N/A	0.136	N/A	N/A

* Akaike’s Information Criterion- a measure of model fit. Smaller values indicate better fit

Analysis of Crying Episodes Only

If we focus on crying events only, the group comparisons play out as shown in the following table. Pairwise comparisons between AVP-923 and DM and AVP-923 and Q are nominally significant in all cases except for the NB2 model which gives a p-value of 0.06. However, in this case the NB1 model fits slightly better than the NB2 model and it agrees with the nonparametric Cochran-Mantel-Haenszel test in terms of the significance of the result. Note that the DM patient with the outlying laughing episode count is included in all cases since that patient's crying count was not an outlier.

Table 13 Study 102: Treatment Effects and Model Fits for Various Models of only Crying Episode Counts

Model	AIC*	Param	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Poisson w/ 08-016	2907.322	dm	0.844	0.065	12.920	0.000	0.716	0.972
		q	1.281	0.058	22.010	0.000	1.167	1.395
NB1 w/ 08-016	881.582	dm	0.631	0.247	2.550	0.011	0.147	1.115
		q	1.283	0.225	5.690	0.000	0.841	1.725
		Delta [#]	24.500	4.557			17.016	35.276
NB2 w/ 08-016	894.951	Dm	0.667	0.355	1.880	0.060	-0.029	1.363
		Q	1.594	0.359	4.440	0.000	0.890	2.298
		Alpha [#]	2.092	0.287			1.598	2.738
CMH nonparametric	N/A	dm	N/A	N/A	N/A	0.040	N/A	N/A
		q	N/A	N/A	N/A	0.000	N/A	N/A

* Akaike's Information Criterion- a measure of model fit. Smaller values indicate better fit

Alpha and Delta are parameters associated with the negative binomial distribution in NB2 and NB1, respectively, alpha=0 or delta=0 suggests a Poisson distribution may be appropriate

Other Secondary Endpoints

The difference between the AVP-923 group and the DM group and the difference between the AVP-923 and Q group on the change from baseline in the VAS-QOL and VAS-QOR as determined from the prespecified ANCOVA model adjusting for baseline score, sites, and treatment groups were nominally significant. These results are shown in Table 14. It is important to point out that there were nominally significant differences between the AVP-923 and Q group mean VAS-QOL and VAS-QOR scores at baseline. The analyses did adjust for the baseline scores but this may not entirely correct for the baseline imbalance, i.e., the treatment group difference estimated by the model may be partially confounded with the baseline imbalance.

Table 14 Study 102: Analyses of Other Secondary Endpoints

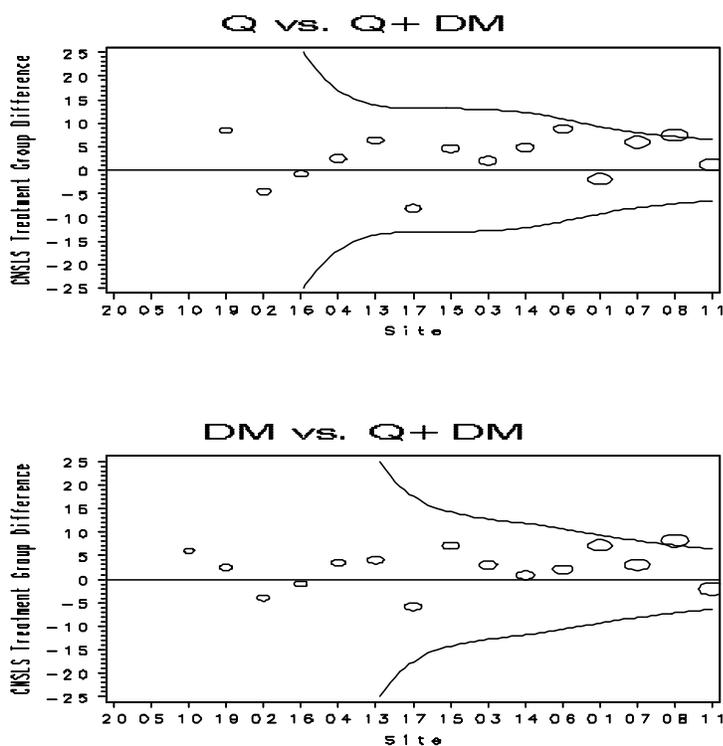
ENDPOINT	POPULATION	GROUP	N	BASELINE	LSMEAN CHANGE (S. E.)	DIFFERENCE LSMEAN (S. E.)	COMPARISON WITH AVP-923 P-VALUE
VAS-Q0L	Last Change	AVP-923	66	34.7	-26.4(2.8)	.	.
		DM	33	44.4	-13.8(3.7)	-12.6(4.5)	0.006
		Q	37	47.3	-13.6(3.5)	-12.8(4.4)	0.004
	Average over Time of Change from Baseline	AVP-923	66	34.7	-24.6(2.4)	.	.
		DM	33	44.4	-13.6(3.3)	-11.0(3.9)	0.006
		Q	37	47.3	-11.1(3.1)	-13.4(3.9)	0.001
	Last Change w/ imputation of no change for dropouts	AVP-923	70	35.1	-23.1(2.8)	.	.
		DM	33	44.4	-12.3(3.9)	-10.8(4.6)	0.020
		Q	37	47.3	-13.3(3.7)	-9.8(4.5)	0.031
VAS-Q0R	Last Change	AVP-923	65	32.0	-24.0(2.5)	.	.
		DM	33	37.8	-8.0(3.5)	-16.0(4.1)	<0.001
		Q	37	42.5	-8.7(3.2)	-15.3(4.0)	<0.001
	Average over Time of Change from Baseline	AVP-923	65	32.0	-23.0(2.3)	.	.
		DM	33	37.8	-7.5(3.2)	-15.5(3.8)	<0.001
		Q	37	42.5	-8.2(3.0)	-14.8(3.7)	<0.001
	Last Change w/ imputation of no change for dropouts	AVP-923	70	31.5	-20.3(2.6)	.	.
		DM	33	37.8	-7.0(3.6)	-13.3(4.3)	0.002
		Q	37	42.5	-8.2(3.4)	-12.1(4.1)	0.004

Effect of Individual Investigators

Excluding site 1 in study 102 results in a p-value of 0.046 for the DM vs. AVP comparison of the Day 29/LOCF changes from baseline in CNSLS. On the basis of the sponsor's pre-specified analysis, of the average of the day 15 and 29 scores, it is 0.013. Site 1 had 15 patients which is 11% of the study total.

A plot of the treatment group differences within each site is presented in Figure 4. Note that the size of the plotting symbol in the plot below indicates the relative size of the site. The sites are sorted on the x-axis according to site size. The upper and lower curves indicate the thresholds for nominal significance of a within site treatment group difference adjusting for the size of the particular site. In about 2/3 of the sites that had patients with post-baseline data in each group the treatment group differences numerically favor AVP-923 over DM and Q. Therefore, the treatment effects seem reasonably consistent across sites.

Figure 4 Study 102: Treatment Group Differences on Change from Baseline in CNSLS (Averaged over Time) within Sites



3.1.2 Study AVR-106

This study was initiated on December 10, 2002 and completed on June 22, 2004.

The objectives of the study were to evaluate and compare to placebo the safety, tolerance, and efficacy of AVP-923 (capsules containing dextromethorphan hydrobromide [30 mg] and quinidine sulfate [30 mg]) for the treatment of pseudobulbar affect in a population of MS patients over a 12-week period.

3.1.2.1 Study Design

This was a multicenter, randomized, double-blind, placebo-controlled study of the treatment of pseudobulbar affect in MS patients with AVP-923 administered orally, two times a day (every 12 hours). Patients were to be randomized to receive either AVP-923 or placebo for 85 days (the last day was to occur anywhere between Day 81 and Day 89). Approximately 25 centers were to be identified. Multiple sclerosis patients thought to exhibit pseudobulbar affect were to be screened for general health (including ECG) within 4 weeks prior to entry into the study. In order to be included in the study, patients must have had clinically diagnosed pseudobulbar affect and have attained a score of 13 or above on the Center for Neurologic Study-Lability Scale (CNS-LS) at the Day 1 clinic visit. The CNS-LS is a 7-item self-report measure that provides a score for total pseudobulbar affect, including assessments of labile laughter and labile tearfulness. The range of possible scores is 7 to 35. On Days 1, 15, 29, 57, and 85, the patients were to be given the CNS-LS to complete and were to be queried regarding any adverse experiences that might have occurred since their prior visit.

Quality of life was to be assessed using two 10-centimeter visual analog scales (Huskisson, 1974). One scale, the VAS-QOL, asked participants to rate how much uncontrollable laughter, tearfulness, or anger had affected the overall quality of their life during the past week, and another scale, the VAS-QOR, asked participants to rate how much uncontrollable laughter, tearfulness, or anger had affected the quality of their relationships with others during the past week. Each scale was to be completed by the patient on the first day of study prior to dosing and on the Day 15, 29, 57, 85 study visits.

For the Pain Intensity Rating Scale, the patients were to indicate the amount of pain experienced within the previous 24 hours using a 5-point Likert scale (none = 0, mild = 1, moderate = 2, severe = 3, extreme = 4). Patients were to complete the Pain Intensity Rating Scale on Days 1, 15, 29, 57, and 85.

The primary efficacy endpoint was to be the CNS-LS score. Secondary efficacy endpoints were patient laughing and crying episode counts, the Visual Analog Scale response for Overall Quality of Life, the Visual Analog Scale response for Quality of Relationships, and the Pain Intensity Rating Scale.

The primary efficacy analysis was to be an intention-to-treat based comparison of average CNS-LS change between the AVP-923 and placebo groups. The primary analysis was to be based on the change in CNS-LS score, where individual change was to be defined as the difference between the baseline scores (Day 1) and the average of the Day 15, Day 29, Day 57 and Day 85 scores (or the non-missing score(s) if one or more was missing). The total CNS-LS score was to be computed if there were at least 5 responses for the 7 items, and missing responses were to be scored as if the patient had answered “1” for those items. If more than two items had a missing response, then the total CNS-LS score was to be considered missing for the visit in question. The statistical analysis plan follows the procedures detailed by Frison and Pocock where the dependent variable is change in CNS-LS (average post-treatment score – baseline score) and the only covariate in the model is the baseline (Day 1) CNS-LS measurement. The model is also adjusted for centers and treatment groups.

