

**CENTER FOR DRUG EVALUATION AND
RESEARCH**

APPLICATION NUMBER:

22-122

STATISTICAL REVIEW(S)



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

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/Serial Number: 22-122

Drug Name: Voltaren® AT (Diclofenac sodium topical gel, 1%)

Indication(s): _____ joints amenable to _____ treatment, such as the
hands and knees

Applicant: Novartis

Date(s): Stamp Date: December 19, 2006
PDUFA Due Date: October 19, 2007

Review Priority: Standard

Biometrics Division: Division of Biometrics II

Statistical Reviewer: Ruthanna Davi

Concurring Reviewers: Dionne Price

Medical Division: Division of Anesthesia, Analgesia and Rheumatology Products

Clinical Team: Neville Gibbs, Medical Reviewer
Mwango Kashoki, Medical Team Leader

Project Manager: Lauren Tornetta

Keywords:
NDA, clinical studies, multiple comparisons, interaction, sensitivity analyses, randomization

Table of Contents

1. EXECUTIVE SUMMARY	3
1.1 CONCLUSIONS AND RECOMMENDATIONS	3
1.2 BRIEF OVERVIEW OF CLINICAL STUDIES	3
1.3 STATISTICAL ISSUES AND FINDINGS	4
2. INTRODUCTION	5
2.1 OVERVIEW.....	5
2.2 DATA SOURCES.....	6
3. STATISTICAL EVALUATION	7
3.1 EVALUATION OF EFFICACY	7
4. FINDINGS IN SPECIAL/SUBGROUP POPULATIONS	30
4.1 GENDER, RACE AND AGE	30
4.2 OTHER SPECIAL/SUBGROUP POPULATIONS	33
5. SUMMARY AND CONCLUSIONS	34
5.1 STATISTICAL ISSUES AND COLLECTIVE EVIDENCE	34
5.2 CONCLUSIONS AND RECOMMENDATIONS	35

1. EXECUTIVE SUMMARY

1.1 Conclusions and Recommendations

In Study 310, in patients with OA of the knee, the Voltaren group had statistically significant better average outcomes in the three primary efficacy endpoints (WOMAC pain index, WOMAC function index, and global rating of disease) than vehicle. These differences were observed in both the MES and ITT analysis groups. The efficacy conclusions are robust against concerns regarding missing data as multiple sensitivity analyses yielded supportive conclusions.

In Study 315, in patients with OA of the hand, there is evidence of an analgesic effect of Voltaren despite the inability of one endpoint to achieve statistical significance. Comparison of the average outcomes for two (OA pain and total AUSCAN score) of the three primary efficacy endpoints at weeks 4 and 6 resulted in p-values less than 0.05. The comparison of the mean global rating of disease at week 4 resulted in a p-value of 0.06 and resulted in a p-value less than 0.05 at week 6. According to the prespecified multiplicity procedures the result for the global rating of disease endpoint at week 4 should have precluded any claims of efficacy and testing the primary efficacy endpoints at week 6. However, due to the borderline nature of this result, the relative clinical importance of the three endpoints, and the fact that the conclusion would be different if a hierarchical multiple comparison procedure were implemented (an approach that would seem reasonable if the protocol were being designed today), this study does provide supportive evidence of efficacy of Voltaren over vehicle despite the failure to satisfy the strict multiple comparison procedure. The study identified a possible treatment-by-OA interaction. The nature of the interaction (if it exists) is that the efficacy of Voltaren relative to vehicle is better in subjects with CMC-1 joint involvement and likely equivalent or worse in subject without CMC-1 joint involvement. In analyses weighted by strata size, inclusion of the treatment-by-OA category interaction did not affect the qualitative conclusion regarding the treatment effect. The primary efficacy conclusions are robust against concerns regarding missing data as multiple sensitivity analyses yielded supportive conclusions.

1.2 Brief Overview of Clinical Studies

The sponsor has submitted the results of two non-identical key phase 3 studies to support the regulatory approval of Voltaren® for _____ joints amenable to _____ treatment, such as the hands and knees.

The study referred to as VOSG-PN-310 or simply study 310 is titled, "A 12-week, randomized, double-blind, multi-center, vehicle-controlled, parallel group study to assess the efficacy and safety of diclofenac sodium gel 1% for the relief of signs and symptoms in patients with osteoarthritis of the knee". The primary objective of study 310 was to compare the efficacy of Voltaren with vehicle in treatment of osteoarthritis of the knee. The protocol-specified three primary efficacy endpoints were (1) pain as assessed by the Western Ontario McMaster Osteoarthritis Index (WOMAC) pain index, (2) functional capacity as assessed by the WOMAC function index, and (3) a global rating of disease assessed by the subject on a visual analog scale (100 mm VAS with 0 representing "very good" and 100 representing "very poor" in response to the question: "Considering all the ways osteoarthritis of the knee affects you, how well are you doing?"), measured at week 12.

The study referred to as VOSG-PE-315 or simply study 315 is titled, "An 8-week, multi-center, randomized, double-blind, placebo-controlled, parallel group trial of Diclofenac Sodium Gel 1% in patients with primary osteoarthritis of the hand". The primary objective of study 315 was to compare the efficacy of Voltaren with vehicle in treatment of osteoarthritis of the hand. The three primary efficacy endpoints, measured on a 100 mm visual analog scale at weeks 4 and 6 were (1) osteoarthritis pain intensity in the dominant (target) hand over the last 24 hours, (2) total Australian/Canadian Osteoarthritis Hand Index (AUSCAN) score for the dominant hand, and (3) global rating of disease activity.

1.3 Statistical Issues and Findings

The following statistical issues and their impact have been described in the context of the review. Please refer to the specified section for details.

Study 310 (OA of the knee)

- ❖ As part of a protocol amendment and prior to unblinding a modified efficacy subpopulation (MES) was defined and designated as the primary efficacy analysis group. The MES was a subset of subjects in the ITT efficacy population excluding all subjects whose pain on movement score in the target knee declined between the screening visit and the baseline visit or with a score of >1 on the abridged pain index for the contralateral knee at the baseline visit. From a statistical standpoint, the random treatment assignment in the MES group is valid. (Section 3.1.1.1)
- ❖ An analysis of variance (ANOVA) model including treatment and center main effects and baseline as a covariate was used. ANOVA models including a treatment-by-center interaction were explored by the sponsor but ultimately the interaction term was dropped from the model due to insignificance of the term and little variation in the least squares means for the main effect of treatment resulting from models with and without the interaction term. Although the conventions for dropping the treatment-by-center interaction term from the model were not completely described in the protocol, fortunately, the interaction term was clearly non-significant and therefore, dropping the interaction term is considered reasonable. (Section 3.1.1.2)
- ❖ The qualitative conclusions from multiple sensitivity analyses are supportive of the protocol-specified primary efficacy analysis minimizing the concern regarding the possible impact of the missing data. (Section 3.1.1.2)

Study 315 (OA of the hand)

- ❖ Prior to database closure and in a blinded fashion (i.e., in a "blind data review" meeting) it was determined that subjects with baseline scores which allowed little or no room for improvement would be excluded from the applicable analysis. While the random treatment assignment is valid regardless of the exclusion of subjects based on pre-randomization characteristics, one may question whether the criterion for exclusion of the subjects was determined in order to yield the most desirable result for the treatment effect since this criterion was developed after the data was collected. This change should have been formally documented as a protocol amendment so that criticism regarding the relative timing of unblinding and development of the criteria could be avoided. Since, only five subjects were affected by this change, there was little impact on the study results. (Section 3.1.2.2)
- ❖ According to the prespecified multiplicity procedures, the sponsor's primary efficacy results fail to demonstrate efficacy in that the by-treatment group comparison for the global rating

of disease endpoint yields a p-value slightly bigger than 0.05 (i.e., $p=0.06$). This statistical result for the global rating of disease endpoint was borderline and could be viewed differently if the current standard was applied. (Section 3.1.2.2)

- ❖ Comparison between treatment groups in each primary outcome was to be done with an ANOVA model including treatment, center, and hand OA category main effects, baseline as a covariate, and the treatment-by-center and treatment-by-OA category interactions. The conventions for dropping the interactions from the model were not explicit in the protocol and the treatment-by-OA category interaction approached or reached statistical significance in some instances suggesting the possibility of heterogeneity of the treatment effect across OA categories. The nature of the interaction (if it exists) is that the efficacy of Voltaren relative to vehicle is better in subjects with CMC-1 joint involvement and likely equivalent or worse in subject without CMC-1 joint involvement. In FDA analyses weighting the OA categories by the number of patients in each category, inclusion of the treatment-by-OA category interaction did not affect the qualitative conclusion regarding the treatment effect. (Sections 3.1.2.2 and 4.2)
- ❖ The sponsor's analyses presented for the OA pain intensity endpoint were conducted using an ANOVA model with main effects of treatment and center and baseline OA pain intensity **and baseline OA pain reported in patient diaries** as covariates. Inclusion of the baseline OA pain reported in patient diaries was not specified in the protocol or study report. FDA analyses excluding this factor are provided. (Section 3.1.2.2)

2. INTRODUCTION

2.1 Overview

The sponsor has submitted the results of two non-identical key phase 3 studies to support the regulatory approval of Voltaren® for _____ joints amenable to _____ treatment, such as the hands and knees.

The study referred to as VOSG-PN-310 or simply study 310 is titled, "A 12-week, randomized, double-blind, multi-center, vehicle-controlled, parallel group study to assess the efficacy and safety of diclofenac sodium gel 1% for the relief of signs and symptoms in patients with osteoarthritis of the knee". The primary objective of study 310 was to compare the efficacy of Voltaren with vehicle in treatment of osteoarthritis of the knee. The three primary efficacy endpoints, measured at week 12 were (1) pain reduction as assessed by the Western Ontario McMaster Osteoarthritis Index (WOMAC) pain index, (2) improvement in function capacity as assessed by the WOMAC function index, and (3) improvement in global rating of disease assessed on a visual analog scale (100 mm VAS).

The study referred to as VOSG-PE-315 or simply study 315 is titled, "An 8-week, multi-center, randomized, double-blind, placebo-controlled, parallel group trial of Diclofenac Sodium Gel 1% in patients with primary osteoarthritis of the hand". The primary objective of study 315 was to compare the efficacy of Voltaren with vehicle in treatment of osteoarthritis of the hand. The three primary efficacy endpoints, measured on a 100 mm visual analog scale at weeks 4 and 6 were (1) osteoarthritis pain intensity in the dominant (target) hand over the last 24 hours, (2) total Australian/Canadian Osteoarthritis Hand Index (AUSCAN) score for the dominant hand, and (3) global rating of disease activity.

Communication with the sponsor regarding these studies is documented under IND 64,334. Pertinent parts of the statistical portion of those communications are summarized below.

- Multiple IND reviews (e.g., submission dated May 26, 2004, March 8, 2005, and May 13, 2005) commented on the issue of missing baseline values. Various proposals had been made by the sponsor for imputation of these values. The Division comments centered on the idea that they believed a process could be put into place so that a patient is not randomized or dosed with study medication without the baseline fully documented. The Division did agree however that if this scenario could not be avoided, missing baseline values should be imputed as the mean values of the respective assessments over all patients with non-missing values at the baseline visit. This is the approach that was taken by the sponsor in the current NDA submission.
- Serial number 172 of IND 64334 (dated October 31, 2005) asks whether if from a statistical point of view, the Division believes it is acceptable to modify the target population for the knee study in light of the fact that a statistically significant efficacy benefit for Voltaren was found in only a subset of subjects in a similar knee study and not in the entire intent-to-treat population for that study. Statistical review of this submission indicates that from a statistical point of view, it is acceptable to modify the target population, stating that the subset of patients now of primary interest would still be a randomized study. The review cautioned however, that an important clinical question would be whether the revised patient population would be sufficient to allow extrapolation of the results to the larger population for which the drug would be likely to be used. A post meeting note added to the minutes for the Pre-NDA meeting occurring on July 21, 2006 indicate that the Division accepted the sponsor's argument that demonstration of efficacy in one knee would be generalizable to the population of OA patients with pain in both knees, and that the modified efficacy population (i.e., the group now of primary interest) is representative of the larger OA population. In the current NDA submission, the sponsor has presented the knee study in accordance with these comments.
- The minutes of the Pre-NDA meeting occurring on July 21, 2006, also indicate that the Division requested that the sponsor provide a continuous responder analysis and a baseline observation carried forward analysis as strategies for handling missing data. In the current NDA submission, the sponsor has provided these analyses.

