CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER: 21-107

STATISTICAL REVIEW(S)

# Statistical Review and Evaluation

NDA#: 21-107\Class 1-P

Name of Drug: Lotronex (alosetron)

Applicant: Glaxo Wellcome, Inc.

Documents Reviewed: Vols 1.138, 1.139, 1.158, 1.167, 1.168, 1.189

Medical Input: Robert Prizont, M.D., HFD-180

## Background

The sponsor has submitted one Phase II dose-ranging trial (S3BA2001) and two Phase III trials with identical protocols (S3BA3001 & SB3A3002) in support of the efficacy and safety of alosetron for diarrhea-predominant IBS Irritable Bowel Syndrome (IBS) in women. In the Phase II trial, the sponsor found that men did not benefit from alosetron therapy. Thus, the subsequent Phase III trials enrolled only women. In one of the Phase III trials (3002), the sponsor concluded that efficacy on the primary endpoint was demonstrated only in the subgroup of women with the diarrhea-predominant type (D-IBS) but *not* in the alternating diarrhea/constipation (A-IBS) or the constipation-predominant types of IBS. The sponsor subsequently performed *post hoc* analyses (not formulated before unblinding the data) on the D-IBS and A-IBS subgroups in study 3001. This review examined the data from only the Phase III studies.

Both trials had identical protocols. They were randomized (4 patients per permuted block), double-blind, multi-center, parallel group, placebo-controlled trials in US centers. The duration of each study was 18 weeks including a two-week screening phase, a 12-week double-blind treatment period, and a 4-week post-treatment follow-up period. Patients were randomized within each center to either 10 mg BID alosetron (A) BID or placebo (P) BID. Key inclusion criteria included an average abdominal pain/discomfort score between 1.0 and 3.3 during the screening phase (where 1=mild, 2=moderate, 3=intense, and 4=severe), and an average stool consistency score of at least 2.5 (where 1=very hard, 2=hard, 3=formed, 4=loose, 5=watery).

According to the protocols, the **primary clinical endpoint** was the patient's weekly response in a diary to the question: "In the past 7 days, have you had adequate relief of your Irritable Bowel Syndrome pain and discomfort (Yes/No)?" The primary analysis compared the number of **"monthly responders"** (patients one who indicated "adequate relief" for at least 2 weeks out of the month). Thus a patient could be a responder for any of months 1, 2, or 3.

Secondary endpoints included a daily pain severity score (0= no pain, 1=mild, 2=moderate, 3=intense, 4=severe), proportion of pain-free days, and evaluations of Lower GI functions such as number of times stool passed/ day and stool consistency using the scale mentioned above in connection with the inclusion criteria. Sense of urgency, bloating , and sense of incomplete evacuation were also evaluated using daily reports of 'Yes/No' to the presence of each symptom. **Amendment 2** contains a "step-down" (closed testing) plan for secondary endpoints where the order of endpoints to be tested would be 1) stool consistency, 2) sense of urgency, 3) stool frequency, 4) sense of incomplete evacuation, and 5) bloating, in that order. **The primary time point for these analyses was to be the change from baseline at month 1, "and if significance is demonstrated for this interval, change from baseline will then be interpreted for each week in the interval..." The protocol does not specify how comparisons of pain scores will control Type I error.**

Key words: clinical studies, NDA review, missing data

The target sample size was 300 patients/group. This number was derived from the protocol-specified definition of adequate relief responders. The sponsor assumed that the P group would have 40% responders and the A group 55%. This results in 90% power at the .05 level.

Due to small numbers of patients in many centers, centers were pooled into 5 *a priori* groupings using groups of states as clusters (strata). The clusters were defined in Amendment #2 for Study 3001 on 11 March 1998.

