CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

21-470

STATISTICAL REVIEW(S)

DEPARTMENT OF HEALTH AND HUMAN SERVICES
PUBLIC HEALTH SERVICE
FOOD AND DRUG ADMINISTRATION
CENTER FOR DRUG EVALUATION AND RESEARCH
OFFICE OF BIOSTATISTICS
DIVISION OF BIOMETRICS III3

STATISTICAL REVIEW AND EVALUATION

DATE REVIEW COMPLETED:	11/18/02
NDA NO.:	21-470
SUPPLEMENT SERIAL NO.:	007
DATE RECEIVED BY THE CENTER:	05/15/02
DRUG NAME:	Finacea™ (Azelaic Acid) 15% gel
INDICATION:	Moderate papulopustular facial rosacea
SPONSOR:	Berlex Laboratories., Inc.
DOCUMENTS REVIEWED:	Fully electronic submission
NAME OF PROJECT MANAGER:	Frank Cross (HFD-540)
NAMES OF CLINICAL REVIEWERS:	Brenda Vaughan, M.D. (HFD-540) Brenda Carr, M.D. (HFD-540)
NAME OF STATISTICAL REVIEWER:	Steve Thomson (HFD-725)
NAME OF STATISTICAL TEAM LEADER:	Mohamed Alosh , Ph.D. (HFD-725)
NAME OF BIOMETRICS DIVISION DIRECTOR:	Mohammed Huque, Ph.D. (HFD-725)

Table of Contents:

1. Executive Summary of Statistical Findings

According to the sponsor, this New Drug Application was submitted to investigate the efficacy and safety of Azelaic Acid (AzA) 15% gel when used for 12 weeks in patients with moderate papulopustular rosacea (stage 2 rosacea). To study the efficacy of AzA 15% gel the sponsor provided results from two virtually identical randomized, double-blind, placebo-controlled, multi-center studies (Studies A03125 and A03126, respectively). With the concurrence of the FDA, inflammatory lesion counts and an Investigator's Global Evaluation were selected as the primary endpoints. However, the FDA analysis differs from the sponsor's analysis of these endpoints in several ways including slightly different analysis populations and the actual definitions of primary endpoints used in the analysis. The sponsor's protocol specified that lesion counts were to be assessed using change from baseline, while the preferred FDA measure was per cent change from baseline. However, the sponsor wins on both measures. In study A03125 lesion count differences between AzA 15% gel and vehicle were highly statistically significant in favor of AzA (for both measures, $p \leq 0.0003$). In study A03126 results were not quite as distinct (for change from baseline $p \leq 0.0077$ while for per cent change $p \leq$

0.0172). For statistical analysis the seven level Investigator's Global Assessment (IGA) was dichotomized into treatment success or treatment failure differently in the FDA analysis and in the sponsor's analysis as explained below. In Study A03125 treatment differences in IGA, using the FDA dichotomization. were highly statistically significant in favor of AzA gel 15% ($p \leq 0.001$), while in Study A03126 results were barely statistically significant ($p \leq 0.044$). In each study there is one center where the difference between treatment groups in lesion counts is especially strong. The effects of centers, skewed response distributions, and the effects of the baseline covariate on lesion counts are all addressed in the report. From a statistical point of view, both in terms of lesion counts and the investigator's global assessment, we would conclude that there were statistically significant differences between AzA and its vehicle, particularly in Study A03125. Results in Study A03126 are somewhat more equivocal, but overall, do tend to favor AzA over its vehicle.

2. Introduction and Background

According to the sponsor, a gel formulation of azelaic acid (AzA), rather than the cream was chosen for developing into a new therapeutic option for topical treatment of rosacea because of the inherent cooling properties of a gel with an added benefit for the inflammation underlying rosacea. To study the efficacy of AzA 15% gel the sponsor provided results from two randomized, double blind. placebo-controlled, multi-center studies (Studies A03125 and A03126, respectively) conducted in the U.S. Data or summaries were provided from two further placebo-controlled studies conducted in Europe and an intra-individual, double blind comparison of AzA to metronidazole 0.75% cream in patients with papulopustular rosacea. However, these studies used a 20% cream formulation, and are not considered to be of much relevance for this submission.

In the two pivotal studies, a total of 27 centers enrolled 333 patients treated with AzA 15% gel and 331 patients treated with vehicle. In Study A03126 two subjects had an Investigator Global Assessment at baseline of "mild." Since this was supposed to correspond with a successful outcome, these two subjects, both in the AzA group at center 02, were deleted from the study. In neither study were there any statistically significant differences between the treatment groups in demographic measures or other baseline characteristics. About 86% of the patients completed Study A03125 and about 88% completed A03126.

The primary assessment of efficacy was based on the intent-to-treat (ITT) population, with missing values imputed by last observation carried forward (LOCF). The sponsor proposed two primary endpoints, the change from baseline in inflammatory lesion counts and a seven ordinal level Investigator Global Assessment (IGA). For the analysis this was to be dichotomized into "responders" versus "non-responders" as defined in section 3.4.1 below. However, this dichotomization is different from that preferred by the Medical Officer and is not used in the FDA analyses below, but is reported in the section 4 about the sponsor's analysis.

For assessing change in lesion counts the DDDDP has emphasized the percent change from baseline, whereas the sponsor's analysis considers this to be a secondary endpoint. In the pre-NDA meeting the sponsor was encouraged to consider both the change and the per cent change as co-primary, and this is done in this report. For the IGA the Medical Officers specified a different dichotomization than the one used by the sponsor. In the FDA analysis a treatment "success" on the IGA is defined as occurring when the subjects had achieved at least a two step reduction in the IGA AND achieved a final score of clear or minimal ("0" or "1"). Results using these endpoints are summarized in the following table:

Table 1: Summary of Efficacy Results Using FDA Primary Endpoints

	Study A03125			Study A03126		
	AzA	Veh	p-value	AzA	Veh	p-value
N	164	165		167	166	
Inflammatory Lesion Count Means						
Change from Baseline	10.7	7.1	0.0001*	9.0	6.4	0.0077*
% Change from Baseline	57.9	39.9	0.0003*	50.0	38.2	0.0172*
Investigators Global Evaluation						
Treatment Success	50 31%	20 12%	0.001†	53 32%	36 22%	0.044†
Failure	114 69%	145 81%		113 68%	131 78%	

*ANCOVA Model: expected response = baseline + center + treatment (interactions not statistically significant)
† Significance level of CMH test of equality of proportions using modified ridit scores.

Thus in Study A03125 all endpoints showed statistically significant differences in favor of AzA 15% gel over vehicle (all three p-values ≤ 0.001). Results are similar, but not as strong in Study A03126 (all three p-values ≤ 0.044). By chance, the dichotomization of the IGA specified by the FDA Medical Officer was actually slightly more favorable to the sponsor than that originally chosen by the sponsor.

As is typical of lesion count data, the values seem to be skewed to the right. However, an ANOVA model using the factors above but with rank transformed data give results that are essentially equivalent. Further in Study A03126 statistical significance of the results are largely driven by one center. However, deleting this center, results were still trending to be in favor the AzA 15% gel treatment over its vehicle (see Appendices 1 and 2 for discussion).

Interactions between treatment and center in the lesion count data were investigated, but were either statistically non-significant or were basically quantitative. The effect of the baseline lesion count covariate was homogeneous over all treatment by center combinations for the percent change, and reasonably close for the absolute change. (For results deleting the covariate from the model see Appendix 3.)

The studies were not powered to detect differences in subgroups, so no definitive conclusions about treatment differences in demographic subgroups are possible. However, it is apparent that for both genders, the superiority of AzA over its vehicle in terms of both lesion counts and the investigators global evaluation are roughly the same. For both endpoints there is some apparent evidence that AzA may be less effective for non-Caucasian patients than among Caucasian patients, although with the small number of non-Caucasian patients this may well be an artifactual result. Finally, there is some apparent evidence that while AzA is superior to its vehicle in the older patients (age 65+), it may be more effective in a younger age group (age <65).

One approach to analyzing safety comparisons is to use the techniques of Westfall and Young (1993) to adjust for the multiple decisions. To limit the number of comparisons, only adverse events occurring in 5 or more patients were considered. Using Westfall and Young techniques we would conclude that among possible adverse events, there is statistically quite significant evidence that the AzA 15% gel treatment is associated with measures of stinging skin and pruritis (all $p \leq 0.0027$, adjusting for multiplicity). Otherwise there was no particular evidence of differences across treatment groups in terms of the listed adverse events. This does not mean there were no differences, just that this relatively conservative procedure did not detect any.

3. Data Analyzed and Sources

3.1 Studies

3.1.1 General:

To study the efficacy of Azelaic Acid (AzA) 15% gel the sponsor provided results from two randomized, double blind, placebo-controlled, multi-center studies (Studies A03125 and A03126, respectively). Data were provided from two further placebo-controlled studies conducted in Europe and an intra-individual, double blind comparison of AzA to metronidazole 0.75% cream in patients with papulopustular rosacea. However, these studies used a 20% cream formulation, and are not considered to be of relevance for this submission.

The two pivotal studies were identically designed multi-center, vehicle-controlled, double blind, randomized, parallel-group studies with the same patient selection criteria and efficacy endpoints. Both studies were designed to enroll 300 patients at up to 15 centers, with at least 20 patients per center. The target study population consisted of male and female patients greater than 18 years of age, with moderate, papulopustular rosacea with a predefined range of the number of inflammatory lesions (papules and/or pustules), persistent erythema, and telangiectasia. A 1:1 randomization between treatment groups was used. Efficacy was to be assessed at Weeks 4, 8, and 12.

The sample size determinations were based on previously observed results. For lesion counts, assuming a common standard deviation of 15 for the number of inflammatory lesions at end of treatment, 98 patients were necessary to detect a difference of 7 in mean lesion count for the two treatments with a power of 90% (vehicle 15 lesions, AzA 15% gel 8 lesions). In the investigator's global assessment: assuming a "responder" (see discussion below) rate of 40% for the vehicle and a 20% difference in favor AzA 15% gel, 140 patients per group were needed to achieve 90% power. To allow for dropouts, 150 randomized patients per treatment group were planned, for a total of 300 patients per study.

Actually, in both studies, a total of 27 centers enrolled 331 patients treated with vehicle and 333 patients treated with AzA 15% gel. However two subjects in the AzA group in center 02 of Study A03126 had a global evaluation of mild at baseline. Since this was supposed to correspond with a successful final outcome these two subjects were deleted from the FDA efficacy analyses, but not from the safety analyses. This gave 331 subjects in the AzA group. In neither study were there any statistically significant differences between the treatment groups in demographic measures or other baseline characteristics. About 86% of the patients completed Study A03125 and 88% completed A03126.

3.1.2 Disposition, Demographics, and Baseline Characteristics:

The disposition of all patients in the two primary studies is presented by study in Table 2. below:

Table 2: Disposition of Patients in Pivotal Studies

	Report A03125			Report A03125		
	AzA 15% gel	Vehicle	Total	AzA 15% gel	Vehicle	Total
Randomized and Treated	164	165	329	167*	166	333
Completed Treatment	133 (81%)	150 (91%)	283 (86%)	148 (89%)	146 (88%)	294 (88%)
Discontinued	31 (19%)	15 (9%)	46 (14%)	19 (11%)	20 (12%)	39 (12%)
Reasons for Discontinuation						
Adverse Events	9	2	11	8	4	12
Lack of Efficacy	1	7	8	0	5	5
Protocol Deviation	6	1	7	1	0	1
Withdrawal of Consent	6	2	8	0	3	3
Death	0	0	0	1	0	1
Other	9	3	12	9	8	17

* Two subjects were deleted due to baseline IGA scores, giving 169 subjects for safety.

In each of the pivotal studies, enrollment was stopped at 300 patients, however, eligible patients who had already begun the screening process were also allowed to enroll, so that by 30

March 2001, a total of 664 patients were randomized to treatment in the 2 studies (giving 662 evaluable patients).

Table 3: Demographics and Baseline Characteristics (ITT Population)

	Report A03125		Report A03126	
	AzA 15% gel (N=164)	Vehicle (N=165)	AzA 15% gel (N=167)	Vehicle (N=166)
Mean Age (years)	48.0 (21-84)	49.2 (24-77)	47.7 (24-86)	47.0 (23-78)
Sex [n (%)]				
Male	40 (24%)	45 (27%)	48 (29%)	46 (28%)
Female	124 (76%)	120 (73%)	119 (71%)	120 (72%)
Race [n (%)]				
Caucasian	159 (97%)	155 (94%)	145 (87%)	153 (92%)
Black	0 (0%)	2 (1%)	2 (1%)	2 (1%)
Hispanic	4 (2%)	7 (4%)	19 (11%)	10 (6%)
Asian	1 (1%)	0 (0%)	0 (0%)	0 (0%)
Other	0 (0%)	1 (1%)	1 (1%)	1 (1%)
Mean height (cm)	167.8	167.9	167.9	167.6
Mean weight (kg)	81.7	81.1	82.0	81.1
Body Mass Index	29.1	28.8	29.1	28.9
Mean Previous Duration of rosacea (months)	100.2	88.5	101.0	103.0
0-6 months	8 (5%)	7 (4%)	8 (5%)	3 (2%)
6 months – 2 years	26 (16%)	38 (23%)	26 (15%)	33 (20%)
2 years – 5 years	50 (30%)	54 (33%)	46 (28%)	54 (33%)
5 years	80 (49%)	66 (40%)	87 (52%)	75 (45%)
Missing	0 (0%)	0 (0%)	0 (0%)	1 (1%)
Mean inflammatory Lesion count	17.5 (8-60)	17.6 (8-52)	17.9 (8-50)	18.5 (8-50)
Investigator global assessment (n [%])				
Clear	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Minimal	0 (0%)	0 (0%)	[2]*	0 (0%)
Mild	25 (15%)	33 (20%)	19 (11%)	23 (14%)
Mild to moderate	67 (41%)	68 (41%)	75 (45%)	80 (48%)
Moderate	57 (35%)	53 (32%)	54 (32%)	42 (25%)
Moderate to severe	14 (9%)	8 (5%)	14 (8%)	16 (10%)
Severe	1 (1%)	3 (2%)	5 (3%)	5 (3%)
Missing	0 (0%)	0 (0%)	0 (0%)	0 (0%)

*These subjects were deleted from the FDA analysis since they did not fit baseline criteria. Note that all other counts in this report except for the sponsor's analysis, safety, and adverse events reflect this deletion, including other counts in this table.

Of the 662 total patients enrolled (deducting the two patients above), 331 (50%) were treated with AzA 15% gel and 331 (50%) were treated with vehicle. Similar percentages of patients completed both studies; 86% in Study A03125 and 88% in Study A03126. A slightly higher percentage of vehicle-treated patients (91% [150/165]) than AzA 15% gel-treated patients (81% [133/164]) completed study A03125.

Within each study, there were no statistically significant differences between the treatment groups in demographics, baseline characteristics, or baseline efficacy measures. In Study A03125 there were slightly more Caucasian patients and correspondingly fewer Hispanic patients than in Study A03126. Note that several of the subjects classified as withdrawals by the sponsor actually had data at the 12^{th} week and are included in the completer population defined below.