Although most aspects of the designs of the studies 310 and 315 are similar, there are differences between the two trials including different target areas (hand and knee) and different primary efficacy endpoints. These studies will be summarized and critiqued separately within this document.

2.2 Data Sources

The sponsor has submitted the results of two key phase 3 studies in support of the efficacy of Voltaren® for the treatment of osteoarthritis of the knee and/or hand. The following data sets were submitted electronically and utilized in the review of these studies.

\\CDSESUB1\NONECTD\N22122\N 000\2006-12-19\DYAD
\\CDSESUB1\NONECTD\N22122\N 000\2006-12-19\QSADI

All submitted data sets were found to be adequately documented and organized.

3. STATISTICAL EVALUATION

3.1 Evaluation of Efficacy

3.1.1 Study 310 (Osteoarthritis of the knee)

3.1.1.1 Study Design (Study 310)

This study was a 12-week, randomized, double-blind, multi-center, vehicle-controlled, parallel group study. The primary objective of the study was to compare the efficacy of Voltaren with vehicle in treatment of osteoarthritis of the knee.

The target population consisted of ambulatory subjects with osteoarthritis of one or both knees. Subjects were to have a history of clinically symptomatic osteoarthritis with moderate to severe pain in one knee only, diagnosed at least six months previously and verified by X-ray. In total, the protocol specified ten inclusion and seventeen exclusion criteria necessary for a subject to be enrolled in the study. Visit 1, the screening visit, was to occur between 14 and 7 days prior to randomization. The protocol required that eligible patients enrolled in the study should undergo a one-week washout of analgesics (or at least five half lives, whichever was longer) prior to random treatment assignment.

At baseline (i.e., Visit 2) subjects were to be randomly assigned (1:1) to one of the following treatment groups:

- Voltaren (diclofenac Na gel, 1%)
- Vehicle

According to the protocol, the vehicle was identical to the active gel in appearance, feel and smell allowing the study to remain double blind. Subjects were to receive their randomly assigned treatment (4 g self-administered topically four times daily) for 12 weeks.

During the 12-week treatment period as well as the screening and washout periods, only acetaminophen tablets were allowed as rescue medication. Rescue medication was supplied to the subjects as part of the study. Rescue medication use was monitored through patient diaries where subjects were asked to record the number of rescue medication tablets taken, the time and the reason for taking them and additionally monitored by the investigator or designee counting the number of remaining rescue medication tablets at each visit. Treatment compliance was also assessed by the investigator or designee by checking the subject diary and by counting the number of used tubes of medication returned at each visit.

Following screening and baseline visits, the subjects were to visit the study site four times (at weeks 1, 4, 8, and 12) for assessment of efficacy, safety and compliance. Patients were contacted by the investigator or designee by phone mid-way between the visits corresponding to weeks 1, 4, 8, and 12 to remind the subjects to apply the study medication four times per day, not to take any disallowed medication, and to indicate any rescue medication usage in the subject diary.

The protocol-specified three primary efficacy endpoints were

- (1) pain as assessed by the Western Ontario McMaster Osteoarthritis Index (WOMAC) pain index,
- (2) functional capacity as assessed by the WOMAC function index, and

(3) a global rating of disease assessed by the subject on a visual analog scale (100 mm VAS with 0 representing “very good” and 100 representing “very poor” in response to the question: “Considering all the ways osteoarthritis of the knee affects you, how well are you doing?”)

The safety population and intent-to-treat (ITT) population were defined identically as all randomized patients who received at least one dose of study medication. Note that the traditional definition of an ITT group includes all subjects randomized regardless of whether or not they received study medication or any other criteria. For practical purposes, the set of subjects included in the sponsor’s protocol defined ITT population will likely be very similar to those that would be in the traditionally defined ITT group, and therefore would likely have only a negligible effect on the generalizability of the by-treatment group comparisons to the larger population. As part of a protocol amendment and prior to unblinding, a modified efficacy subpopulation (MES) was defined and designated as the primary efficacy analysis group. The MES was a subset of subjects in the ITT efficacy population excluding all subjects whose pain on movement score in the target knee declined between the screening visit and the baseline visit or with a score of >1 on the abridged pain index for the contralateral knee at the baseline visit. This definition was determined based on the results of another similar phase 3 study in which, according to the sponsor, positive efficacy findings were achieved only in this subset of subjects rather than the larger ITT group. Note that from a statistical standpoint, the random treatment assignment (allowing one to legitimately attribute differences between the treatment groups to a treatment effect and not an imbalance in some other covariate) in the MES group is valid regardless of the exclusion of subjects based on pre-randomization characteristics such as these; however, the generalizability of the conclusions drawn from this subpopulation to the larger population may be questioned. A clinical assessment of the appropriateness of such an extrapolation should be assessed.

The primary efficacy analyses were conducted on the three primary endpoints measured at week 12 using the MES group. Statistical significance on all three endpoints was considered necessary for demonstration of efficacy so appropriately, no adjustments for multiple comparisons were made. Each analysis was conducted at a two-sided alpha level of 0.05. The comparison between treatment groups in each primary outcome was assessed with an analysis of variance (ANOVA) model including treatment and center main effects, treatment-by-center interaction, and baseline as a covariate. The protocol indicated that if inclusion of the treatment-by-center interaction had a “meaningful impact on the statistical significance of the main effect”, sensitivity analyses would be conducted to address this finding. No significance level for testing the interaction was provided in the protocol.

The protocol provides the following conventions for imputation of missing data relevant to the primary efficacy comparisons.

I. For the case where the entire post-baseline visit(s) is (are) missing:

- Missing data resulting from a visit or consecutive visits being skipped by a subject in the middle of the study will be replaced by the average of the outcome for the latest non-missed visit before the missed visit(s) and the earliest non-missed visit after the missed visit.
- Missing data resulting from a visit or consecutive visits being skipped by a subject when no earliest non-missed visit after the missed visit is available (i.e., early termination) will be replaced by the outcome for the latest non-missed visit before the missed visit(s), i.e., last-observation carried forward.

- Exception: If a subject discontinues because of lack of efficacy, the imputed value will be the maximum of the value from the last non-missed visit and the baseline value.

II. For the case where only a subset of the questions needed to compute a WOMAC index are missing:

- A value for each missing question will be imputed as described in the preceding case with the additional stipulation that if a subset of questions in a WOMC index is not answered at the **baseline** visit, the score for that patient on the index will be computed and analyzed at all visits as if the question(s) did not exist.

III. For cases where a subject is determined to be a “treatment failure” (Treatment failures are defined as cases where there are a series of four or more consecutive days (after day 7) in which a subject takes either (a) all eight tablets of rescue medication or (b) the maximum daily OTC dose of a NSAID or (c) one or more single prescription strength doses of a nonselective or COX-2 selective NSAID, specifically to treat osteoarthritis in the target knee)

- Efficacy data beginning at the first of the four days leading to designation of a subject as a treatment failure will be ignored and will be replaced as if the subject had prematurely discontinued for lack of efficacy (i.e., the imputed value will be the maximum of the value from the last non-missed visit and the baseline value).

IV. For cases where a subject is determined to be a “treatment confounder” (Treatment confounders are defined as cases where there are a series of four or more consecutive days in which a subject takes either (a) any rescue medication or (b) any other analgesic, including nonselective or COX-2 selective NSAIDs at any dose, specifically to treat osteoarthritis pain in the contralateral knee or (c) applies study medication to the contralateral knee)

- Efficacy data recorded at the visit immediately following the episode leading to designation of a subject as a treatment confounder will be replaced with the corresponding efficacy data from the immediately preceding visit. This will apply to every immediately following visit for as long (and only for as long) as the subject continues to satisfy the definition of a treatment confounder.
 - In addition, a sensitivity analysis will be conducted in which the criteria of the treatment confounder with respect to NSAID and APAP use are modified to match those of treatment failure.

The protocol also specifies the following sensitivity analyses for the primary efficacy outcomes.

- (1.) An analysis substituting the mean of all non-missing values in the *active group* at visit X for any missing values in the *active group* at visit X and correspondingly, substituting the mean of all non-missing values in the *placebo group* at visit X for any missing values at visit X in the *placebo group*.
- (2.) An analysis substituting the mean of all non-missing values in the *active group* at visit X for any missing values in the *placebo group* at visit X and correspondingly, substituting the mean of all non-missing values in the *placebo group* at visit X for any missing values at visit X in the *active group*.

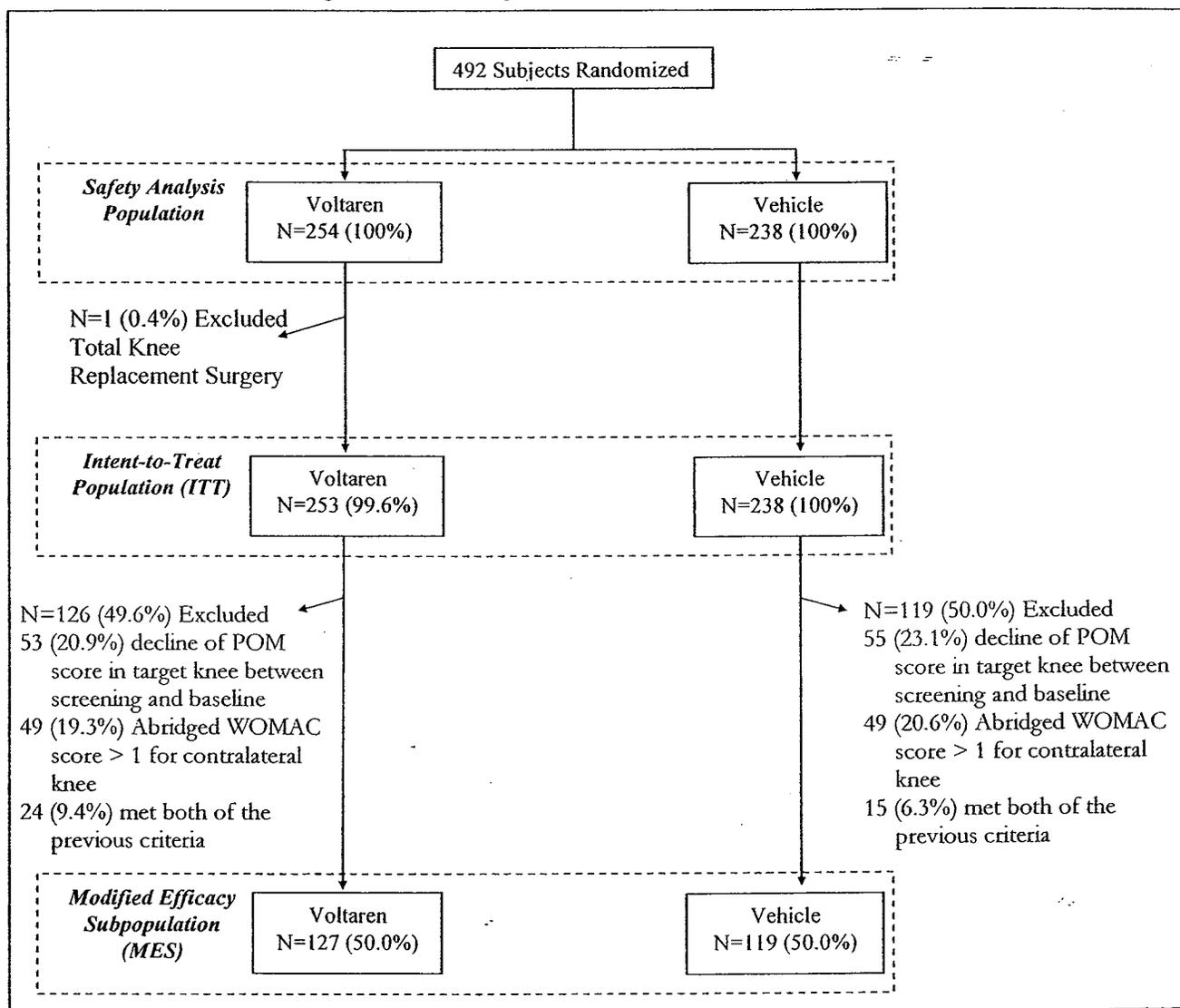
Based on data from three previous studies, the sponsor assumed a true difference between the active and vehicle groups of 6.4 with a standard deviation of 25 for the global rating of disease endpoint and determined that 240 subjects per group would provide 80% power for detecting a difference between treatment groups. Stating that the correlations between the three primary efficacy outcomes are expected to be quite high, the protocol suggests that the power for

achieving statistically significant results for all three primary efficacy comparisons is “in the vicinity of 80%”.

