CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

21-812

STATISTICAL REVIEW(S)

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

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/Serial Number:	21-812 / 000
Drug Name:	Men's Rogaine® Extra Strength Minoxidil 5% Topical Foam
Indication(s):	Androgenetic Alopecia
Applicant:	Pfizer Consumer Healthcare
Date(s):	Received 3/24/2005, user fee (10 months) 1/23/2006
Review Priority:	Standard
Biometrics Division:	Division 3, HFD-725
Statistical Reviewer:	Steve Thomson, HFD-725
Concurring Reviewers:	Team Leader: Mohamed Alosh, Ph. D., HFD-725
Medical Division:	Division of Over-the-Counter Drug Evaluation, HFD-560 Dermatology and Dental Products, HFD-540
Clinical Team:	Efficacy Reviewer: P. Huene, M.D.,HFD-540 Safety Reviewer: D. Shetty, M.D., HFD-560 Team Leader: M. Luke, M.D., Ph.D.,HFD-540
Project Manager:	T. Frazier, HFD-560
Keywords:	Analysis of covariance, Cochran-Mantel-Haenszel

Table of Contents

LIST OF TABLES:

Note Sponsor tables in Appendix 6 are not listed.

LIST OF FIGURES

1. EXECUTIVE SUMMARY

In 1997, Rogaine Extra Strength (5% minoxidil) solution was approved for over the counter use for baldness in males. The current submission is for a 5% foam formulation. The Sponsor submitted three pharmacokinetic or sensitization studies and one Phase 3, Efficacy Study in males to support their claims. Only the latter study was reviewed here.

1.1 Conclusions and Recommendations

The two primary endpoints in the study were 1) the change from baseline in hair count in a small area on the vertex of scalp described above, and 2) a subject rating of overall treatment benefit scored on a seven point ordinal scale. For both endpoints differences in favor of the 5% Foam were very highly statistically significant (both $p < 0.0001$). This was true both in the intent to treat (ITT) and the per protocol (PP) populations. However, despite the fact these are both measures of hair growth, as discussed in Appendix 4, these measures have a surprisingly low association.

From a statistical perspective, there are two primary issues with this submission. The first issue with this submission is whether or not one of the primary endpoint, change from baseline in magnified hair counts in a 1 cm^2 target area at the vertex of the scalp at nominal Week 16, is an appropriate measure. In particular, under magnification, the hair counts may be biased by the inclusion of small, so-called "vellus", hairs. It was felt that such "peach fuzz" hairs would not appreciably contribute to the appearance of the hair and should not be included in hair counts. The Sponsor provided an argument that their technology is consistent with previous submissions and only counts a relatively small number of such hairs. This argument is summarized in Section 2.1.3 below. A sensitivity analysis to the proportion of vellus hairs is given in Appendix 1. In addition the endpoints are assessed at Week 16, which may be too early to assess any long term effect of treatment. However, at the End of Phase 2 (EOP 2) meeting, held on January 16, 2003, the Division accepted this study length but did note it might be reflected in labeling.

The second issue with the submission is whether or not a single study is adequate to demonstrate efficacy. At the EOP 2 meeting the Sponsor was informed that "a single . . . study in which the new product is compared to its vehicle might be acceptable in the study of androgenic alopecia on the vertex." At that meeting, the Sponsor was also reminded, for a single study, of the need to achieve very small significance levels for tests of the primary endpoints, plus consistency across centers and study subgroups. As discussed in Section 1.3 below, one could argue that a 0.0006 level of significance for any particular endpoint could be used to define a conclusively small p-value. These issues, and the analysis of study subgroups, were addressed below.

Grouping ITT patients into two roughly equally sized groups of those aged less than 40 and those 40-49, for both primary endpoints treatment differences in each group were still highly

statistically significant in favor of minoxidil foam (all $p \leq 0.0001$). However, at the EOP 2 meeting the Division recommended that, since the condition was common in men aged over 49, "there be no upper age limit." One might argue that the Sponsor's restriction to subjects aged 49 or less implies an insufficient analysis in all relevant demographic subgroups.