Results for lesion counts were extreme in one center in each study. This issue is addressed in Appendices 1and 2.

3.2 Data Sources:

According to the sponsor: "The data for Reports A03125 and A03126 were monitored at each center to compare case report form (CRF) entries with the source data and completed CRF pages were collected. The data were computerized and stored in an Oracle version 7.3.3.0.0 database. The data entered on the CRF were first subjected to computer-aided verification and checked for completeness. After double-data entry, the data sets were manually reviewed for illogical entries and violations of the study protocols. Data corrections were documented according to good clinical practices (GCP). Complete case listings of the data used for statistical analysis and generation of the tables were compiled. The general decoding of the blind was done after the studies had been completed, and after all data had been electronically recorded, validated and checked, and the databases locked." (page 19 of sponsor's Integrated Summary of Efficacy report)

The data used for the FDA analyses were some 25 SAS 6.12 data sets per study, transferred from the FDA electronic data room.

3.3 Randomization:

For both studies, the study report indicates that randomization was 1:1 between AzA and vehicle, conducted in blocks of 10. This was visually verified by inspection of within center treatment allocation by date. There seemed to be no apparent problems with the randomization.

3.4 Endpoints

3.4.1 Primary Endpoints:

The protocol specified two primary efficacy response measures: inflammatory lesion counts (sum of counts of papules and pustules) and an ordinal scale Investigator Global Assessment. For analysis the inflammatory lesion count was defined in the protocol as the change in inflammatory lesion count from baseline to the Week 12/last available visit. At each assessment lesion counting was performed under constant lighting and all make-up was removed prior to counting. The sponsor notes that "To provide consistency, each patient was assessed by the same investigator over the entire treatment period, when possible." (page 30 of Study Report A03125). In the sponsor's analysis the change from baseline was considered to be a secondary endpoint, while it is considered to be co-primary in the FDA analysis.

The second primary response measure was the Investigator Global Assessment (IGA) evaluated on a 7-point scale as described in Table 4 below. The sponsor particularly noted that the IGA score "was used only for the description of papulopustular rosacea, not for rosacea in general. For example, a score of 6 = severe did not implicate a stage 3 rosacea with nodules and rhinophyma, but rather a severe form of stage 2, papulopustular rosacea." (page 16 of sponsor's Integrated Summary of Efficacy report)

Table 4. Investigator Global Assessment

Score	Label	Description
0	Clear	No papules and/or pustules; no or residual erythema; no or mild to moderate telangiectasia
1	Minimal	Rare papules and/or pustules; residual to mild erythema; mild to moderate telangiectasia
2	Mild	Few papules and/or pustules; mild erythema; mild to moderate telangiectasia
3	Mild to moderate	Distinct number of papules and/or pustules; mild to moderate erythema; mild to moderate telangiectasia
4	Moderate	Pronounced number of papules and/or pustules; moderate erythema; mild to moderate telangiectasia
5	Moderate to severe	Many papules and/or pustules, occasionally with large inflamed lesions; moderate erythema; moderate degree of telangiectasia
6	Severe	Numerous papules and/or pustules, occasionally with confluent areas of inflamed lesions; moderate or severe telangiectasia

According to the sponsor's protocol for the actual analysis this IGA was to be dichotomized into "responders" and "non-responders" based on the IGA score at baseline and at the end of treatment. In particular, a patient was classified as a responder if they had achieved either a score of 0 or 1 on the IGA, and had a decrease of at least one unit from baseline, or had

achieved a score of 2 with a decrease of two units from baseline. Otherwise a patient was classified as a non-responder.

However, this dichotomization was not used in the FDA analyses. Instead, in the FDA analyses the IGA was dichotomized in terms of success and failure at end of treatment. The Medical Officers determined that an appropriate measure of treatment success would be defined as those subjects who had achieved at least a two step reduction in the IGA AND achieved a final score of clear or minimal ("0" or "1"). In a Pre-NDA meeting the FDA recommended a 5 point IGA, to be analyzed by dichotomizing either at clear ("0") or clear or minimal ("0" or "1"). To match the 7 point scale used by the sponsor, a classification where a success was defined as a response of clear, minimal, or mild ("0", "1", or "2") was also included. However, the dichotomization at treatment success as defined above is the primary endpoint. The other three dichotomizations are secondary. A fifth dichotomization, not analyzed in the FDA analysis is the responder/ nonresponder grouping used by the sponsor.

The primary endpoints are these lesion count measures (i.e., change from baseline and percent change from baseline) and the treatment success in the investigator's global assessment evaluated at nominal week 12 (end of treatment) in the ITT population (with last observation carried forward, possibly from baseline). Other response measures are provided for information purposes only.

Profiles over time of the primary response measures are given in Appendix 4.0.

3.4.2 Secondary Endpoints:

Two further secondary lesion count variables that were proposed by the sponsor are nodule counts and total lesion count (sum of papules, pustules, and nodules). Total lesion count is actually the sum of inflammatory lesion count and nodule count, and largely follows the same distribution as the inflammatory lesion count. It is ignored in this report. An analysis of nodule counts is included in Appendix 5.

For the Investigator's Global Assessment of rosacea the sponsor also specified the change from baseline in the IGA as a response. Other secondary endpoints defined by the sponsor included assessments of erythema and telangiectasia as measured on the following scales:

The severity of erythema was rated as follows:

1. None	Either no visible erythema or minimal residual erythema
2. Mild	Slight erythema either centrofacial or generalized to whole face
3. Moderate	Pronounced erythema either centrofacial or generalized to whole face
4. Severe	Severe erythema/red to purple hue, either centrofacial or generalized to whole face

The severity of telangiectasia was rated as follows:

1. None	No telangiectasia
2. Mild	Only few fine vessels discernible, involving 10% or less of the facial area
3. Moderate	Multiple fine vessels and/or few large vessels discernible, involving 10% to 30% of the facial area
4. Severe	Many fine vessels and/or large vessels discernible, involving more than 30% of the facial area

The protocol indicated that these were to be analyzed through change from baseline scores. The statistical analysis plan is not very explicit, but can be read to suggest that the analysis should be based on the original scores. However, the protocol would take precedence, and the change from baseline in these variables is used for analysis. For comparison and completeness, a comparison based on the original scores is also given. For erythema the Medical Officers preferred an analysis based on a variable labelled as "success in response", i.e., the proportion of subjects who had a baseline score of moderate or severe and achieved a final score of none or, alternatively, none or mild.

At the completion of the treatment period, both the investigator and the patient rated their subjective impression of the change from baseline in rosacea severity.

- Investigator overall ratings: 1 = complete remission, 2 = marked improvement, 3 = moderate improvement, 4 =no improvement, 5 = deterioration.
- Patient overall ratings: 1 = excellent improvement, 2 = good improvement, 3 = moderate improvement, 4 = no improvement, 5 = worse.

Cosmetic acceptance of the topical preparation:
At the end of the study, patients were asked their opinion of the cosmetic acceptance of the topical preparation: 1 = very good, 2 = good, 3 = satisfactory, 4 = poor, 5 = no opinion (treated as missing in the analysis of this endpoint).

Scores in erythema, telangiectasia, investigator rating of improvement, and the nodule count were considered of clinical relevance. Results for these secondary endpoints appear in sections 6.1.2 and 6.2.2 below and in Appendix 5.0. Scores in patient rating of improvement and cosmetic acceptance were considered to be of little to no regulatory utility and are given in Appendix 6.0.

3.5 Patient Populations:

Three different end of study populations were defined for the FDA analysis:

Intent-to-Treat Population: The ITT population consisted of all patients who were randomized and dispensed study medication, and is the primary analysis population..

Completers Population: All patients with data at the nominal 12 week endpoint.

Sponsor defined Per-Protocol Population: The PP population consisted of all completers who met the following criteria:

- At least 1 primary efficacy measure (lesion count or investigator's global assessment of rosacea) collected both at baseline and at least 1 post-baseline visit
- No major violations of any inclusion/exclusion criteria at screening.
- Week 12 Visit falling between 77 and 98 days post-randomization
- Compliant with study medication (did not miss more than 7 doses during the 12 weeks of treatment)
- Did not use any prohibited medications during the course of the study

Usually the completers and the Per Protocol populations are nearly the same, and are not distinguished. But in these studies these appear to be quite different populations and so both populations are considered. However, results in these two populations should be considered as supportive only.

The sponsor labeled the ITT population above as a "revised ITT", and defined the ITT population with the restriction that any patients with no follow-up after baseline were deleted from the study. That definition is not used in this report. In response to a suggestion from the FDA about a longer washout period for prior drug use the sponsor proposed a Modified intent-to-treat (MITT) population. However, at the recommendation of the Medical Officer results this proposed MITT population is also ignored.

4. Sponsor's Analyses of the Phase 3 Pivotal Studies:

As noted above, originally the sponsor proposed an ITT type analysis where only the observations after baseline were carried forward. After the sponsor met with the FDA on 30 August 2001, the sponsor was informed of the preferred DDDDP definition where even baseline values were carried forward. The sponsor considered this an alternative analysis, but it is the current standard in the DDDDP, and is used in the FDA summary of the sponsor's results presented in the table below. Further, the sponsor's analyses included the two subjects with a baseline Investigator Global Assessment (IGA) deleted from Study A03126 in the FDA analyses. Finally the dichotomization of the IGA used for the sponsor's analysis is the responder/ nonresponder split described above, rather than the FDA split.

The two primary endpoints were the inflammatory lesion counts and the IGA at study termination. For the sponsor's analysis the IGA scores were dichotomized to classify patients as responders or non-responders. Again, responders were defined either as patients who had achieved a clear or minimal final global assessment with a decrease of at least 1 unit from baseline, or who had achieved a clinically favorable mild final global assessment with a decrease of at least 2 units from baseline. Otherwise they were classed as non-responders. As a second

dichotomization, the sponsor assessed responses where a score of clear or minimal (0, 1) on the IGA was considered a success.

Results from both of these dichotomizations of the IGA are presented in Table 5 below. The sponsor reported change in lesion counts in terms of difference from baseline, so that a decrease is a negative number. To make the sponsor's results consistent with the convention in this report, lesion counts are given in terms of reduction, i.e. a positive number. (Hence a negative number corresponds to an increase.)

Finally the sponsor reported values for lesion counts in terms of least squares means from a model with baseline, treatment, and center as factors (no interaction). These are adjusted for baseline values and unbalanced treatment allocation. By comparison, for simplicity, in the FDA review only simple means are reported. However, the models used in the FDA analysis retain the term for interaction between treatment and center. Usually simple means and least squares means will be close, but they can be disparate.

Table 5: Summary of Sponsor's Results for Primary Endpoints

	Study A03125			Study A03126		
	AzA	Veh	p-value	AzA	Veh	p-value
N	164	165		169	166	
LS Means Inflammatory Lesion Count						
Change from Baseline	10.8	7.2	<0.0001*	9.2	6.5	0.0144*
% Change from Baseline	58.1	40.2	0.0001*	50.7	38.7	0.0208*
Investigators Global Evaluation						
Responder #	60 37%	32 19%	0.0002†	66 39%	48 29%	0.0551†
Nonresponder	104 63%	133 81%		103 61%	118 71%	
IGA Clear or Minimal (0,1)	73 45%	39 24%	<0.0001†	77 46%	58 35%	0.0529†
Other (>1)	91 55%	126 76%		92 54%	108 65%	

*Model: expected response = baseline + center + treatment (interactions were not statistically significant)
† Significance level of CMH test of equality of proportions using modified ridit scores.
# responders were defined as patients who had achieved a clear or minimal final global assessment with a decrease of at least 1 unit from baseline, or who achieved a score of at least mild with a decrease of at least 2 units from baseline.

In both studies, the percent change from baseline and the simple change from baseline show that there are statistically significant reductions in lesion counts (all $p \leq 0.0208$). Further, in terms of both the sponsor's proposed dichotomization into responders and non-responders, or in terms of clear or minimal versus other, in Study A03125 Azelaic Acid 15% gel is statistically better than its vehicle (both $p \leq 0.0002$). Treatment differences on the IGA are almost statistically

significant at the usual level in Study A03126 (both $p \leq 0.0551$). With small caveats due to differences in definition and implementation the sponsor's results are consistent with those in this FDA review below.

5. Statistical Methodologies

Again, the original protocols for the Phase 3 studies listed inflammatory lesion counts and the Investigator's Global Assessment (IGA) as the primary endpoints. At the August 21, 2000, pre-NDA meeting, the sponsor was encouraged to assess the percent change from baseline on inflammatory lesions as a measure and to score the IGA on a 5 point 0-4 scale. Initially it was suggested this 5 point scale be dichotomized at 0 (i.e., none) or 0 or 1 for analysis. However, the sponsor decided emphasize the change from baseline in lesion counts and to use a 7 point, 0-6 scale for the IGA. As explained above, for the sponsor's analysis, the IGA was to be dichotomized on a "responder/non-responder" scale. However, the FDA Medical Officers preferred definition of treatment success on the IGA was those subjects who had achieved at least a two step reduction in the IGA AND achieved a final score of clear or minimal ("0" or "1"). This is the primary analysis variable in FDA analyses of the IGA, and is different from the sponsors proposed dichotomization into responder/nonresponder. For comparison with earlier proposed dichotomizations, results for both the dichotomization at 0,1, or 2 and at 0 or 1 are reported here, as well as for clear ("0") versus other. The 0, 1, or 2 dichotomization and the 0 or 1 dichotomization are supposed to be roughly equivalent to the 0 or 1 dichotomization on the 5 point scale.

Thus, there are four dichotomizations of the IGA in the FDA analysis. However, only the success/fail dichotomization defined by the FDA Medical Officer is a primary endpoint. The others are supporting and tests based on these can be considered to be different sensitivity analyses of the final results. This reviewer would argue that these dichotomizations can be reasonably ordered in importance, and thus can be cast as a sequence of nested tests. Using the nested test formulation error rate for complete or partial null hypotheses would be controlled. Still, this was not proposed in the original protocol and is a post hoc adjustment only used for error control. Nonetheless, this reviewer would claim that this formalism in controlling overall error may be useful.

The change from baseline and the percent change from baseline were analyzed using standard linear models, including terms for baseline count, treatment, center, and interaction. The significance levels reported come from using so-called Type 3 sums of squares. Note the sponsor reports so-called least squares means. Least squares means are particularly helpful in interpreting the results of an ANOVA or ANCOVA using Type 3 sums of squares, since Type 3 tests are based on simple contrasts of these least squares means. However, for simplicity, the FDA results report simple unadjusted means, not least squares means.

Treatment by center interactions in the lesion count data were investigated, but were

either statistically non-significant or were basically quantitative. Tests for heterogeneity of the baseline covariate were statistically non-significant in both studies for the percent change from baseline. Results were more problematic for the change from baseline. For several populations in each study tests of heterogeneity were statistically significant, but in each case the effect (with a number of degrees of freedom) was much smaller than the effect due to the covariate (with only one degree of freedom).