3.1.1.2 Results (Study 310)

Study 310 randomized 492 subjects, 254 to the Voltaren group and 238 to the vehicle group. Patient inclusion in and exclusion from the safety analysis population, intent-to-treat population (ITT), and modified efficacy subpopulation (MES) analysis data sets are described in Figure 1.

Figure 1: Patient Disposition and Analysis Groups



Source: Clinical Study Report, Study 310, Table 7-3 (with modifications in format)

As per-protocol, the safety analysis population includes all randomized subjects. (According to the protocol, subjects who did not receive study medication were to be excluded from this group; however, there were no subjects randomized who did not receive at least one dose of study medication.) One subject in the safety analysis population was excluded from the ITT

group. This subject was excluded because the subject had undergone total knee replacement surgery. This was not in accordance with the protocol; however, in a relatively large study such as study 310, the exclusion of one subject from the analysis is not expected to have greatly impacted the study results.

One hundred twenty six subjects in the Voltaren group and 119 subjects in the vehicle group were excluded from the MES population, as indicated in Figure 1.

The proportions of subjects in the Voltaren and vehicle groups who discontinued from the study early were 17.7% and 25.2%, respectively. The reasons for these discontinuations are tabulated in Table 1. Specifically in trials assessing pain, patients may experience relief from pain due to the treatment but may also experience intolerable side effects caused by the treatment. These patients may have a good score at the time of withdrawal that does not reflect the unfavorable outcome. Thus, particular attention is given to discontinuations due to adverse events. As shown in Table 1, subject discontinuations due to adverse events were relatively low and balanced among treatment groups, 5.1% and 3.8% for the Voltaren and vehicle groups, respectively. Concerns regarding possible bias in the efficacy analyses from imputation of missing data may be somewhat mitigated due to these findings.

	ITT Population	
	Voltaren N=254	Vehicle N=238
Total Discontinued	45 (17.7%)	60 (25.2%)
Adverse Event	13 (5.1%)	9 (3.8%)
Unsatisfactory therapeutic effect	10 (3.9%)	16 (6.7%)
Protocol deviation	1 (0.4%)	5 (2.1%)
Subject withdrew consent	15 (5.9%)	16 (6.7%)
Lost to follow-up	5 (2.0%)	12 (5.0%)
Administrative problems	1 (0.4%)	2 (0.8%)

Source: Clinical Study Report, Study 310, Table 7-1 (with modifications in format)

Demographic and background characteristics for the safety analysis population and MES population were provided in the submission and are summarized in Table 2. The difference between treatment groups in body mass index in the MES population revealed a nominal p-value less than 0.05 ($p=0.019$); however from a statistical perspective, this may be a spurious finding and is not considered a significant detriment to the study or an indication that the random treatment assignment was inadequate. No other differences between treatment groups were noted in demographic and background characteristics in the safety analysis population or MES populations.

Table 2: Baseline Demographics	Safety Analysis Population		MES Population	
	Voltaren (N=254)	Vehicle (N=238)	Voltaren (N=127)	Vehicle (N=119)
Gender – n (%)				
Male	83 (32.7%)	82 (34.5%)	44 (34.6%)	40 (33.6%)
Female	171 (67.3%)	156 (65.5%)	83 (65.4%)	79 (66.4%)
Race – n (%)				
Caucasian	191 (75.2%)	191 (80.3%)	101 (79.5%)	99 (83.2%)
Black	34 (13.4%)	31 (13.0%)	9 (7.1%)	10 (8.4%)
Asian	6 (2.4%)	1 (0.4%)	4 (3.1%)	0 (0.0%)
Other	23 (9.1%)	15 (6.3%)	13 (10.2%)	10 (8.4%)
Age in years				
Mean ± SD	59.7 ± 10.5	59.2 ± 10.6	59.7 ± 10.8	58.4 ± 10.4
Range	36 – 90	35 – 92	36 – 90	35 – 82
Height in cm				
N*	252	236	127	117
Mean ± SD	168.2 ± 10.6	168.3 ± 11.6	168.0 ± 10.3	167.5 ± 12.2
Range	132 – 198	123 – 196	132 – 198	123 – 193
Weight in kg				
N*	250	235	127	116
Mean ± SD	87.5 ± 18.8	90.0 ± 20.6	85.2 ± 19.6	89.9 ± 20.9
Range	49 – 159	50 – 159	49 – 159	54 – 159
BMI in kg/m²				
N*	250	235	127	116
Mean ± SD	30.9 ± 6.2	31.8 ± 7.0	30.1 ± 6.4	32.2 ± 7.1
Range	18.4 – 53.2	18.5 – 57.7	20.0 – 53.2	18.5 – 54.9

*Sample sizes for summary statistic calculation are given at the top of the table unless otherwise indicated.

Source: Clinical Study Report, Study 310, Table 7-4 (with modifications in format)

By-treatment group comparisons of the endpoints resulting from the baseline knee examinations and osteoarthritis assessments were also provided by the sponsor and are given in Tables 3 and 4, respectively. Although the sponsor cites two analyses with nominal p-values less than 0.05 in the MES group (i.e., p=0.041 for joint space medially and p=0.026 for joint space laterally), there is no indication in either the safety analysis population or the MES population that the treatment randomization was inadequate or that the by-treatment group differences should not be attributed to a treatment effect.

**APPEARS THIS WAY
ON ORIGINAL**

Table 3: Baseline Knee Exam	Safety Analysis Population		MES Population	
	Voltaren (N=254)	Vehicle (N=238)	Voltaren (N=127)	Vehicle (N=119)
Affected Knee – n (%)				
Left	113 (44.5%)	105 (44.1%)	59 (46.5%)	55 (46.2%)
Right	141 (55.5%)	133 (55.9%)	68 (53.5%)	64 (53.8%)
Receiving physical therapy at visit 1 – n (%)				
No	247 (97.2%)	235 (98.7%)	123 (96.9%)	117 (98.3%)
Yes	7 (2.8%)	3 (1.3%)	4 (3.1%)	2 (1.7%)
Periarticular pain				
No	75 (29.5%)	57 (23.9%)	38 (29.9%)	30 (25.2%)
Yes (caused by OA)	179 (70.5%)	181 (76.1%)	89 (70.1%)	89 (74.8%)
Range of Motion				
Extension (degrees)				
N*	249	234	123	116
Mean ± SD	1.1 ± 11.4	1.1 ± 11.9	1.0 ± 4.7	1.6 ± 16.0
Range	-10 – 170	-15 – 170	-10 – 35	-10 – 170
Neutral (degree)				
N*	249	234	123	116
Mean ± SD	4.4 ± 9.0	5.1 ± 8.6	3.7 ± 8.9	5.9 ± 8.9
Range	0 – 77	-8 – 40	0 – 77	0 – 40
Flexion (degree)				
N*	249	232	123	114
Mean ± SD	112.9 ± 19.9	112.8 ± 19.4	114.8 ± 20.3	110.2 ± 19.6
Range	60 – 150	56 – 150	60 – 150	56 – 150
Tenderness on pressure				
Joint space medially				
0=none	72 (28.5%)	55 (23.1%)	40 (31.7%)	27 (22.7%)
1=mild	92 (36.4%)	85 (35.7%)	47 (37.3%)	43 (36.1%)
2=moderate	78 (30.8%)	82 (34.5%)	36 (28.6%)	41 (34.5%)
3=severe	11 (4.3%)	16 (6.7%)	3 (2.4%)	8 (6.7%)
N*	253	238	126	119
Mean ± SD	1.1 ± 0.9	1.2 ± 0.9	1.0 ± 0.8	1.3 ± 0.9
Joint space laterally				
0=none	110 (43.5%)	92 (38.7%)	62 (49.2%)	46 (38.7%)
1=mild	87 (34.4%)	81 (34.0%)	46 (36.5%)	41 (34.5%)
2=moderate	50 (19.8%)	62 (26.1%)	17 (13.5%)	32 (26.9%)
3=severe	6 (2.4%)	3 (1.3%)	1 (0.8%)	0 (0.0%)
N*	253	238	126	119
Mean ± SD	0.8 ± 0.8	0.9 ± 0.8	0.7 ± 0.7	0.9 ± 0.8
Patella medially				
0=none	114 (45.1%)	101 (42.4%)	65 (51.6%)	54 (45.4%)
1=mild	79 (31.2%)	72 (30.3%)	37 (29.4%)	34 (28.6%)
2=moderate	57 (22.5%)	57 (23.9%)	24 (19.0%)	27 (22.7%)
3=severe	3 (1.2%)	8 (3.4%)	0 (0.0%)	4 (3.4%)
N*	253	238	126	119
Mean ± SD	0.8 ± 0.8	0.9 ± 0.9	0.7 ± 0.8	0.8 ± 0.9
Patella laterally				
0=none	130 (51.4%)	134 (56.3%)	79 (62.7%)	71 (59.7%)
1=mild	72 (28.5%)	62 (26.1%)	31 (24.6%)	28 (23.5%)
2=moderate	51 (20.2%)	37 (15.5%)	16 (12.7%)	18 (15.1%)
3=severe	0 (0.0%)	5 (2.1%)	0 (0.0%)	2 (1.7%)
N*	253	238	126	119
Mean ± SD	0.7 ± 0.8	0.6 ± 0.8	0.5 ± 0.7	0.6 ± 0.8
Joint capsule swelling – n (%)				
0=none	109 (42.9%)	112 (47.1%)	59 (46.5%)	54 (45.4%)
1=mild	106 (41.7%)	94 (39.5%)	47 (37.0%)	48 (40.3%)
2=moderate	37 (14.6%)	31 (13.0%)	20 (15.7%)	16 (13.4%)
3=severe	2 (0.8%)	1 (0.4%)	1 (0.8%)	1 (0.8%)
Mean ± SD	0.7 ± 0.7	0.7 ± 0.7	0.7 ± 0.8	0.7 ± 0.7
Joint effusion – n (%)				
No	189 (74.4%)	189 (79.4%)	99 (78.0%)	95 (79.8%)
Yes	65 (25.6%)	49 (20.6%)	28 (22.0%)	24 (20.2%)

*Sample sizes for summary statistic calculation are given at the top of the table unless otherwise indicated.

Source: Clinical Study Report, Study 310, Table 7-3 (with modifications in format)

Table 4: Baseline Osteoarthritis Assessment	Safety Analysis Population		MES Population	
	Voltaren (N=254)	Vehicle (N=238)	Voltaren (N=127)	Vehicle (N=119)
Global rating of disease¹				
Mean ± SD	62.3 ± 19.4	61.9 ± 19.5	61.6 ± 20.2	63.9 ± 21.0
Range	9 – 100	1 – 100	9 – 98	1 – 100
Spontaneous pain²				
Mean ± SD	59.2 ± 22.7	59.2 ± 24.4	58.9 ± 23.6	60.9 ± 26.7
Range	0 – 100	1 – 100	0 – 98	1 – 100
Pain on movement²				
Target knee				
Mean ± SD	71.3 ± 11.8	71.4 ± 12.7	72.9 ± 11.5	74.3 ± 13.0
Range	47 – 99	40 – 101	50 – 99	50 – 100
Contralateral knee				
N*	253	238	127	119
Mean ± SD	6.2 ± 5.9	6.2 ± 5.7	4.7 ± 5.7	4.4 ± 4.5
Range	0 – 24	0 – 21	0 – 20	0 – 20
WOMAC pain index³				
Mean ± SD	11.67 ± 2.43	11.72 ± 2.45	11.47 ± 2.51	12.06 ± 2.71
Range	5 – 18	5 – 18	5 – 18	5 – 18
WOMAC stiffness index³				
Mean ± SD	4.77 ± 1.47	4.91 ± 1.46	4.67 ± 1.58	4.87 ± 1.62
Range	0 – 8	0 – 8	0 – 8	0 – 8
WOMAC physical function index³				
Mean ± SD	38.0 ± 10.0	37.9 ± 10.7	37.2 ± 10.9	37.5 ± 12.0
Range	4 – 65	9 – 64	4 – 61	9 – 64

*Sample sizes for summary statistic calculation are given at the top of the table unless otherwise indicated.

1. 0=very good, 100=very poor

2. 0=no pain, 100=unbearable pain

3. pain: 0-20; stiffness:0-8; physical function: 0-68

Source: Clinical Study Report, Study 310, Table 7-6 (with modifications in format)

All primary efficacy analyses were conducted using the statistical procedures specified in the protocol. An analysis of variance (ANOVA) model including treatment and center main effects and baseline as a covariate was used. ANOVA models including a treatment-by-center interaction were explored by the sponsor but ultimately the interaction term was dropped from the model due to insignificance of the term and little variation in the mean treatment effect resulting from models with and without the interaction term. Although the conventions for dropping the treatment-by-center interaction term from the model were not completely described in the protocol, fortunately, the non-significance of the interaction term was clear cut and therefore, dropping the interaction term is considered reasonable. Missing data was imputed as described in the protocol and in this document (see section 3.1.1.1).