For the primary analysis, Last Observation Carried Forward (LOCF) was used "whereby months with all missing weeks of adequate relief are replaced by the number of weeks with relief in the previous non-missing month". Since there were 3 months of evaluation, the sponsor proposed a multiple endpoint adjustment using O'Brien's global testing approach. If the global test was significant at the .05 level, Koch and Gansky's strategy was used: viz., each month was analyzed separately for treatment effect at the nominal .05 level using the CMH test using geographic clusters as strata. In addition to monthly responders, a full trial responder was defined as anyone who completed the study and reported adequate relief for at least 6 of the study's 12 weeks.

Pain-free Days would be analyzed by defining a "monthly responder to be one who reported at least 50% pain/discomfort-free days in a month with at least 14 daily pain assessments".

Results

**3001**

Out of one hundred fifteen (115) centers slated to enroll patients 104 did so, ultimately randomizing a total of 626 patients {N=317 (P), N=309 (A)} after the 2-week Screening Phase. The number of patients randomized among the centers ranged from a low of 1 to a high of 23. The five (5) geographical clusters contained between 3 and 15 states. **There were no imbalances in important baseline characteristics between the two groups. In all 4 randomized groups (between the two trials), the mean baseline quantities were: pain score 2.0, stool consistency 3.4, stool frequency 2.7/day, sense of urgency 69% days/week, abdominal bloating 77% days/week, and incomplete evacuation 70% days/week.**

According to the sponsor, 77% of the (A) patients and 78% of the (P) patients "completed the study". See Table 1 for patient flow throughout the trial. The numbers quoted by the sponsor may not always match those used in calculations made by the reviewer due to the difference between the number of subjects who had a measurement at any particular time on the data set used for the reviewer's analyses, and the sponsor's definition (unknown) of what constitutes a 'completer'.

There were a total of 446 patients with diarrhea-predominant IBS (D-IBS) and 169 patients with alternating diarrhea/constipation IBS (A-IBS). The remaining 11 patients had constipation-predominant IBS. These categories were listed on the case report form for the investigator to assign when first evaluating the patient.

There were 469 patients who had a 'adequate response' evaluation at week 12. Thus there were 626-469= 157 patients who did not have a week-12 evaluation for the primary endpoint. The table below displays the number of patients whose last 'adequate response' evaluation occurred at a particular week, i.e., the number of patients in each 'dropout cohort', where **'dropout' refers not to a formal exit from the trial as might be defined by the sponsor, but to the last week at which the patient was evaluated for the primary endpoint.**

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # patients | 12 | 28 | 13 | 14 | 19 | 12 | 10 | 4 | 14 | 7 | 4 | 20 | 469 | 626 |

Key words: clinical studies, NDA review, missing data

One hundred twenty-five (125) centers randomizing a total of 647 patients {N=323 (P), N=324 (A)} after the 2-week Screening Phase. The number of patients randomized among the centers ranged from a low of 1 to a high of 33. **There were no apparent imbalances in baseline characteristics between the two groups.**

According to the sponsor, 76% of the (A) patients and 84% of the (P) patients "completed the study". See Table 2. The numbers quoted by the sponsor may not always match those used in calculations made by the reviewer due to the difference between the number of subjects who had a measurement at any particular time on the data set used for the reviewer's analyses, and the sponsor's definition (unknown) of what constitutes a 'completer'.

There were a total of 458 patients with diarrhea-predominant IBS (D-IBS) and 180 patients with alternating diarrhea/constipation IBS (A-IBS). The remaining 9 patients had constipation-predominant IBS.

There were 499 patients who had a 'adequate response' evaluation at week 12. Thus there were 647-499= 148 patients who did not have a week-12 evaluation for the primary endpoint. The table below displays the number of patients whose last 'adequate response' evaluation occurred at a particular week, i.e., the number of patients in each 'dropout cohort'.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # patients | 17 | 18 | 19 | 19 | 13 | 4 | 6 | 5 | 9 | 6 | 7 | 25 | 499 | 647 |