Secondary endpoints were to be analyzed using statistical models with a parallel model structure to the one used for the primary endpoint. Episode counts were to include laughing, crying, and laughing plus crying. Episode counts were to be reported and analyzed as a rate, expressed as episodes per week, that is, total number of episodes divided by total number of weeks on treatment. Previous experience (study 102) strongly suggested that a negative binomial regression model with constant dispersion a) would fit the data well b) account for between-subject variability in rates, and c) allow for testing treatment effects with controls for study center differences. The negative binomial model accounts for Poisson overdispersion by allowing the variance of the episode counts to be $\lambda(1+\delta)$, where λ is the expected number of episodes and δ measures the degree of overdispersion due to between-subject variability. Note that if $\delta=0$ the model is close to a Poisson model, like the one proposed for the earlier study.

As a supplement to the secondary analyses reported above, adjustment for multiple comparisons due to multiple endpoints in the secondary analyses were to employ the nonparametric method of O’Brien (1984). The method combines the endpoints into one by summing the ranks on each of the endpoints for each patient. In the sponsor’s opinion this method maintains overall Type I error rates, but has greater power for detecting effects than the Bonferroni method when the treatment affects more than one of the secondary endpoints.

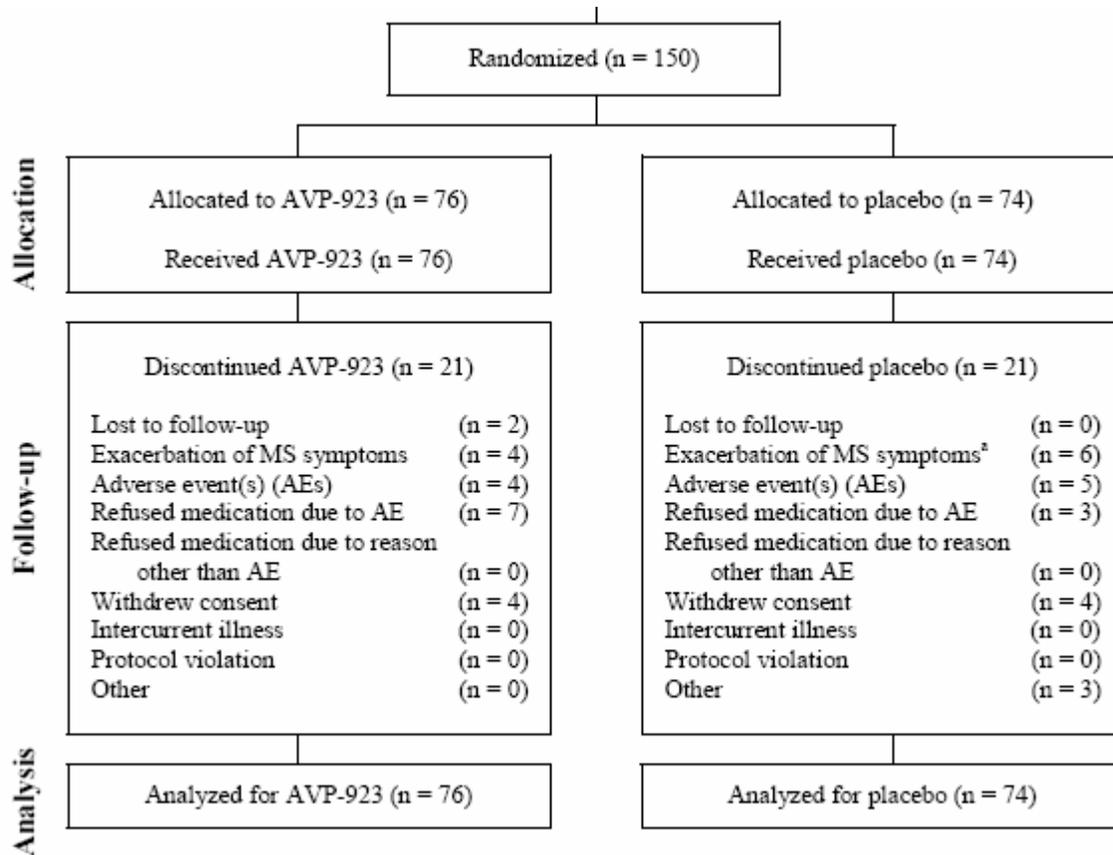
Sample Size Calculations

A sample size of 48 in each randomized treatment group (96 total) is sufficient to detect a difference of 3 points in the primary endpoint (CNS-LS change score, adjusted for baseline and center) with 90% power. If only 36 in each group (72 total) were able to complete the protocol, the study would still have 80% power. These calculations were based on an observed difference of 3.4 units in the adjusted average improvement in CNS-LS scores, with a residual standard deviation of 4.5, drawn from a randomized study of AVP-923 in ALS patients.

3.1.2.2 Patient Disposition

A total of 150 subjects were randomized to treatment, 76 in the AVP-923 group and 74 in the placebo group. The number of subjects who discontinued was 21 (27.6%) in the AVP-923 group and 21 (28.4%) in the placebo group. Reasons for discontinuation are shown in Figure 5.

Figure 5 Study 106: Patient Disposition



3.1.2.3 Patient Demographics

As shown in Table 15 the AVP-923 and placebo groups were comparable with respect to race and gender. The mean age of the AVP-923 group was slightly higher than that of the placebo group but the difference did not reach nominal significance.

Table 15 Study 106: Patient Demographics

Characteristic (unit)	Category or Statistic	AVP-923 (N=76)	Placebo (N=74)	P-value^a
Age (years)	n	76	74	0.1033
	Mean	46.3	43.7	
	SD ^b	9.78	9.95	
	Median	49.0	45.0	
	Min/Max	25/68	21/71	
Gender, n (%)	Male	14 (18.4)	12 (16.2)	0.7214
	Female	62 (81.6)	62 (83.8)	
Race, n (%)	Caucasian	68 (89.5)	68 (91.9)	0.7275
	Black	5 (6.6)	5 (6.8)	
	Asian	1 (1.3)	0 (0.0)	
	Hispanic	2 (2.6)	1 (1.4)	
	Other	0 (0.0)	0 (0.0)	

^a P-values to compare means for continuous variables were computed by using t-tests. P-values for categorical variables were computed by using chi-square tests.

^b SD = Standard deviation.

The treatment groups were similar with regard to number of years with MS ($p = 0.5751$) and frequency of laughing and/or crying episodes ($p = 0.4048$) prior to the study.

Table 16 Study 106: Disease Characteristics

Characteristic	Statistic	AVP-923 (N=76)	Placebo (N=74)	P-value^a
Years with Multiple Sclerosis	n	76	74	0.5751
	Mean	10.3	9.6	
	SD ^b	8.59	7.36	
	Median	7.5	8.0	
	Min/Max	1/40	1/31	
Weekly Episodes of Pathological Laughing and/or Crying	n	75	74	0.4048
	Mean	14.1	17.3	
	SD	20.36	25.24	
	Median	7.0	9.5	
	Min/Max	1/140	1/140	

^a P-values to compare means for continuous variables were computed by using t-tests. P-values for categorical variables were computed by using chi-square tests.

^b SD = Standard deviation.

A total of 11 (4 AVR and 7 placebo) subjects had MS exacerbations during the study and received intravenous steroid treatment. All except for 1 of these subjects were withdrawn from

the study. The remaining subject, in the placebo group (Subject 1901), completed study treatment in violation of the protocol.

The treatment groups were comparable on baseline scores of the various efficacy measures as seen in Table 17.

Table 17 Study 106: Baseline scores on Efficacy Measures

Scale	Statistic	AVP-923 (N=76)	Placebo (N=74)	P-value ^b
Screening CNS-LS ^a	n	71	71	0.3581
	Mean	21.1	22.0	
	SD ^c	5.90	5.18	
	Median	20.0	21.0	
	Min/Max	13/35	13/35	
Day 1 CNS-LS	n	76	74	0.1683
	Mean	20.3	21.4	
	SD	5.02	5.09	
	Median	20.0	22.0	
	Min/Max	13/35	13/35	
Day 1 VAS, ^d Overall Quality of Life	n	76	74	0.4206
	Mean	50.4	54.1	
	SD	28.40	27.49	
	Median	50.0	57.0	
	Min/Max	0/100	2/98	
Day 1 VAS, Quality of Relationships	n	76	74	0.4233
	Mean	45.6	49.2	
	SD	28.76	27.49	
	Median	46.5	48.5	
	Min/Max	2/98	0/100	
Day 1 Pain Intensity Rating Scale	n	76	74	0.8206
	Mean	1.4	1.4	
	SD	1.02	0.99	
	Median	1.0	2.0	
	Min/Max	0/3	0/4	

^a CNS-LS = Center for Neurologic Study - Lability Scale

^b P-values to compare means were computed by using t-tests.

^c SD = Standard deviation.

^d VAS = Visual analog scale.

3.1.2.4 Sponsor's Results

3.1.2.4.1 Primary Analysis

The primary analysis result is shown in Table 18. Subjects who received AVP-923 had a significantly greater decrease in CNS-LS score than subjects who received placebo ($p < 0.0001$). Subjects treated with AVP-923 had an adjusted mean decrease in CNS-LS score more than twice as great as that of subjects on placebo (7.7 points versus 3.3 points, respectively).

Table 18 Study 106: Change in Center for Neurological Study-Lability Scale (CNS-LS) Score— ITT Population

Characteristic	Statistic	AVP-923 (N=76)	Placebo (N=74)	P-value ^b
Change in CNS-LS Score ^a	n	73	74	
	Mean	7.9	4.3	
	SD ^c	5.32	5.26	
	Median	6.5	2.6	
	Min/Max	-1/25	-4/18	
	Adjusted mean ^d	7.7	3.3	< 0.0001
	SE ^e	0.57	0.58	

^a Change in CNS-LS was defined as baseline CNS-LS minus the mean of the scores on Day 15, Day 29, Day 57, and Day 85.

^b P-value was computed by using linear regression according to Frison and Pocock's ANCOVA method and adjusting for baseline CNS-LS and center.

^c SD = Standard deviation.

^d Least-squares means were computed from a regression model for an individual with a CNS-LS of 20 at baseline and the average of center effects.

^e SE = Standard error.

3.1.2.4.2 Secondary Analyses

Subjects treated with AVP-923 experienced approximately half as many episodes of inappropriate laughing, crying, and laughing and crying as subjects receiving placebo; the number of these episodes was significantly lower in the AVP-923 group than the placebo group (all $p \leq 0.0077$).