The primary efficacy comparisons for both the MES and ITT populations are displayed in Table 5. The Voltaren group demonstrated statistically significantly better average results than the vehicle group for all three endpoints in both the MES and ITT populations. In general, larger by-treatment group differences were observed in the MES population than in the ITT population; however, statistical significance was still demonstrated in the ITT group due in part to the larger sample size available for analysis.

Table 5: Primary Efficacy Comparisons at Week 12						
	ITT Population			MES Population		
	Voltaren N=253	Vehicle N=238	Difference (vehicle - drug)	Voltaren N=127	Vehicle N=119	Difference (vehicle - drug)
WOMAC Pain Index (scale = 0 to 20)						
Least squares mean ²	6.49	7.34	0.86	5.95	7.29	1.34
95% CI for difference			(0.09, 1.62)			(0.18, 2.49)
p-value			0.028			0.023
WOMAC Function Index (scale = 0 to 68)						
Least squares mean ²	22.4	26.2	3.8	20.2	25.9	5.7
95% CI for difference			(1.4, 6.3)			(2.0, 9.4)
p-value			0.002			0.003
Global Rating of Disease (100 mm VAS)						
Least squares mean ²	33.8	41.7	8.0	34.1	42.6	8.5
95% CI for difference			(3.2, 12.8)			(1.5, 15.6)
p-value			0.001			0.018

1. Missing data imputed as specified in the protocol.

2. Least squares mean calculated using protocol specified ANOVA model with main effects of treatment and center and baseline as a covariate.

Source: Clinical Study Report, Study 310, Tables 9-1 and 9-3 (with modifications in format)

Several sensitivity analyses to address the robustness of these results to the protocol-specified missing data conventions were provided by the sponsor using the MES population. Selections of these analyses are compared to the protocol-specified primary analysis and are provided in Tables 6 through 8. Tables 6 and 7 contain the protocol-specified sensitivity analyses. An analysis substituting the mean of all non-missing values in the *active group* at visit X for any missing values in the *active group* at visit X and correspondingly, substituting the mean of all non-missing values in the *placebo group* at visit X for any missing values at visit X in the *placebo group* is referred to as “Same Mean” and is shown in Table 6. An analysis substituting the mean of all non-missing values in the *active group* at visit X for any missing values in the *placebo group* at visit X and correspondingly, substituting the mean of all non-missing values in the *placebo group* at visit X for any missing values at visit X in the *active group* is referred to as “Alternate Mean” and is shown in Table 7. An analysis substituting the subject’s baseline observation for any missing value(s) was requested by the Division at the pre-NDA meeting, is referred to as “Baseline-Observation-Carried-Forward” (BOCF) and is shown in Table 8.

Table 6: Least Squares Means by Treatment Group – Sensitivity Analyses								
	Protocol Specified				Same Mean			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
WOMAC Pain	5.9	7.3	1.3	0.02	5.5	5.9	0.5	0.37
WOMAC Function	20.2	25.9	5.7	0.003	18.8	22.2	3.4	0.04
Global Rtg of Disease	34.1	42.6	8.5	0.02	30.2	35.7	5.6	0.07

Source: Clinical Study Report, Appendix 5 (Statistical Methods, Randomization Scheme, and Codes), Section 3.4

Table 7: Least Squares Means by Treatment Group – Sensitivity Analyses								
	Protocol Specified				Alternate Mean			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
WOMAC Pain	5.9	7.3	1.3	0.02	5.6	5.6	0.0	>0.99
WOMAC Function	20.2	25.9	5.7	0.003	19.4	20.9	1.5	0.37
Global Rtg of Disease	34.1	42.6	8.5	0.02	31.2	33.5	2.3	0.46

Source: Clinical Study Report, Appendix 5 (Statistical Methods, Randomization Scheme, and Codes), Section 3.3

Table 8: Least Squares Means by Treatment Group – Sensitivity Analyses								
	Protocol Specified				BOCF			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
WOMAC Pain	5.9	7.3	1.3	0.02	6.4	7.9	1.5	0.02
WOMAC Function	20.2	25.9	5.7	0.003	21.2	27.4	6.2	0.001
Global Rtg of Disease	34.1	42.6	8.5	0.02	35.4	44.1	8.6	0.02

Source: Clinical Study Report, Appendix 5 (Statistical Methods, Randomization Scheme, and Codes), Section 3.5

The analyses incorporating the “same mean” imputation scheme are supportive of the conclusions from the protocol-specified primary efficacy analyses. As expected, the least squares mean for each endpoint and each treatment group is numerically higher in the primary efficacy analysis; however, the numerical differences between treatment groups continue to favor the Voltaren group. The analyses of the WOMAC function and global rating of disease endpoints are fairly robust against this imputation ($p=0.04$ and $p=0.07$, respectively). The significance of the by-treatment group difference in the WOMAC pain endpoint is not maintained ($p=0.37$).

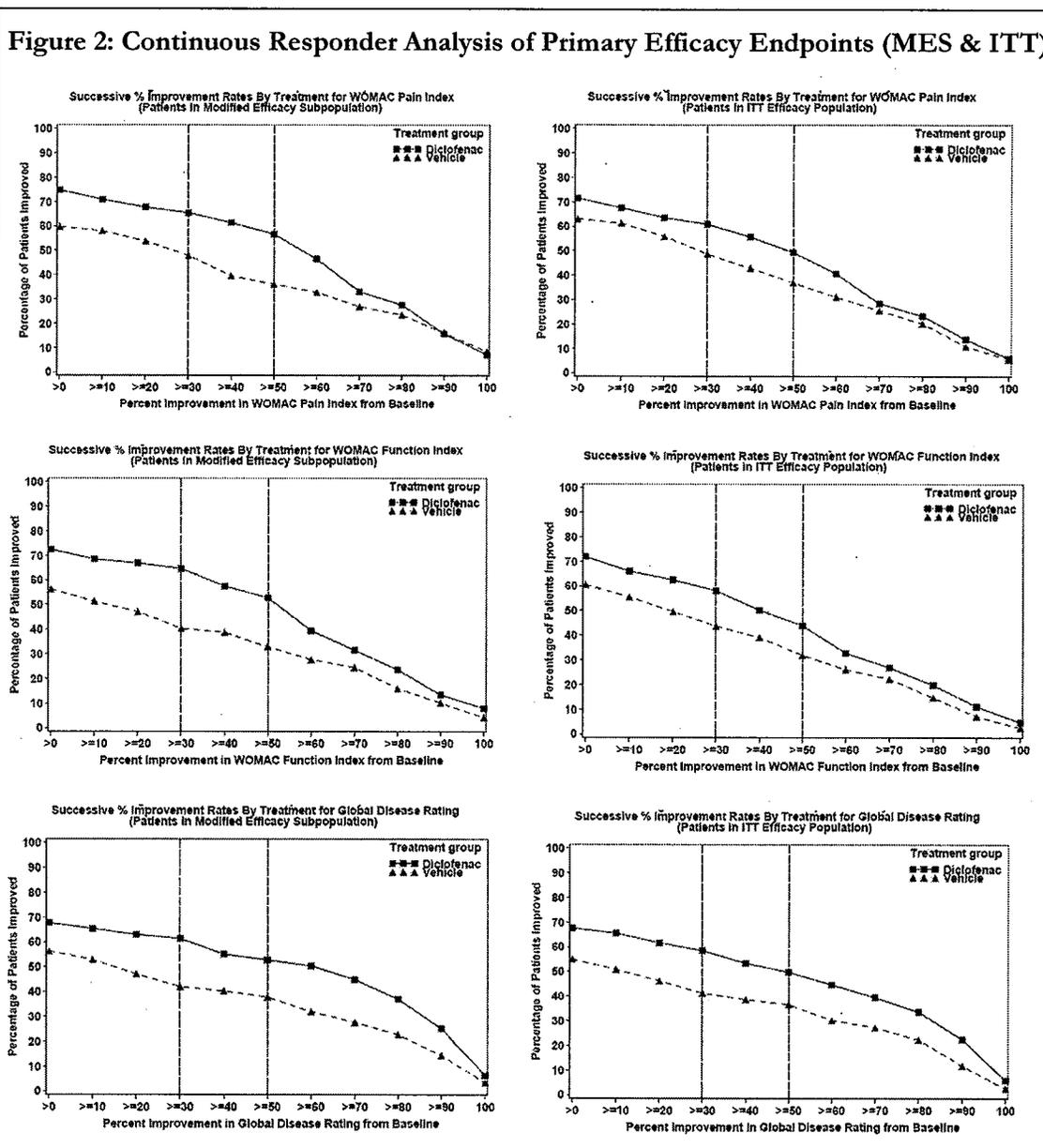
Statistically significant by-treatment group differences are not achieved in the analyses incorporating the “alternate mean” imputation scheme ($p>0.99$, $p=0.37$, and $p=0.46$ for the WOMAC pain, WOMAC function, and global rating of disease endpoints, respectively). However, as highlighted by the sponsor, in this case, this analysis is conservative and biased in favor of the vehicle group since more favorable results are imputed for the missing data in the vehicle group than in the Voltaren group. The lack of statistically significant findings in this analysis is not unexpected.

The results of the analyses when imputing according to the BOCF approach support the primary efficacy analyses for all three primary efficacy endpoints ($p=0.02$, $p=0.001$, and $p=0.02$ for the WOMAC pain, WOMAC function, and global rating of disease endpoints, respectively). Since the BOCF strategy helps to address the concern regarding the potential for good scores to be assigned to subjects who had unfavorable outcomes (such as dropping out due to an adverse event), this analysis provides robust support for the primary efficacy analyses and minimizes the concern regarding the possible impact of the missing data.

In summary, the qualitative conclusions from the missing data sensitivity analyses are supportive of the protocol-specified primary efficacy analysis. With the exception of the “same mean” analysis of the WOMAC pain endpoint, by-treatment group differences that are clearly statistically non-significant occur only in analyses that are designed to be very conservative.

At the request of the division, the sponsor provided cumulative distribution plots (i.e., a continuous responder analyses) for the primary efficacy endpoints using the MES and ITT populations. All discontinuations are classified as treatment failures in the formulation of the

plots. The descriptive conclusions from these plots are supportive of the conclusions from the primary efficacy analyses and are provided in Figure 2.



Source: Clinical Study Report, Post-text Supplement 3, Figure 9.15.1 through 9.15.3a

3.1.2 Study 315 (Osteoarthritis of the hand)

3.1.2.1 Study Design (Study 315)

The design of study 315 is largely the same as study 310 with the following exceptions noted.

- The duration of the study was 8 weeks.
- The target population consisted of subjects who were at least 40 years old with a diagnosis of OA of their dominant hand as defined by The American College of

Rheumatology (ACR) criteria. Eligible subjects could have a diagnosis of OA in their non-dominant hand but the symptoms were to be of lesser intensity than those of the dominant hand. The protocol specified a total of 15 inclusion and 22 exclusion criteria necessary for a subject to be enrolled.