It is apparent that both trials suffered from a substantial number of patients who exited prematurely, approximately 25% in each trial. An examination of the numbers and timing of dropouts in each group should help to assess the any difficulty in drawing conclusions about efficacy. The reasons for dropping out, together with their frequency in each treatment group over time, are displayed in the table below. The columns 'W4', 'W8', and 'W11' refer to separate epochs during which patients dropped out, i.e., 'W4' includes patients who dropped out during the first 4 weeks of the trial, 'W8' includes patients who dropped out between weeks 5 and 8, and 'W11' includes patients who dropped out between weeks 9 and 11. AE= Adverse Event, CW= Consent Withdrawn, LOE= Lack of Efficacy, LTF= Lost to Follow-up, and PV= Protocol Violation. The "miscellaneous" category is not the sponsor's. It accounts for patients who were not evaluated for the primary endpoint but were not accounted for by the sponsor.

| | | 3001 | | | 3002 | | |
|---|---|---|---|---|---|---|---|
| | | W4 | W8 | W11 | W4 | W8 | W11 |
| AE | (P) | 8 | 7 | 6 | 10 | 2 | 1 |
| | (A) | 36 | 11 | 1 | 36 | 10 | 3 |
| CW | (P) | 16 | 8 | 1 | 6 | 1 | 0 |
| | (A) | 4 | 1 | 1 | 7 | 0 | 0 |
| LOE | (P) | 3 | 4 | 0 | 8 | 5 | 1 |
| | (A) | 4 | 2 | 1 | 4 | 1 | 1 |
| LTF | (P) | 4 | 4 | 2 | 5 | 1 | 1 |
| | (A) | 4 | 1 | 0 | 6 | 1 | 2 |
| PV | (P) | 0 | 0 | 0 | 0 | 1 | 1 |
| | (A) | 0 | 1 | 0 | 1 | 0 | 0 |

Key words: clinical studies, NDA review, missing data

| | | 3001 | | | 3002 | |
|---|---|---|---|---|---|---|
| | W4 | W8 | W11 | W4 | W8 | W11 |
| OTHER (P) | 5 | 0 | 1 | 1 | 0 | 1 |
| (A) | 1 | 1 | 2 | 0 | 1 | 0 |
| MISC. (P) | 0 | 0 | 6 | 0 | 1 | 13 |
| (A) | 1 | 0 | 10 | 2 | 0 | 14 |
| TOTAL | 86 | 40 | 31 | 86 | 24 | 38 |

These tables indicate that a substantial number of dropouts occurred within the first 4 weeks. It is clear that, initially, there were patients who were assigned to alosetron who felt they could not continue in the trial due to adverse events (partly constipation) and that placebo patients left for a variety of reasons. Dropouts tended to taper off after 4 weeks. The reason for the substantial number of "Withdrawn Consents" in the placebo group in study 3001 has not been determined. In order to assess the extent to which dropouts contributed to the achievement of adequate relief response, this reviewer tabulated the number of dropouts achieving adequate relief in each group:

In Study 3001, there were a total of 40 dropouts who happen to be monthly responders for adequate relief for at least one month during the trial, 18 (P) and 22 (A). Of the 22 (A) dropouts, 9 did so due to adverse events.

In Study 3002, there were a total of 38 dropouts who happen to be monthly responders for adequate relief for at least one month during the trial, 16 (P) and 22 (A). Of the 22 (A) dropouts, 7 did so due to adverse events.

Reviewer's Comments

Dropouts are always a concern in clinical trials. They can occur randomly or non-randomly. If non-randomly, they can complicate the interpretation of any statistic which purports to represents some "overall effect" at a specified time in the trial, usually the end. However, it is often possible to make a judgement about whether the drug is active compared to placebo in the presence of dropouts. One consideration is whether the number and pattern of dropouts is such that any attempt to distinguish drug from placebo is hopelessly compromised due to bias introduced in the analysis. That is, is there any substantial reason in these trials that dropouts would lead one to suspect that any comparison of the drug to placebo would be so biased in favor of the drug as to preclude a useful statistical analysis? Of all the reasons for withdrawal, the only one which is specific enough to likely affect the comparison of the two arms is "adverse events." Clearly those dropouts are not random. It could be that patients who drop out on alosetron contribute more adequate relief responses than placebo dropouts. But this does not appear to be the case as shown immediately above. In addition, the bulk of non-random dropouts occurred within the first 4 weeks of the trial, thus leaving the remaining cohort relatively free of non-random dropouts. Since there were non-random dropouts, it is not possible to estimate a "true" treatment difference at any particular time. However, using all the data in the trial, one can ask the global question: "Is there convincing evidence that the distribution of responses on the drug is different from that on placebo, given the pattern of dropouts?" If the pattern and number of dropouts is judged not to have overwhelmingly determined the result of the treatment comparison, then a statistical analysis may be reasonable. Similar results of analyses using the 75% of the initial cohorts who completed the trial are useful as a way to check that dropouts did not unduly affect the evidence which will lead to an inference concerning the activity of the drug.