In each study there was one center where treatment differences in favor of AzA 15% gel were especially large. Specifics are given in Appendix 2. Even deleting or adjusting for the discrepant center results in Study A03125 are still statistically significant in favor of AzA over its vehicle. In Study A03126 results are somewhat more problematic, but at least trend in favor of AzA 15% gel over its vehicle.

The dichotomized IGA scores were analyzed using a Mantel-Haenszel test stratified on center. The protocols specified that modified ridit (i.e., standardized midrank) scores were to be used in the Mantel-Haenszel tests. These tend to lead to more powerful tests than those using integer or table scores, but for interpretability this reviewer would usually choose the latter. In particular, the actual modified ridit scores used for testing differences with a dichotomous response will generally vary across strata, whereas this reviewer would recommend that actual scores should be the same for each stratum. However since the sponsor's protocol specified modified ridits they are used in all Cochrane-Mantel-Haenszel (CMH) tests cited in this report.

The ordinal responses of the secondary endpoints were also tested using CMH tests, again using including modified ridits scores. In addition a Bayesian analysis of a zero inflated Poisson regression is used to assess treatment differences in nodule counts, and is included in Appendix 5.0.

Profiles of lesion counts and IGA scores over time are presented in Appendix 4. Again only the ITT results are primary. Others are only supporting. Here, as above, this reviewer would argue that it would be reasonable to control error across time points, by structuring these as a sequence of nested hypotheses. That is, as an alternative analysis, first test the primary endpoint, the ITT population at nominal week 12. If that is statistically significant then test the Per Protocol (PP) population at week 12. If in turn the comparison in the PP populations is statistically significant then test the completers at week 12. If the test on completers at week 12 is statistically significant test then test week 8 completers. Finally, if the comparison among the week 8 completers is statistically significant test the week 4 completers. Stop testing and ignore latter comparisons the first time a non-significant result is obtained. This procedure applies to profiles over time, or just to the ITT population, followed by the Per Protocol Population, followed by the Week 12 population or to other possible sequences of tests. Again, it should be noted that this procedure was not included in the sponsor's protocol and is a post hoc adjustment, but still could be applied to maintain error rate. However, again, in this particular NDA results other than with the ITT population are considered as secondary, and the formalism of the nested

tests above is considered to be optional.

As is typical of lesion count data, the inflammatory lesion count values seem to be skewed to the right. However, results from an equivalent rank transformed analysis are essentially equivalent to the results on the untransformed data reported here. This reviewer has some concerns about the rank transform and suspects that in some circumstances it may be anti-conservative. So an analysis was also performed on Winsorized data for both the change from baseline and the percent change from baseline. First a cutpoint within each center was computed as the highest quartile plus 1.5 times the interquartile range. This was the value specified by Tukey to determine if a point was an outlier in his EDA boxplot. Values greater than this cutpoint value were recoded to this value. The recoded (i.e., Winsorized) data were then analyzed and gave essentially the same results as before for each study.

The usual analysis of ANOVA models treats centers as fixed. We also usually impute data for the ITT using LOCF. A corresponding mixed model analysis treating centers as random allows for responses correlated within center, may be more powerful, seems to test hypotheses that are arguably more relevant, and only requires that missing data be missing at random. Appendix 8.0 provides mixed model repeated measures analyses for both the change from baseline in inflammatory lesion counts and the corresponding per cent change. Responses were modeled at weeks 4, 8, and 12. Unlike a traditional repeated measures analysis no restrictions were placed on the between measures covariance. Note these results, presented in the Appendix 8.0, are generally quite consistent with the results at each individual time point. However, again, these were not specified in the protocol and are only considered to be supporting analyses.

All analyses were conducted using SAS 6.12 or WINBUGS 1.3.

6. Reviewer's Analysis of the Phase 3 Pivotal Studies

The protocols for both studies were virtually identical, and the study designs are described above.

6.1 Results for Study A30125

6.1.1 Primary Endpoints

The following table gives week 12 means for the three patient populations, and the significance levels of the test for treatment differences from an ANCOVA model with classification effects for treatment, center, and interaction, and baseline lesion count as a covariate.

Table 6: Study A30125 Means and Tests of Treatment Differences in Inflammatory Lesion Counts

		ITT			Completers			Per Protocol		
Response	Treat	N	Mean	p-value	N	Mean	p-value	N	Mean	p-value
Change	Veh	165	7.1	0.0001	151	7.7	<0.0001	125	8.3	0.0010
	AzA	164	10.7		134	12.7		114	12.4	
% Change	Veh	165	39.9	0.0003	151	42.5	<0.0001	125	44.7	0.0016
	AzA	164	57.9		134	67.2		114	66.5	

Model: expected response = baseline + center + treatment + interaction

For all three populations, but particularly the ITT population, whether we use the change from baseline as specified in the protocol, or the percent change from baseline, differences are statistically significant (all $p \leq 0.0010$). Results from deleting the interaction term, as proposed in the sponsor's protocol, are given in Appendix 3.0, and are virtually identical to those above. From ANOVA results not displayed here, in the ITT population interaction between center and treatment is highly non-significant ($p \geq 0.6391$ for both measures). In the completer population the test for interaction is statistically significant for the change from baseline, and close for the percent change ($p \leq 0.0271$ and $p \leq 0.0931$, respectively). Results are somewhat more problematic for the per protocol population ($p \leq 0.06124$ and $p \leq 0.1082$, respectively). Interestingly, without the baseline, all interactions become statistically non-significant (all $p \geq 0.1792$). This is largely due to the large sums of squares attributable to the baseline covariate. For percent change from baseline the test of heterogeneity in covariate slope was highly statistically non-significant for all three endpoints (all $p \geq 0.84$). For simple change from baseline in inflammatory lesion count tests of heterogeneity of slope were statistically significant or close to significant for each population ($p \leq 0.0004$, $p \leq 0.0013$, and $p \leq 0.1579$, respectively). However in each case, the sums of squares due to heterogeneity in slopes were always a small fraction of the sums of squares due to the baseline (about 1% or 1.5%).

As noted before, in center 07 results are especially favorable to the sponsor. This issue is addressed in Appendices 1 and 2. However, in Appendix 2 evidence is presented that this center has no particular effect on conclusions.

The following table, Table 7, summarizes responses to the Investigator's Global Assessment at nominal Week 12 or at study endpoint. For comparison results for the dichotomizations into treatment success as defined by the Medical Officer (see above) and by dichotomizing at 0,1, or 2 or at 0 or 1 are reported here, as well as for clear ("0") versus other. These other dichotomizations should be considered as secondary to treatment success as defined by the Medical Officer.

Table 7: Study A30125 Investigator Global Evaluation at End of Study

		ITT			Completers			Per Protocol		
Treat		AzA	Veh	p-value	AzA	Veh	p-value	AzA	Veh	p-value
Success	N	50	20	0.001	49	19	0.001	39	17	0.001
	%	30.5%	12.1%		36.6%	12.6%		34.2%	13.6%	
0 (clear)	N	12	5	0.039	12	5	0.017	9	4	0.028
	%	7.3%	3.0%		9.0%	3.3%		7.9%	3.2%	
0,1 (≤Min)	N	60	32	0.001	57	31	0.001	45	24	0.001
	%	36.6%	19.4%		42.5%	20.5%		39.5%	19.2%	
0-2 (≤Mild)	N	100	67	0.001	93	62	0.001	76	51	0.001
	%	61.0%	40.6%		69.4%	41.1%		66.7%	40.8%	
3-6	N	64	98		41	89		38	74	
	%	39.0%	59.4%		30.6%	58.9%		33.3%	59.2%	
All	N	164	165		134	151		114	125	

*Significance level of CMH test of equality of mean proportions using modified ridit scores.

Besides treatment success, the rows labelled "0 (clear)" give counts and percentages of subjects evaluated as clear on the investigator's global assessment in the three populations. The rows labelled "0,1 (≤Min)" give counts and percentages of subjects evaluated as clear or minimal, and the rows labelled rows labelled "0-2 (≤Mild)" give counts and percentages of subjects evaluated as clear, minimal, or mild. The rows labelled "3-6" gives the counts and percentages of the remaining subjects. Finally row labelled the "All" gives the total number of subjects in each population at nominal week 12. Of course, the results on treatment success should be considered as primary. The rest are supporting. Results deleting center 07 are consistent with these (see Appendix 2).

6.1.2 Secondary Endpoints

As noted in section 3 above, of the secondary endpoints only the scores in erythema, telangiectasia, investigator rating of improvement, and the nodule count were considered to be of clinical relevance. According to the protocol the first two were to be evaluated as change from baseline, but for completeness the original scores are also analyzed below. Results for the investigator rating of improvement and nodule count are given in Appendix 5.0.

Table 8: Study A03125 Decrease From Baseline in Erythema Ratings at End of Study

Decrease †		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
2	N	14	7	13	7	10	6
	%	8.5%	4.2%	9.7%	4.6%	8.8%	4.8%
1	N	58	41	55	38	48	28
	%	35.4%	24.9%	41.0%	25.2%	42.1%	22.4%
0	N	85	105	61	95	51	82
	%	51.8%	63.6%	45.5%	62.9%	44.7%	65.6%
-1	N	7	12	5	11	5	9
	%	4.3%	7.3%	3.7%	7.3%	4.4%	7.2%
p-value*		0.016		0.003		0.001	

*Significance level of CMH test of equality of mean proportions using modified ridit scores.
†Note that a negative decrease is an increase.

The protocol indicates a comparison of mean change using modified ridit scores as defined above. However, at the direction of the Medical Officer this is analyzed using variables treatment success, where "success (1)" denotes the proportion of subjects who had a baseline score of moderate or severe and achieved a final score of none, while "success (1,2)" denotes the proportion of subjects who had a baseline score of moderate or severe and achieved a final score of none or mild. These are post hoc definitions of response, and are not given in the protocol.

Table 9: Study A03125 Decrease From Baseline in Erythema Medical Officer Definitions of Success

Decrease		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
Success (1)#	N	7	3	7	3	4	3
	%	4.3%	1.8%	5.2%	2.0%	3.5%	2.4%
p-value‡		0.166		0.114		0.433	
Success (1,2)#	N	56	40	52	37	45	30
	%	34.2 %	24.2%	38.8 %	24.5%	39.5 %	24.0%
p-value‡		0.041		0.011		0.007	
All	N	164	165	134	151	114	125

#Success(1) and Success(1,2) denote the proportion of subjects who have a baseline score of 3 or 4 and whose final erythema score is either 1 or is 1 or 2, respectively.
‡Significance level of MH test of equality of proportions in success in erythema using modified ridit scores.

Results Success(1) and Success(1,2) in the table above are given for the populations as randomized, i.e., including subjects who had a score of 2, i.e., "Mild", at baseline. Note that by the definition of treatment success in erythema, these latter subjects are defined as failures in

each population. For Successs(1) no treatment differences are statistically significant. However, whether we look at the ITT population, the Per Protocol population, or the completers there are statistically significant differences in treatment ($p \leq 0.041$, $p \leq 0.007$, and $p \leq 0.011$, respectively).. The patient populations above could be modified to delete those subjects the subjects with a baseline score of 2. The corresponding tests using these modified populations are also all statistically significant (from results not presented here: $p \leq 0.035$, $p \leq 0.004$, $p \leq 0.004$ respectively).

Using the mean modified ridit score over the differences from baseline as implied by the protocol there are statistically significant differences in treatment ($p \leq 0.016$, $p \leq 0.001$, and $p \leq 0.003$ for the ITT, PP, and completer populations respectively). The actual erythema scores are given in table 10 below.

Table 10: Study A03125 Erythema Ratings

Response		Baseline		End of Study					
				ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
1. None	N	0	0	17	7	17	7	11	4
	%	-	-	10.4%	4.2%	12.7%	4.6%	9.7%	3.2%
2. Mild	N	53	53	89	76	78	69	67	57
	%	32.3%	32.1%	54.3%	46.1%	58.2%	45.7%	58.8%	45.6%
3. Moderate	N	93	101	49	77	35	70	33	59
	%	56.7%	61.2%	29.9%	46.7%	26.1%	46.4%	29.0%	47.2%
4. Severe	N	18	11	9	5	4	5	3	5
	%	11.0%	6.7%	5.5%	3.0%	3.0%	3.3%	2.6%	4.0%
p-value*		0.345		0.002		0.001		0.004	
All	N	164	165	164	165	134	151	114	125

*Significance level of CMH test of equality of proportions using modified ridit scores.

Again using the mean unmodified original erythema scores as seem to be suggested in the Statistical Analysis Plan, there are statistically significant differences in treatment ($p \leq 0.002$, $p \leq 0.004$, and $p \leq 0.001$ for the ITT, PP, and completer populations respectively).

From Table 11 below we see there are no statistically significant comparisons based on the change from baseline in telangiectasia or the original telangiectasia scores:

Table 11: Study A03125 Decrease From Baseline in Telangiectasia Ratings at End of Study

Decrease †		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
2	N	2	3	2	3	1	2
	%	1.2%	1.8%	1.5%	2.0%	0.9%	1.6%
1	N	27	21	21	21	18	17
	%	16.5%	12.7%	15.7%	13.9%	15.8%	13.6%
0	N	127	132	105	119	91	100
	%	77.4%	80.0%	78.4%	78.8%	79.8%	80.0%
-1	N	8	8	6	7	4	6
	%	5.0%	4.9%	4.5%	4.6%	3.5%	4.8%
-2	N	0	1	0	1	0	0
	%	-	0.6%	-	0.7%	-	-
All	N	164	165	134	151	114	125
p-value*		0.743		0.884		0.745	

*Significance level of CMH test of equality of mean proportions using modified ridit scores.
†Note that a negative decrease is an increase.

Table 12: Study A03125 Telangiectasia Ratings at End of Study

Response		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
1. None	N	8	4	8	4	4	3
	%	4.9%	2.4%	6.0%	2.6%	4.9%	2.4%
2. Mild	N	110	111	90	102	77	83
	%	67.1%	67.3%	67.2%	67.6%	67.1%	66.4%
3. Moderate	N	45	48	36	43	33	38
	%	27.4%	29.1%	26.9%	28.5%	27.4%	30.4%
4. Severe	N	1	2	0	2	0	1
	%	0.6%	1.2%	-	1.3%	-	0.8%
All	N	164	165	134	151	114	125
p-value*		0.598		0.213		0.468	

*Significance level of CMH test of equality of mean proportions using modified ridit scores.

Again, whether one uses change from baseline or the original scores, there were no statistically significant differences in telangiectasia (all $p \geq 0.213$).

Thus we see that in Study A03125 the erythema ratings, and from Appendix 5.0, table A.7 the investigators rating of improvement all show statistically significant differences in favor of Azelaic Acid (AzA) over its vehicle. In addition the patient's rating of improvement also shows statistically significant differences in favor of AzA.