- Subjects were to self-administer 2 g of their randomly assigned treatment to the dominate hand and 2 g to the non-dominant hand four times daily.
- Subjects were to visit the study site five times (i.e., at weeks 1, 2, 4, 6, and 8). The sites were not required to contact the subjects by phone between visits.
- The three primary efficacy endpoints, measured at weeks 4 and 6 were
 - (1) OA pain intensity in the dominant (target) hand over the last 24 hours measured on a 100 mm visual analog scale with 0 representing “no pain” and 100 representing “unbearable pain”,
 - (2) a global rating of disease assessed by the subject on a 100 mm visual analog scale with 0 representing “very good” and 100 representing “very poor” in response to the question: “Considering all the ways osteoarthritis of the hand affects you, how well are you doing?”
 - (3) Total Australian/Canadian Osteoarthritis Hand Index (AUSCAN) score in the target hand designed to assess functional status measured as an unweighted sum of the scores on 15 questions.
- The protocol-specified primary efficacy analysis was to be conducted using the intent-to-treat population. However, the sponsor indicated that prior to database closure and in a blinded fashion (i.e., in a “blind data review” meeting), it was determined that subjects with baseline scores which allowed little or no room for improvement would be excluded from the applicable analysis. As a result, five subjects (3 Voltaren and 2 placebo) were excluded from the analysis of one of the primary efficacy endpoints. Sensitivity analyses addressing this change were provided by the sponsor and yielded similar results to the analyses excluding these subjects.
- The primary efficacy analyses were conducted on the three primary endpoints measured at week 4 and week 6 using the ITT group but excluding subjects with baseline scores that allowed little or no room for improvement in that primary efficacy endpoint. Statistical significance on all three endpoints was considered necessary for demonstration of efficacy at a particular time point so appropriately, no adjustments for multiple endpoints were made. To control for the issue of multiple efficacy time points, the statistical analysis was to proceed in a stepwise fashion. All primary outcomes were to be analyzed at week 4. If a statistically significant difference was demonstrated on all three primary endpoints then the analysis of the week 6 data was to be conducted. Otherwise it was to be concluded that the efficacy of Voltaren had not been demonstrated regardless of any results at week 6. If a statistically significant difference was demonstrated on all three co-primary endpoints at week 6 then it was to be concluded that the efficacy of Voltaren had been demonstrated at both week 4 and week 6, otherwise it was to be concluded that efficacy had been demonstrated at week 4 but not week 6. Comparison between treatment groups in each primary outcome was to be done with an ANOVA model including treatment, center, and hand OA category main effects, baseline as a covariate, and the treatment-by-center and treatment-by-OA category interactions. As with study 310, the conventions for dropping the interactions from the model were not explicit in the protocol.
- The sponsor indicated that prior to database closure and in a blinded fashion (i.e., in a “blind data review” meeting) the following additional changes were made to the statistical analysis plan. First, the OA category variable was collapsed from three

categories (“only CMC-1 is painful”, “mixed OA”, and “painful joints do not include CMC-1”) to two categories (“only CMC-1 is painful or mixed OA” and “painful joints do not include CMC-1”) since only a few subjects fell into the “only CMC-1 is painful” category. Second, the sponsor indicated that subsequent to discussions with the FDA, instead of using a regression approach to impute the missing baseline assessments, the mean values over all subjects in the ITT efficacy population with non-missing values at baseline were used to impute missing baseline assessments.

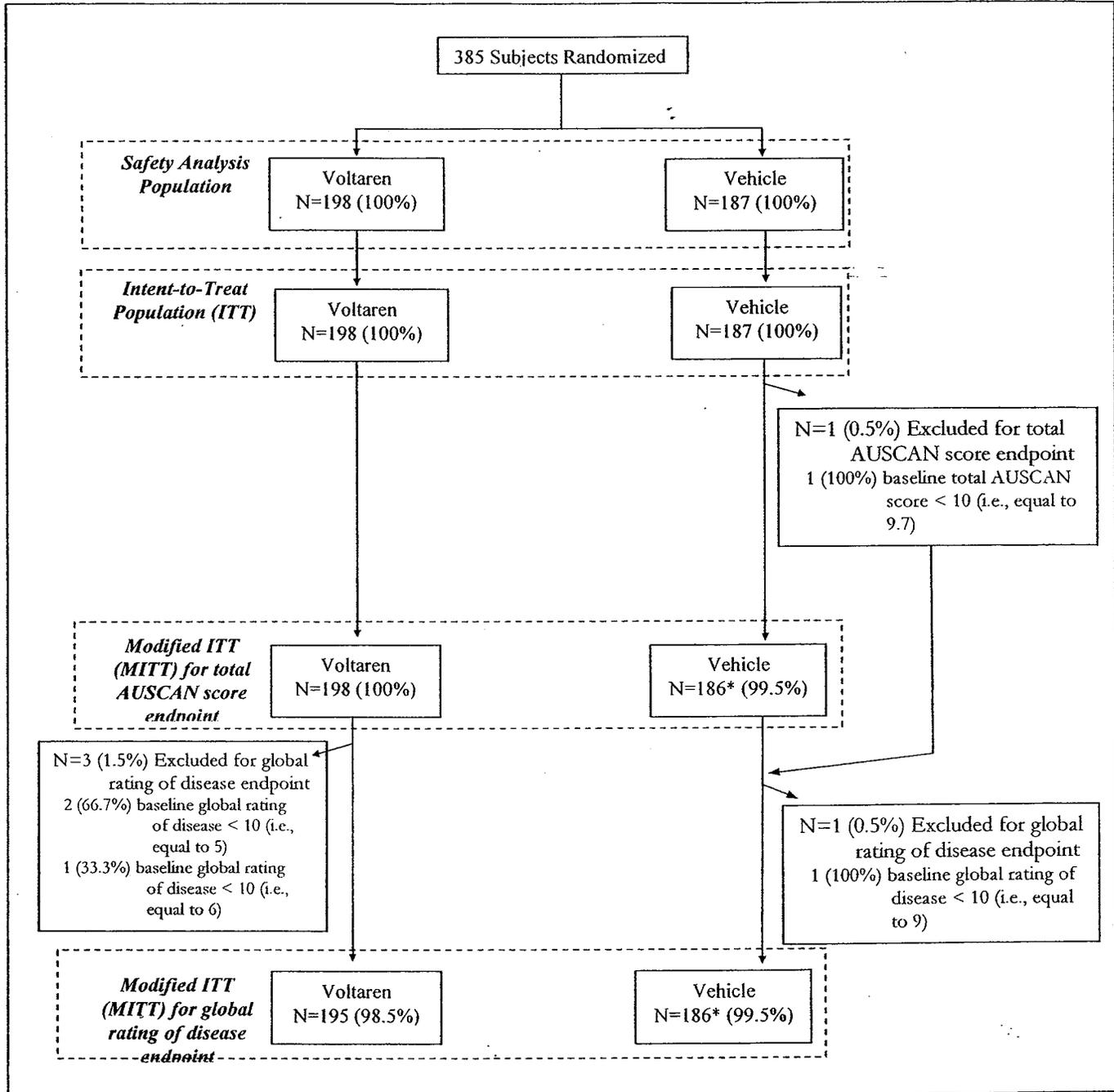
- Based on data from two previous studies, the sponsor assumed a true difference between the active and vehicle groups of 7 mm with a standard deviation of 20.5 mm for the OA pain intensity endpoint and determined that 180 subjects per group would provide 90% power for detecting a difference between treatment groups. The other two endpoints were presumed to have smaller standard deviations and thus higher power for detecting a difference of 7 mm on a 100 point scale. The sponsor also stated that with a correlation among each pair of co-primary endpoints of 0.6, the power to reject all three null hypotheses simultaneously was approximately 80%. Finally, the sponsor concluded that conditional on success at week 4, the same power would be expected for the week 6 analyses.

3.1.2.2 Results (Study 315)

Study 315 randomized 385 subjects, 198 to the Voltaren group and 187 to the vehicle group. Patient inclusion in and exclusion from the safety analysis population, intent-to-treat population (ITT), and modified ITT (MITT) populations are described in Figure 3.

All randomized subjects were treated and were included in the safety and ITT analysis populations. However, the sponsor indicated that prior to database closure and in a blinded fashion (i.e., in a “blind data review” meeting) it was determined that subjects with baseline scores which allowed little or no room for improvement would be excluded from the applicable analysis. As a result, five subjects (3 Voltaren and 2 placebo) were excluded from the analysis of one of the primary efficacy endpoints. Note that from a statistical standpoint, the random treatment assignment (allowing one to legitimately attribute differences between the treatment groups to a treatment effect and not an imbalance in some other covariate) in these modified ITT groups is valid regardless of the exclusion of subjects based on pre-randomization characteristics such as their measurement at baseline. However, in principle, these exclusions could give rise to other legitimate concerns. First, one may question whether the criterion for exclusion of the subjects was determined in order to yield the most desirable result for the treatment effect since this criterion was developed after the data was collected. The sponsor’s claim that this change was made prior to unblinding helps to alleviate this concern but ideally this change should have been formally documented as a protocol amendment so that criticism regarding the relative timing of unblinding and development of the criteria could be avoided. Secondly, the generalizability of the conclusions drawn from this subpopulation (i.e., patients with baseline scores greater than 10) to the larger population (i.e., patients with any baseline score) may be questioned and a clinical assessment of the appropriateness of such an extrapolation may be considered. Note though that in a relatively large study such as study 315, the exclusion of five subjects from the analysis is not expected to have greatly impacted the legitimacy of the study results.

Figure 3: Patient Disposition and Analysis Groups



*Although the sample sizes for the analyses of the global rating of disease endpoint and total AUSCAN score endpoint in the vehicle group are both 186, the groups do not consist of the identical set of subjects. The vehicle subject excluded from the analysis of the global rating of disease endpoint and the vehicle subject excluded from the analysis of the total AUSCAN score endpoint are different subjects.

Source: Clinical Study Report, Study 315, Table 7-2 and text from section 6.1.1 (with modifications in format)

The proportions of subjects in the Voltaren and vehicle groups who discontinued from the study early were 12.6% and 13.9%, respectively. The reasons for these discontinuations are tabulated in Table 9. Specifically in trials assessing pain, patients may experience relief from pain due to the treatment but may also experience intolerable side effects caused by the treatment. These patients may have a good score at the time of withdrawal that does not reflect the unfavorable outcome. Thus, particular attention is given to discontinuations due to adverse events. As shown in Table 9, subject discontinuations due to adverse event were relatively low and balanced among treatment groups, 5.1% and 2.1% for the Voltaren and vehicle groups, respectively. Concerns regarding possible bias in the efficacy analyses from imputation of missing data may be somewhat mitigated due to these findings.

Table 9. Subject Discontinuations		
	ITT Population	
	Voltaren N=198	Vehicle N=187
Total Discontinued	25 (12.6%)	26 (13.9%)
Adverse Event	10 (5.1%)	4 (2.1%)
Unsatisfactory therapeutic effect	8 (4.0%)	13 (7.0%)
Protocol deviation	1 (0.5%)	1 (0.5%)
Subject withdrew consent	4 (2.0%)	6 (3.2%)
Lost to follow-up	2 (1.0%)	1 (0.5%)
Administrative problems	0 (0.0%)	1 (0.5%)

Source: Clinical Study Report, Study 315, Table 7-1 (with modifications in format)

Demographic and background characteristics for the ITT population were provided in the submission and are summarized in Table 10. No differences between treatment groups were noted in demographic and background characteristics.

**APPEARS THIS WAY
ON ORIGINAL**

Table 10: Baseline Demographics (ITT Analysis Population)		
	Voltaren (N=198)	Vehicle (N=187)
Gender – n (%)		
Male	46 (23.2%)	43 (23.0%)
Female	152 (76.8%)	144 (77.0%)
Race – n (%)		
Caucasian	173 (87.4%)	170 (90.9%)
Black	11 (5.6%)	4 (2.1%)
Asian	3 (1.5%)	0 (0.0%)
Other	11 (5.6%)	13 (7.0%)
Age in years		
Mean ± SD	63.6 ± 10.3	64.7 ± 9.6
Range	40 – 92	40 – 87
Height in cm		
N*	196	187
Mean ± SD	165.0 ± 9.8	164.7 ± 10.1
Range	132 – 191	132 – 193
Weight in kg		
N*	198	186
Mean ± SD	76.6 ± 18.4	77.8 ± 20.0
Range	42 – 138	44 – 166
BMI in kg/m²		
N*	196	186
Mean ± SD	28.0 ± 6.3	28.6 ± 6.5
Range	17.6 – 55.0	17.5 – 49.8

*Sample sizes for summary statistic calculation are given at the top of the table unless otherwise indicated.
Source: Clinical Study Report, Study 315, Table 7-3 (with modifications in format)

By-treatment group comparisons of the endpoints resulting from the baseline hand examinations and osteoarthritis assessments were also provided by the sponsor and are given in Tables 11 and 12, respectively. Although the sponsor cites two analyses with nominal p-values close to 0.05 (i.e., p=0.06 for OA hand category and p=0.05 for AUSCAN stiffness), there is no indication that the treatment randomization was inadequate or that the by-treatment group differences should not be attributed to a treatment effect.