Key words: clinical studies, NDA review, missing data

Primary Endpoint-Sponsor's Analyses

According to the sponsor, the "O'Brien test" referred to in the protocol was intended to be essentially equivalent to a Mantel- Haentszel test comparing the number of monthly adequate responders. For Study 3001, the p-value was .001 and that for Study 3002 was .012 using LOCF for adequate response data. Thus they concluded that testing at each month was justified. **Since non-random dropouts are of concern, this reviewer repeated the sponsor's analysis deleting patients with adverse events. P-values of .001 resulted in both trials. Figure 1** displays this reviewer's plot of the cumulative percent of the number of weeks of adequate response for each group in study 3001 without carrying response data forward when it is missing; i.e., times at which the patient was not in the trial are counted as times of non-response. The figure indicates that subjects receiving alosetron tended to have more weeks of adequate relief. The median for placebo was approximately 2-3 weeks and the median for alosetron patients is approximately 5-6 weeks.

**Tables 3a, 3b, and 3c** illustrate the percentage of patients who achieved adequate relief at each month of the trial for the ITT, D-IBS and A-IBS cohorts, respectively using the sponsor's LOCF strategy in Study 3001. **Tables 4a, 4b, and 4c** display the respective results for Study 3002. In both studies statistical significance was reported in both the ITT and D-IBS populations, but not in the A-IBS population.

**Tables 5 and 6** display the percentages of patients in each group who achieved various numbers of months of adequate relief in the ITT cohort and the 2 subgroups for Studies 3001 and 3002, respectively. Note that in **Table 4c** (Study 3002) and **Table 6** (Study 3002), placebo patients in the Alternating Constipation/Diarrhea group tended to do better than alosetron patients. This is the finding which led the sponsor to narrow the target patient population to diarrhea-predominant women. The sponsor used a proportional odds model to analyze the four possible outcomes: 0, 1, 2, or 3 months adequate relief. Not surprisingly, the term modeling the interaction between IBS type and treatment was significant ($p<.001$).

Interaction Between Treatment and Menses?

The sponsor reports that in Study 3001 patients who had a menstrual cycle at any time during the study showed greater frequencies of adequate relief compared to placebo than those patients who did not menstruate during the study did. That is, a proportional odds model produced a p-value of .023 for a treatment by menstruation interaction. Specifically, although about 25% of the placebo patients achieved adequate relief in all 3 months in both menstruating and non-menstruating subgroups, 50% of the alosetron patients who menstruated achieved 3 months of adequate relief, whereas only 34% of alosetron patients in the non-menstruating category did. See **Figure 2** which illustrates the distributions for each group and menstrual status. The 'right hand' bars in each block represent the percentage of patients who achieved adequate relief in all 3 months. There was no indication of a similar quantitative (as opposed to qualitative) heterogeneity of treatment difference in Study 3002.