6.2 Results for Study A30126

6.2.1 Primary Endpoints

As before, the following table gives week 12 means in inflammatory lesion counts for the three patient populations, and the significance levels of the test for treatment differences from an ANCOVA model with classification effects for treatment, center, and interaction, and baseline lesion count as a continuous covariate.

Table 13: Study A30126 Means and Tests of Treatment Differences in Inflammatory Lesion Counts

Response		ITT			Completers			Per Protocol		
	Treat	N	Mean	p-value	N	Mean	p-value	N	Mean	p-value
Change	Veh	166	6.4	0.0077	146	7.3	0.0012	121	6.6	0.0012
	AzA	167	9.0		148	10.2		110	10.2	
% Change	Veh	166	38.2	0.0172	146	42.8	0.0079	121	41.3	0.0124
	AzA	167	50.0		148	55.8		110	56.1	

ANCOVA Model: expected response = baseline + center + treatment + interaction

As in Study A03125, for all three populations differences, whether we use the change from baseline as specified in the protocol, or the percent change from baseline, treatment differences are statistically significant (all $p \leq 0.0172$), although results are much less extreme than in the other study. From results not displayed here we find that the only interaction that is close to statistical significance is for the change from baseline in the population of completers ($p \leq 0.0687$). All others were greater than 0.17, usually much greater. Even among the population of completers, the interactions were only quantitative. Again, results deleting the interaction term are given in appendix 4.0, and are virtually identical to those above. Deleting the baseline covariate makes all interactions become statistically non-significant (all $p \geq 0.14$). Results on this covariate parallel those in Study A03125. That is, tests of homogeneity of covariates are either statistically non-significant or have much smaller sums of squares than those attributable to the baseline covariate.

The following table, Table 14, summarizes responses to the Investigator's Global Assessment at nominal Week 12 or at study endpoint. As above, the rows labelled success give numbers and percentages of patients who fit the treatment success category defined by the Medical Officer. The entry "0(clear)" gives counts and percentages of subjects evaluated as clear on the investigator's global assessment. The rows labelled "0,1(≤Min)" give counts and percentages of subjects evaluated as clear or minimal, and the rows labelled rows labelled "0-2(≤Mild)" give counts and percentages of subjects evaluated as clear, minimal, or mild. The rows labelled "3-6" gives the counts and percentages of the remaining subjects. Finally row labelled the "All" gives the total number of subjects in each population at nominal week 12.

Table 14: Study A30126 Investigator Global Evaluation at End of Study

Treat		ITT			Completers			Per Protocol		
		AzA	Veh	p-value*	AzA	Veh	p-value*	AzA	Veh	p-value*
Success	N	53	36	0.044	51	36	0.061	40	25	0.009
	%	31.7%	21.7%		34.5%	24.7%		36.4%	20.7%	
0 (clear)	N	11	10	0.889	11	10	0.819	9	5	0.164
	%	6.6%	6.0%		7.4%	6.8%		8.2%	4.1%	
0,1 (≤Min)	N	64	48	0.078	62	46	0.065	49	34	0.010
	%	38.3%	28.9%		41.9%	31.5%		44.5%	28.1%	
0-2 (≤Mild)	N	102	79	0.016	97	75	0.013	74	59	0.006
	%	61.1%	47.6%		65.5%	51.4%		67.3%	48.8%	
3-6	N	65	87		51	71		36	62	
	%	38.9%	52.4%		34.5%	48.6%		32.7%	51.2%	
All	N	167	166		148	146		110	121	

*Significance level of CMH test of equality of mean proportions u

Again, as discussed in the statistical analysis section, the results on treatment success should be considered primary. In the ITT populations results are statistically significant (p≤ 0.044). Results in the per protocol group are also statistically significant (p ≤ 0.009), and are close to significance in the completer population (p ≤ 0.061). These latter tests, as well as other dichotomizations, are considered to be secondary.

6.2.2 Secondary Endpoints

Scores in erythema, telangiectasia, investigator rating of improvement, and the nodule count were considered to be of clinical relevance. Table 15 below presents results on change from baseline in erythema as specified by the protocol. Using the mean modified ridit score over treatments as given in the protocol there are statistically significant differences in treatment (p≤0.006, p≤0.001, and p≤0.003 for the ITT, PP, and completer populations respectively).

However, in the analysis directed by the Medical Officer erythema is analyzed using variables denoting treatment success, where "success (1)" and "success (1,2)" denote the proportions of subjects who had a baseline score of moderate or severe and achieved either a final score of none or a final score of none or mild, respectively. These results appear in Table 16 below. For Successs(1) no treatment differences are statistically significant. However, for Success (1,2) all three populations have statistically significant differences in treatment (all p≤0.001). Again modifying these populations to delete the subjects with an erythema score of 2 at baseline we still have statistically significant differences from baseline (from results not presented here: p≤0.003, p≤0.005, and p≤0.003 for the ITT, PP, and completer populations respectively).

Table 15: Study A03126 Decrease From Baseline in Erythema Ratings at End of Study

Decrease †		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
2	N	8	6	8	6	5	6
	%	4.8%	3.6%	5.4%	4.1%	4.6%	5.0%
1	N	69	40	65	37	52	27
	%	41.3%	24.1%	43.9%	25.3%	47.3%	22.3%
0	N	80	102	68	88	48	74
	%	47.9%	61.5%	46.0%	60.3%	43.6%	61.2%
-1	N	10	18	7	15	5	14
	%	6.0%	10.8%	4.7%	10.3%	4.6%	11.6%
p-value*		0.006		0.003		0.001	

*Significance level of CMH test of equality of mean proportions using modified ridit scores.
†A negative decrease is an increase.

Table 16: Study A03126 Decrease From Baseline in Erythema Ratings at End of Study

Decrease		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
Success(1)#	N	4	4	4	4	3	4
	%	2.4%	2.4%	2.7%	2.7%	2.7%	3.3%
p-value‡		0.936		0.950		0.783	
Success (1,2)#	N	61	31	57	30	46	26
	%	36.5 %	18.7%	38.5 %	20.6%	41.8 %	21.5%
p-value‡		0.001		0.001		0.001	
All	N	167	166	148	146	110	121

#Success(1) and Success(1,2) denote the proportion of subjects who have a baseline score of 3 or 4 and whose final erythema score is either 1 or is 1 or 2, respectively.
‡Significance level of MH test of equality of proportions in success in erythema using modified ridit scores.

Results for the actual erythema scores are given in Table 17 below. Using the mean modified ridit score results are close to statistical significance ($p \leq 0.079$, $p \leq 0.055$, and $p \leq 0.069$ for the ITT, PP, and completer populations respectively).

Table 17: Study A03126 Erythema Ratings

Response		Baseline		End of Study					
				ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
1. None	N	0	0	9	9	9	9	7	7
	%	-	-	5.4%	5.4%	6.1%	6.2%	6.4%	5.8%
2. Mild	N	43	57	90	67	85	63	63	50
	%	25.7%	34.3%	53.9%	40.4%	57.4%	43.2%	57.3%	41.3%
3. Moderate	N	100	85	54	72	42	58	30	51
	%	59.9%	51.2%	32.3%	43.4%	28.4%	39.7%	27.3%	42.2%
4. Severe	N	24	24	14	18	12	16	10	13
	%	14.4%	14.5%	8.4%	10.8%	8.1%	11.0%	9.1%	10.7%
p-value*		0.194	[redacted]	0.079	[redacted]	0.069	[redacted]	0.055	[redacted]
All	N	167	166	167	166	148	146	110	121

*Significance level of CMH test of equality of proportions using modified ridit scores.

Table 18 below gives change from baseline in telangiectasia:

Table 18: Study A03126 Decrease From Baseline in Telangiectasia Ratings at End of Study

Decrease †		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
2	N	2	3	2	2	1	2
	%	1.2%	1.8%	1.4%	1.4%	0.9%	1.7%
1	N	26	23	22	21	19	18
	%	15.6%	13.9%	14.9%	14.4%	17.3%	14.9%
0	N	122	129	109	114	81	95
	%	73.1%	77.7%	73.7%	78.1%	73.6%	78.5%
-1	N	16	11	14	9	9	6
	%	9.6%	6.6%	9.5%	6.2%	8.2%	5.0%
-2	N	1	0	1	0	0	0
	%	0.6%	-	0.7%	-	-	-
All	N	167	166	148	146	110	121
p-value*		0.691	[redacted]	0.645	[redacted]	0.523	[redacted]

*Significance level of CMH test of equality of mean proportions using modified ridit scores.
†Note that a negative decrease is an increase.

Table 19: Study A03126 Telangiectasia Ratings at End of Study

Response		ITT		Completers		Per Protocol	
		AzA	Veh	AzA	Veh	AzA	Veh
1. None	N	8	9	7	8	4	6
	%	4.8%	5.4%	4.7%	5.5%	3.6%	5.0%
2. Mild	N	96	98	85	83	66	70
	%	57.5%	59.0%	57.4%	56.9%	60.0%	57.9%
3. Moderate	N	57	51	50	47	35	39
	%	34.1%	30.7%	33.8%	32.2%	31.8%	32.2%
4. Severe	N	6	8	6	8	5	6
	%	3.6%	4.8%	4.1%	5.5%	4.5%	5.0%
All	N	167	166	148	146	110	121
p-value		0.900	[redacted]	0.920	[redacted]	0.924	[redacted]

*Significance level of CMH test of equality of proportions using modified ridit scores.

Again, whether one uses change from baseline or the original scores, there were no statistically significant differences in telangiectasia (all $p \geq 0.523$).

Thus, as before, we see that in Study A03126 the erythema ratings (whether the Medical Officer's or the sponsor's), and from Appendix 5.0, table A.7 the investigators rating of improvement all show statistically significant differences in favor of Azelaic Acid 15% gel over its vehicle. In addition the patient's rating of improvement also shows statistically significant differences in favor of AzA.

7. Supporting Studies

Not applicable.

8. Safety Variables / Adverse Events

Assessing adverse events is primarily a matter of clinical judgement. Further, since there are large numbers of adverse events and the study was not powered to test for adverse events, any statistical analysis should be interpreted as a "post hoc" analysis, with all the possible problems associated with such post hoc analyses. However, it was felt that a multiplicity adjusted test of differences between AzA 15% gel and its vehicle for the various adverse events collected over the studies might be useful.

These analyses are based on the pooled adverse event data from both studies cited above, and after the baseline visit. Since this ignores dropouts these percentages should be considered

as underestimates of the proportions of adverse events. Of course, as noted before the percentage of dropouts is not particularly large, so the impact should not be too extreme.

To test for the statistical significance of any differences in reported adverse events between azelaic acid and its vehicle, the adverse events were first screened for those with five or more subjects experiencing the event. The number five was arbitrary, but reduces the number of adjustments required, and hence should increase power in the tests adjusted for multiplicity. Twenty adverse events met this criterion in the pooled data set. However, only the following were close to statistically significant (prior to adjusting for multiplicity of tests):

	Incidence		Unadjusted	Adjusted
Description	Vehicle	Azelaic Acid	p-value	p-value
	N=331	N=333		
Pain in Skin	14	74	< 0.0001	< 0.0001
Paresthesia	3	40	< 0.0001	< 0.0001
Pruritis	15	41	0.0004	0.0027

The unadjusted p-value is the p-value from a Fisher Exact test of differences between AzA and vehicle. All other unadjusted p-values were greater than 0.12. Adjusting the tests for the 20 potential comparisons using the techniques of Westfall and Young (1993) gives the "Adjusted p-value" cited above. In this particular case, the adjustments were done by resampling 664 vectors of adverse events from the permutation distribution corresponding to the adverse events. This gave 20 Fisher exact tests for each complete replicate from the permutation distribution. This was repeated 20,000 times to get 20,000 p-values for each Fisher exact test. The adjusted p-value is the proportion of time more extreme results than the original critical value were obtained. By sampling from the subject vectors of reported adverse events features of the distribution and inter-test correlations are incorporated into the analysis.

Thus, we would conclude that after adjusting for the multiplicity of possible adverse events, there is statistically quite significant evidence that the AzA treatment is associated with stinging skin and pruritis. Otherwise there was no particular evidence of differences across treatment groups in terms of listed adverse events. This does not mean there no differences, just that this relatively conservative procedure did not detect any.

To investigate the effect of these over time, the following table gives the numbers and percentages of patients with valid data who report one of the events above. Using the subjects with valid data at nominal weeks 4, 8 and 12 as the risk set (i.e. the denominator in computed proportions), we get the values below:

Table 20: Pooled Studies Incidence of Adverse Events Over Time

AE	Treatment		Week		
			04	08	12
Pain Skin	AzA	N	67 / 323	33 / 295	16 / 284
		%	20.7 %	11.2 %	5.6 %
	Vehicle	N	12 / 319	7 / 303	5 / 297
		%	3.8 %	2.3 %	1.6 %
Parathesia	AzA	N	39 / 323	26 / 295	18 / 284
		%	12.1 %	8.8 %	6.3 %
	Vehicle	N	12 / 318	7 / 303	5 / 297
		%	3.8 %	2.3 %	1.7 %
Pruritis	AzA	N	39 / 324	26 / 295	16 / 284
		%	12.0 %	8.8 %	5.6 %
	Vehicle	N	12 / 318	5 / 303	6 / 297
		%	3.8 %	1.7 %	2.0 %

So, for all three adverse events, while the proportions are statistically significantly higher in the Azelaic Acid 15% gel group than in the vehicle group, these proportions do tend to become smaller over time.

9. Subgroup Analyses:

The studies were not powered to detect differences in subgroups. So no statistical tests are provided. However profiles over time of responses in subgroups may be helpful in describing results. For convenience, and to exaggerate possible differences, the two studies were pooled.

The following table displays mean change from baseline in lesion counts for various subgroups of the ITT population in the two pooled Phase 3 Studies.

Table 21: Pooled Studies Lesion Counts by Demographic Subgroup

		Examination Visit									
Gender	Week:	04		08		12		Per Protocol		ITT	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
	Change from baseline										
Female	Mean	6.2	4.8	9.5	6.4	10.9	7.5	10.5	7.5	9.4	6.6
	Std Dev	10.0	8.5	10.2	8.7	10.1	8.7	10.1	8.4	11.0	8.9
Male	Mean	7.7	5.0	10.5	7.0	12.9	7.4	14.0	7.2	11.0	7.2
	Std Dev	9.7	7.7	10.6	8.4	10.8	10.2	11.1	10.3	11.2	10.0
	% Change from baseline										
Female	Mean	35.8	24.5	53.6	35.2	60.4	43.1	59.8	44.7	53.4	38.7
	Std Dev	42.0	48.6	42.9	45.9	40.9	41.6	43.0	41.2	45.9	43.4
Male	Mean	37.5	27.5	51.6	37.1	63.3	41.5	66.3	38.7	55.3	40.0
	Std Dev	39.2	35.4	42.2	42.5	37.8	45.3	33.7	45.1	42.8	44.9

Note that for both genders the differences between treatment means are roughly the same. That is, for both genders the superiority of AzA over its vehicle in terms of lesion counts is roughly the same.

Patients were dichotomized into two "race" groups, Caucasian versus non-Caucasian.