**APPEARS THIS WAY
ON ORIGINAL**

Table 11: Baseline Hand Examination (ITT Analysis Population)		
	Voltaren (N=198)	Vehicle (N=187)
Target (dominant) hand – n (%)		
Left	14 (7.1%)	20 (10.7%)
Right	184 (92.9%)	167 (89.3%)
Hand OA category – n (%)		
With CMC-1 joint involvement	133 (67.2%)	142 (75.9%)
Without CMC-1 joint involvement	65 (32.8%)	45 (24.1%)
X-ray Evaluations* – n (%)		
Sclerosis	89 (44.9%)	93 (49.7%)
Subchondral cysts	68 (34.3%)	63 (33.7%)
Joint space narrowing	151 (76.3%)	141 (75.4%)
Osteophytes	167 (84.3%)	164 (87.7%)
Kellgren-Lawrence Grade		
1	38 (19.2%)	27 (14.4%)
2	55 (27.8%)	63 (33.7%)
3	105 (53.0%)	97 (51.9%)
Current treatment* – n (%)		
NSAIDs	108 (54.5%)	91 (48.7%)
Acetaminophen (paracetamol)	55 (27.8%)	42 (22.5%)
Aspirin	5 (2.5%)	7 (3.7%)
Glucosamine and/or chondroitin	24 (12.1%)	21 (11.2%)
Other	5 (2.5%)	10 (5.3%)
Receiving physical therapy at Visit 1 – n (%)		
Yes	3 (1.5%)	2 (1.1%)
No	195 (98.5%)	185 (98.9%)

*Subjects could be in multiple categories.

Source: Clinical Study Report, Study 315, Table 7-4 (with modifications in format)

**APPEARS THIS WAY
ON ORIGINAL**

Table 12: Baseline Hand OA Assessments (ITT Analysis Population)		
	Voltaren (N=198)	Vehicle (N=187)
Global Rating of Disease ^a		
Mean ± SD	57.6 ± 19.0	56.5 ± 19.9
Range	5 – 97	9 – 97
OA Pain Intensity ^b (dominant hand)		
Mean ± SD	73.6 ± 15.6	73.6 ± 14.2
Range	40 – 100	41 – 100
Total AUSCAN Index ^c (dominant hand)		
Mean ± SD	67.2 ± 17.4	66.7 ± 16.8
Range	13 – 96	10 – 98
AUSCAN Pain Index ^c (dominant hand)		
Mean ± SD	66.3 ± 17.9	66.8 ± 16.2
Range	12 – 98	11 – 99
AUSCAN Stiffness Index ^c (dominant hand)		
Mean ± SD	66.0 ± 22.8	66.6 ± 23.9
Range	1 – 98	4 – 100
AUSCAN Physical Function Index ^c (dominant hand)		
Mean ± SD	67.9 ± 18.8	66.7 ± 18.4
Range	9 – 99	8 – 99
OA Pain Intensity ^b (non-dominant hand)		
Mean ± SD	27.8 ± 17.9	30.2 ± 18.2
Range	1 – 77	0 – 78
AUSCAN Pain Index ^c (non-dominant hand)		
Mean ± SD	31.3 ± 19.7	33.9 ± 20.3
Range	0 – 98	0 – 94
AUSCAN Stiffness Index ^c (non-dominant hand)		
Mean ± SD	32.4 ± 22.8	37.1 ± 23.7
Range	0 – 98	1 – 98
FIHOA Index ^d (non-dominant hand)		
Mean ± SD	12.8 ± 4.4	12.5 ± 4.6
Range	1 – 22	1 – 25

a. 100 mm visual analogue scale: 0 = very good, 100 = very poor

b. 100 mm visual analogue scale: 0 = no pain, 100 = unbearable pain

c. Average over multiple questions: 0 = no pain / stiffness / difficulty, 100 = extreme pain / stiffness / difficulty

d. FIHOA index: 0 – 30

Source: Clinical Study Report, Study 315, Table 7-6 (with modifications in format)

The primary efficacy results as presented by the sponsor are included in Table 13. Taken at face value and in a strict statistical sense, the analyses in Table 13 fail to demonstrate efficacy in that the by-treatment group comparison for the global rating of disease endpoint yields a p-value slightly bigger than 0.05 (i.e., p=0.06). According to the pre-specified multiplicity procedures, this should preclude any claims of efficacy at week 4 (as efficacy in all three endpoints was required to make efficacy claims). In addition, strictly speaking, this should also preclude testing of the data for any of the co-primary endpoints at week 6 due to the pre-specified sequential procedure to account for the multiple time points. However, additional consideration is warranted regarding the results as the thinking of the Division of Anesthesia, Analgesia and Rheumatology Products (DAARP) has evolved since the inception of the submitted studies. In

trials of OA, DAARP assesses the analgesic effect of a treatment via a measurement of pain intensity. The global rating of disease assesses patients' overall satisfaction with the treatment and may provide supportive evidence of the analgesic effect. The current standard within DAARP does not require success on multiple endpoints. Instead if multiple endpoints are evaluated, a hierarchical approach is recommended. Statistical significance for the pain endpoint would be required before analyzing endpoints measuring global satisfaction or function. Thus when considering the relative importance of each endpoint as viewed currently and the borderline statistical result for the global rating of disease endpoint, some evidence of an analgesic effect does exist. Limited efficacy claims for Voltaren based on the OA pain intensity and total AUSCAN endpoints for weeks 4 and 6 could be supported by the sponsor's analyses.

Separate from the multiplicity issue, the analyses in Table 13 are subject to ambiguity in the model selection in terms of whether or not to include the treatment-by-OA category interaction. Procedures for inclusion/exclusion of the treatment-by-center and/or treatment-by-OA category interactions in the ANOVA model for the primary efficacy analysis were not explicit. The analyses presented in Table 13 are results of ANOVA models which do not include either of the interaction terms. The protocol stated that if inclusion of the treatment-by-center and/or treatment-by-OA category interactions had a "meaningful impact on the statistical significance of the main effect of treatment", sensitivity analyses would be conducted to address the finding. No significance level for testing the interactions was provided in the protocol. A commonly used significance level for testing interactions is $\alpha=0.10$. Although not prespecified, with this standard, the treatment-by-OA category interaction approached or reached statistical significance in some instances indicating that this interaction should be included in the analysis and suggesting possible heterogeneity of the effect across OA categories. However in FDA analyses, inclusion of the treatment-by-OA category interaction did not affect the qualitative conclusion regarding the treatment effect. Table 14 includes FDA analyses of the co-primary efficacy endpoints that appropriately incorporate the treatment-by-OA category interaction. The statistical methodology used to produce the results weights the OA categories by the number of patients in each category (i.e. type II sums of squares in SAS). This methodology is utilized due to the unequal strata sizes observed in this study (i.e., approximately 70% of subjects had CMC-1 involvement and 30% did not). In contrast, the sponsor employs methodology which in essence averages results across strata while ignoring the strata sizes (i.e. type III sums of squares in SAS). There are two instances displayed in Table 14 where the p-values for the treatment-by-OA category interaction are less than 0.10 indicating that a treatment-by-OA category interaction may exist. The nature of the interaction (if it exists) is that the efficacy of Voltaren relative to vehicle is better in subjects with CMC-1 joint involvement and likely equivalent or worse in subject without CMC-1 joint involvement. The reader may refer to section 4.2 for display of the primary efficacy endpoints by OA category. However, given the borderline nature of the significance of the interaction term and the indication from the clinical team that the efficacy of Voltaren relative to vehicle is not expected to be different depending on the involvement of the CMC-1 joint, it is likely that the borderline significance of the interaction is not reliable and the overall analysis is the most appropriate descriptor of the study results. Those results are included in Table 14 and provide evidence of efficacy for Voltaren over placebo.

Separate from the multiple comparison issue and the issue of a treatment-by-OA-category interaction, the analyses presented in Table 13 for the OA pain intensity endpoint were conducted using an ANOVA model with main effects of treatment and center and baseline OA pain intensity **and baseline OA pain reported in patient diaries** as covariates. Inclusion of the baseline OA pain reported in patient diaries was not specified in the protocol. Additionally

in the study report, the footnote to the sponsor’s table displaying this data does not indicate inclusion of the additional term. The footnote states, “Analysis is analysis of covariance (ANCOVA) with main effects of treatment, center, hand OA category and baseline covariate” and refers to all three primary endpoints. While the sponsor did not provide a rationale for inclusion of this additional baseline term., using the sponsor’s methodology (i.e., type III sums of squares in SAS), evidence of a treatment effect was apparent when the additional term was included in the model with a treatment-by-OA category term; however, the treatment effect was not apparent when the additional baseline term was not included. Post-hoc manipulation of the terms in the ANOVA model (especially without explicitly indicating that the model has been changed) is not appropriate. Table 14 includes FDA analyses of the OA pain endpoint without inclusion of the baseline OA pain reported in patient diaries covariate.

In summary, Table 13 provides the sponsor’s analysis of the three primary endpoints; however, these analyses can be criticized. Table 14 includes the analogous results addressing the problems with the sponsor’s analyses through the following modifications in the statistical approach: (1) by appropriately incorporating the treatment-by-OA category interaction and employing methodology which weights the OA categories by the number of patients in each category and (2) by analyzing the OA pain intensity endpoint without inclusion of the baseline OA pain reported in patient diaries.

Table 13: Primary Efficacy Comparisons as Presented by the Sponsor (MITT Analysis Population) ¹						
	Week 4			Week 6		
	Voltaren	Vehicle	Difference (vehicle - drug)	Voltaren	Vehicle	Difference (vehicle - drug)
OA Pain Intensity (100 mm VAS) ²						
N	198	187		198	187	
Least squares mean	43.3	49.3		41.1	47.4	
95% CI for difference			(1.1, 11.0)			(0.9, 11.7)
p-value			0.018			0.023
Total AUSCAN (100 mm VAS) ³						
N	198	186		198	186	
Least squares mean	44.4	50.7		42.5	49.6	
95% CI for difference			(1.5, 11.2)			(2.1, 12.2)
p-value			0.011			0.006
Global Rating of Disease (100 mm VAS) ³						
N	195	186		195	186	
Least squares mean	38.3	43.2		35.5	41.5	
95% CI for difference			(-0.2, 9.9)			(0.8, 11.2)
p-value			0.060			0.023

1. Missing data imputed as specified in the protocol.

2. Analyses conducted using ANOVA model with main effects of treatment, center, and OA category and baseline OA pain intensity and baseline OA pain reported in patient diaries as covariates. No interactions are included in this model.

3. Analyses conducted using ANOVA model with main effects of treatment, center, and OA category and baseline (of the relevant endpoint) as a covariates. No interactions are included in this model.

Source: Clinical Study Report, Study 315, Tables 9-1 and 9-2 (with modifications in format)

Table 14: Primary Efficacy Comparisons as Conducted by FDA (MITT Analysis Population) ¹						
	Week 4			Week 6		
	Voltaren	Vehicle	Difference (vehicle - drug)	Voltaren	Vehicle	Difference (vehicle - drug)
OA Pain Intensity (100 mm VAS)^{2,3}						
N	198	187		198	187	
Unadjusted mean	42.6	49.7		39.9	46.9	
Weighted mean			6.9			7.0
95% CI for difference			(1.6, 12.1)			(1.4, 12.6)
p-value for trt. effect			0.0106			0.0144
p-value for trt-by-OA category interaction			0.0783			0.2946
Total AUSCAN (100 mm VAS)³						
N	198	186		198	186	
Unadjusted mean	43.7	50.2		41.4	48.5	
Weighted mean			6.3			7.1
95% CI for difference			(1.5, 11.1)			(2.1, 12.2)
p-value for trt. effect			0.0110			0.0061
p-value for trt-by-OA category interaction			0.3524			0.4570
Global Rating of Disease (100 mm VAS)³						
N	195	186		195	186	
Unadjusted mean	37.5	41.9		35.2	40.4	
Weighted mean			4.9			6.0
95% CI for difference			(-0.2, 9.9)			(0.9, 11.2)
p-value for trt. effect			0.0593			0.0230
p-value for trt-by-OA category interaction			0.0837			0.1082

1. Missing data imputed as specified in the protocol.

2. As protocol specified (but unlike the analyses presented by the sponsor – see Table 13), model does not include baseline OA pain reported in the patient diaries.

3. Analyses conducted using ANOVA model with main effects of treatment, center, OA category and baseline (of the relevant endpoint) as a covariates and the treatment-by-OA category interaction. Differences between treatment groups (and the confidence intervals) are weighted in accordance with SAS “type II sums of squares”.

Source: FDA analyses

Several sensitivity analyses to address the robustness of the primary efficacy results to the protocol-specified missing data conventions were provided by the sponsor. Some of these sensitivity analyses are compared to the protocol-specified primary analysis and are provided in Tables 15 through 17. Tables 15 and 16 contain the protocol-specified sensitivity analyses. These analyses mimic those discussed in Section 3.1.1.1 of my review.