Most patients were Caucasian, and among this group treatment differences were still highly statistically significant in favor of minoxidil (for both change from baseline in hair count and the subject rating $p \leq 0.0001$). Although actual success proportions among the non-Caucasian patients were similar to those among the Caucasian, there were only a relatively small number of patients in this subgroup (44) and treatment differences were barely statistically significant for change from baseline in hair counts ($p \leq 0.0475$) and non-significant for the subject rating ($p \leq 0.2109$). For the Norwood Baseline Hair Loss Patterns IIIv, IV, and V treatment differences in the change from baseline were all statistically significant ($p \leq 0.0001$, $p \leq 0.0184$, and $p \leq 0.0020$, respectively). Results for the subject rating over the hair loss patterns were also all statistically significant ($p \leq 0.0001$, $p \leq 0.0026$, and $p \leq 0.0006$, respectively).

An expert panel review was a secondary endpoint, and had overall similar results although the consistency between raters was generally low.

1.2 Brief Overview of Clinical Studies

The Sponsor submitted a single Phase 3, double-blind, randomized trial in 352 subjects comparing 5% minoxidil foam with vehicle. Male subjects with androgenetic alopecia with vertex pattern IIIv, IV, and V on the Norwood-Hamilton Scale and no known sensitivity to minoxidil were to be randomized to treatment in 14 centers. The primary efficacy endpoints were the Week 16/early termination mean change from Baseline in visualized hair counts in a $1 cm^2$ target region at the vertex of the scalp and a subject rating of overall treatment benefit, ranked on a score from -3 to 3. The secondary efficacy variable was the median score of an expert panel review of hair regrowth. The Division was concerned that vellus hairs (i.e. "peach fuzz") may have been included in the total hair counts, and thus in the change from baseline.

1.3 Statistical Issues and Findings

Statistical Issues

1. At the End of Phase 2 (EOP 2) meeting on January 16, 2003, the Division encouraged the Sponsor to either conduct two placebo controlled Phase 3 trials, or to conduct a 3-arm Phase 3 trial including an arm with the currently labeled ointment. However, the Agency also stated that "a single . . . study in which the new product is compared to its vehicle might be acceptable in the study of androgenic alopecia on the vertex." With a single study, either 2-arm or 3-arm, the Sponsor was told they need very small significance levels for tests of the primary endpoints, plus consistency across centers and study subgroups. The Sponsor has submitted a single placebo-controlled, 2-arm study of androgenetic alopecia on the vertex. The adequacy of this study and the choice of endpoints seem to be the key points in determining that the results in this trial are

conclusive. The discussion with the Sponsor is summarized below in the history section 2.1.3 of this report and in more detail in the Sponsor's responses included in Appendix 6.

2. At the EOP 2 meeting and later, the Division also requested that only non-vellus hairs be included in the hair count total. In response to an information request (received September 2, 2005), the Sponsor stated that "Pfizer believes that the magnification level of 5.7 fold yields visualization and thereby counting of non-vellus hairs (i.e. those ≥ 0.03 mm in diameter) and adequately filters out vellus and insignificant miniaturized non-vellus hairs (i.e. those < 0.03 mm in diameter). This argument is also summarized in the history section 2.1.3 of this report.

3. Despite some problems related to actual interpretation, in many circumstances p-values can be useful as rough measures of the weight of evidence. To show superiority the Agency generally recommends two-sided statistical tests with a significance level of 0.05 or smaller. But since the drug would not be rejected if it was considerably superior to its comparator, in effect, as evidence of efficacy, the Agency requests one-sided tests with statistical significance levels at or below 0.025 in two studies. Then with independent studies the probability of a type I error in both studies is $(0.025)^2 = 0.000625$. To achieve an equivalent level of significance with a single study, we could use a 0.000625 level. Note that the Sponsor did generally achieve this level of significance in tests on the primary endpoints, so we would conclude that they achieved the "small" significance levels.