Table 19 Study 106: Number of Episodes of Inappropriate Laughing and/or Crying —ITT Population

Type of Episode ^a	Statistic	AVP-923 (N=76)	Placebo (N=74)	P-value ^b
Laughing	n	75	73	0.0077
	Mean	2.5	4.8	
	SD ^c	8.36	13.39	
	Median	0.1	0.8	
	Min/Max	0/65	0/106	
Crying	n	75	73	< 0.0001
	Mean	2.2	6.7	
	SD	4.94	10.81	
	Median	0.6	2.8	
	Min/Max	0/34	0/52	
Laughing and Crying	n	75	73	0.0002
	Mean	4.7	11.5	
	SD	10.93	19.43	
	Median	1.3	5.0	
	Min/Max	0/80	0/130	

^a Episode rates were reported as episodes per week, computed as the total number of episodes divided by the total number of weeks on treatment (weeks were computed to the nearest day).

^b P-values were computed by using negative binomial regression.

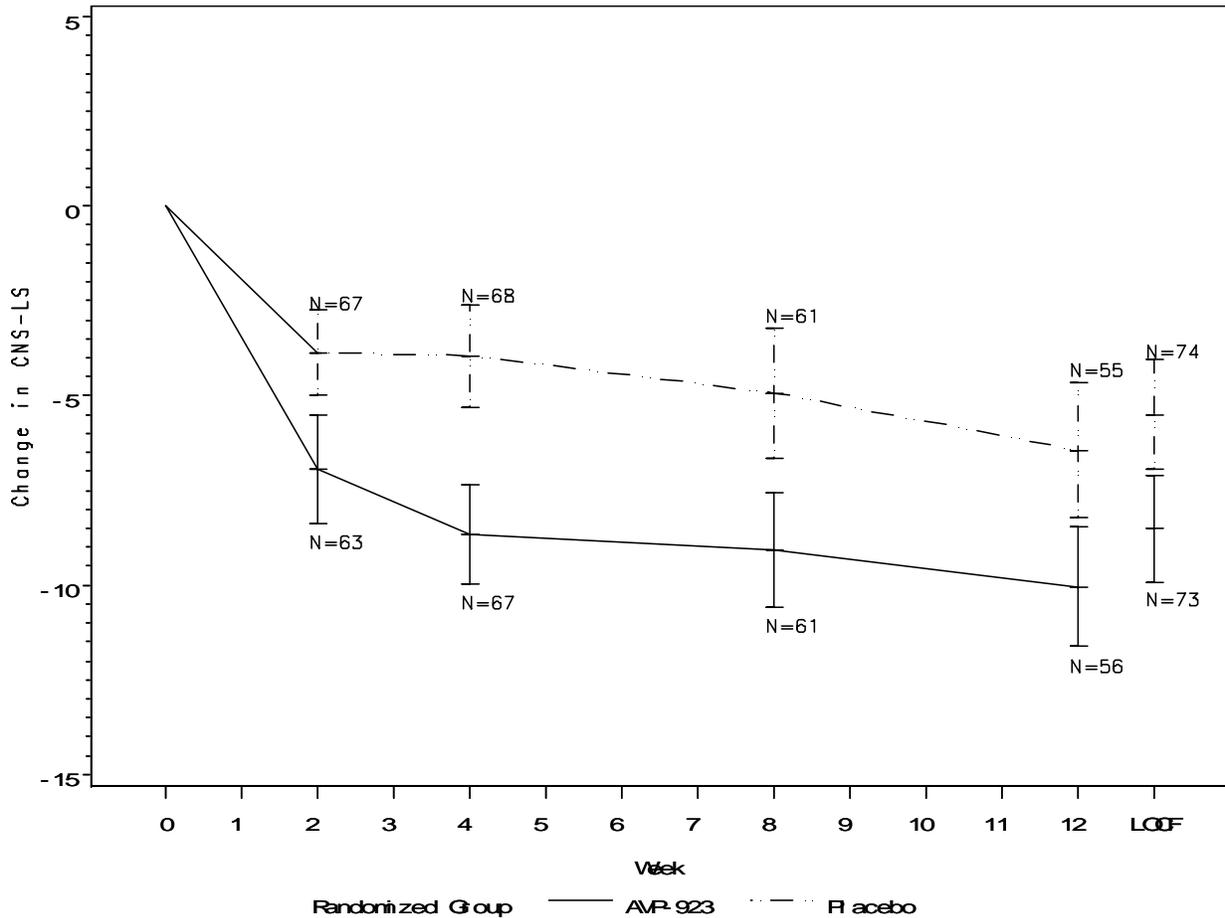
^c SD = Standard deviation.

3.1.2.5 Reviewer's Results

3.1.2.5.1 Primary Analysis

The primary analysis compared the change from baseline averaged over the double blind treatment period between the treatment groups using an ANCOVA model. For the MITT population the difference in group least squares mean changes from baseline in CNSLS averaged over the study period was estimated to be 4.4 points (+/- .74 S.E., $p<0.0001$), using the available period for patients that did not complete the study. For the completers population the difference in group least squares means was estimated to be 4.8 points (+/- .86 S.E., $p<0.0001$). It is more common in clinical trials reviewed by the division of Neurology to base the treatment group comparison on the change from baseline at the end of the study (or last follow-up) instead of averaging the change over the entire double blind treatment period. In this trial the difference in group least squares means based on the change from baseline at the end of the study is estimated to be 3.9 points (+/- .86 S.E., $p<0.0001$) in the MITT population and 4.1 points (+/- .96 S.E., $p<0.0001$) in the completers. Although there is no difference in the significance of the group difference between the change averaged over the entire period and the change at the last measurement, the group difference is slightly larger in the averaged changes than the last changes. A repeated measures model (MMRM) using all the available post-baseline CNSLS scores from each patient also suggests that the difference at the end is smaller than the difference in the averages over the whole period but that it is still significant. The MMRM model estimates the group difference in least squares mean changes at week 12 to be 3.6 points (+/- .89 S.E., $p<0.0001$). Figure 6 shows the observed mean change from baseline in CNSLS scores by visit week for the observed cases.

Figure 6 Study 106: Change from Baseline in CNSLS scores by Visit Week



3.1.2.5.2 Assessment of Sensitivity to Dropouts

Between 25 and 30 percent of patients in each group failed to complete the study. Most of the remaining patients had assessments out to at least day 70. Table 20 shows the primary analysis result and additional analyses for assessing sensitivity to dropouts. Since the conclusions from the MITT-LOCF and Observed case populations, as well as additional analyses, agree the primary analysis result seems reasonably robust to the missing data.

Table 20 Study 106: Primary Analysis and Additional Analyses for Assessing Sensitivity to Dropouts

Analysis Population/Endpoint	AVP-923			PLACEBO			DIFFERENCE (S. E.)	P-value
	N	Baseline Mean	LSMEAN Change (S. E.)	n	Baseline Mean	LSMEAN Change (S. E.)		
MITT-Average Change*	73	20.1	-8.1 (0.6)	74	21.4	-3.7 (0.6)	-4.4 (0.7)	<0.0001
MITT-Last Change (Day 85/ LOCF)	73	20.1	-8.7 (0.6)	74	21.4	-4.8 (0.6)	-4.0 (0.8)	<0.0001
ITT- Impute No change where last assessment before day 70	76	20.3	-7.4 (0.7)	74	21.4	-4.0 (0.7)	-3.4 (1.0)	0.0005
ITT- Impute Worst Change where last assessment before day 70	76	20.3	-5.8 (0.9)	74	21.4	-2.3 (0.9)	-3.5 (1.2)	0.0046
Completers - Change Averaged over Time	55	20.7	-9.7 (0.7)	53	21.5	-4.9 (0.8)	-4.8 (0.9)	<0.0001
Completers - Change (Day 85)	55	20.7	-11.0 (0.8)	53	21.5	-6.8 (0.9)	-4.2 (1.0)	<0.0001

*Difference between Baseline and Average of non-missing CNSLS scores from Days 15, 29, 57, 85
MITT- randomized patients with at least one post-baseline CNSLS measurement

Table 21 shows the mean change from baseline in CNSLS scores at day 85 or last assessment according to reason for dropout. Overall, mean changes in CNSLS were more comparable between the groups for dropouts than for completers. The mean change for AVP-923 dropouts was almost 7 points higher than for AVP-923 completers and the mean change for placebo dropouts was 4 points higher than the mean for placebo completers.

Table 21 Study 106: Mean CNSLS Change at Day 85 or Last Assessment by Termination Reason

COMPLETER	REASON	N	PLACEBO		N	AVP-923	
			Baseline score Mean (S. D.)	Change Mean (S. D.)		Baseline score Mean (S. D.)	Change Mean (S. D.)
NO	LOST TO FOLLOW UP	.	.	.	1	21.0 (2.8)	-1.0 (.)
NO	EXACERBATION OF MS SYMPTOMS	6	24.5 (4.3)	-3.7 (5.6)	4	21.8 (6.3)	-4.0 (2.8)
NO	AE or REFUSAL DUE TO TOXICITY	8	19.9 (5.7)	-1.5 (3.0)	10	18.2 (5.7)	-3.3 (2.2)
NO	WI THDREW CONSENT	4	17.8 (2.6)	-0.8 (2.5)	3	19.0 (5.9)	-3.7 (4.0)
NO	OTHER	3	24.0 (1.0)	-5.0 (1.0)	.	.	.
NO	ALL	21	21.4 (5.0)	-2.5 (3.8)	18	19.3 (5.5)	-3.4 (2.5)
YES	N/A	53	21.5 (5.2)	-6.7 (6.6)	55	20.7 (4.8)	-10.2 (5.9)

3.1.2.5.3 Secondary Analyses

Analysis of Laughing and Crying Episode Counts

According to the negative binomial model pre-specified for the analysis of the sum of the episode counts of the laughing or crying type the AVP-923 group average was estimated to be 54 % of the placebo group average with a 95% C.I. of (38, 75), $p < 0.001$. The coefficient column in Table 22 gives the estimated difference in the natural logarithms of the expected counts. Exponentiating the coefficient for AVP-923 gives the estimated ratio of the AVP-923 rate to the placebo episode rate to be 0.54, as described above.