Reviewer's comments and analyses

**Adequate Relief**

The sponsor's LOCF strategy for filling in data on 'adequate relief 'monthly responders for the purpose of an all patients-randomized analysis could be misleading because it reports percentages of patients who were responders at month 3 who were not in the trial at that time. As an alternative, one can analyze the *patterns* of response over the 3 months. For example, consider the following orderings where R= responder, NR= non-responder due to either not responding for the required 2 weeks during the month on study, or due to having left the trial and hence not being available for future evaluations. Note that, like the sponsor's LOCF analysis, this one incorporates all patients, but **does not carry forward the last response evaluation of a dropout.**

| | Month 1 | Month 2 | Month 3 |
|---|---|---|---|
| (1) | NR | NR | NR |
| (2) | R | NR | NR |
| (3) | NR | R | NR |
| (4) | NR | NR | R |
| (5) | R | R | NR |
| (6) | R | NR | R |
| (7) | NR | R | R |
| (8) | R | R | R |

The rationale behind this particular ordering is that it is better to have a response later in the trial if at all and, of course, the more responses the better. It is not meant to be the only proper ordering.

The distribution of these 8 possible patterns of data incorporates all 626 randomized patients. The percentages of patients who fell into each pattern is displayed below.

**3001**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| (P) | 43 | 7 | 5 | 6 | 6 | 2 | 7 | 24 |
| (A) | 32 | 5 | 4 | 5 | 4 | 3 | 9 | 38 |

**3002**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| (P) | 40 | 6 | 4 | 5 | 6 | 4 | 8 | 27 |
| (A) | 33 | 6 | 5 | 3 | 2 | 5 | 6 | 39 |

The CMH test using modified ridit scores (essentially a Wilcoxon Rank Sum test) yielded a p-value of <.001 in 3001 and p-value equal to .008 in 3002, indicating that the (A) group has a more favorable adequate relief profile than placebo group's. Note that the difference between the two distributions appears to be the fact that more patients were *never* responders in the placebo group, and that more patients responded at all 3 months in the (A) group. Since there is little difference between the groups in the middle 6 categories, the ordering of the 8 categories is not critical. There was no statistical evidence of interaction between treatment and either baseline pain, pre-study symptom duration, or geographical cluster. The results for the D-IBS subgroup were similar. However, there is no evidence of efficacy among those who did not have a 12-week evaluation. See table below:

**3001**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| (P) | 75 | 11 | 0 | 0 | 6 | 1 | 3 | 4 |
| (A) | 72 | 13 | 4 | 0 | 6 | 0 | 1 | 4 |

Key words: clinical studies, NDA review, missing data

3002

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| (P) | 63 | 12 | 3 | 0 | 10 | 2 | 2 | 8 |
| (A) | 67 | 13 | 4 | 0 | 2 | 3 | 3 | 8 |

Another analysis is to simply compute the percentage of patients who were monthly responders at month 3, among all patients who were 'candidates' to be responders at months 3, viz. patients who had at least one non-missing value (0 or 1) at either weeks 10, 11, or 12. In that case, there were a total of 493 patients in 3001, 70% (A) were responders at month 3 and 49% (P) were responders at month 3. In 3002, there were 531 candidates where the respective percentages were 68% and 51%.

Finally, as previously noted, the difference between the treatment arms' efficacy lies in the number of patients who responded for all 3 months. The number of patients who discontinued prematurely who were adequate relief responders for 3 months was 2 in the placebo arm and 3 in the alosetron arm. In study 3002, the respective numbers were also 2 and 3, respectively.

Secondary Endpoints

Mean Weekly Pain Score

It should be pointed out that even though a patient may not have called in an adequate relief evaluation at the *end* of a particular week, she likely had some pain score diary data *during that week*. Thus, patients with no 'adequate relief' evaluation at the end of a particular week (the dropout week) are nevertheless likely have a mean pain score for that week.

Patients averaged their daily pain scores (1=mild, 2=moderate, 3=intense, 4=severe), for however many days were filled out during the week in order to produce an weekly pain score (the denominator being the number of days during the week that the form was filled out).