Table 21 (cont.): Pooled Studies Lesion Counts by Demographic Subgroup

Race		Examination Visit Week: 04		08		12		Per Protocol		ITT	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
	Change from baseline										
Caucasian	Mean	6.9	4.5	9.8	6.4	11.7	7.3	11.6	7.4	10.2	6.6
	Std Dev	8.9	8.1	10.4	8.4	10.2	8.8	10.3	8.4	10.6	8.8
Other	Mean	3.2	9.0	9.0	9.2	7.8	9.6	8.9	7.9	5.5	8.8
	Std Dev	17.8	9.1	9.7	10.3	11.5	13.7	12.6	14.8	15.2	13.0
	% Change from baseline										
Caucasian	Mean	37.3	24.2	53.2	35.1	62.8	41.8	62.4	42.6	55.8	38.3
	Std Dev	38.6	45.6	43.1	44.6	38.7	42.3	40.2	41.9	43.7	43.6
Other	Mean	23.5	41.5	51.5	44.8	41.8	54.2	48.0	48.8	33.2	49.5
	Std Dev	64.8	37.6	38.8	49.1	50.2	46.3	49.0	49.1	54.8	45.6

Most subjects were Caucasian (see Table 3, page 7). There is some apparent evidence that AzA may be less effective for non-Caucasian patients than among Caucasian patients, although with the small number of non-Caucasian patients this may well be an artifactual result. In an ANOVA (not shown in this review) with terms for baseline, race, treatment, and interaction, the interaction was found to be statistically non-significant. Due to the vagaries of hypothesis testing this does not show that there is no difference. However, any difference is too small to be statistically significant.

Age group		Week: 04		08		12		Per Protocol		ITT	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
	Change from baseline										
age 65+	Mean	6.9	5.5	9.8	6.9	10.5	9.3	11.1	9.8	9.8	8.1
	Std Dev	7.5	6.8	7.9	7.4	7.8	8.1	7.9	8.4	7.6	8.0
age<65	Mean	6.6	4.8	9.7	6.5	11.5	7.3	11.4	7.1	9.8	6.6
	Std Dev	10.2	8.4	10.6	8.7	10.6	9.3	10.8	9.0	11.5	9.3
	% Change from baseline										
age 65+	Mean	42.2	30.8	60.4	39.5	64.7	50.6	64.7	49.6	62.7	45.6
	Std Dev	34.7	29.9	31.5	37.2	34.3	33.0	28.9	33.5	34.5	34.7
age<65	Mean	35.4	24.6	52.1	35.2	60.7	41.7	60.9	42.2	52.8	38.2
	Std Dev	42.0	46.9	43.9	45.8	40.8	43.6	42.4	43.3	46.2	44.8

Nearly 90% of the patients were aged below 65. It may appear that while AzA is superior to its vehicle in the older patients, it seems to be more effective in the younger patients. However, as above, in a separate ANOVA with terms for baseline, age group, treatment, and interaction, the interaction was found to be statistically generally quite non-significant. Again, this suggests that any difference is too small to be statistically significant.

The following table displays the pooled results from the investigator global evaluation. Recall that a "success" was defined as a subject who if they were mild at baseline (i.e. a score of 2) achieved a clear (i.e. score of 0) at the point of measurement or achieved a score of clear or minimal (i.e., 0 or 1) at the end of the study with a baseline score of 3-6, i.e. "Mild to moderate to severe." More complete tables of the IGA are given in Appendix 7.0.

Table 22: Pooled Studies Investigator's Global Evaluation by Demographic Subgroup

		Examination Visit									
Gender	Week:	04		08		12		Per Prot.		ITT	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Female											
Success	N	25	10	56	29	72	41	58	36	74	42
	%	10.6	4.4	25.8	13.4	34.6	19.5	34.3	20.2	30.5	17.5
Overall	N	235	228	217	217	208	210	169	178	243	240
Male											
Success	N	8	6	17	10	28	14	21	6	29	14
	%	9.4	6.7	22.4	11.6	37.8	16.1	38.2	8.8	33.0	15.4
Overall	N	85	90	76	86	74	87	55	68	88	91

Observe that for both genders the superiority of Azelaic Acid over its vehicle in terms of the IGA scores is roughly the same.

Race	Week:	04		08		12		Per Prot.		ITT	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Caucasian											
Success	N	30	15	65	35	93	48	74	37	95	49
	%	10.2	5.1	24.1	12.4	35.8	17.3	35.6	16.1	31.3	15.9
Overall	N	295	297	270	283	260	277	208	230	304	308
Other											
Success	N	3	1	8	4	7	7	5	5	8	7
	%	12.0	4.8	34.8	20.0	31.8	35.0	31.3	31.3	29.6	30.4
Overall	N	25	21	23	20	22	20	16	16	27	23

Unlike for gender, there is apparent evidence that Azelaic acid 15% gel may be less effective for non-Caucasian patients than among Caucasian patients. However, the number of patients is so small that is may well be an artifact of the experiment.

Age Group	Week:	04		08		12		Per Prot.		ITT	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Age<65											
Success	N	28	14	60	36	86	51	69	38	88	51
	%	10.0	5.0	23.3	13.4	34.7	19.2	35.2	17.4	30.1	17.4
Overall	N	281	281	258	269	248	265	196	218	292	293
Age 65+											
Success	N	5	2	13	3	14	4	10	4	15	5
	%	12.8	5.4	37.1	8.8	41.2	12.5	35.7	14.3	38.5	13.2
Overall	N	39	37	35	34	34	32	28	28	39	38

Observe that for both age groups the superiority of Azelaic Acid 15% gel over its vehicle in terms of the IGA scores is fairly consistent. Again more extensive tables are given in Appendix 7.

10. Statistical and Technical Issues:

Appendix 1 displays box plots of per treatment means in the ITT population for each center in each study, both at baseline and at the end of each study (usually nominal week 12). As discussed in Appendix 2, in each study, one center seems to have especially high efficacy relative to baseline. In Study A03125 center 07 would seem to be especially efficacious, but deleting that center or down-weighting it with the rank analysis or the Winsorized analysis does not change results. However, Study A03126 is somewhat more problematical. While deleting this center does generally have an effect on final conclusions we can say that results would still trend to be in favor of Azelaic Acid 15% gel.

The sponsor proposed no method of error control for the multiplicity of tests performed in the analysis. This reviewer would argue that error control over the numerous tests performed might be useful. However, for this NDA only tests on the primary endpoints using the ITT population are considered as primary. All other results are considered as secondary. Since there is no error control any p-values associated with these secondary endpoints should be considered as supporting, not conclusive. That is, the secondary analyses can be viewed as sensitivity analyses.

Besides the earlier manifestations of the primary endpoints, the sponsor proposes seven other secondary endpoints. Only four of these are considered to be clinically relevant, namely erythema, telangiectasia, investigator rating of improvement, and nodule counts. One of these four, nodule counts, is analyzed using Bayesian techniques, where typically multiplicity problems are not considered relevant (see Appendix 5). Typically no adjustment for multiplicity is applied for a few endpoints. Optionally, these three could be analyzed using a Bonferroni correction. That is, a comparison is only labelled as statistically significant if it is nominally significant at a .05/3=.0167 level. Other secondary endpoints are considered as tertiary in the FDA analysis (see Appendix 6).

The ANCOVA models used to analyze the lesion counts included terms for treatment, center (investigator), interaction, and the baseline covariate. These were analyzed using SAS® Type 3 sums of squares. Tests based on Type 3 sums of squares are not as powerful as those based on sequential sums of squares, but unless the center by treatment cells are extremely disparate in size, the loss in power is not large. The advantage is that Type 3 sums of squares are based on simple contrasts in least squares means, which enhances interpretability. The sponsor's procedure is to delete statistically non-significant interaction terms from the model. However this tends to attenuate the most extreme least squares means and can mask extreme treatment by center interaction cells as are observed here in both studies. Thus for small models, this reviewer

prefers to retain interaction terms in the model. However, it can be argued that this is largely a matter of taste.

11. Statistical Evaluation of Evidence:

Deleting the two subjects with IGA values of "minimal" at baseline, in the two studies a total of 27 centers enrolled 331 patients treated with Azelaic Acid (AzA) 15% gel and 331 patients treated with vehicle. Neither study had any statistically significant differences between the treatment groups in demographic measures or other baseline characteristics. About 86% of the patients completed Study A03125 and about 88% completed Study A03126.

The primary assessment of efficacy was based on the intent-to-treat (ITT) population, with missing values imputed by last observation carried forward (LOCF). The sponsor proposed two primary endpoints, the change from baseline in inflammatory lesion counts and a measure based on an investigator global assessment (IGA). The IGA was measured on a seven point ordinal scale, 0-6.

Table 23: Summary of Results Using FDA Primary Endpoints

	Study A03125			Study A03126		
	AzA	Veh	p-value	AzA	Veh	p-value
N Subjects	164	165		167	166	
Inflammatory Lesion Count Means						
Change from Baseline	10.7	7.1	0.0001*	9.0	6.4	0.0077*
% Change from Baseline	57.9	39.9	0.0003*	50.0	38.2	0.0172*
Investigators Global Evaluation						
Treatment Success	50 31%	20 12%	0.001†	53 32%	36 22%	0.044†
Failure	114 69%	145 88%		113 68%	131 78%	

*Model: expected response = baseline + center + treatment (interactions not statistically significant)
† Significance level of CMH test of equality of proportions using modified ridit scores.

In communication with the sponsor the DDDDP recommended both the simple change from baseline and the percent change from baseline as co-primary endpoints. This differs from the sponsor's protocol where the simple change from baseline is primary and the percent change from baseline is considered as secondary. Further, the Medical Officers specified a different dichotomization of the investigator global evaluation than that given in the sponsor's protocol. In the FDA analysis a treatment success is defined as occurring when the subjects had achieved at least a two step reduction in the IGA AND achieved a final score of clear or minimal ("0" or "1"). Results using these endpoints are summarized in the table above.

Thus in Study A03125 all endpoints showed statistically significant differences in favor of AzA over vehicle (all three p-values ≤ 0.001). Results are similar, but not as strong in Study A03126 (all three p-values ≤ 0.044). By chance, the dichotomization of the Investigators Global Evaluation specified by the FDA Medical Officer was slightly more favorable to the sponsor than that originally chosen by the sponsor.

As is typical of lesion count data, the values seem to be skewed to the right. However, results from an equivalent rank transformed analysis are essentially equivalent to the results on the untransformed data reported here. Further, to some extent in Study A03126 statistical significance of the results are largely driven by one center. However, deleting this center, results were still trending to be in favor the AzA over its vehicle (see Appendix 2 for details.).

One approach to analyzing safety comparisons is to use the techniques of Westfall and Young (1993). To limit the size of the family this was restricted to adverse events occurring 5 or more times in both studies. Using these techniques we would conclude that of possible adverse events, there is statistically quite significant evidence that the AzA 15% gel treatment is associated with measures of stinging skin and pruritis (all $p \leq 0.0027$, adjusting for multiplicity). Otherwise there was no particular evidence of differences across treatment groups in terms of listed adverse events. This does not mean there no differences, just that this relatively conservative procedure did not detect any.

12. Conclusions

According to the sponsor, this New Drug Application was submitted to investigate the efficacy and safety of Azelaic Acid (AzA) 15% gel when used for 12 weeks in patients with moderate papulopustular rosacea (stage 2 rosacea). To study the efficacy of AzA 15% gel the sponsor provided results from two virtually identical randomized, double-blind, placebo-controlled, multi-center studies (Studies A03125 and A03126, respectively). With the concurrence of the FDA, inflammatory lesion counts and an Investigator's Global Evaluation were the primary endpoints. However, the FDA analysis of these endpoints differs from the sponsor's analysis in several ways including slightly different analysis populations and the actual definitions of primary endpoints. The sponsor's protocol specified that lesion counts were to be assessed using change from baseline, while the preferred FDA measure was per cent change from baseline. However, the sponsor wins on both measures. In study A03125 lesion count differences between AzA15% gel and vehicle were highly statistically significant in favor of AzA (for both measures, $p \leq 0.0003$). In study A03126 results were not quite as distinct (for change from baseline $p \leq 0.0077$ while for per cent change $p \leq 0.0172$). For statistical analysis the seven level Investigator's Global Assessment (IGA) was dichotomized into treatment success or treatment failure as explained below. For Study A03125 treatment differences in favor of AzA gel 15% on the IGA were highly statistically significant ($p \leq 0.001$), while for Study A03126 results were barely statistically significant ($p \leq 0.044$). By chance, the dichotomization

of the Investigators Global Evaluation specified by the FDA Medical Officer was slightly more favorable to the sponsor than that originally chosen by the sponsor. In each study there is one center where the difference between treatment groups is especially strong. These and other questions about the distributions of the data were addressed in the report or the associated appendices.

To summarize, restricting attention to the primary analysis population, i.e., the ITT subjects with LOCF imputation, we find that in Study A03125 the change from baseline in inflammatory lesions, the percent change from baseline in inflammatory lesion counts, and the investigator's global assessment, all showed statistically significant differences in favor of Azelaic Acid over vehicle (all three p-values ≤ 0.001). Results are somewhat similar, but not as strong in Study A03126 (all three p-values ≤ 0.044).

Steve Thomson
Mathematical Statistician, Biometrics III

concur: Mohamed Alosh, Ph.D.
Team Leader, Biometrics III

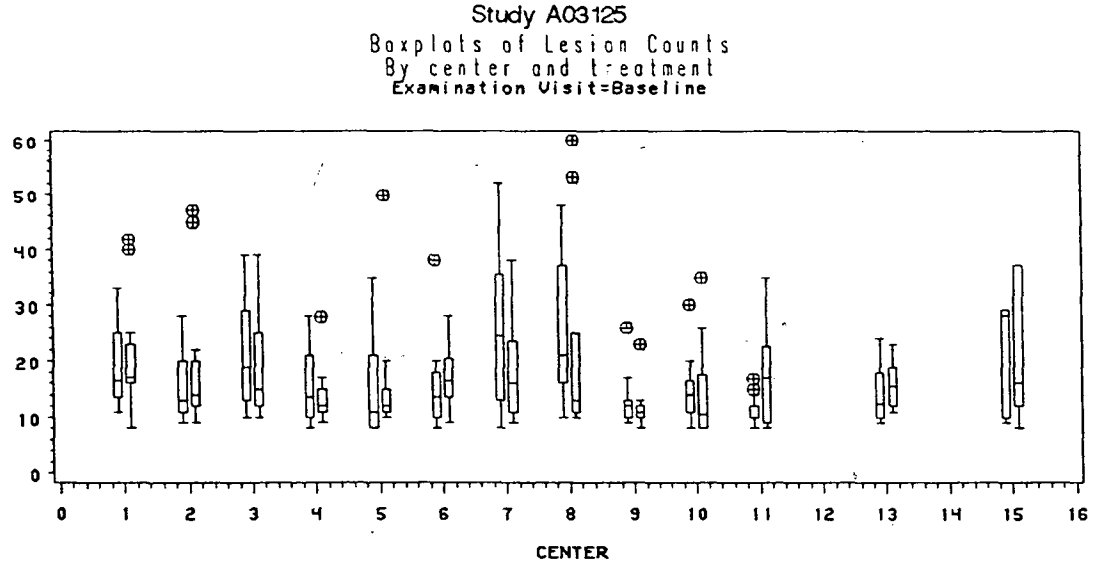
cc:
Archival NDA: 21,470 Finevin 15% gel
HFD-540/Division File
HFD-540/Dr. Wilkin
HFD-540/Dr. Vaughan
HFD-540/Dr. Luke
HFD-540/Ms. Wright
HFD-725/Dr. Huque
HFD-725/Dr. Anello
HFD-725/Dr. Alosh
HFD-725/Mr. Thomson
This review has 34 pages, plus a 17 page appendix.
Chron.