Table 22 Study 106: Analysis of the Number of Episodes of the Laughing or Crying Type

Model	AIC (Model Fit -smaller is better)	Parameter	Coefficient	Std. Err.	z	P>z	[95% Conf.]	Interval
NB2	1460.870	AVP-923	-0.947	0.237	-4.000	<0.001	-1.411	-0.483
		Alpha*	1.498	0.160			1.215	1.846
NB1(sponsor's primary)	1528.395	AVP-923	-0.623	0.171	-3.650	<0.001	-0.957	-0.289
		Delta*	120.951	18.514			89.601	163.270

* Alpha and Delta are parameters associated with the negative binomial distribution in NB2 and NB1, respectively, $\alpha = 0$ or $\delta = 0$ suggests a Poisson distribution may be appropriate

Recall that two different negative binomial models were evaluated for the study 102 data. The sponsor believed that model NB1, for which the variance is equal to the mean plus a constant (i.e., $V = \mu(1 + \delta)$), was a better fit to the study 102 data. Therefore, they prespecified this model for the analysis of episode counts in study 106. The second negative binomial model (NB2) with variance depending on the mean and the square of the mean (i.e., $V = \mu + \alpha\mu^2$) appears to fit the study 106 data somewhat better than the first negative binomial model as evidenced by its smaller values for the Akaike's Information Criterion (AIC: 1461 vs. 1528) and Bayesian Information Criterion (BIC: 1533 vs. 1600), the negative of twice the log likelihood with a penalty for the number of parameters in the model. However, both negative binomial models agree on the significance of the treatment group difference in favor of AVP-923 ($p < 0.001$).

A nonparametric comparison of the number of laughing and crying episodes per day averaged over the entire study period also yielded a significant result ($p < 0.0001$) in favor of the AVP-923 group. This was true for both the Wilcoxon rank sum test and the center adjusted Cochran Mantel Haenszel test with modified ridit scores (i.e., Van Elteren test). Therefore, the significance of the treatment group difference in the episode counts in study 106 seems to be relatively insensitive to the analysis method.

Other Secondary Endpoints

This reviewer verified the sponsor’s results on the visual analog scale for quality of life (VAS-QOL) and visual analog scale for quality of relationships (VAS-QOR). The results are shown in Table 23. The sponsor’s analysis of the secondary endpoint, pain intensity, which compared the groups on the difference between the baseline score and the average of the post-baseline scores yielded a nominally significant group difference favoring DM+Q (p=0.024). However, this reviewer’s analysis of the change in pain scores at day 85 (or LOCF) did not yield a nominally significant result (p=0.119). In addition to the ANCOVA analysis used by the sponsor this reviewer also investigated a nonparametric, center-stratified, Cochran-Mantel-Haenszel test because the pain score can only take on integer values between 0 and 4 and therefore it (or its change from baseline) may be far from normally distributed. As seen in the following table the p-values for the Cochran-Mantel-Haenszel tests (shown in parentheses) are slightly higher than those based on ANCOVA. Of note, the p-value for the group difference between the average of the post-baseline scores and the baseline score which was nominally significant based on ANCOVA is not for the Cochran-Mantel-Haenszel test.

Table 23 Study 106: Analyses of Other Secondary Endpoints

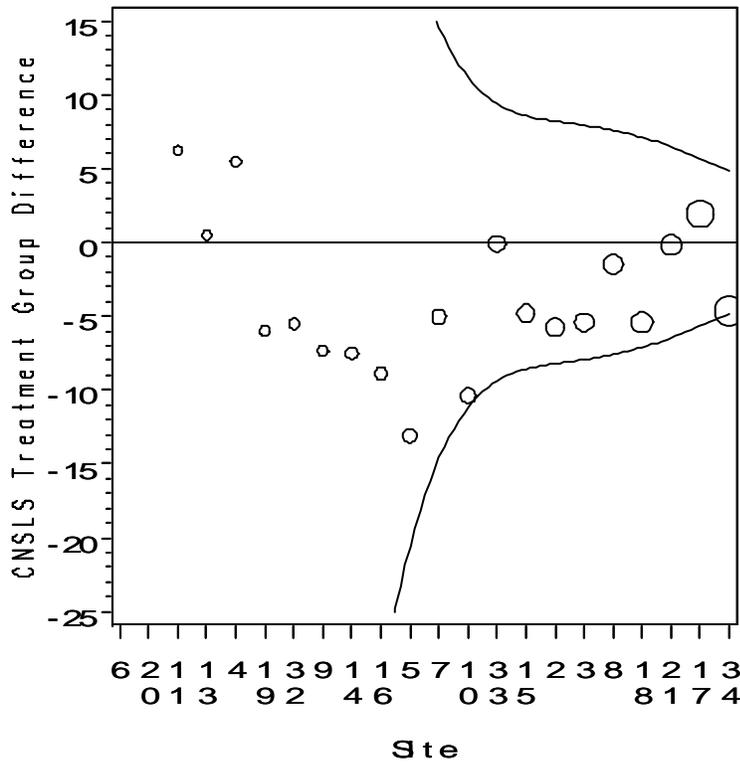
		AVP-923			PLACEBO			DIFFERENCE	
	Analysis Population /Endpoint	N	Baseline Mean	LSMEAN Change	n	Baseline Mean	Lsmean Change (S. E.)	LSMEAN DIFFERENCE (S. E.)	ANCOVA P-value (CMH P-value)
VAS-QOL	Change-Time Average	73	50.4	-33.9(2.7)	74	54.1	-16.8(2.7)	-17.0(3.5)	<0.001
	Change-Day 85 or LOCF	73	50.4	-34.9(3.2)	74	54.1	-20.8(3.2)	-14.1(4.2)	0.001
VAS-QOR	Change-Time Average	73	46.0	-28.9(2.8)	74	49.2	-14.2(2.8)	-14.7(3.6)	<0.001
	Change-Day 85 or LOCF	73	46.0	-29.4(3.2)	74	49.2	-19.8(3.2)	-9.6(4.2)	0.023
Pain	Change-Time Average	73	1.40	-0.46(0.09)	74	1.42	-0.20(0.09)	-0.26(0.11)	0.024 (0.065)
	Change-Day 85 or LOCF	73	1.40	-0.46(0.11)	74	1.42	-0.24(0.11)	-0.22(0.14)	0.119 (0.137)

Effect of Individual Investigators

Centers ranged in size from 1 patient to 22 patients in study 106. The primary analysis result was robust to the exclusion of individual centers.

Figure 7 is a plot of the treatment group differences on the change from baseline in CNSLS within each site. Note that the size of the plotting symbol in the plot below indicates the relative size of the site. The sites are sorted on the x-axis according to site size. The upper and lower curves indicate the thresholds for nominal significance of a within site treatment group difference adjusting for the size of the particular site. Negative differences in the plot favor the AVP-923 group. In the majority of sites the treatment group difference favored the AVP-923 group over placebo.

Figure 7 Study 106: Treatment Group Differences in Change in CNSLS (Averaged over Time) by Site



3.2 Evaluation of Safety

Safety is not evaluated in this review. Please see the clinical review(s) for the evaluation of safety.

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

4.1 Gender, Race and Age

This section contains this reviewer’s summary statistics for gender, race, and age subgroups. The studies were not adequately powered to estimate treatment differences in subgroups precisely and reported p-values should be interpreted cautiously as they have not been adjusted for multiple comparisons.

Gender

As shown in Table 24 there was no compelling evidence in study 102 of a difference in treatment effects as a function of gender. Thirty nine percent of patients were female. The small sample sizes limit the reliability of the estimates.

Table 24 Study 102: Analysis of Last Change in CNSLS by Gender

	MALE			FEMALE			ALL	
	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)
AVP	41	-7.2 (0.9)	.	25	-7.9 (1.1)	.	66	-7.5 (1.0)
DM	18	-4.2 (1.2)	0.032	15	-5.1 (1.3)	0.079	33	-4.6 (1.3)
Q	22	-3.9 (1.1)	0.012	15	-4.0 (1.4)	0.017	37	-3.9 (1.3)

In study 106, 83% of patients were female. Although in study 106 the difference between the AVP-923 group and the placebo group was smaller and not nominally significant in males, as displayed in Table 25, this may be due to the smaller sample size.

Table 25 Study 106: Analysis of Last Change in CNSLS by Gender (LSMean)

TREAT	MALE			FEMALE			ALL	
	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)
PLACEBO	12	-6.6 (1.6)	.	62	-4.7 (0.8)	.	74	-5.0 (1.0)
AVP	14	-8.5 (1.5)	0.351	59	-9.1 (0.8)	<0.001	73	-9.0 (1.0)

Age

In study AVR102 the median age was about 55 and about 25% of patients were 65 or older and 25% were 45 or younger. Table 26 shows results for the mean change from baseline in CNSLS score at day 29 or LOCF by age group (Age < 65 vs. Age ≥ 65). There was no compelling evidence that the treatment effects depended on age. If one assumes a constant slope for the relationship between change in CNSLS and age and one tests for unequal group slopes the resulting p-values for slope differences between the AVP-923 and DM and AVP-923 and Q slopes are p=0.45 and p=0.34, respectively. This provides further evidence that the group differences do not depend on age.

Table 26 Study 102: Analysis of Last Change in CNSLS by Age Group (<65 vs. ≥ 65)

Group	AGE < 65			AGE ≥ 65			ALL	
	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)
AVP-923	48	-6.5 (0.8)	.	18	-8.7 (1.2)	.	66	-7.1 (0.9)
DextroMethorphan	25	-4.4 (1.0)	0.067	8	-3.9 (1.8)	0.021	33	-4.3 (1.3)
Quinidine	30	-2.9 (1.0)	0.001	7	-7.0 (1.9)	0.438	37	-3.7 (1.3)

The average age in study 106 was 45 and ages ranged from 21 to 71. Study 106 provides no compelling evidence that the treatment group difference depends on age. Table 27 shows results for the mean change from baseline in CNSLS score at day 85 or LOCF by age group (Age ≤ 45 vs. Age > 45). Since less than 5% of the patients were age 65 or older no meaningful estimates of the treatment difference in such patients can be obtained. If one assumes a constant slope for the change in CNSLS age relationship and tests for unequal group slopes the resulting p-value is 0.90. This provides further evidence that the group differences do not depend on age.

Table 27 Study 106: Analysis of Last Change in CNSLS by Age Group (LSMeans)

TREAT	AGE ≤ 45			AGE > 45			ALL	
	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)	Comparison w/ AVP-923 P-value	N	LSMEAN (SE)
Placebo	38	-5.3 (0.9)	.	36	-4.8 (1.0)	.	74	-5.1 (1.0)
AVP-923	30	-9.6 (1.1)	0.001	43	-8.4 (1.0)	0.004	73	-8.9 (1.1)

Race

In study AVR102 89% of patients were White, 8% were Hispanic, and others (Asians, Blacks, and others) accounted for 3%. Sample sizes were too small for estimates of treatment effects to be reliable in non-white races. Nevertheless, mean changes from baseline in CNSLS score at Day 29 or LOCF are shown in Table 28.