Table 15a: Least Squares Means by Treatment Group – Sensitivity Analyses (Week 4)								
	Protocol Specified				Same Mean			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
OA Pain Intensity	43.3	49.3	6.0	0.02	41.6	46.4	4.8	0.048
Total AUSCAN	44.4	50.7	6.3	0.01	43.2	47.6	4.4	0.06
Global Rating of Disease	38.3	43.2	4.9	0.06	37.2	40.8	3.6	0.14

Table 15b: Least Squares Means by Treatment Group – Sensitivity Analyses (Week 6)								
	Protocol Specified				Same Mean			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
OA Pain Intensity	41.1	47.4	6.3	0.02	38.7	42.6	3.9	0.13
Total AUSCAN	42.5	49.6	7.1	0.006	40.2	45.2	4.9	0.04
Global Rating of Disease	35.5	41.5	6.0	0.02	33.8	37.5	3.7	0.13

Source: Clinical Study Report, Appendix 5 (Statistical Methods, Randomization Scheme, and Codes), Section 5.1.2

Table 16a: Least Squares Means by Treatment Group – Sensitivity Analyses (Week 4)								
	Protocol Specified				Alternate Mean			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
OA Pain Intensity	43.3	49.3	6.0	0.02	42.1	46.0	3.9	0.11
Total AUSCAN	44.4	50.7	6.3	0.01	43.5	47.3	3.8	0.11
Global Rating of Disease	38.3	43.2	4.9	0.06	37.5	40.5	3.0	0.23

Table 16b: Least Squares Means by Treatment Group – Sensitivity Analyses (Week 6)								
	Protocol Specified				Alternate Mean			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
OA Pain Intensity	41.1	47.4	6.3	0.02	39.1	42.0	2.9	0.27
Total AUSCAN	42.5	49.6	7.1	0.006	40.7	44.6	3.9	0.11
Global Rating of Disease	35.5	41.5	6.0	0.02	34.1	37.2	3.1	0.21

Source: Clinical Study Report, Appendix 5 (Statistical Methods, Randomization Scheme, and Codes), Section 5.1.2

Table 17a: Least Squares Means by Treatment Group – Sensitivity Analyses (Week 4)								
	Protocol Specified				BOCF			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
OA Pain Intensity	43.3	49.3	6.0	0.02	44.3	49.7	5.5	0.03
Total AUSCAN	44.4	50.7	6.3	0.01	45.2	50.6	5.4	0.03
Global Rating of Disease	38.3	43.2	4.9	0.06	39.3	43.8	4.5	0.08

Table 17b: Least Squares Means by Treatment Group – Sensitivity Analyses (Week 6)								
	Protocol Specified				BOCF			
	Voltaren	Vehicle	Diff	p-value	Voltaren	Vehicle	Diff	p-value
OA Pain Intensity	41.1	47.4	6.3	0.02	42.6	48.1	5.6	0.045
Total AUSCAN	42.5	49.6	7.1	0.006	43.9	49.8	6.0	0.02
Global Rating of Disease	35.5	41.5	6.0	0.02	36.7	42.4	5.7	0.03

Source: Clinical Study Report, Appendix 5 (Statistical Methods, Randomization Scheme, and Codes), Section 5.1.2

The analyses incorporating the “same mean” imputation scheme are somewhat supportive of the conclusions from the protocol-specified primary efficacy analyses. As expected, the least squares mean for each endpoint and each treatment group is numerically higher in the primary efficacy analysis; however, the numerical differences between treatment groups continue to favor the Voltaren group.

Statistically significant by-treatment group differences are not achieved in the analyses incorporating the “alternate mean” imputation scheme. However, as highlighted by the sponsor, this analysis is conservative and biased in favor of the vehicle group in this case since more favorable results are imputed for the missing data in the vehicle group than in the Voltaren group. The lack of statistically significant findings in this analysis is not unexpected.

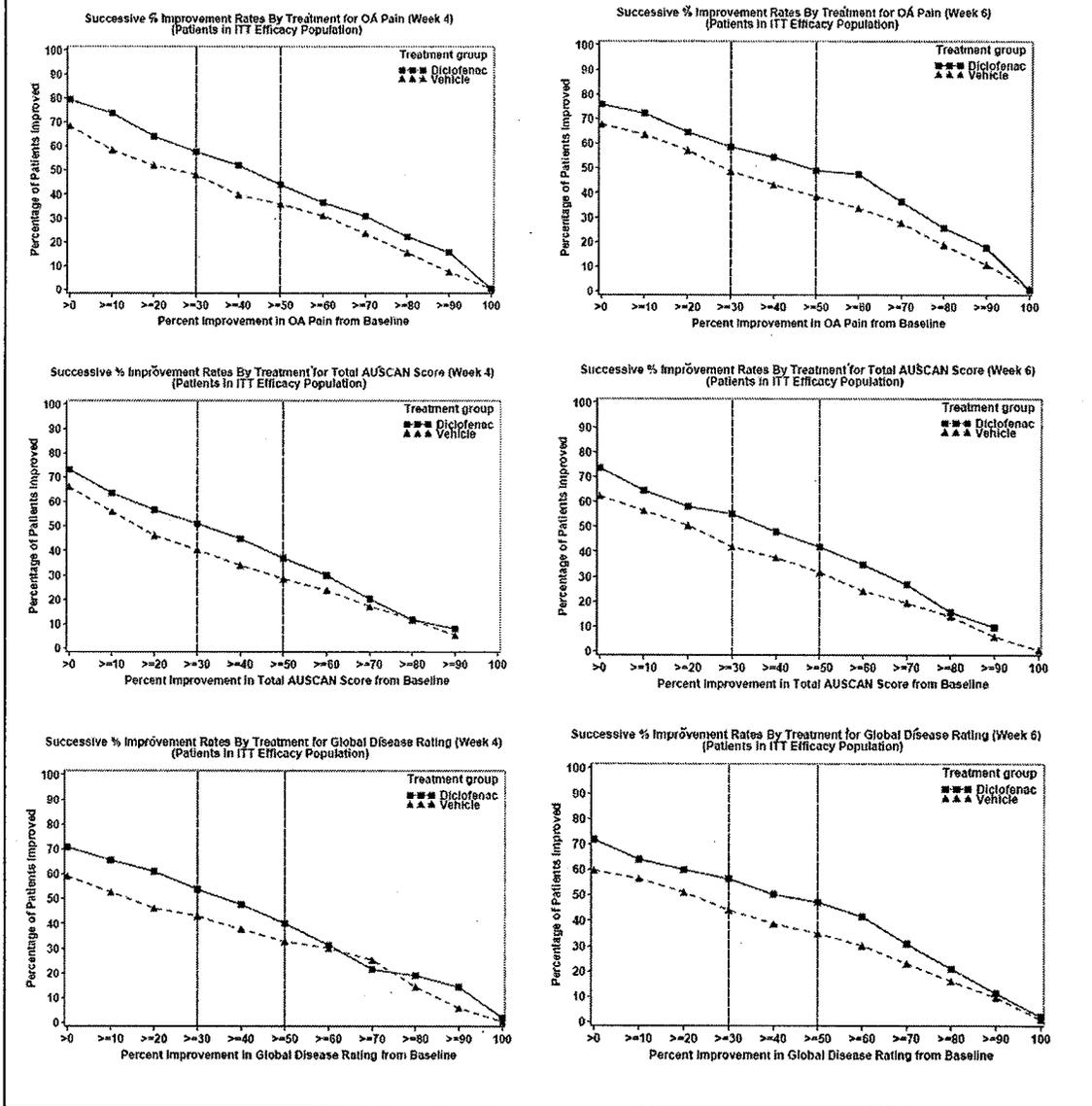
The results of the analyses when imputing according to the BOCF approach yield similar conclusions to the primary efficacy analyses and therefore, provide support for the primary efficacy analyses and minimize the concern regarding the possible impact of the missing data.

In summary, the qualitative conclusions from the missing data sensitivity analyses are supportive of the protocol-specified primary efficacy analysis. Several instances with p-values greater than 0.05 do occur but the numerical by-treatment group differences continue to favor Voltaren.

At the request of the division, the sponsor provided cumulative distribution plots (i.e., a continuous responder analyses) for the primary efficacy endpoints for weeks 4 and 6. All discontinuations are classified as treatment failures in the formulation of the plots. The descriptive conclusions from these plots are supportive of the efficacy of Voltaren over vehicle for the primary efficacy endpoints and are provided in Figure 4.

**APPEARS THIS WAY
ON ORIGINAL**

Figure 4: Continuous Responder Analysis of Primary Efficacy Endpoints (Weeks 4 & 6)



Source: Clinical Study Report, Post-text Supplement 3, Figure 9.18.1.1 through 9.18.3.2

4. FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

4.1 Gender, Race and Age

Summaries of the primary efficacy endpoints are given in Tables 18 through 20 for studies 310 using the MES population and 315 using the ITT populations.

Table 18: Primary Efficacy Endpoints by Gender										
		Study 310 (Week 12, MES)			Study 315 (Week 4, ITT)			Study 315 (Week 6, ITT)		
		Voltaren	Vehicle	Diff. (vehicle - drug)	Voltaren	Vehicle	Diff. (vehicle - drug)	Voltaren	Vehicle	Diff. (vehicle - drug)
WOMAC Pain Index (scale = 0 to 20) / OA Pain Intensity (scale = 0 to 100)										
Males	N	44	40		46	43		46	43	
	Mean	5.61	6.53	0.92	44.2	48.0	3.8	39.5	48.7	9.2
	95% CI for diff			(-0.91, 2.75)			(-8.16, 15.76)			(-2.98, 21.38)
Females	N	83	79		152	144		152	144	
	Mean	5.63	7.81	2.18	42.1	50.2	8.1	40.1	46.4	6.3
	95% CI for diff			(0.60, 3.76)			(1.26, 14.94)			(-0.83, 13.43)
WOMAC Function Index (scale = 0 to 68) / Total AUSCAN (scale = 0 to 100)										
Males	N	44	40		46	43		46	43	
	Mean	18.55	22.68	4.13	43.0	45.7	2.7	39.6	46.6	7.0
	95% CI for diff			(-2.16, 10.4)			(-8.49, 13.89)			(-4.35, 18.35)
Females	N	83	79		152	143		152	143	
	Mean	20.27	27.27	7.00	43.9	51.5	7.6	41.9	49.1	7.2
	95% CI for diff			(1.72, 12.28)			(1.20, 14.00)			(0.61, 13.79)
Global Rating of Disease (100 mm VAS)										
Males	N	44	40		46	43		46	43	
	Mean	29.7	37.0	7.3	41.3	40.7	-0.6	37.6	42.6	5.0
	95% CI for diff			(-4.56, 19.2)			(-11.72, 10.52)			(-6.18, 16.18)
Females	N	83	79		149	143		149	143	
	Mean	32.7	43.8	11.1	36.4	42.2	5.8	34.5	39.7	5.2
	95% CI for diff			(1.64, 20.6)			(-0.22, 11.82)			(-0.95, 11.35)

Source: Clinical Study Report, Study 310 and 315, Appendix 8 (Additional data and information), Tables 9.2-2 through 9.4-2(with modifications in format)