\Thomson\MyDocuments\x 7-2044\November 18, 2002\c:\mydocuments\N21470FinaceaBerlex.doc

Appendix 1.0: Box Plots of Inflammatory Lesion Counts

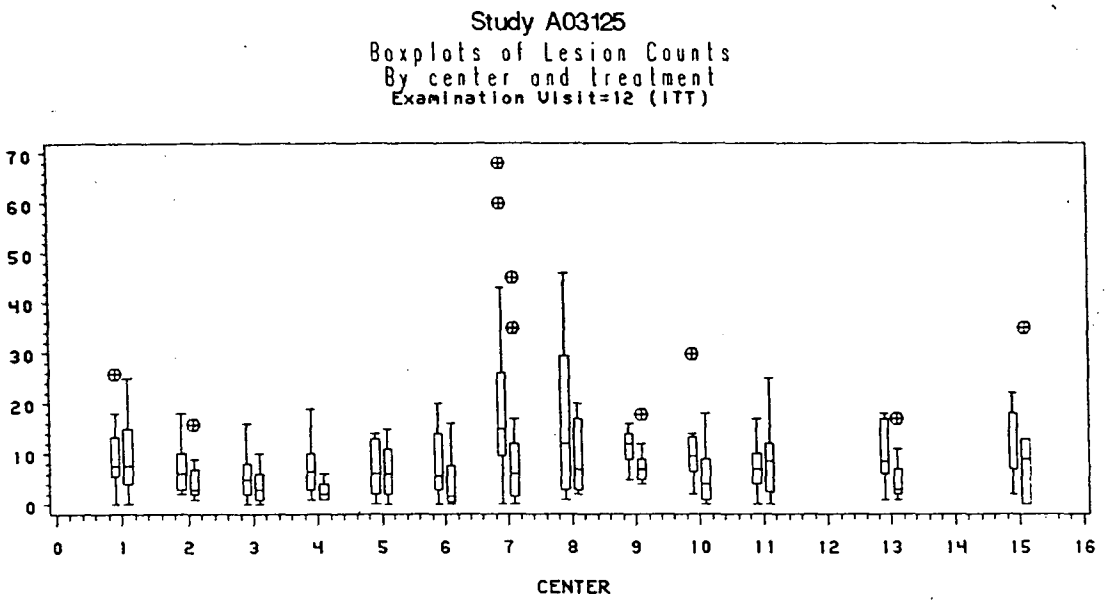
Boxplots give reasonable summary pictures of a distribution. The first and third sample quartiles form the box (see plots following). The line within the box is the median. The "whiskers" extend out from the quartiles to the maximum (or minimum) of the last data point whose distance from the box is less than or equal to 1.5 times the interquartile range. Points outside these limits are generally considered as outliers. In the following plots each center has a separate boxplot for vehicle and for Azelaic acid treatment. Centers are listed on the x-axis. Within each center the boxplot for the Azelaic acid group appears on the right. The plot for the vehicle group appears on the left.

Figure A1. Study A03125: Box Plots for Baseline Inflammatory Lesion Counts:



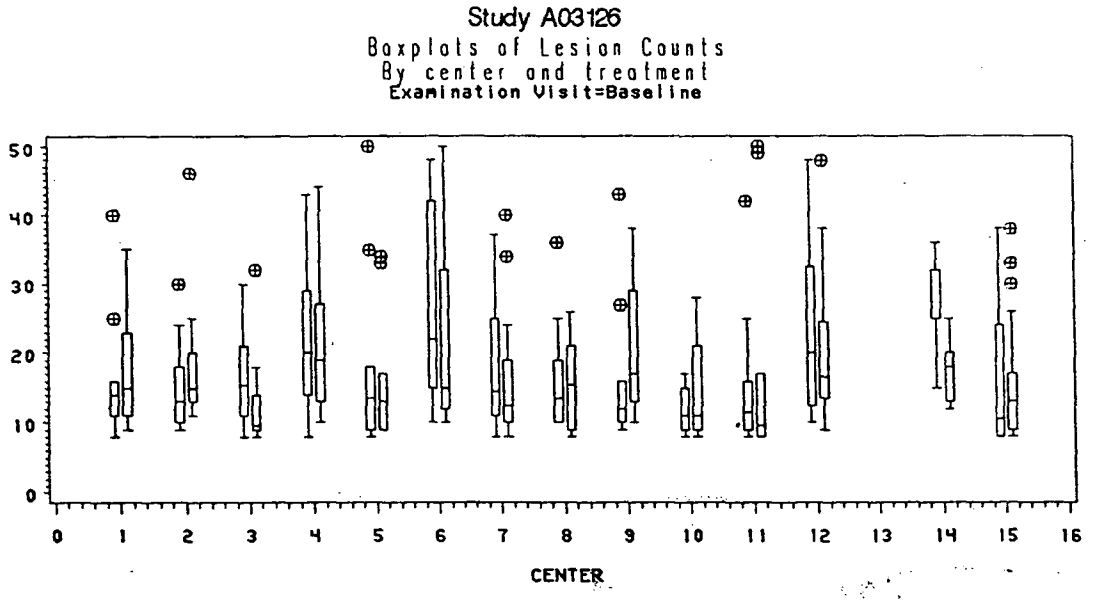
Smaller numbers indicate fewer lesions at baseline. Thus, in some centers the baseline allocation appears to favor AzA, in others it favors vehicle.

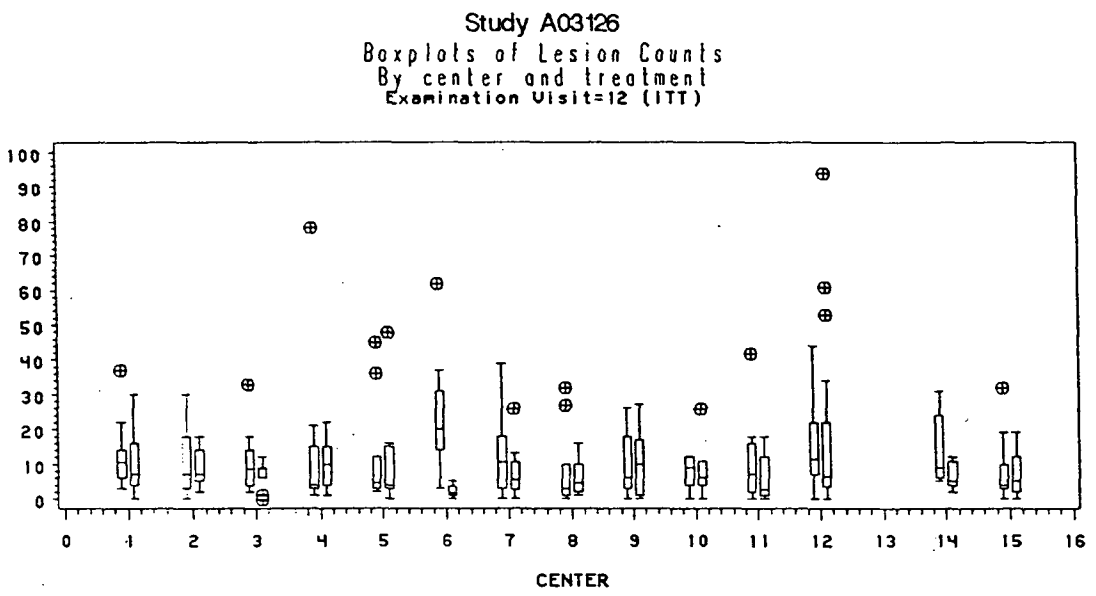
At the end of the study we get the following plot:



Observe that in the preceding plots there is only one center whose mean clearly favors the vehicle, and that is center 11, where the baseline scores strongly favored vehicle. This difference vanishes when looking at least squares means (i.e., adjusted for the baseline)

Figure A2. Study A03126: Box Plots for Inflammatory Lesion Counts:





Note again, vehicle box plots are to the left, treatment boxplots to the right.

In Study A03126 note the extreme compression of lesion counts in the Azelaic Acid 15% gel relative to the vehicle.

Appendix 2.0: Sensitivity to Centers

For both studies there were relatively few dropouts, but in the Study A03125 they are largely concentrated in the Azelaic Acid group in three centers.

Appendix Table A.1: Subject Counts by Center

Study A03125 (protocol 304342)
Number of Subjects by Center: Week 12 Count/ Baseline Count

Center	01	02	03	04	05	06	07	08	09	10	11	13	15
Veh	18/ 20	10/ 11	15/ 15	10/ 10	9/ 10	13/ 14	19/ 20	5/ 8	8/ 9	17/ 20	12/ 13	10/ 10	5/ 5
AzA	11 /20	11/ 13	15/ 15	10/ 10	8/ 9	11/ 12	14/ 20	6/ 9	8/ 9	19/ 20	7/ 12	10/ 10	4/ 5

Note that centers 01, 07, and 11 all seem to have a relatively large number of dropouts in the AzA group versus the Vehicle group. But, as will be seen later, except for center 07, the estimated least squares means do not show any unusual discrepancy between the AzA estimates and the vehicle estimates.

Study A03126 (protocol 304344)
Number of Subjects by Center: Week 12 Count/ Baseline Count

Center	01	02	03	04	05	06	07	08	09	10	11	12	14	15
Veh	8/ 10	8/ 10	8/ 10	15/ 15	9/ 10	7/ 10	17/ 20	9/ 10	9/ 11	5/ 7	8/ 10	20/ 20	5/ 5	18/ 18
AzA	9/ 10	6/ 9	9/ 10	17/ 19	7/ 9	8/ 8	20/ 20	9/ 10	10/ 11	5/ 7	9/ 10	18/ 20	5/ 5	16/ 19

Unlike the previous study the numbers of dropouts are roughly the same for each treatment group.

The following displays least squares means for the two studies, where the least squares means are from a model of the ITT population at week 12, with missing values imputed using LOCF. The model has terms for baseline lesion count, treatment, center, and interaction. Least squares means are presented here since the ANOVA tests for treatment are actually simple contrasts in these least squares means.

Appendix Table A.2: Least Square Means and Differences Between AzA and Vehicle

Study A03125

	Change from Baseline			Percent Change from Baseline		
Center	AzA LSM	Veh LSM	diff	AzA LSM	Veh LSM	diff
01	8.9	8.0	-0.9	47.0	45.8	-1.2
02	13.5	10.1	-3.4	72.1	54.2	-18.0
03	14.4	13.7	-0.7	77.2	71.8	-5.4
04	13.7	9.8	-3.9	83.3	51.4	-32.0
05	10.4	9.7	-0.7	55.2	52.4	-2.8
06	13.4	8.7	-4.7	71.5	47.0	-24.5
07	8.5	-0.3	-8.8	48.2	15.1	-33.1
08	10.7	3.7	-7.0	38.6	41.6	3.0
09	7.4	4.5	-2.9	36.3	10.3	-26.0
10	10.5	6.1	-4.4	60.1	30.4	-29.6
11	8.8	7.6	-1.2	44.7	35.3	-9.4
13	11.6	7.0	-4.6	60.9	31.1	-29.7
15	7.9	5.5	-2.4	58.0	39.6	-18.4

In Study A03125 Center 07 seems to be somewhat discrepant from the others. This center does seem to drive part of the overall efficacy results. However, even deleting this center, for the percent change in lesion counts at 12 weeks, treatment differences for each of the ITT, Completer, and Per Protocol populations remain statistically significant (from ANOVAs, not displayed here, deleting this center: $p \leq 0.0001$, $p \leq 0.0001$, and $p \leq 0.0006$ respectively). Similarly MH tests comparing success rates in the IGA are all statistically significant for all three populations $p \leq 0.001$). Results are similar for the change from baseline in lesion counts.

Study A03126

	Change from Baseline			Percent Change from Baseline		
Center	AzA LSM	Veh LSM	diff	AzA LSM	Veh LSM	diff
01	8.0	4.2	-3.8	47.2	24.9	-22.3
02	10.1	6.4	-3.8	40.8	28.2	-12.7
03	8.0	6.9	-1.1	38.9	37.1	-1.7
04	10.2	8.7	-1.5	48.9	53.3	4.4
05	5.9	6.1	0.1	42.1	40.2	-1.9
06	18.6	0.3	-18.3	86.0	17.1	-68.9
07	9.7	5.4	-4.4	53.8	32.8	-21.0
08	10.5	8.7	-1.7	56.1	58.9	2.8
09	9.7	7.2	-2.5	56.4	33.5	-23.0
10	7.5	6.7	-0.8	41.8	33.2	-8.6
11	12.2	6.2	-6.0	70.7	42.6	-28.1
12	2.0	4.6	2.6	27.3	33.2	5.9
14	11.0	8.4	-2.6	56.4	46.8	-9.6
15	10.2	9.1	-1.1	54.3	50.4	-3.9

Even more than any center in Study A03125, in terms of lesion counts, center 06 in Study A03126 was apparently discrepant from the others. The large difference is accentuated by the fact that at the end of the study in this center the AzA 15% gel group shows the highest overall efficacy while the corresponding vehicle group shows the lowest efficacy. Deleting this center, for the percent change in lesion counts at 12 weeks, treatment differences for each of the patient populations vary in statistical significance (from ANOVAs, not displayed here, deleting this

center: p≤ 0.1029, p≤ 0.0637, and p≤ 0.0343, respectively). Similarly MH tests comparing success rates in the IGA seem to show a somewhat similar pattern for each of the ITT, Completer, and Per Protocol populations (p≤ 0.170, p≤ 0.190, and p≤ 0.044, respectively). Again, results are similar for the change from baseline in lesion counts.

Interestingly again, in both studies deletion of the discrepant centers renders treatment by center interactions statistically not significant for all week 12 endpoints. That is, any apparent interaction in both studies is largely driven by these two centers.

Appendix 3.0: Alternative Lesion Count Model (No Baseline Covariate)

It has been observed from other studies that inclusion of a baseline covariate can have a dramatic effect on the conclusions. That is not true for the analyses in these studies. (Note that the following include all centers, including the problematic ones discussed above).