Table 28 Study 102: Analysis of Last Change in CNSLS by Race

TREAT	WHI TE			HI SPANIC			OTHER			ALL	
	N	LSMEAN (SE)	Compari son w/ AVP-923 P-val ue	N	LSMEAN (SE)	Compari son w/ AVP-923 P-val ue	N	LSMEAN (SE)	Compari son w/ AVP-923 P-val ue	N	LSMEAN (SE)
AVP-923	59	-7.9 (0.7)	.	5	-3.0 (2.3)	.	2	-6.2 (3.4)	.	66	-7.5 (1.0)
DextroMethorphan	28	-4.3 (1.0)	0.001	3	-5.2 (2.8)	0.546	2	-9.3 (3.4)	0.517	33	-4.7 (1.4)
Qui ni di ne	34	-4.0 (0.9)	<0.001	3	-4.0 (2.8)	0.782	.	.	.	37	-4.0 (1.2)

In study AVR106 91% were white so there is not sufficient data to reliably estimate the treatment effect for other races or contrast the treatment effects among the races. Nevertheless, mean changes from baseline in CNSLS score at Day 29 or LOCF are shown in Table 29.

Table 29 Study 106: Analysis of Last Change in CNSLS by Race

TREAT	WHI TE			OTHER			ALL	
	N	LSMEAN (SE)	Compari son w/ AVP-923 P-val ue	N	LSMEAN (SE)	Compari son w/ AVP-923 P-val ue	N	LSMEAN (SE)
Pl acebo	68	-5.2 (0.7)	.	6	-3.8 (2.3)	.	74	-5.1 (0.9)
AVP-923	66	-9.0 (0.8)	<0.001	7	-8.1 (2.3)	0.142	73	-8.9 (1.0)

4.2 Other Special/Subgroup Populations

No other special/subgroup populations were investigated.

5 SUMMARY AND CONCLUSIONS

5.1 Statistical Issues and Collective Evidence

In study 102 the AVP-923 group had 8 (11%) patients drop out due to toxicity before the week 2 assessment, whereas the DM and Q groups had no dropouts before week 2. Four of these 8 AVP-923 patients had an early post-baseline assessment and 4 did not. The latter 4 patients were the only patients who did not have any post-baseline CNSLS measures but all of them were members of the AVP-923 group (4/70=5.7%). Six other AVP-923 patients dropped out due to toxicity before the week 4 assessment and one died due to ALS complications, according to the sponsor. Note that 4 other AVP-923 patients and 1 Quinidine patient were not considered to be completers by the sponsor, despite having CNSLS assessments at or near days 15 and 29, because they refused to take the medication due to toxicity.

Average baseline scores on the primary efficacy measure, the CNSLS, were 20 for the AVP-923 group, 21 for DM, and 22 for Q (possible range is 0 to 28). Both single component groups had slightly worse scores at baseline and the AVP-923 vs. Q comparison of the baseline CNSLS scores approached nominal significance ($p=0.065$). A similar trend was observed for the Visual Analog Scale quality of life (VAS QOL) and quality of relationships (VAS QOR) ratings at baseline. The global test for any differences among the three VAS QOL means was nominally significant ($p=0.024$). The Q group was 12.2 points higher than the AVP-923 group on the quality of life VAS ($p=0.011$) and 11 points higher on the quality of relationships VAS ($p=0.039$). The Q group also had a higher percentage of patients with the bulbar (as opposed to limbic) type of ALS than the AVP-923 group (62% vs. 43% $p=0.057$). In the presence of baseline differences on variables associated with an efficacy measure the reported treatment group differences on that measure may not be due to the treatment alone.

Based on the primary analysis which was a site, treatment group, and baseline adjusted ANCOVA of the difference between the baseline and the average of the day 15 and 29 CNSLS scores the comparison between the AVP-923 group and the DM group is significant ($p=0.001$) as is the AVP-923 vs. Q group comparison ($p<0.0001$). The primary analysis utilized the last observation carried forward for those patients with only one post-baseline efficacy assessment. A mixed model analysis of repeated measures using all observed post-baseline CNSLS data and an analysis restricted to the completers population supported the primary analysis results.

Carrying baseline forward is usually discouraged as a method for imputing missing data in the division of Neurologic drugs because it can lead to underestimating the variance of the group difference and thus to a biased test. In study 102 there were 4 patients in the combination group with no post-baseline primary efficacy measures as compared to 0 in the other groups. Usually one focuses on the ITT population modified to exclude these patients as long as they are few in number and not all in one group. Since they are all in one group and the sample sizes are small in this case it is important to assess their potential impact on the results and carrying their baseline

scores forward is one way to accomplish this. In study 102 if we impute no change, i.e., carry the baseline forward, for those who were last assessed on the CNSLS before day 23, i.e., more than a week before the intended final assessment time, and for those who had no post-baseline CNSLS assessments (4 patients - all DM/Q) the p-value for the DM/Q vs. DM comparison increases to 0.083 and that for the AVP-923 vs. Q comparison increases to 0.005. Therefore, the significance of the primary analysis result may be affected by changing assumptions regarding the dropouts. For the sake of completeness, if we focus on the usual MITT population where these 4 AVP-923 patients are excluded then the 0.083 p-value for the DM comparison reduces to 0.042. Instead of carrying the baseline forward we could use the more traditional approach of carrying the last observation forward for dropouts with some post-baseline CNSLS scores and examine the effect of a worst case like imputation for the 4 patients with no post-baseline CNSLS scores. In particular, if we impute a change from baseline of +5 for the 4 AVP-923 dropouts with no post-baseline CNSLS scores, which is one point worse than the worst observed change, then the resulting p-values are 0.056 for the AVP-923 vs. DM comparison and <0.05 for the AVP-923 vs. Q comparison. In this reviewer's opinion considering that it is a p-value from a worst-case analysis the AVP-923 vs. DM p-value of 0.056 may be close enough to 0.05 in this case. Therefore, the primary analysis result in study 102 doesn't seem too sensitive to several reasonable assumptions regarding the missing data.

In study 102 the sponsor excluded patients that were randomized but were poor metabolizers from the primary analysis, as stipulated in the statistical analysis plan. There were 5 (7%) AVP-923 patients, 3 (9%) DM, and 3 (8%) Q patients that were determined to be poor metabolizers of cytochrome P450 2D6. The primary analysis result is not sensitive to the inclusion of these patients.

The results for the analysis of the counts of all episodes of the laughing or crying type based on the sponsor's prespecified analysis of the episodes are not robust and there is evidence that the assumptions of the model are not satisfied. The observed distribution of the number of episodes does not fit the Poisson distribution proposed by the sponsor for the analysis of episodes in study 102. The sponsor acknowledged this and prespecified a more appropriate negative binomial model instead of the Poisson model for the analysis of episodes in the following study (106). Numerous alternatives to the Poisson model fail to detect a group difference between AVP-923 and DM in the average number of laughing and crying episodes per week in study 102. This endpoint was designated as secondary by the sponsor but the division indicated a preference for it being primary in meetings with the sponsor. There is no established precedent for a primary endpoint because this is a new indication. If the division still held its initial preference for the episode count endpoint over the CNSLS then the interpretation of the study outcome could be quite different.

Although the CNSLS, the primary endpoint, contains both laughing and crying items if one considers only the laughing items of the CNSLS (4 of the 7 CNSLS items) in study 102 then the group differences between Avanir and Quinidine and Avanir and Dextromethorphan are not statistically significant. Also, analyses of the episodes of laughing recorded in patients' diaries fail to detect a significant difference between Avanir and Quinidine or Avanir and

Dextromethorphan. Results from the analysis of change from baseline in the sum of the CNSLS crying items and the analysis of crying episode counts were also concordant, but in contrast to the laughing results the crying results reached nominal significance. In the placebo controlled study Avanir was significantly better than placebo on the change in the sum of the CNSLS laughing items and the laughing episodes counts, but since this is a combination drug the placebo-controlled trial result doesn't rule out the possibility that one of the components of the combination is enough for laughing episodes.

The sponsor used a non-parametric O'Brien test in an effort to control the type I error for the secondary endpoints. The O'Brien test combines the patient's ranks on each of the endpoints into a single measure (sum of the patient's ranks on each endpoint) and thus requires only one test. The problem with the O'Brien test is that it doesn't indicate which secondary endpoints are significant, only that some combination or composite of them is. There is also a question of whether or not all of the secondary endpoints provide information that is distinct enough from the primary efficacy measure. Secondary endpoints include number of episodes of crying and number of episodes of laughing, Visual Analog scale score for Quality of Life, and Visual Analog scale score for Quality of Relationships. The Pain intensity rating scale was an additional secondary endpoint in study 106 only. The sponsor reported that the O'Brien test was significant for both studies. However, it doesn't indicate which endpoints are significant so it doesn't really avoid the multiplicity problem. In fact, in study 102 this reviewer found a lack of clear significance between the AVP-923 and DM groups in terms of the sums of the episode counts of the laughing or crying type and in study 106 this reviewer found that the AVP-923 vs. placebo comparison on the change from baseline to the end of the study on the pain intensity rating scale was not nominally significant. So the significance of the secondary endpoints depends on the multiplicity adjustment method and the sponsor did not choose an appropriate one.

In study 106, 74 patients were randomized to AVP-923 and 76 were randomized to placebo. There were 21 dropouts in each group (about 28% for each groups). The average Baseline CNSLS score was 21 (the possible range is 7 to 35). For the ITT population, excluding those with no post-baseline CNSLS scores, the difference in group least squares mean changes from baseline in CNSLS score averaged over all available post-baseline visits was estimated to be 4.4 points (+/- .74 S.E., $p < 0.0001$). If the comparison was based on the change from baseline at day 85 (or LOCF), instead of averaging over the entire period as prespecified by the sponsor, the group difference was slightly smaller but still statistically significant: 3.9 points (+/- .86 S.E., $p < 0.0001$). These results seem to be robust to several reasonable assumptions regarding missing data since analysis of the completers population and a mixed model repeated measures analysis still resulted in nominally significant p-values. Therefore, study 106 seems to support the superiority of AVP-923 to placebo for treating pseudobulbar affect in MS patients.

Although this drug is a combination of two drugs the sponsor has conducted only one study comparing the combination to each of the single components. Ideally, a drug combination should be demonstrated statistically significantly superior to each of its components in two studies.