**APPEARS THIS WAY
ON ORIGINAL**

Table 19: Primary Efficacy Endpoints by Race										
		Study 310 (week 12, MES)			Study 315 (week 4, ITT)			Study 315 (week 6, ITT)		
		Voltaren	Vehicle	Diff. (vh-dr)	Voltaren	Vehicle	Diff. (vh-dr)	Voltaren	Vehicle	Diff. (vh-dr)
WOMAC Pain Index (scale = 0 to 20) / OA Pain Intensity (scale = 0 to 100)										
Caucasian	N	101	99		173	170		173	170	
	Mean	5.59	7.13	1.54	41.2	49.6	8.4	39.4	46.2	6.8
	95% CI for diff			(0.18, 2.90)			(2.20, 14.60)			(0.35, 13.25)
Black	N	9	10		11	4		11	4	
	Mean	5.67	10.40	4.73	56.4	51.0	-5.4	52.1	57.3	5.2
	95% CI for diff			(0.25, 9.21)			(-48.16, 37.36)			(-40.54, 50.94)
Asian	N	4	0		3	0		3	0	
	Mean	5.00	NA	NA	30.3	NA	NA	22.7	NA	NA
	95% CI for diff			NA			NA			NA
Other	N	13	10		11	13		11	13	
	Mean	6.00	6.80	0.80	53.6	50.5	-3.1	40.8	52.8	12.0
	95% CI for diff			(-2.91, 4.51)			(-27.87, 21.67)			(-13.40, 37.40)
WOMAC Function Index (scale = 0 to 68) / Total AUSCAN (scale = 0 to 100)										
Caucasian	N	101	99		173	169		173	169	
	Mean	19.75	24.53	4.78	42.6	49.6	7.0	40.7	47.5	6.8
	95% CI for diff			(0.23, 9.33)			(1.16, 12.84)			(0.81, 12.79)
Black	N	9	10		11	4		11	4	
	Mean	20.33	38.30	17.97	56.3	59.4	3.1	54.8	63.5	8.7
	95% CI for diff			(2.69, 33.25)			(-32.58, 38.78)			(-30.21, 47.61)
Asian	N	4	0		3	0		3	0	
	Mean	20.50	NA	NA	25.0	NA	NA	34.5	NA	NA
	95% CI for diff			NA			NA			NA
Other	N	13	10		11	13		11	13	
	Mean	19.23	25.00	5.77	49.2	39.9	-9.3	40.0	57.0	17.0
	95% CI for diff			(-6.79, 18.33)			(-31.91, 13.31)			(-5.15, 39.15)
Global Rating of Disease (100 mm VAS)										
Caucasian	N	101	99		170	169		170	169	
	Mean	32.4	42.4	10.00	36.3	41.9	5.6	34.7	39.8	5.1
	95% CI for diff			(1.68, 18.32)			(0.07, 11.13)			(-0.51, 10.71)
Black	N	9	10		11	4		11	4	
	Mean	27.7	40.8	13.1	49.4	47.8	-1.6	46.5	54.3	7.8
	95% CI for diff			(-13.77, 39.97)			(-37.95, 34.75)			(-31.36, 46.96)
Asian	N	4	0		3	0		3	0	
	Mean	17.3	NA	NA	25.0	NA	NA	20.0	NA	NA
	95% CI for diff			NA			NA			NA
Other	N	13	10		11	13		11	13	
	Mean	32.7	32.8	0.1	49.2	39.9	-9.3	36.8	44.2	7.4
	95% CI for diff			(-24.30, 24.50)			(-31.91, 13.31)			(-16.56, 31.36)

Source: Clinical Study Report, Study 310 and 315, Appendix 8 (Additional data and information), Tables 9.2-4 through 9.4-4 (with modifications in format)

Table 20: Primary Efficacy Endpoints by Age										
		Study 310 (week 12, MES)			Study 315 (week 4, ITT)			Study 315 (week 6, ITT)		
		Voltaren	Vehicle	Diff. (vehicle - drug)	Voltaren	Vehicle	Diff. (vehicle - drug)	Voltaren	Vehicle	Diff. (vehicle - drug)
WOMAC Pain Index (scale = 0 to 20) / OA Pain Intensity (scale = 0 to 100)										
< 65 yrs	N	86	85	1.43 (-0.07, 2.93)	109	100	4.9 (-3.51, 13.31)	109	100	5.0 (-3.51, 13.51)
	Mean	5.57	7.00		43.6	48.5		40.8	45.8	
	95% CI for diff									
≥ 65 yrs	N	41	34	2.59 (0.53, 4.65)	89	87	9.8 (1.47, 18.13)	89	87	9.4 (0.48, 18.32)
	Mean	5.73	8.32		41.3	51.1		38.9	48.3	
	95% CI for diff									
WOMAC Function Index (scale = 0 to 68) / Total AUSCAN (scale = 0 to 100)										
< 65 yrs	N	86	85	6.30 (1.26, 11.34)	109	99	3.8 (-4.07, 11.67)	109	99	9.4 (0.48, 18.32)
	Mean	18.36	24.66		44.5	48.3		38.9	48.3	
	95% CI for diff									
≥ 65 yrs	N	41	34	5.97 (-0.93, 12.87)	89	87	9.5 (1.70, 17.30)	89	87	8.8 (0.70, 16.90)
	Mean	22.41	28.38		42.8	52.3		40.4	49.2	
	95% CI for diff									
Global Rating of Disease (100 mm VAS)										
< 65 yrs	N	86	85	8.20 (-0.52, 16.92)	108	99	2.0 (-5.44, 9.44)	108	99	5.2 (-2.35, 12.75)
	Mean	29.4	37.6		38.5	40.5		36.1	41.3	
	95% CI for diff									
≥ 65 yrs	N	41	34	15.0 (1.31, 28.69)	87	87	7.0 (-0.50, 14.50)	87	87	5.3 (-2.37, 12.97)
	Mean	36.2	51.2		36.4	43.4		34.1	39.4	
	95% CI for diff									

Source: Clinical Study Report, Study 310 and 315, Appendix 8 (Additional data and information), Tables 9.2-3 through 9.4-3 (with modifications in format)

4.2 Other Special/Subgroup Populations

Table 21 presents the primary efficacy endpoints by OA category for study 315. This subgroup analysis is of particular importance due to a possible interaction between treatment and OA category. The nature of the interaction (if it exists) is that the efficacy of Voltaren relative to vehicle is better in subjects with CMC-1 joint involvement and likely equivalent or worse in subject without CMC-1 joint involvement. Refer to section 3.1.1.2, specifically Table 14, for further discussion.

APPEARS THIS WAY
ON ORIGINAL

Table 21: Primary Efficacy Endpoints in MITT by OA Category (Study 315)							
		Week 4			Week 6		
		Voltaren	Vehicle	Diff. (vehicle - drug)	Voltaren	Vehicle	Diff. (vehicle - drug)
OA Pain Intensity (100 mm VAS)							
With CMC-1 joint involvement	N	133	142	6.6	133	142	6.7
	Mean	35.4	42.0		33.5	40.2	
	95% CI for difference				(0.4, 12.8)		
Without CMC-1 joint involvement	N	65	45	-0.1	65	45	-1.2
	Mean	40.8	40.7		39.2	40.3	
	95% CI for difference				(-10.3, 10.2)		
Total AUSCAN (100 mm VAS)							
With CMC-1 joint involvement	N	133	141	6.9	133	141	7.0
	Mean	35.4	42.3		33.5	40.5	
	95% CI for difference				(0.7, 13.1)		
Without CMC-1 joint involvement	N	65	45	-0.1	65	45	1.2
	Mean	40.8	40.7		39.2	40.3	
	95% CI for difference				(-10.3, 10.2)		
Global Rating of Disease (100 mm VAS)							
With CMC-1 joint involvement	N	131	141	6.5	131	141	7.3
	Mean	35.7	42.3		33.2	40.4	
	95% CI for difference				(0.3, 12.8)		
Without CMC-1 joint involvement	N	64	45	-0.6	64	45	0.9
	Mean	41.3	40.7		39.5	40.3	
	95% CI for difference				(-10.8, 9.7)		

Source: FDA analyses

5. SUMMARY AND CONCLUSIONS

5.1 Statistical Issues and Collective Evidence

The following statistical issues and their impact have been described in the context of the review. Please refer to the specified section for details.

Study 310 (OA of the knee)

- ❖ As part of a protocol amendment and prior to unblinding, a modified efficacy subpopulation (MES) was defined and designated as the primary efficacy analysis group. The MES was a subset of subjects in the ITT efficacy population excluding all subjects whose pain on movement score in the target knee declined between the screening visit and the baseline visit or with a score of >1 on the abridged pain index for the contralateral knee at the baseline visit. From a statistical standpoint, the random treatment assignment in the MES group is valid. (Section 3.1.1.1)
- ❖ An analysis of variance (ANOVA) model including treatment and center main effects and baseline as a covariate was used. ANOVA models including a treatment-by-center interaction were explored by the sponsor but ultimately the interaction term was dropped from the model due to insignificance of the term and little variation in the least squares means for the main effect of treatment resulting from models with and without the interaction term. Although the conventions for dropping the treatment-by-center interaction term from the model were not completely described in the protocol, fortunately, the interaction term was clearly non-significant and therefore, dropping the interaction term is considered reasonable. (Section 3.1.1.2)

- ❖ The qualitative conclusions from multiple sensitivity analyses are supportive of the protocol-specified primary efficacy analysis minimizing the concern regarding the possible impact of the missing data. (Section 3.1.1.2)

Study 315 (OA of the hand)

- ❖ Prior to database closure and in a blinded fashion (i.e., in a “blind data review” meeting), it was determined that subjects with baseline scores which allowed little or no room for improvement would be excluded from the applicable analysis. While the random treatment assignment is valid regardless of the exclusion of subjects based on pre-randomization characteristics, one may question whether the criterion for exclusion of the subjects was determined in order to yield the most desirable result for the treatment effect since this criterion was developed after the data was collected. This change should have been formally documented as a protocol amendment so that criticism regarding the relative timing of unblinding and development of the criteria could be avoided. Since only five subjects were affected by this change, there was little impact on the study results. (Section 3.1.2.2)
- ❖ According to the prespecified multiplicity procedures, the sponsor’s primary efficacy results fail to demonstrate efficacy in that the by-treatment group comparison for the global rating of disease endpoint yields a p-value slightly larger than 0.05 (i.e., $p=0.06$). This statistical result for the global rating of disease endpoint was borderline and could be viewed differently if the current standard was applied (Section 3.1.2.2)
- ❖ Comparison between treatment groups in each primary outcome was to be done with an ANOVA model including treatment, center, and hand OA category main effects, baseline as a covariate, and the treatment-by-center and treatment-by-OA category interactions. The conventions for dropping the interactions from the model were not explicit in the protocol and the treatment-by-OA category interaction approached or reached statistical significance in some instances suggesting the possibility of heterogeneity of the treatment effect across OA categories. The nature of the interaction (if it exists) is that the efficacy of Voltaren relative to vehicle is better in subjects with CMC-1 joint involvement and likely equivalent or worse in subject without CMC-1 joint involvement. In FDA analyses weighting the OA categories by the number of patients in each category, inclusion of the treatment-by-OA category interaction did not affect the qualitative conclusion regarding the treatment effect. (Sections 3.1.2.2 and 4.2)
- ❖ The sponsor’s analyses presented for the OA pain intensity endpoint were conducted using an ANOVA model with main effects of treatment and center and baseline OA pain intensity **and baseline OA pain reported in patient diaries** as covariates. Inclusion of the baseline OA pain reported in patient diaries was not specified in the protocol or study report. FDA analyses excluding this factor are provided. (Section 3.1.2.2)

5.2 Conclusions and Recommendations

In Study 310, in patients with OA of the knee, the Voltaren group had statistically significant better average outcomes in the three primary efficacy endpoints (WOMAC pain index, WOMAC function index, and global rating of disease) than vehicle. These differences were observed in both the MES and ITT analysis groups. The efficacy conclusions are robust against concerns regarding missing data as multiple sensitivity analyses yielded supportive conclusions.

In Study 315, in patients with OA of the hand, there is evidence of an analgesic effect of Voltaren despite the inability of one endpoint to achieve statistical significance. Comparison of the average outcomes for two (OA pain and total AUSCAN score) of the three primary efficacy

endpoints at weeks 4 and 6 resulted in p-values less than 0.05. The comparison of the mean global rating of disease at week 4 resulted in a p-value of 0.06 and resulted in a p-value less than 0.05 at week 6. According to the prespecified multiplicity procedures, the result for the global rating of disease endpoint at week 4 should have precluded any claims of efficacy and testing the primary efficacy endpoints at week 6. However due to the borderline nature of this result, the relative clinical importance of the three endpoints and the fact that the conclusion would be different if a hierarchical multiple comparison procedure were implemented (an approach that would seem reasonable if the protocol were being designed today), this study does provide supportive evidence of efficacy of Voltaren over vehicle despite the failure to satisfy the strict multiple comparison procedure. The study identified a possible treatment-by-OA interaction. The nature of the interaction (if it exists) is that the efficacy of Voltaren relative to vehicle is better in subjects with CMC-1 joint involvement and likely equivalent or worse in subject without CMC-1 joint involvement. In analyses weighted by strata size, inclusion of the treatment-by-OA category interaction did not affect the qualitative conclusion regarding the treatment effect. The primary efficacy conclusions are robust against concerns regarding missing data as multiple sensitivity analyses yielded supportive conclusions.

The following recommendations are being made for the Clinical Studies section of the Voltaren labeling.

- Inclusion of ~~_____~~, it is recommended that this information not be included in the label.
- The following recommendations are made for the Table displaying the efficacy data from studies 310 and 315.
 - The sample size in each treatment group should be provided.
 - ~~_____~~ need not be displayed.
 - Heading for the column including the difference between Voltaren and placebo should indicate that the difference was calculated as placebo minus Voltaren.

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

/s/

Ruth Davi
9/5/2007 11:24:15 AM
BIOMETRICS

Dionne Price
9/5/2007 01:33:36 PM
BIOMETRICS
Concur