Appendix Table A.3: Study A03125 Means and Tests of Treatment Differences in Inflammatory Lesion Counts

		ITT			Completers			Per Protocol		
Response	Treat	N	Mean	p-value*	N	Mean	p-value*	N	Mean	p-value*
Change	Veh	165	7.1	0.0012	151	7.7	<0.0001	125	8.3	0.0006
	AzA	164	10.7		134	12.7		114	12.4	
% Change	Veh	165	39.9	0.0003	151	42.5	<0.0001	125	44.7	0.0007
	AzA	164	57.9		134	67.2		114	66.5	

*ANOVA Model: expected response = center + treatment + interaction

Appendix Table A.4: Study A03126 Means and Tests of Treatment Differences in Inflammatory Lesion Counts

		ITT			Completers			Per Protocol		
Response	Treat	N	Mean	p-value*	N	Mean	p-value*	N	Mean	p-value*
Change	Veh	166	6.4	0.0213	146	7.3	0.0081	121	6.6	0.0022
	AzA	167	9.0		150	10.2		110	10.2	
% Change	Veh	166	38.2	0.0157	146	42.8	0.0028	121	41.3	0.0125
	AzA	167	50.0		150	55.8		110	56.1	

*ANOVA Model: expected response = center + treatment + interaction

For both studies, tests on treatment differences have virtually the same results as those for the models including the baseline score as a covariate (See Tables 6 and 13 of the report).

Appendix 4.0: Response Profiles Over Time

The following tables give profiles of lesion counts over time. Data from visit 00 to visit 12 use subjects who completed up to that point, while the ITT population imputes missing data using last observation carried forward techniques. The models used for time points after the first include terms for baseline count, treatment, center, and interaction. At visit 00 the model used has terms only for treatment, center, and interaction. This is a test of differences at baseline.

Note that since these are not pre-planned comparisons and are only exploratory, a reasonable argument can be made that no family-wise error control is necessary. However, this can be done at little cost, and one way to maintain appropriate error control at later endpoints would be to start testing from the right most entry, and proceed to the left, stopping when the first comparison is statistically non-significant. But again, such an approach is optional.

Appendix Table A.5: Profiles of Lesion Counts

Study A03125	Completers								Per Protocol		ITT	
Examination Visit	00		04		08		12					
Lesion Count	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Change from Baseline												
Mean	17.5	17.6	7.7	4.6	11.1	7.2	12.7	7.7	12.4	8.3	10.7	7.1
Std Dev	9.6	9.2	8.9	7.6	9.6	7.9	9.9	8.7	9.8	8.7	10.4	8.6
p-value	0.8663		0.0008		0.0005		<0.0001		0.0010		0.0001	
% Change From Baseline												
Mean	.	.	41.3	24.3	60.4	38.6	67.2	42.5	66.5	44.7	57.9	39.9
Std Dev	.	.	37.1	39.7	35.6	39.5	33.8	39.8	33.5	39.1	41.9	40.5
p-value	.		0.0002		<0.0001		<0.0001		0.0016		0.0003	

Study A03126	Completers								Per Protocol		ITT	
Examination Visit	00		04		08		12					
Lesion Count	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Change from Baseline												
Mean	17.9	18.5	5.5	5.1	8.5	6.0	10.2	7.3	10.2	6.6	9.0	6.4
Std Dev	10.0	10.5	10.8	8.9	10.9	9.2	10.6	9.7	11.0	9.2	11.7	9.8
p-value	0.5845		0.3937		0.0043		0.0012		0.0012		0.0077	
% Change From Baseline												
Mean	.	.	31.1	26.4	46.3	32.9	55.8	42.8	56.1	41.3	50.0	38.2
Std Dev	.	.	44.6	50.4	47.5	49.6	44.3	45.6	47.0	45.5	46.9	47.7
p-value	.		0.4010		0.0116		0.0079		0.0124		0.0172	

The following tables give similar profiles of the investigator global evaluation over time. Interest is primarily in terms of the dichotomization labelled treatment success.

Appendix Table A.6: Profiles of Investigator's Global Evaluation

Study A03125		Completers								Per Protocol		ITT	
Visit		00		04		08		12					
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Treat Success	N	0	0	14	6	36	16	49	19	39	17	50	20
	%	.	.	8.7	3.8	25.5	10.7	36.6	12.6	34.2	13.6	30.5	12.1
0(Clear)	N	0	0	3	0	4	6	12	5	9	4	12	5
	%	.	.	1.9	0.0	2.8	4.0	9.0	3.3	7.9	3.2	7.3	3.0
0,1(≤Minimal)	N	0	0	18	12	43	22	57	31	45	24	60	32
	%	.	.	11.2	7.5	30.5	14.7	42.5	20.5	39.5	19.2	36.6	19.4
0-2(≤Mild)	N	25	33	70	48	88	57	93	62	76	51	100	67
	%	15.2	20.0	43.5	30.0	62.4	38.0	69.4	41.1	66.7	40.8	61.0	40.6
3-6	N	139	132	91	112	53	93	41	89	38	74	64	98
	%	84.8	80.0	56.5	70.0	37.6	62.0	30.6	58.9	33.3	59.2	39.0	59.4
All	N	164	165	161	160	141	150	134	151	114	125	164	165
p-value(success)		.		0.040		0.001		0.001		0.001		0.001	
p-value(0,1)		.		0.192		0.001		0.001		0.001		0.001	

Study A03126		Completers								Per Protocol		ITT	
Visit		00		04		08		12					
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Treat Success	N	0	0	19	10	37	23	51	36	40	25	53	36
	%	.	.	11.9	6.3	24.3	15.0	34.5	24.7	36.4	20.7	31.7	21.7
0(Clear)	N	0	0	2	0	2	1	11	10	9	5	11	10
	%	.	.	1.3	0.0	1.3	0.7	7.4	6.8	8.2	4.1	6.6	6.0
0,1(≤Minimal)	N	0	0	24	17	44	30	62	46	49	34	64	48
	%	.	.	15.1	10.8	28.9	19.6	41.9	31.5	44.5	28.1	38.3	28.9
0-2(≤Mild)	N	19	23	73	59	88	61	97	75	74	59	102	79
	%	11.4	13.9	45.9	37.3	57.9	39.9	65.5	51.4	67.3	48.8	61.1	47.6
3-6	N	148	143	86	99	64	92	51	71	36	62	65	87
	%	88.6	86.1	54.1	62.7	42.1	60.1	34.5	48.6	32.7	51.2	38.9	52.4
All	N	167	166	159	158	152	153	148	146	110	121	167	166
p-value(success)		.		0.110		0.050		0.063		0.009		0.044	
p-value(0,1)		.		0.290		0.072		0.066		0.011		0.079	

For both studies results show a clear trend over time. However it is worth recalling that this analysis is exploratory only and is not pre-planned in the protocol.

Appendix 5.0: Clinically Relevant Secondary Endpoints

The following tests all use modified ridit scores for the Cochrane-Mantel-Haenszel tests, as were specified in the sponsor's protocol. For interpretability of results, this reviewer has some preference for consecutive integer scores, but not sufficient to override the fact that this specification was pre-planned in the protocol. Note this family of comparisons specified by the Medical Officer as clinically relevant includes comparisons of erythema, telangiectasia, investigator rating of improvement, and nodule counts. The latter is analyzed using Bayesian techniques, where multiplicity is usually ignored. For the three remaining endpoints, no adjustment for multiplicity was deemed necessary. If one wishes to adjust for multiplicity in the three endpoints an easy procedure would be to apply a Bonferroni correction. That is, the comparison is statistically significant only if it is nominally significant at a .05/3=.017 level. Using that bound, in both studies the erythema rating and the investigators rating of improvement showed statistically significant differences in favor of Azelaic Acid (AzA) over its vehicle.

Appendix Table A.7: Investigator Rating of Improvement

		Study A03125				Study A03126			
		ITT		Per Protocol		ITT		Per Protocol	
Response		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
1. Complete Remission	n	18	5	13	4	11	13	8	6
	%	11.6%	3.1%	11.4%	3.2%	7.0%	8.3%	7.3%	5.0%
2. Marked Improvement	n	60	39	52	32	63	36	48	32
	%	38.7%	24.2%	45.6%	26.6%	39.9%	19.0%	44.0%	22.8%
3. Moderate Improvement	n	44	56	31	45	49	46	34	36
	%	28.4%	34.8%	27.2%	36.0%	31.0%	29.1%	31.2%	29.8%
4. No Improvement	n	29	53	17	41	24	45	14	37
	%	18.7%	32.9%	14.9%	32.8%	15.2%	28.5%	12.8%	30.6%
5. Deterioration	n	4	8	1	3	11	18	5	10
	%	2.6%	5.0%	0.9%	2.4%	7.0%	11.4%	4.6%	8.3%
All	n	155	161	114	125	159	158	110	121
p-value*		0.001		0.001		0.003		0.005	

*Significance level of CMH test of equality of proportions using modified ridit scores.

The sponsor's protocol also specified an analysis of nodule count.

Appendix Table A.8a: Study A03125 Distribution of Nodules

#	Visit 04		08		12		LOCF	
Nodules	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
0	151	153	130	138	132	138	157	147
1	8	4	10	5	2	10	3	10
2	1	1	1	7	0	2	0	4
3	1	2	0	0	0	1	1	1

Appendix Table A.8b: Study A03126 Distribution of Nodules

#	Visit 04		08		12		LOCF	
Nodules	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
0	150	141	142	139	138	134	148	146
1	4	11	7	10	7	6	7	7
2	3	3	1	3	1	2	2	2
3	1	1	2	0	0	2	0	2
4	0	1	0	0	0	1	0	0
>4	1	1	0	1	2	1	0	0

A natural model for such data is a zero-inflated Poisson model, i.e. a mixture of Poisson distributions: θ Poisson (μ_i) + (1-θ) Poisson (0), where the Poisson (0) distribution puts probability point mass at 0. Here $\log(\mu_i)$ is the linear predictor involving treatment differences:

$$\log(\mu_i) = b(1) + b(2)*\text{treatment},$$

where treatment = 0 for vehicle and treatment=1 for Azelaic acid.

Such models are naturally fit using an EM algorithm, but due to time constraints, it was decided to use a simple Bayesian analysis in BUGS. For convenience only the LOCF data were modelled and assessed for treatment differences, but it is apparent that results would be similar at each time point.

All analyses were conducted using WINBUGS 1.3. For technical reasons involving copying BUGS graphics, illustrative posterior density plots, autocorrelation plots, traces, etc. are not displayed. However, these were reviewed and seemed to show no problems with the parameters in the model. Only summary statistics from the probability distributions for the parameters were readily copied. Note that 2.5% and 97.5% denote the empirical quantiles of the posterior distribution, and define credibility intervals, as given below:

Study A03125:

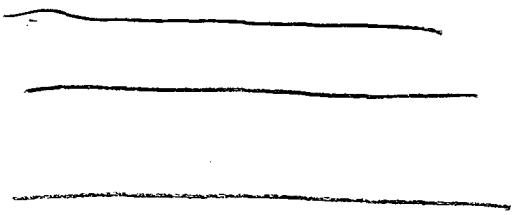
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b[1]	0.9155	0.2761	0.01038	0.4592	0.8827	1.522	10001	30000
b[2]	-0.4869	0.3124	0.009501	-1.098	-0.4884	0.1751	10001	30000
theta	0.14	0.04729	0.001341	0.0697	0.1326	0.2519	10001	30000

Study A03126:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b[1]	2.56	0.4249	0.01087	1.792	2.542	3.441	10001	30000
b[2]	-0.8483	0.5959	0.01499	-1.985	-0.8556	0.3379	10001	30000
theta	0.09248	0.01833	1.475E-4	0.06019	0.09133	0.1318	10001	30000

Treatment effect is reflected in b[2] above. Theta denotes the proportion in the latent class with the linear predictor (so the proportion with point mass at zero is 1-theta). In both studies the symmetric credibility interval for b[2] includes zero. That is, the distribution of our knowledge about the parameter is not bounded away from zero. So in this model there is no particular evidence of a difference between Azelaic acid and vehicle in terms of nodule counts.

The following coded BUGS program was used:



Appendix 6.0: Other Secondary Endpoints

Note that in both studies there statistically significant differences in the patient rating of improvement.

Appendix Table A.7: Patient Rating of Improvement

		Study A03125				Study A03126			
		ITT		Per Protocol		ITT		Per Protocol	
Response		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
1. Excellent Improvement	N	50	20	41	14	47	24	37	16
	%	32.3%	12.4%	36.0%	11.2%	29.6%	15.2%	33.6%	13.2%
2. Good Improvement	N	44	49	38	42	45	30	33	25
	%	28.4%	30.4%	33.3%	33.6%	28.3%	19.0%	30.0%	20.7%
3. Moderate Improvement	N	32	41	21	34	38	53	25	45
	%	20.7%	25.5%	18.4%	27.2%	23.9%	33.5%	22.7%	37.2%
4. No Improvement	N	24	42	13	32	21	39	11	30
	%	15.5%	26.1%	11.4%	25.6%	13.2%	24.7%	10.0%	24.8%
5. Aggravation	N	5	9	1	3	8	12	4	5
	%	3.3%	5.6%	8.8%	2.4%	5.0%	7.6%	3.6%	4.1%
All	N	155	161	114	125	159	158	110	121
p-value*		0.001		0.001		0.003		0.001	

*Significance level of CMH test of equality of proportions using modified ridit scores.

Appendix Table A.9: Cosmetic Acceptance by Patient

		Study A03125				Study A03125			
		ITT		Per Protocol		ITT		Per Protocol	
Response		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
1. Very Good	N	60	51	51	40	52	52	37	41
	%	38.7%	31.7%	44.7%	32.0%	32.7%	32.9%	33.6%	33.9%
2. Good	N	46	61	37	49	59	43	47	33
	%	29.7%	37.9%	32.5%	39.2%	37.1%	27.2%	42.7%	27.3%
3. Satisfactory	N	29	37	19	28	28	38	19	33
	%	18.7%	23.0%	16.7%	22.4%	17.6%	24.1%	17.3%	27.3%
4. Poor	N	15	9	5	6	15	19	3	10
	%	9.6%	5.6%	4.4%	4.8%	9.4%	12.0%	2.7%	8.3%
5. No Opinion	N	5	3	2	2	5	6	4	4
	%	3.2%	1.8%	1.8%	1.6%	3.1%	3.8%	3.6%	3.3%
All	N	155	161	114	125	159	158	110	121
p-value*		0.130		0.134		0.186		0.018	

*Significance level of CMH test of equality of proportions deleting the "No Opinion" Group, using modified ridit scores.

Appendix 7.0: Detailed Subgroup Analyses of Investigator Global Assessment

The following table displays results of the IGA for demographic subgroups. For convenience data were pooled across studies. Again, a "success" was defined as a subject who if they were mild at baseline (i.e. a score of 2) achieved a clear (i.e. score of 0) at the point of measurement or achieved a score of clear or minimal (i.e., 0 or 1) at the end of the study with a baseline score of 3-6. Other endpoints were described in section 3.4.1 of the text. Recall that the ITT and Per Protocol results are measured at week 12, where the ITT population uses LOCF to impute missing values.