5.2 Conclusions and Recommendations

Efficacy data on pseudobulbar affect from a study in Multiple Sclerosis (MS) patients showed that the AVP-923 combination of 30 mg Dextromethorphan and 30 mg Quinidine was significantly better than placebo in treating pseudobulbar affect in the study. An earlier study conducted in Amyotrophic Lateral Sclerosis (ALS) patients with pseudobulbar affect compared the same combination of Dextromethorphan and Quinidine to the individual components of the combination. By design this study had a shorter follow-up (1 month) than what is normally expected in ALS patients and the company did not follow the division's advice to lengthen the follow-up. In addition, while the combination was significantly better than the components on the primary efficacy measure, change from baseline in Center for Neurologic Study-Lability Scale (CNS-LS) score, it was not clearly significantly better in terms of the analysis of the laughing and crying episode counts which the agency had encouraged the company to use as the primary efficacy measure. The sponsor's statistician correctly reported that an assumption underlying the sponsor's prespecified method for the analysis of the episode counts (sponsor designated as a secondary endpoint) was not supported by the study data and that it is well known that ignoring this fact would lead to p-values that are misleadingly small. No back-up analysis method was specified in the protocol. Several reasonable alternatives to the prespecified method failed to find a significant difference while one other method advocated by the sponsor did. There are no precedents for primary endpoints in pseudobulbar affect because it is a new indication. If one deems the sponsor's pre-specified primary endpoint as a valid endpoint for the indication then the ALS study suggests that the combination is superior in terms of efficacy to each of its individual components for pseudobulbar affect in ALS patients after up to one month of treatment. However, the p-value of 0.001 for the primary analysis seems to be optimistic since it excludes 4 patients with no post-baseline efficacy measures all of whom were in the combination group and some sensitivity analyses including these patients result in p-values greater than 0.05 (see section 1.3 for details). Therefore, while the study is considered positive it may not have the strength and robustness one would expect in the case where there is only one study comparing the combination to each of its components. The placebo controlled study in MS patients with pseudobulbar affect lends some support to the efficacy of the drug combination but only relative to placebo, i.e., not relative to the individual components of the combination because they were not included in the design.

I. Appendix

The Effect of Overdispersion on the Poisson Model for the Episode Counts

If the episode count, Y , has a Poisson distribution then the expected count $E(Y)=\text{Var}(Y)=\mu$. The Poisson model is given by $\log(E(Y))=X\beta$ where X is a matrix of explanatory variables and β is a vector of associated parameters.

The Pearson Chi square is a measure of the fit of the model. It is given by $\sum\{(y_i - \mu_i)^2\}/V(\mu_i)$ where i is the observation number, μ_i is the expected episode count for patient number i and $V(\mu_i)$ is the dependence of the variance of y_i on μ_i . The Variance of Y equals $\phi*V(\mu_i)$. Thus, the Pearson Chi square divided by the degrees of freedom estimates ϕ . When ϕ is greater than 1 so that the counts are more variable than predicted by the Poisson model the data is said to be overdispersed. The jk -th entry of the information matrix, I , the inverse of which gives the covariance matrix of the parameter estimates, is given by $I_{jk} = \sum_{i=1}^n x_{ji} x_{ik} / (\phi V(\mu_i) g'(\eta_i)^2)$. Here x_{ji} and x_{ik} are the values of explanatory variables j and k for the i -th patient and $g'(\eta_i)$ is the derivative of the link function, which links the expected value of Y to the linear predictor, $\eta=X\beta$. Because ϕ is in the denominator the information, I , decreases as the scale, ϕ , increases and thus the variance (inverse of I) increases as the scale increases. Thus, if the scale which is assumed to be 1 for the Poisson model is actually greater than 1 then the standard errors will be too small for the Poisson model. This leads to oversignificance of tests since the test statistics are proportional to $1/\text{standard error}$ of the relevant parameter. For the Poisson model of the episodes of the laughing or crying type prespecified by the sponsor the scale, estimated on the basis of the Pearson chi square, is 208 which means that standard errors reported for the Poisson model are as much as 14 times too small.

**This is a representation of an electronic record that was signed electronically and
this page is the manifestation of the electronic signature.**

/s/

Tristan Massie
8/31/2006 02:10:40 PM
BIOMETRICS

Kun Jin
8/31/2006 03:11:43 PM
BIOMETRICS

James Hung
8/31/2006 05:59:39 PM
BIOMETRICS



U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Pharmacoepidemiology and Statistical Science
Office of Biostatistics

Statistical Review and Evaluation

CARCINOGENICITY STUDIES

IND/NDA Number: NDA 21-879

Drug Name: Dextromethorphan hydrobromide and Quinidine sulphate
(Neurodex)

Indication(s): 26 Week Carcinogenicity in Transgenic Mice

Applicant: Avanir Pharmaceuticals
11388 Sorrento Valley Road, Suite 200
San Diego, California 92121

Documents Reviewed: Submission submitted electronically
Data submitted electronically

Review Priority: Standard

Biometrics Division: Division of Biometrics VI

Statistical Reviewer: Mohammad Atiar Rahman, Ph.D.

Concurring Reviewer: Karl Lin, Ph.D.

Medical Division: Division of Neurology Products

Reviewing Pharmacologist: Kathleen Young, Ph.D.

Project Manager: Melina Griffis

Keywords: Carcinogenicity, Dose-Response

1. Background

In this submission the sponsor included a report of an animal carcinogenicity study in hemizygous Tg.rasH2 mice. This study was intended to assess the carcinogenic potential of dextromethorphan hydrobromide (DM) and quinidine sulfate (Q) dosed individually or in combination for 26 weeks. The results of this review have been discussed with the reviewing Pharmacologist Dr. Young.

2. Design

Two separate experiments, one in males and one in females were conducted. In each of these two experiments there were five treated groups along with a positive control and a vehicle control group. The treated groups were (1) Low dose DM/Q combination (25/50 mg/kg/day), (2) Mid dose DM/Q combination (50/50 mg/kg/day), (3) High dose DM/Q combination (100/100 mg/kg/day), (4) High dose DM only (100 mg/kg/day), (5) High dose Q only (100 mg/kg/day). The selected doses were administered by oral gavage once daily for 7 days per week for 26 weeks in 10 mL vehicle/kg body weight. The positive control group received three (Day 1, 2, and 5) intraperitoneal injection of urethane (1000 mg/kg) in saline at a dosage volume of 10 ml/kg during study week 1, while the vehicle control received the vehicle (1% methyl cellulose) by gavage. The purpose of the positive control was to assess the sensitivity of the study. One hundred and seventy five Tg.rasH2 mice of each sex were randomly allocated to the control and treated groups of equal size of 25 animals. Animals were housed two per cage while on quarantine and individually while on the test.

The animals were observed twice daily for mortality and palpation, and were examined weekly for clinical signs of toxicity. A complete histopathological examination was performed on all animals found dead, killed moribund, or sacrificed during or at the end of the experiment. Body weights were determined weekly for the first 13 weeks and then once in every two weeks thereafter.

2.1 Sponsor's analyses

Survival analysis: The sponsor presented the summary tables of animals died during the study and at terminal sacrifice, and performed pairwise comparisons of proportions of death in the treated groups with vehicle control using the Fisher's Exact Test.

The sponsor's analyses showed that the spontaneous mortality in the treated groups was limited to 1 to 3 animals per group in both sexes. As a result of the increased sensitivity of Tg.rasH2 mice to carcinogenicity, treatment with positive control article resulted in increased tumor related morbidity. The surviving positive control animals were sacrificed on Day 120 once the majority of the animals exhibited clinical signs (rapid or shallow breathing) associated with target organs (lungs and spleen). Tests did not show statistically significant differences in mortality between any of the treated groups compared to the vehicle control in either sex.

Tumor data analysis: The sponsor performed pairwise comparisons of proportions of animals with tumor in the treated groups with vehicle control group using the Fisher's Exact Test. No adjustment procedure for multiple testing was mentioned in the sponsor's analysis. The test did not show statistically significant differences (at $\alpha=0.05$) in the incidence of any tumor type in the treated group compared to the vehicle control group in either sex. However, statistically significant increased incidences of adenoma and carcinoma in lung, and hemangiosarcomas in spleen were shown in the positive control compared to the vehicle control in both sexes.

2.2 Reviewer's analyses

This reviewer independently performed survival and tumor data analyses. Survival data were analyzed using the log-rank (Cox, Regression models and life tables, *Journal of the Royal Statistical Society*, B, 34, 187-220, 1972) and Wilcoxon (Wilcoxon, A generalized Wilcoxon test for comparing arbitrarily singly censored samples, *Biometrika*, 52, 203-223, 1965) tests. The tumor data analyses were performed using the Poly-K method (Bailer and Portier, Effects of Treatment-Induced Mortality and Tumor-Induced Mortality on Test for Carcinogenicity on Small Samples, *Biometrics* 44, 417-431, June 1988). Data used in this reviewer's analyses were provided by the sponsor electronically.

Survival analysis: The intercurrent mortality data are given in Tables 1A and 1B for males and females, respectively. The Kaplan-Meier survival curves are given in Figures 1A and 1B for males and females, respectively. The homogeneity of survival distributions of the vehicle control, the three combination dose groups, and the two mono dose groups was tested separately for males and females using the log-rank and the Wilcoxon tests. Results of the tests are given in Tables 2A and 2B for males and females, respectively. The tests did not show statistically significant (at 0.05 level) differences in survival distribution across treatment groups in either sex.

Tumor data analysis: There were two controls in this study namely, vehicle and positive controls. The role of the outcomes of the positive control was to assess the sensitivity of the study and not to be compared with the outcomes from the treated groups. Outcomes from the vehicle control group were intended to be compared with those from the treated groups. Therefore, in this reviewer's analysis all tests were performed with respect to the vehicle control only.

There were three increasing doses of DM/Q combination namely, DM/Q 25/50 mg/kg/day, DM/Q 50/50 mg/kg/day, and DM/Q 100/100 mg/kg/day. A dose-response analysis of these treatment groups may be of interested. Therefore, this reviewer performed a dose-response analysis for these three combination doses along with the vehicle control using the Poly-k method. This reviewer also made pairwise comparisons of all treated groups with the vehicle control using the Fisher's exact test.

One critical point for Poly-k test is the choice of the appropriate value of k. For long term 104 week standard rat and mouse studies, a value of k=3 is suggested in the literature. However, the present submission is a 26 week study in transgenic mouse. Unlike the 104 week standard rat and mouse studies, there is no suggested appropriate value of k available in the literature for 26 weeks study in transgenic mice. Because of this situation, in this analysis this reviewer tried multiple values of k namely, k=1, 3, and 6. Besides these three values of k, another analysis was also performed, where the value of k is chosen by the program using a boots trap technique. Since, the calculated p-values from k=1, 3 and 6 were approximately same (at least up to two decimal points) in this review this reviewer reported only p-value for k=3. The p-values from the boots trap method were considerably different from p-values with k=1, 3 or 6. P -values from bootstrap method were also were reported.