The table below displays the average pain score for each dropout cohort at the week that patients had their last observation:

3001

| | Week | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | Average Pain Score | | | | | | | | | | |
| (P) | 1.5 | 1.3 | 1.1 | 1.9 | 1.5 | 1.5 | 2.0 | 1.2 | 3.0 | 1.9 | 1.2 |
| (A) | 2.1 | 2.3 | 2.1 | 1.6 | 1.9 | 2.6 | 3.0 | 0.5 | 1.0 | 0.8 | 0.9 |
| N= | 28 | 13 | 14 | 19 | 12 | 10 | 4 | 14 | 7 | 4 | 489 |

Note that at several times, placebo patients had less pain on average than the corresponding (A) patients when they dropped out. The sponsor reports a p-value of .06 for the month 3 treatment comparison using a stratified Wilcoxon test. However, this reviewer used ANOVA to compare changes from baseline using LOCF (p-value .10) and the 481 patients with week 12 observations (p-value .0001) with a mean change from baseline of –0.7 for (P) and –1.1 for (A) in the latter analysis. Both groups' mean baseline pain score was 2.0 in this cohort of patients. Numbers (N) at the bottom represent the number of patients who were in the 'adequate response' evaluation cohort for that week. The number of patients who are represent in the computation of the mean pain scores from week to week may be slightly less than the numbers shown. The table below presents the average weekly pain score by dropout cohort for study 3002.

**3002**

| | Week | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | Average Pain Score | | | | | | | | | | |
| (P) | 0.2 | 2.5 | 2.2 | 1.0 | 4.0 | 0.0 | 2.7 | 4.0 | 3.0 | 0.0 | 1.1 |
| (A) | 2.0 | 1.7 | 2.1 | 0.4 | 0.9 | 1.3 | 1.7 | 0.3 | 0.4 | 1.6 | 0.9 |
| N= | 18 | 19 | 19 | 13 | 4 | 6 | 5 | 9 | 6 | 7 | 524 |

The sponsor reports a p-value of .036 using a stratified Wilcoxon test for the month 3 LOCF analysis. This reviewer's analysis of variance on the change from baseline using LOCF produced a p-value of .01. Comparing treatments after 12 weeks using the 511 patients with week 12 observations yielded a p-value of .0003 for change from baseline with a change from baseline of -0.8 for (P) and –1.1 for (A). Numbers (N) at the bottom represent the number of patients who were in the 'adequate response' evaluation cohort for that week. The number of patients who are represented in the computation of the mean pain scores from week to week may be slightly less than the numbers shown.

Relation Between 'Adequate Response' and Pain Scores

Having discussed the results of the "adequate relief" and "pain diary" analyses, it is natural to ask whether or not the declaration of adequate relief in fact parallels the pattern of diary weekly pain scores. For example, the sponsor has supplied week by week Spearman correlations of the secondary efficacy endpoints (pain, stool consistency, etc) with the occurrence or non-occurrence of adequate relief ( 0 or 1). Table 7 presents these results. The p-values refer to the comparison of the average change from baseline pain scores between responders and non-responders. The sponsor conducted similar analyses correlating adequate response with other secondary endpoints. Although statistical significance ($H_0$: correlation=0) is not indicated in the table, the correlations do reach nominal significance at the 5% level.

Another approach is to simply look at the distribution of pain scores at (for instance) week 12 (for "completers") and then compute the percentage of patients who were week 12 adequate relief responders. Pain scores at 12 weeks were divided into quartiles and the percentages of patients with adequate relief at week 12 were assessed. Nineteen percent (75%) in the 1st quartile responded (0-0.3), 60% in the 2nd (0.3-0.9), 48% in the 3rd (0.9-1.6), and 19% in the 4th (1.6-4.0). This trend indicates that perception of pain was at least grossly consistent with the perception of adequate relief of abdominal pain/discomfort. Logistic

regression analyses at each week confirm a relation between the weekly occurrence of adequate relief and lower weekly pain scores.