4. The protocol specifies that both the change from baseline in hair count and the subject hair loss condition rating are to be analyzed with an analysis of variance (ANOVA) model with factors for treatment and center and covariates age and duration of hair loss. The protocol also specified that if the residuals are not normal a Wilcoxon test is to be used to compare median scores of the treatment group. However, the actual test used by the Sponsor was a Wilcoxon test stratified on center, i.e., a Van Elteren test. Even with skewed data as here, cell means are roughly normally distributed and ANOVA would still be appropriate. However, since the hair loss condition rating is ordinal, one might, in principle, prefer a Cochran-Mantel-Haenszel (CMH) test of treatment differences. Using modified ridit scores in the CMH test leads to the Van Elteren test. Results from both statistical tests are reported here, and are always essentially equivalent.

5. Several centers only recruited a small number of patients into the study. Pooling of subjects for the analysis was specified in the analysis plan, but not in the Study protocol. Since the dating of the analysis plan was not apparent, this pooling may be post hoc. However, this pooling (see 2.2 below) was deemed to be acceptable, and for convenience was followed in the Agency analysis. Note that endpoints, methods of analysis, etc. were specified in the protocol.

6. The Review team wanted to investigate the consistency of ratings across the three raters for the expert panel review of hair growth. In response, the Sponsor provided kappa statistics, including a pooled kappa across the three raters. One possible problem with using kappa statistics to assess consistency is that one can argue that kappa statistics actually measure lack of independence in the ratings, not consistency. But since raters evaluate the same subjects we

expect some lack of independence in the ratings. This issue is addressed in more detail in Appendix 5.

Statistical Findings

For both mean change from Baseline in visualized hair counts in the target region and the subject rating of treatment benefit, treatment differences in favor the 5% Foam were highly statistically significant ($p < 0.0001$). This was true both in the intent to treat (ITT) population using ANOVA, van Elteren, or Cochran-Mantel Haenszel tests, and in the per protocol population. However, as detailed in Appendix 4, the association between the primary endpoints was generally low. Grouping ITT patients into two roughly equally sized groups of those aged less than 40 and those 40-49, for both primary endpoints treatment differences in each group were still highly statistically significant in favor of minoxidil foam (all $p \leq 0.0001$). Most patients were Caucasian, and among them all differences were still highly statistically significant in favor of Minoxidil (both $p \leq 0.0001$). Although actual success proportions among the non-Caucasian patients were similar to those among the Caucasian, there were only a relatively small number of patients in this subgroup (44) and treatment differences were only barely statistically significant for change from baseline in hair counts ($p \leq 0.0475$) and non-significant for the subject rating ($p \leq 0.2109$). For the Norwood Baseline Hair Loss Patterns IIIv, IV and V treatment differences in the change from baseline were all statistically significant ($p \leq 0.0001$, $p \leq 0.0184$, and $p \leq 0.0020$, respectively). Results for the subject rating were similar ($p \leq 0.0001$, $p \leq 0.0110$, and $p \leq 0.0002$, respectively). An expert panel review found consistent results.

2. INTRODUCTION

2.1 Overview

The Sponsor submitted four studies in support off this application. Studies MINOB-9410-001 and MINOB-9410-005 were pharmacokinetic studies in healthy subjects. Study MINOB-9410-004 was a sensitization study in healthy subjects. These studies are not addressed further in this review. Finally, Study MINOB-9410-006 was a Phase 3 study titled:

> A Double-Blind, Randomized, Placebo-Controlled Trial of the Efficacy and Safety of 5% Minoxidil Foam in the Treatment of Androgenetic Alopecia in Males.

This was a 1:1 randomized trial comparing 5% Minoxidil foam to vehicle in 352 male subjects with androgenetic alopecia. Fourteen centers across the U.S. were used. This single Phase 3 trial is the target of this review.

2.1.1 Design

The protocol for the proposed Phase 3 study was submitted to the Agency for comments on September 24, 2003, but the study was initiated shortly thereafter on October 8, 2003. The

Sponsor reports that the trial was completed on July 29, 2004. Please see 2.1.2, below, for details on the regulatory history.