Sponsor's analysis showed that in both males and females there were significant imbalances in animals' body weight gains. It was suspected by the reviewing pharmacologist that this imbalance in body weight gain could have impacted the initiation of tumors. The heavier animals might have a higher chance than the lighter animals. To address this concern of the pharmacologist, this reviewer reanalyzed data of some selected tumor types after adjusting for the body weight gain. In this re-analysis pairwise comparisons of treated groups with the vehicle control were performed by stratifying the body weight gain into 4 strata namely, (i) body weight gain ≤ 0 gm,

(ii) $0 \text{ gm} < \text{body weight gain} \leq 5 \text{ gm}$, iii) $5 \text{ gm} < \text{body weight gain} \leq 10 \text{ gm}$, and iv) $10 \text{ gm} < \text{body weight gain} \leq 15 \text{ gm}$. P-values were calculated combining data from all strata using the exact permutation test.

Multiple testing adjustments: For the adjustment of multiple testing this reviewer used the Hochberg procedure. In this method the largest p-value from all tested tumor types is first compared to $\alpha=0.05$. If this test is found to be significant (i.e. $p < \alpha$) then results of all tested tumor types are considered to be significant. If this test is found to be not significant then the second largest p-value from all tested tumor types is compared to $\alpha=0.05/2$. If the test is found to be significant then the results of all tested tumor types except the tumor type already tested for significance are considered to be significant. This process is continued stepwise for the next ordered p-values with the k^{th} largest p-value from all tested tumor types being compared to $\alpha=0.05/K$. This method of multiple testing is applied separately to tests for dose-response and pairwise comparisons and also separately for each gender.

This reviewer's analyses results for dose-response and body weight unadjusted pairwise comparisons are given in Tables 3A and 3B for males and females, respectively. Results from the body weight adjusted pairwise comparisons are given in Table 4A and 4B. Based on this reviewer's analyses results and using the Hochberg's method of adjusting for multiplicity of testing, the dose-response of the three combination dose groups in none of the tested tumor type was found to be statistically significant. Also none of the body weight unadjusted or body weight adjusted pairwise comparisons of the treated groups with the vehicle control was found to be statistically significant.

4. Summary

In this submission the sponsor included a report of an animal carcinogenicity study in hemizygous Tg.rasH2 mice. This study was intended to assess the carcinogenic potential of dextromethorphan hydrobromide (DM) and quinidine sulfate (Q) dosed individually or in combination for 26 weeks.

In this review, the phrase "Dose-response" refers to the linear component of the effect of treatment, and not necessarily to a strictly increasing or decreasing mortality or tumor rates as dose increases.

This study had five treated groups along with a positive control and a vehicle control group. The treated groups were (1) Low dose DM/Q combination (25/50 mg/kg/day), (2) Mid dose DM/Q combination (50/50 mg/kg/day), (3) High dose DM/Q combination (100/100 mg/kg/day), (4) High dose DM only (100 mg/kg/day), (5) High dose Q only (100 mg/kg/day).

The tests on survival data did not showed statistically significant difference in survival distributions across treatment groups in males or females. Based on this reviewer's analyses results and using the Hochberg's method of adjusting for multiplicity of testing, the dose-response of the three combination dose groups in none of the tested tumor type was found to be statistically significant. Also none of the body weight unadjusted or body weight adjusted pairwise comparisons of the treated groups with the vehicle control was found to be statistically significant.

M. Atiar Rahman, Ph.D.
Mathematical Statistician

Concur: Karl Lin, Ph.D.
Team Leader, Biometrics VI

cc:

Archival NDA 21-879

Dr. Freed

Dr. Katz

Dr. Bryan

Dr. Farkas

Ms. Griffis

Dr. Machado

Dr. Huque

Dr. Lin

Dr. Rahman

Dr. Jin

Dr. O'Neill

Ms. Patrician

**Table 1B: Intercurrent Mortality Rate
Female Mice**

Week	Dose group								Total	
Frequency	Vehicle	Positive, DM/Q	DM/Q	DM/Q 100, DM	Q					
Col Pct	Control	Control	25/50 mg.	50/50 mg.	/100mg	100 mg	100 mg			
0<=Week<10	0	0	0	0	1	0	0	0.00	0.00	1
10<=Week<15	0	2	0	0	0	0	1	0.00	8.00	3
15<=Week<20	0	23	1	0	0	1	1	0.00	92.00	26
20<=Week<26	1	0	0	1	0	0	0	4.00	0.00	2
Ter. Sac.	24	0	24	24	24	24	23	96.00	0.00	143
Total	25	25	25	25	25	25	25			175

**Table 2A: Intercurrent Mortality Comparison
Male Mice (Without the Positive Control)**

Method	Test	Statistic	P-value
Log-Rank	Homogeneity	7.89	0.1619
Wilcoxon	Homogeneity	7.86	0.1636

**Table 2B: Intercurrent Mortality Comparison
Female Mice (Without the Positive Control)**

Method	Time adjusted Trend test	Statistic	P-value
Log-Rank	Homogeneity	0.91	0.9696
Wilcoxon	Homogeneity	0.87	0.9727

Table 3A
Tumor Rates, Dose Response, and Pairwise Comparisons of Tested Tumors
Male Mouse

ORGANNAME	TUMORNAME	Treatment Groups*							Dose-Response**		Pairwise Comparisons***				
		_1	_2	_3	_4	_5	_6	_7	Boots	K=3	_1vs._3	_1vs._4	_1vs._5	_1vs._6	_1vs._7
Adrenal glands	Adenoma	1/24	0/25	2/23	2/24	1/24	1/25	0/25	0.468	0.460	1.000	1.000	1.000	1.000	1.000
Bone, mandibular	Hemangiosarcoma	0/25	1/25	0/25	0/25	0/25	0/25	0/25
Cavity, nasal	Hemangiosarcoma	0/25	0/25	0/25	0/25	0/25	1/25	0/25
Eyes	Hemangiosarcoma	0/25	0/25	0/25	0/25	0/25	1/25	0/25
Harderian glands	Adenoma	0/25	0/25	0/25	2/23	0/25	0/25	0/25	0.094	0.252	.	0.500	.	.	.
Intestine, ileum	Hemangiosarcoma	0/25	0/25	1/24	0/25	0/25	0/25	0/25	0.958	0.571	1.000
Liver	Hemangiosarcoma	0/25	0/25	1/25	0/25	0/25	0/25	0/25	0.722	0.572	1.000
Lungs with bronchi	Adenoma	1/24	24/24	5/23	1/25	1/24	4/23	0/25	0.804	0.612	0.190	1.000	1.000	0.349	1.000
Lungs with bronchi	Carcinoma	0/25	8/25	1/25	1/24	0/25	1/25	0/25	0.547	0.420	1.000	1.000	1.000	.	.
Pancreas	Hemangioma	0/25	0/25	1/25	0/25	0/25	0/25	0/25	0.962	0.571	1.000
Prostate gland	Transitional cell car	0/25	1/25	0/25	0/25	0/25	0/25	0/25
Salivary glands	Adenocarcinoma	0/25	1/25	0/25	0/25	0/25	0/25	0/25
Skin	Hemangiosarcoma	0/25	1/25	0/25	0/25	0/25	0/25	0/25
Spleen	Hemangiosarcoma	0/25	22/25	0/25	0/25	1/25	1/24	3/22	0.124	0.115	.	.	1.000	1.000	0.235
Spleen	Leukemia	0/25	0/25	0/25	0/25	1/24	0/25	0/25	0.124	0.115	.	.	1.000	.	.
Stomach	Squamous cell car	0/25	3/25	0/25	0/25	1/24	0/25	0/25	0.130	0.115	.	.	1.000	.	.
Testes	Hemangiosarcoma	1/24	0/25	0/25	0/25	0/25	1/25	0/25	1.000	0.936	1.000	1.000	1.000	1.000	1.000
Thymus	Thymoma	0/25	0/25	0/25	0/25	0/25	2/25	0/25	0.490	.

* In treatment groups _1=Vehicle control, _2 =Positive control, _3 = DM/Q combination of 25/50 mg/kg/day, _4= DM/Q combination of 50/50 mg/kg/day, _5 =DM/Q combination of 100/100 mg/kg/day, _6 =DM only with 100 mg/kg/day, and _7 =Q only with 100 mg/kg/day

**Dose-Responses were tested for increased doses of combined treatment with DM and Q along with the vehicle control i.e. among Vehicle control, DM/Q combination of 25/50 mg/kg/day, DM/Q combination of 50/50 mg/kg/day, and DM/Q combination of 100/100 mg/kg/day. Dose-Response P-Values were calculated using the Poly-K method. The first column represents Poly_K P-Value using the bootstrap method, and the second column represents Poly_K P-Value using K=3

** *Pairwise comparisons were performed using the Fisher's exact test

Table 3B
Tumor Rates, Dose Response, and Pairwise Comparisons of Tested Tumors
Female Mouse

ORGANAM	TUMORNAM	Treatment Groups*							Dose-Response**		Pairwise Comparisons***				
		_1	_2	_3	_4	_5	_6	_7	Boots	K=3	1 vs. _3	_1 vs. _4	_1 vs. _5	_1 vs. _6	_1 vs. _7
Adrenal glands	Adenoma	0/25	0/25	2/24	0/25	0/25	0/25	1/24	0.939	0.596	0.490	.	.	.	1.000
Harderian glands	Carcinoma	0/25	0/25	1/25	0/25	1/24	0/25	0/25	0.155	0.236	1.000	.	1.000	.	.
Kidneys	Hemangiosarcoma	0/25	0/25	1/24	0/25	0/25	0/25	0/25	0.955	0.567	1.000
Lungs with bronchi	Adenoma	2/23	24/25	4/22	1/24	3/22	0/25	2/24	0.517	0.469	0.667	1.000	1.000	0.490	1.000
Lungs with bronchi	Carcinoma	0/25	17/25	0/25	1/25	0/25	0/25	0/25	0.241	0.303	.	1.000	.	.	.
Ovaries	Hemangiosarcoma	0/25	1/25	0/25	0/25	0/25	0/25	0/25
Salivary glands	Adenocarcinoma	0/25	1/25	0/25	0/25	0/25	0/25	0/25
Skin	Hemangiosarcoma	1/25	0/25	0/25	0/25	0/25	0/25	0/25	1.000	0.935	1.000	1.000	1.000	1.000	1.000
Spleen	Hemangiosarcoma	1/24	25/25	3/23	2/25	0/25	0/25	1/24	0.934	0.665	0.609	1.000	1.000	1.000	1.000
Spleen	Leukemia	1/25	0/25	0/25	0/25	0/25	0/25	0/25	1.000	0.935	1.000	1.000	1.000	1.000	1.000
Stomach	Papilloma	0/25	0/25	1/25	0/25	0/25	0/25	0/25	0.952	0.567	1.000
Thymus	Lymphoma	0/25	0/25	0/25	0/25	0/25	1/24	0/25
Thymus	Thymoma	0/25	0/25	2/23	0/25	0/25	1/24	2/24	0.941	0.596	0.490	.	.	1.000	0.490
Uterus	Hemangiosarcoma	2/23	0/25	1/24	5/23	1/24	3/22	1/24	0.452	0.429	1.000	0.417	1.000	1.000	1.000
Vagina	Hemangiosarcoma	0/25	0/25	0/25	1/25	0/25	0/25	0/25	0.241	0.303	1.000