On the other hand, the methods above do not use the multivariate response profile for each patient. They break that vector into 12 components and analyze each week separately. In order to preserve the information contained in the entire vector, we can do the following: Define a concordant pair of observations at each week within a patient as the following: either there *was* an adequate response accompanied a pain score *below* a specified cutoff **OR** *no* adequate response accompanied by a pain score *above* a specified cutoff. Then count the number of concordant pairs within each patient and produce the distribution of the number of patients with the different possible numbers of concordant pairs (0-12). Do this for several values of the cutoff. If adequate response within an individual is insensitive to pain, then the distribution of concordant pairs should be essentially 'random' over the range of cutoffs. Figure 3 displays the distributions for six (6) values of the cutoff (.3, .6, 1.0, 1.6, 2.0, 2.5). The 13 possible numbers of concordant pairs have been categorized for ease of display. Note that in the 0-3 range of concordant pairs, the number of patients decreases as the cutoff increases to 1.0 and then increases. Simultaneously, at the upper end of the range, the number of concordant pairs increases as the cutoff increases to 1.0 and then decreases. This general pattern of increasing and decreasing agreement of adequate response and pain with changing cutoff suggests that pain and response have some association.

Monthly Pain/Discomfort-Free Responders

The sponsor defined a **pain/discomfort free monthly responder** as one who reported at least 50% pain-free days out of at least 14 days in a given month. There was no statistical significance between the treatment groups with respect to the percentage of these monthly responders using LOCF for any month, notwithstanding the borderline nominal result (p= .031) at month 3 in the ITT cohort in Study 3001.

An analysis similar to the 'response pattern' approach used above with 'adequate relief' confirms a trend without statistical significance. See table below displaying the percentages of patients in each group with the patterns defined above.

**3001**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| (P) | 63 | 4 | 2 | 3 | 1 | 1 | 9 | 17 |
| (A) | 56 | 1 | 2 | 7 | 2 | 1 | 16 | 15 |

The p-value using a CMH test with modified ridit scores is .11.

**3002**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| (P) | 57 | 3 | 3 | 4 | 2 | 2 | 10 | 19 |
| (A) | 58 | 2 | 2 | 5 | 3 | 1 | 12 | 17 |

The p-value using a CMH test with modified ridit scores is .71.

Other secondary endpoints included daily scores which were then averaged over each of the 12 weeks: stool consistency (1=very hard, 2=hard, 3=formed, 4=loose, 5=watery), sense of urgency (Yes or No), bloating or abdominal distention (Yes or No), daily stool frequency, and sense of incomplete evacuation (Yes or No).

In both studies, the step-down procedure produced statistically significant results in favor of alosetron for urgency, stool frequency, and stool consistency, only, using the step-down procedure. P-values below .01 were maintained through 3 months. In order to give some sense of a "clinically interpretable" result, this reviewer has tabled below the proportion of patients in each treatment group who experienced at least a 50% change from baseline as of the last observed value for each patient. However, if a patient did not have an adequate relief evaluation at week 6 (half-way through the trial), then the patient was treated as a "non-responder".

| | | All Patients | |
|---|---|---|---|
| | | 3001 | 3002 |
| Pain | Alosetron | 47% | 49% |
| | Placebo | 38% | 45% |
| Stl Freq | Alosetron | 20% | 24% |
| | Placebo | 11% | 11% |
| Stl Con | Alosetron | 10% | 13% |
| | Placebo | 3% | 5% |
| Urgency | Alosetron | 49% | 48% |
| | Placebo | 35% | 34% |

To What Extent do "Hard Stools" Contribute to Efficacy Assessments?

It is of some interest to assess the degree to which patients whose weekly stool consistency scores were in the "hard" range (score=2), contributed to the efficacy as measured by weekly adequate relief. In order to address this issue, this reviewer tabulated the distribution of the number of 'adequate relief weeks' per patient *which were also accompanied by a weekly average score of below 2*, that is, in the 'hard' to 'very hard' region of the stool consistency scale. Only patients without missing weekly data were used and only alosetron patients were pooled from the two trials. However, we must also ask whether these 'hard stool' adequate responses were all that patients experienced. In other words, even if a patient had 4 of this kind of response, were these the only adequate responses that the patient got? Were there other responses associated with *higher* stool consistency scores ? To that end, the proportion of weekly adequate responses associated with weekly stool consistencies of less than 2 were computed for each patient (denominator is the number of adequate response for the patient, not the total number of weeks (12) on trial) for each numerical category of the number of weekly adequate responses (i.e., 1, 2 ... 12). Those proportions were averaged within each of the 12 categories and then illustrated in Figure 4. Not shown is the fact that, of the total of 437 alosetron patients in the analysis 217, (50%) had no cases of simultaneous adequate response and stool consistency below 2. The proportions in Figure 4 estimate the probability of having a pain score below 2.0 given that one has had an adequate response at some time. There appears to be a pattern in which, as the number of weekly adequate responses in a patient increases, the proportion contributed by "hard" stools increases. However, numbers of patients in the higher numerical categories become quite small (5-15). Thus, for the bulk of the patients in the trial who had at least 1 adequate relief