The Sponsor's protocol specified that male subjects, aged 49 or less, with androgenic alopecia with vertex pattern IIIv, IV, and V on the Norwood-Hamilton Scale and having no known sensitivity to minoxidil were to be randomized 1:1 to receive either 5% minoxidil foam or vehicle BID for 16 weeks. Efficacy endpoints were based on assessments of the (1) mean change in visualized hair count in the target region; (2) subject's assessment of treatment benefit and (3) panel review of treatment benefit by three experts.

Polaroid photographs for use with the subject self-assessment of treatment benefit were taken at baseline and Week 16. At the screening visit and at visits on weeks 8, 12 and 16, 35mm "global" photographs were also taken (also as needed at baseline and week 1). These were used for global assessments by a panel of three experts.

The primary efficacy endpoints for the analysis were defined as follows:
"a. Mean change in visualized hair count in the target region between Baseline and Week 16 as determined by a validated computer assisted dot mapping technique.
b. Subject rating of treatment benefit via use of global photographs of the vertex region assessed as an overall change from baseline, collected on subject questionnaire." (Protocol, page 21)

Secondary Efficacy Variable:
"c. Expert panel review of hair regrowth when comparing global photographs of Baseline to Week 16." (Protocol, page 21)

In the event of early termination an attempt was made to assess the Week 16 endpoints.

2.1.2 Regulatory History:

1. End of Phase 2 Meeting (January 16, 2003):
According to the FDA minutes (faxed to the Sponsor on March 24, 2003), the Sponsor at that time, Pharmacia Consumer Healthcare, was informed that: "A) The recommended co-primary efficacy endpoints are nonvellus hair counts and subject assessment of treatment benefit. B) Generally, line extensions are based on either two clinical trials in which the new product is compared to vehicle, or a single, three-armed study in which the new product is compared to the currently-marketed product and the new vehicle. However, a single, adequate and well-controlled study in which the new product is compared to its vehicle might be acceptable in the study of androgenic alopecia on the vertex. ... C) A trial of 16 weeks duration is acceptable in the study of androgenetic alopecia on the vertex, if the Sponsor were to agree to include in the label a discussion of the diminution of treatment effect seen in the clinical trials with their currently-marketed 5% MTS."

2. Clinical comments faxed to the Sponsor, (March 19, 2004) included the following (among other comments):
a. The Sponsor, Pharmacia & Upjohn, was reminded that "vellus hairs which are visible in photographs should not be included in hair counts. Increased numbers of vellus hairs may not result in a clinically meaningful improvement for this cosmetic indication."
b. "As indicated at the end-of-phase 2 meeting, to evaluate the relative efficacy of this product compared to 5% minoxidil topical solution, the Sponsor should consider a three arm safety and efficacy trial comparing the new formulation to the already approved 5% minoxidil topical solution and placebo in men. If such a three-armed study is not conducted, it is strongly recommended that two clinical trials be performed.
c. "The inclusion criteria for age need to be revised. The Sponsor has not justified the exclusion of males over age 49 with androgenetic alopecia who otherwise meet eligibility criteria. This indication is common in men over age 49. It is recommended that there be no upper age limit."
d. "Polaroid photographs used to assist subjects in rating treatment benefit may introduce bias based on photographic variation and does not reflect actual use conditions. It is recommended that mirrors be used to assist subjects in rating treatment benefit.
e. "The study appears too short to provide sufficient safety information of this new topical formulation if the product is intended for long-term use. The Sponsor is encouraged to follow the ICH E1A document guidelines."

3. Teleconference between the Division and the Sponsor (July 26, 2004) in response to the Sponsor's submission (IND 50,063/SN 23 dated July 16, 2004), FDA minutes faxed September 23, 2004:
a. First, the Division Director noted that the protocol for the proposed study was submitted to the Agency for comments on September 24, 2003, and that the study was initiated shortly thereafter in October of 2003. Thus the Sponsor was not seeking the Division's comments prior to study initiation.

b. "In response, the Sponsor indicated that they thought that all aspects of the study were acceptable to the Division and consequently they proceeded with the study shortly after the submission of the protocol to the Division. The Division reminded the Sponsor that unresolved issues related to primary endpoints of the study were raised at the End of Phase 2 meeting but were not resolved prior to protocol submission and study initiation."

c. "The Division's comments strongly encouraged a 3-arm Phase 3 trial in order to ensure robustness of study findings and consequently study interpretability, as an alternative approach to replication of study findings based on 2 trials. It should be noted however, that a single study submission for this indication would be acceptable to provide such assurances if results from this single study had the following properties:
(i) Very small p-values, as one would be looking for a much smaller p-value than the 0.05 required for 2 studies.
(ii) Consistency in efficacy results across study subgroups
(iii) Consistency of efficacy across study centers, as efficacy results driven by a few centers would not provide assurance of robustness in a single multi-center trial."