* In treatment groups _1=Vehicle control, _2 =Positive control, _3 = DM/Q combination of 25/50 mg/kg/day, _4= DM/Q combination of 50/50 mg/kg/day, _5 =DM/Q combination of 100/100 mg/kg/day, _6 =DM only with 100 mg/kg/day, and _7 =Q only with 100 mg/kg/day

**Dose-Responses were tested for increased doses of combined treatment with DM and Q along with the vehicle control i.e. among Vehicle control, DM/Q combination of 25/50 mg/kg/day, DM/Q combination of 50/50 mg/kg/day, and DM/Q combination of 100/100 mg/kg/day. Dose-Response P-Values were calculated using the Poly-K method. The first column represents Poly_K P-Value using the bootstrap method, and the second column represents Poly_K P-Value using K=3

** *Pairwise comparisons were performed using the Fisher's exact test

Table 4A

**Pairwise Comparisons of Selected Tumors Using the Permutation Test
After Adjusting for Body Weight Gain
Male Mouse**

ORGANNAM	TUMORNAM	Pairwise Comparisons***				
		_1vs._3	_1vs._4	_1vs._5	_1vs._6	_1vs._7
Lungs with bronchi	Adenoma	0.203	1.000	1.000	0.349	0.487
Lungs with bronchi	Carcinoma	1.000	1.000	.	1.000	.
Spleen	Hemangiosarcoma	.	.	1.000	0.441	0.231
Spleen	Leukemia	.	.	1.000	.	.
Thymus	Thymoma	.	.	.	0.560	.

Treatment group _1=Vehicle control, _2 =Positive control, _3 = DM/Q combination of 25/50 mg/kg/day, _4= DM/Q combination of 50/50 mg/kg/day, _5 =DM/Q combination of 100/100 mg/kg/day, _6 =DM only with 100 mg/kg/day, and _7 =Q only with 100 mg/kg/day

** Pairwise comparisons were performed using the Permutation test

Table 4B

**Pairwise Comparisons of Selected Tumors Using the Permutation Test
After Adjusting for Body Weight Gain
Female Mouse**

ORGANNAM	TUMORNAM	Pairwise Comparisons***				
		_1vs._3	_1vs._4	_1vs._5	_1vs._6	_1vs._7
Lungs with bronchi	Adenoma	0.668	1.000	1.000	0.224	1.000
Lungs with bronchi	Carcinoma
Spleen	Hemangiosarcoma	0.674	1.000	0.489	0.479	1.000
Spleen	Leukemia	.	1.000	.	.	1.000
Thymus	Thymoma	0.490	.	.	1.000	0.489

Treatment group _1=Vehicle control, _2 =Positive control, _3 = DM/Q combination of 25/50 mg/kg/day, _4= DM/Q combination of 50/50 mg/kg/day, _5 =DM/Q combination of 100/100 mg/kg/day, _6 =DM only with 100 mg/kg/day, and _7 =Q only with 100 mg/kg/day

** Pairwise comparisons were performed using the Permutation test

Figure 1A

Survival Function Including Positive Control

Male Tg.rasH2 mouse with Positive Control

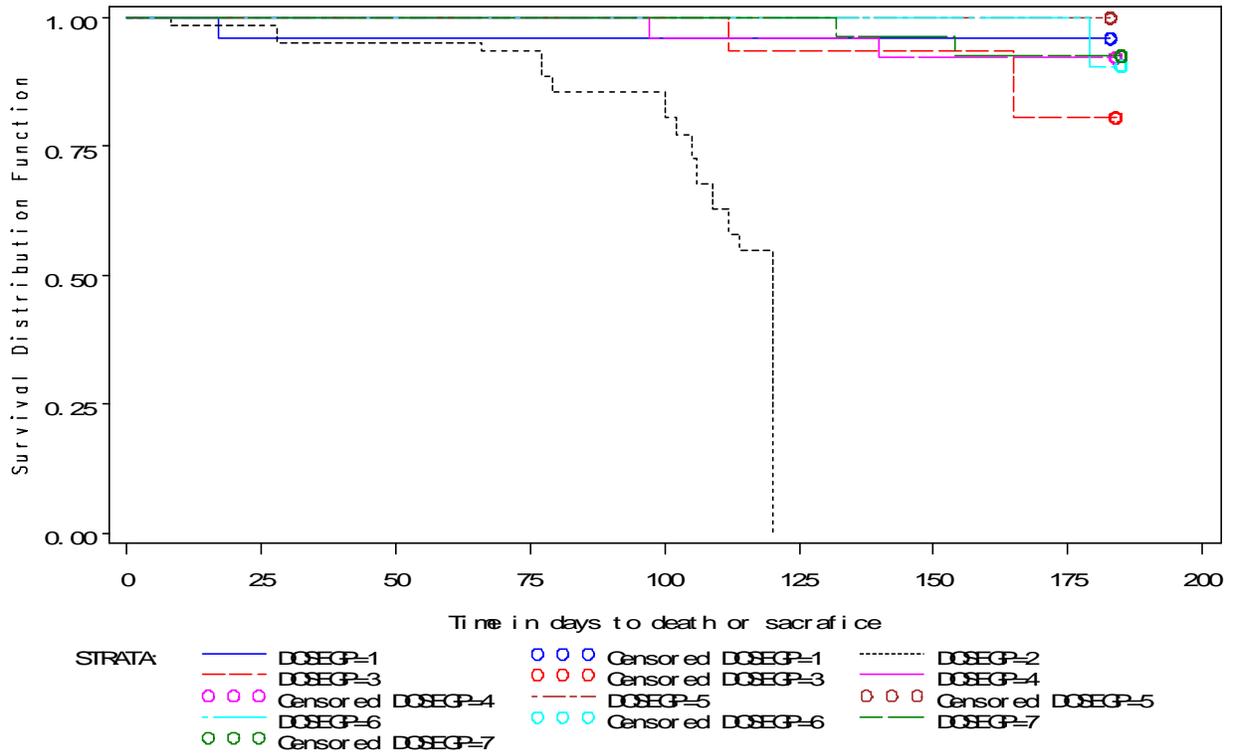
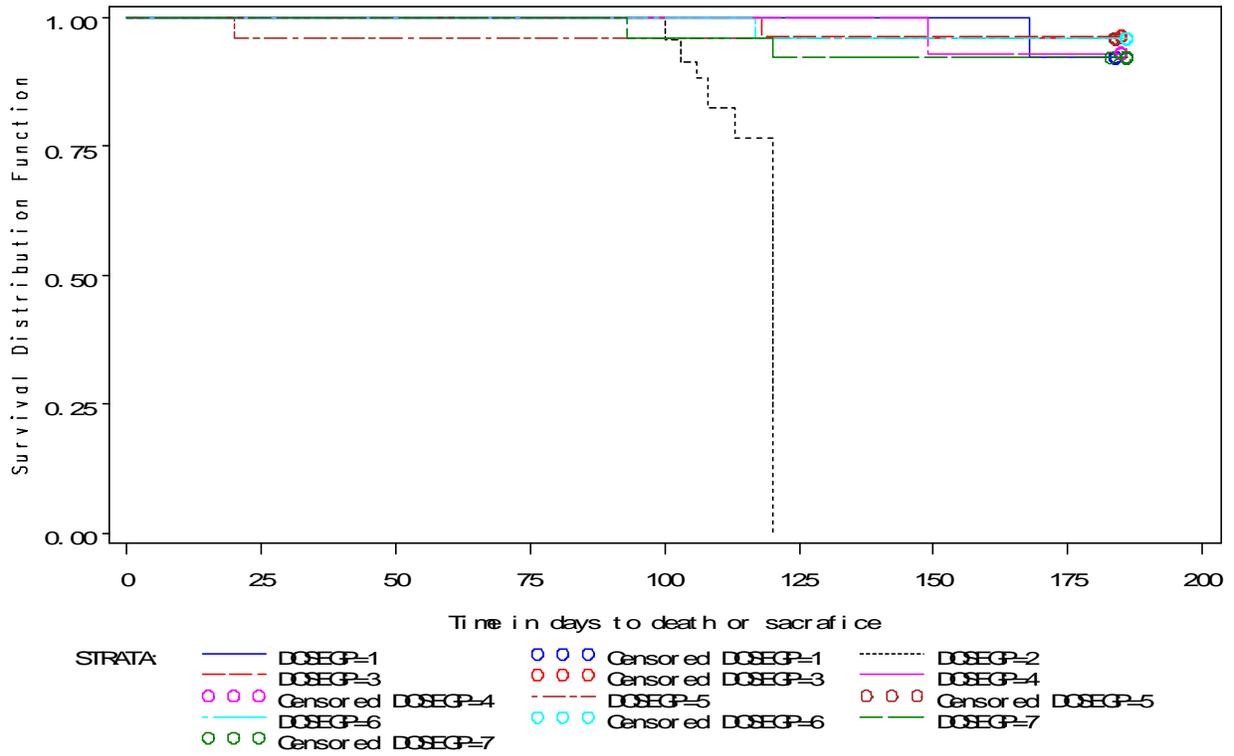


Figure1B

Survival Function Including Positive Control

Female Tg.rasH2 mouse with Positive Control



**This is a representation of an electronic record that was signed electronically and
this page is the manifestation of the electronic signature.**

/s/

Atiar Rahman
1/26/2006 11:49:08 AM
BIOMETRICS

Karl Lin
1/27/2006 09:44:28 AM
BIOMETRICS
Concur with review