response accompanied by a stool consistency of less than 2, adequate relief with stool consistency less than 2 accounted for approximately 50% of a patient's adequate responses (ignoring categories 6-12 and with the exception of having only 1 adequate response with stool consistency less than 2). These results could be construed to suggest that hardening of the stool was not the predominant determinant of patients' adequate relief from week to week.

The Potential of Pain Confounding in Patients who Menstruated During the Trial

Since abdominal pain could be associated with menstrual pain, it is important to ensure that the relief attributed to the drug is not confined to menstrual pain. To that end, the sponsor has conducted an exploratory analysis which compares the groups with respect to the proportion of weeks with 'adequate relief **overlapping with menstrual cycle,** and separately, **not overlapping with menstrual cycle.** *Note that the analysis includes only those patients who menstruated during the trial.* Tables 8a and 8b displays the results for **Studies 3001 and 3002,** respectively. They indicate that, on average, patients had adequate responses for more weeks on (A) then on (P), **regardless of whether the weeks of evaluation overlapped weeks of menstruation. See explanation at the bottom of the tables.**

The sponsor has also responded to an FDA request to use patients who menstruated during the trial to determine whether their daily pain scores before menstruation tended to be less than during menstruation. If the average 'before' was less than 'during', it is possible that some pain diminution attributed to the drug could be for menstrual pain and not IBS pain. The sponsor averaged the 3 days before and 3 days during menses over each patient and then averaged these differences over all the patients. Approximately 130 women were used in each trial. Point estimates and t-tests indicated no difference between the average pain scores before and during menses (p=.477 in Study 3001 and p=.976 in Study 3002).

Discussion and Conclusions

Trials 3001 and 3002 provide sufficient statistical evidence that alosetron is active in the relief of IBS in pain/discomfort in women (primary endpoint). Both trials suggest that the treatment difference is manifested by the number of patients who responded in all three months. Approximately 40% of patients who started the trial on alosetron achieved adequate relief for all three months, whereas for patients assigned to placebo the proportion was 25%.

The sponsor has raised the question of the effect of alosetron in alternating diarrhea/constipating patients. It is true that study 3002 found no effect on 'adequate relief' in these patients with a numerical difference favoring placebo. However they did seem to benefit on some secondary endpoints. The lack of such finding in 3001 means that there is no satisfactory statistical answer to this question at this time.

Finally, there was some discussion at the advisory committee meeting on November 16, 1999 concerning the chances of "responding" to therapy if one had not responded early on. Recall that the sponsor's definition of the primary endpoint 'monthly adequate relief' was at least two weeks of adequate relief during that month. Combining both trials yields the result that, of the patients who did not have adequate relief in the first month of alosetron treatment, 67% did not achieve adequate relief for the rest of the 3 month trial.

/S/

David Hoberman, Ph.D.
Mathematical statistician

Key words: clinical studies, NDA review, missing data

Concur: Dr. Flyer /S/ 12/7/99

Dr. Nevius /S/ 12/7/99

Cc:
Arch NDA# 21-107
HFD-180
HFD-180/RPrizont, JSenior, HGallo-Torres,LTalarico
HFD-715/DHoberman, PFlyer, DOB2, Chron

APPEARS THIS WAY
ON ORIGINAL

Key words: clinical studies, NDA review, missing data