Appendix Table A.10: Pooled Studies Investigator's Global Evaluation by Demographic Subgroup

Gender		Examination Visit: Week 04 AzA	04 Veh	08 AzA	08 Veh	12 AzA	12 Veh	Per Prot. AzA	Per Prot. Veh	ITT AzA	ITT Veh
Female											
Success	N	25	10	56	29	72	41	58	36	74	42
	%	10.6	4.4	25.8	13.4	34.6	19.5	34.3	20.2	30.5	17.5
0 (Clear)	N	3	0	3	6	13	10	10	8	13	10
	%	1.3	0.0	1.4	2.8	6.3	4.8	5.9	4.5	5.3	4.2
0,1(≤Minimal)	N	31	19	67	37	88	57	71	49	92	60
	%	13.2	8.3	30.9	17.1	42.3	27.1	42.0	27.5	37.9	25.0
0-2(≤ Mild)	N	107	82	134	86	143	105	115	90	154	114
	%	45.5	36.0	61.8	39.6	68.8	50.0	68.0	50.6	63.4	47.5
3-6	N	128	146	83	131	65	105	54	88	89	126
	%	54.5	64.0	38.2	60.4	31.3	50.0	32.0	49.4	36.6	52.5
All	N	235	228	217	217	208	210	169	178	243	240
Male											
Success	N	8	6	17	10	28	14	21	6	29	14
	%	9.4	6.7	22.4	11.6	37.8	16.1	38.2	8.8	33.0	15.4
0 (Clear)	N	2	0	3	1	10	5	8	1	10	5
	%	2.4	0.0	3.9	1.2	13.5	5.7	14.5	1.5	11.4	5.5
0,1(≤Minimal)	N	11	10	20	15	31	20	23	9	32	20
	%	12.9	11.1	26.3	17.4	41.9	23.0	41.8	13.2	36.4	22.0
0-2(≤ Mild)	N	36	25	42	32	47	32	35	20	48	32
	%	42.4	27.8	55.3	37.2	63.5	36.8	63.6	29.4	54.5	35.2
3-6	N	49	65	34	54	27	55	20	48	40	59
	%	57.6	72.2	44.7	62.8	36.5	63.2	36.4	70.6	45.5	64.8
All	N	85	90	76	86	74	87	55	68	88	91

Observe that for both genders the superiority of Azelaic Acid over its vehicle in terms of the IGA scores is roughly the same.

Race		Week 04 AzA	04 Veh	08 AzA	08 Veh	12 AzA	12 Veh	Per Prot. AzA	Per Prot. Veh	ITT AzA	ITT Veh
Caucasian											
Success	N	30	15	65	35	93	48	74	37	95	49
	%	10.2	5.1	24.1	12.4	35.8	17.3	35.6	16.1	31.3	15.9
0 (Clear)	N	4	0	5	6	21	11	17	7	21	11
	%	1.4	0.0	1.9	2.1	8.1	4.0	8.2	3.0	6.9	3.6
0,1(≤Minimal)	N	38	28	78	48	112	70	89	53	116	73
	%	12.9	9.4	28.9	17.0	43.1	25.3	42.8	23.0	38.2	23.7

0-2(≤ Mild)	N	133	99	163	108	178	123	140	99	189	131
	%	45.1	33.3	60.4	38.2	68.5	44.4	67.3	43.0	62.2	42.5
3-6	N	162	198	107	175	82	154	68	131	115	177
	%	54.9	66.7	39.6	61.8	31.5	55.6	32.7	57.0	37.8	57.5
All	N	295	297	270	283	260	277	208	230	304	308
Other											
Success	N	3	1	8	4	7	7	5	5	8	7
	%	12.0	4.8	34.8	20.0	31.8	35.0	31.3	31.3	29.6	30.4
0 (Clear)	N	1	0	1	1	2	4	1	2	2	4
	%	4.0	0.0	4.3	5.0	9.1	20.0	6.3	12.5	7.4	17.4
0,1(≤Minimal)	N	4	1	9	4	7	7	5	5	8	7
	%	16.0	4.8	39.1	20.0	31.8	35.0	31.3	31.3	29.6	30.4
0-2(≤ Mild)	N	10	8	13	10	12	14	10	11	13	15
	%	40.0	38.1	56.5	50.0	54.5	70.0	62.5	68.8	48.1	65.2
3-6	N	15	13	10	10	10	6	6	5	14	8
	%	60.0	61.9	43.5	50.0	45.5	30.0	37.5	31.3	51.9	34.8
All	N	25	21	23	20	22	20	16	16	27	23

As noted in the report, unlike for gender, there is apparent evidence that Azelaic acid 15% gel may be less effective for non-Caucasian patients than among Caucasian patients. However, again, the number of patients is so small that is may well be an artifact of the experiment.

Age Group	Week	04		08		12		Per Prot.		ITT	
		AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh	AzA	Veh
Age<65											
Success	N	28	14	60	36	86	51	69	38	88	51
	%	10.0	5.0	23.3	13.4	34.7	19.2	35.2	17.4	30.1	17.4
0 (Clear)	N	3	0	5	7	19	13	15	7	19	13
	%	1.1	0.0	1.9	2.6	7.7	4.9	7.7	3.2	6.5	4.4
0,1(≤Minimal)	N	36	23	74	46	103	70	82	53	106	71
	%	12.8	8.2	28.7	17.1	41.5	26.4	41.8	24.3	36.3	24.2
0-2(≤ Mild)	N	126	93	153	106	167	124	132	99	177	131
	%	44.8	33.1	59.3	39.4	67.3	46.8	67.3	45.4	60.6	44.7
3-6	N	155	188	105	163	81	141	64	119	115	162
	%	55.2	66.9	40.7	60.6	32.7	53.2	32.7	54.6	39.4	55.3
All	N	281	281	258	269	248	265	196	218	292	293
Age 65+											
Success	N	5	2	13	3	14	4	10	4	15	5
	%	12.8	5.4	37.1	8.8	41.2	12.5	35.7	14.3	38.5	13.2
0 (Clear)	N	2	0	1	0	4	2	3	2	4	2
	%	5.1	0.0	2.9	0.0	11.8	6.3	10.7	7.1	10.3	5.3
0,1(≤Minimal)	N	6	6	13	6	16	7	12	5	18	9
	%	15.4	16.2	37.1	17.6	47.1	21.9	42.9	17.9	46.2	23.7
0-2(≤ Mild)	N	17	14	23	12	23	13	18	11	25	15
	%	43.6	37.8	65.7	35.3	67.6	40.6	64.3	39.3	64.1	39.5
3-6	N	22	23	12	22	11	19	10	17	14	23
	%	56.4	62.2	34.3	64.7	32.4	59.4	35.7	60.7	35.9	60.5
All	N	39	37	35	34	34	32	28	28	39	38

So again, for all endpoints, for both age groups the superiority of Azelaic Acid 15% gel over its vehicle in terms of the IGA scores is fairly consistent.

Appendix 8.0: Mixed Model Repeated Measures Analysis of Lesion Counts

The following summarize the results of mixed model/repeated measures analyses of the profile of the three post baseline lesion count endpoints. No subjects are deleted and no restriction is placed on the covariance matrix of the three response times. For study A03125 both center and treatment by center interaction are treated as random. This model results in inadmissable variance estimates in study A03126. Hence only the center term is included in the analysis of that study. These analyses only need assume that dropouts are random to remain valid.

A.6.1 Study A03125 Multivariate Analyses

Change from Baseline

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
Center		3.35
Treat*Center		1.91
UN(1,1)	RANDID	44.26
UN(2,1)	RANDID	21.58
UN(2,2)	RANDID	40.42
UN(3,1)	RANDID	25.75
UN(3,2)	RANDID	28.68
UN(3,3)	RANDID	47.26

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
Base Inflam	1	219	256.15	0.0001
Treatment	1	13	15.17	0.0017
Visit	2	287	43.75	0.0001
Treat*Visit	2	287	1.09	0.3367

Least Squares Means

Effect	Treatment	LSMEAN	Std Error	DF	t	Pr > |t|
Treat	AzA	10.01	0.79	11.7	12.66	0.0001
Treat	Veh	6.70	0.79	11.3	8.49	0.0001

Percent Change from Baseline

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
Center		0.011
Treat*Center		0.003
UN(1,1)	RANDID	0.132
UN(2,1)	RANDID	0.062
UN(2,2)	RANDID	0.131
UN(3,1)	RANDID	0.066
UN(3,2)	RANDID	0.073
UN(3,3)	RANDID	0.131

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
Base Inflam	1	220	6.48	0.0116
Treatment	1	13	21.90	0.0004
Visit	2	286	46.62	0.0001
Treat*Visit	2	286	0.65	0.5219

Least Squares Means

Effect	Treatment	LSMEAN	Std Error	DF	t	Pr > |t|
Treat	AzA	0.54	0.041	10.7	13.18	0.0001
Treat	Veh	0.35	0.041	10.3	8.58	0.0001

For both endpoints treatment differences are statistically significant in favor of Azelaic Acid 15% gel. ($p \leq 0.0017$ and $p \leq 0.0004$, respectively).

A.6.2 Study A03126 Multivariate Analyses

Change from Baseline

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
Center		1.40
UN(1,1)	RANDID	88.31
UN(2,1)	RANDID	55.42
UN(2,2)	RANDID	84.88
UN(3,1)	RANDID	52.28
UN(3,2)	RANDID	62.81
UN(3,3)	RANDID	88.60

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
Base Inflam	1	270	63.75	0.0001
Treatment	1	281	4.61	0.0326
Visit	2	288	20.49	0.0001
Treat*Visit	2	288	3.59	0.0288

Least Squares Means

Effect	Treatment	LSMEAN	Std Error	DF	t	Pr > |t|
Treat	AzA	7.99	0.74	39.3	10.86	0.0001
Treat	Veh	6.00	0.73	39.9	8.20	0.0001

Percent Change from Baseline

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
Center		0.0003
UN(1,1)	RANDID	0.2264
UN(2,1)	RANDID	0.1259
UN(2,2)	RANDID	0.2378
UN(3,1)	RANDID	0.1190
UN(3,2)	RANDID	0.1372
UN(3,3)	RANDID	0.2078

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
Base Inflam	1	299	0.16	0.6900
Treatment	1	302	5.11	0.0246
Visit	2	300	26.59	0.0001
Treat*Visit	2	300	1.55	0.2147

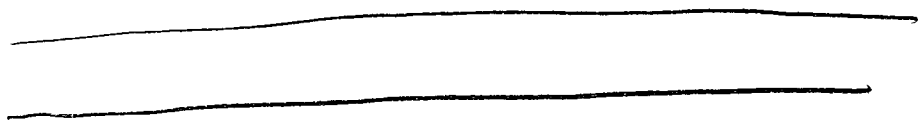
Least Squares Means

Effect	Treatment	LSMEAN	Std Error	DF	t	Pr > |t|
Treat	AzA	0.44	0.03	56.8	13.60	0.0001
Treat	Veh	0.34	0.03	59.6	10.48	0.0001

And, again, for both endpoints treatment differences are statistically significant in favor of Azelaic Acid. ($p \leq 0.0326$ and $p \leq 0.0246$, respectively).

A.6.3 SAS Program

The following program for the change from baseline in Study A03125 was typical of the SAS programs used.



For both endpoints the variance component due to center by treatment interaction (i.e., invest*rxgrp) was relatively small. In the A03126 study for both endpoints the estimate of this component was not feasible, so, as was noted above, to maintain positive definiteness of the estimate of the covariance matrix, the component was deleted from the model (i.e. RANDOM invest; was specified in the A03126 analyses.)

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

/s/

Steven Thomson
11/18/02 02:30:09 PM
BIOMETRICS

Mohamed Alosh
11/19/02 09:29:38 AM
BIOMETRICS

I concur with the primary reviewer's conclusion. A related discussion to the reviewer's nested hypothses approach for the Intent-to-Treat, population, Per-Protocol population and completers is given in my secondary review dated 11/19/02.

Secondary Statistical Review
Addendum to Primary Statistical Review for NDA 21-470 (Finevin),
Berlex Laboratories, Inc., signed on 11/19/02)

The primary review carried out by Mr. Thomson is a comprehensive review as it included several analyses based on frequentists and Bayesian approaches to check robustness of statistical findings under different statistical methodologies. Furthermore, he carried out efficacy evaluation over time as an exploratory analysis, as this was not specified in the protocol, to provide information, presumably helpful to clinician, about how early significant difference can be detected.

Mr. Thomson, however, went further to carry out statistical testing as a nested hypothesis on various study populations (Intent to treat, Per protocol and completers) and combine this with testing over time. Also, he carried out similar analysis for different success criteria according to the Investigator Global Evaluation. As Mr. Thomson pointed out such analyses were not pre-specified in the protocol, yet he sees utility for them. He formulated these as nested hypotheses as, he pointed out in his review, to control Type I error rates (see p.14 and p.15 of the primary review). Specifically, for nesting across patient populations, he will start testing for the ITT at Week 12, if this is significant then will test for the per protocol at Week 12, and if the last is significant he will test for completers at week 12. Then, if the last is significant he will test for completers at week 8 and if this is significant he will test for the completers at week 4.

While the proposed testing approach would not have impact on the efficacy results as all testing are done at the 0.05 level, however, personally I do not see the utility of such testing because:

(a) Multiple testing is carried out normally for an additional claim, which could be related to an additional endpoint or an early efficacy (time point) where normally there is cost for that in term of adjustment. In contrast, testing for efficacy results across patient populations is intended to check the impact of data imputation for missing observations on the efficacy findings. Usually, the primary analysis is carried out on the ITT population and for checking robustness of findings with respect to handling dropout we check results for per-protocol population to examine consistency. This in my view might explain why the nested hypothesis proposed here for different patient populations was not to my knowledge addressed in any other statistical review in the agency.

(b) The claim that such testing approach controls Type I error rates is at best debatable, since it is carried out after seeing the efficacy results which might impact the way the hypothesis are nested (ordered). Furthermore, selection of the sequence of nested hypothesis (ordering) could be debatable as well. For example, with respect to the Investigator Global Evaluation the reviewer consider 3 nested hypotheses, one based on the Medical Officer (MO) definition of treatment success, and if this is significant consider testing with success dichotomization at 0 and 1 categories, and if this is significant consider success dichotomization based on the 0, 1 and 2 categories (see page 14 of the primary review). While, it is reasonable to take the MO definition as the primary hypothesis, choice of the proposed sequence for the next two hypotheses is debatable. In fact, one might argue to start with the larger set (success defined on categories 0,1,2) then go to the smaller set (success defined on categories 0,1).

(c) If there are no added benefits from the nested approach, I would prefer simplicity of statistical analyses instead of formalism of hypotheses, which the utility of this formalism is at best debatable. The fact that one does not loose anything in terms of the significance level does not in my opinion justify unnecessary formalism.

It should be noted however that the above minor comments do not reduce the depth of the statistical analyses in the primary review.

Mohamed Alosh, Ph.D.
Biostatistics Team Leader.

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

/s/

Mohamed Alosh
11/19/02 10:05:52 AM
BIOMETRICS