4. To summarize, in one or more communications, the Division had recommended that the Sponsor (1) ensure that only non-vellus hair counts be included in the count total, (2) include an arm comparing the minoxidil foam to the 5% minoxidil solution, and (3) include patients older than 49 years. In particular, in internal meetings the Division expressed concern that the magnification used would tend to cause the inclusion of vellus hairs in count totals.

5. According to the FDA reviews of NDA 20-834, Rogaine Extra Strength (5% minoxidil) topical solution was statistically significantly better than its vehicle it terms of mean change from baseline in nonvellus hair counts assessed in a 1 cm^2 target area at weeks 16 (mean 36 versus mean 4). Results at Week 32 were similar. The reviews do not indicate whether or not the technology used for the hair counts involved magnification. Moreover, the label for the minoxidil solution in the Physician's Desk Reference (PDR) does not include any description of the clinical studies. However the corresponding label for Finasteride does state that "Hair counts were assessed by photographic enlargements of a representative area of active hair loss." In the paper: Kaufman, K. D. et al, (1998): "Finasteride in the treatment of men with andorgenetic alopecia," *J. Am. Acad Dermatol* **39**: 578-89, available on the internet, the authors note the finasteride studies used a final magnification of 5.84:1.

6. On August 25, 2005, the Agency requested a teleconference, partly to discuss the whether or not vellus hairs were included in the hair counts. At that teleconference, Pfizer presented their position that both the dot mapping technique used for counting hair and the 5.7:1 photographic magnification used for visualized hair counts were valid and were the same as hair count methods used in prior submissions. The Sponsor's complete responses to information requests on the hair measurement are included Appendix 6 below.

To summarize their responses, the Sponsor stated that hair diameters were not measured in this study. However, the Sponsor noted that 0.03 mm is accepted as the upper limit diameter for vellus hairs, and provided references supporting this claim. The Sponsor claimed that over

the past twenty years, hair counts using photographs magnified to 5.7 power were sufficient to exclude the visualization, and thereby counting, of vellus hairs. Further, in their complete response the Sponsor stated that " [redacted] . has confirmed that a final magnification of 5.7 fold was used for both Pfizer's 006 and . . . Finasteride studies." (please see Appendix 6) The Sponsor did agree that without a diameter measurement of each individual hair, some vellus and miniaturized non-vellus hairs might be counted, and stated that technology for assessing hair diameter was being investigated for future use. Further, the Sponsor summarized the results of a small study that "compared the number of non-vellus hairs ($\geq$ 0.03 mm)-counted with the magnification technique to the actual number of hairs with a diameter of $\geq$ 0.03 mm as measured by the newer technology using digital techniques and image analysis. The results of this study showed the mean target area hair count using the 5.7 fold magnification technique was 169.1 and the number of hairs in the same target area with a measured diameter of $\geq$ 0.03 mm was 166.6. Results of this study lend support to 5.7 fold as the magnification level which yields visualization and counting of non-vellus hairs in the photographic magnification technique." (poster presented as a poster at the European Hair Research Society meeting July, 2005.) Note the Sponsor's communication did not indicate the sample size used in this study.

2.2 Data Sources

Data for the pivotal study was downloaded from the FDA Electronic Data Room as SAS transport files, located in the following link:

\\CDSESUB1\N21812\N_000\2005-03-23

3. STATISTICAL EVALUATION

3.1 Evaluation of Efficacy

This is based on the data from a single study:
MINOB-9410-006:
A Double-Blind, Randomized, Placebo-Controlled Trial of the Efficacy and Safety of 5% Minoxidil Foam in the Treatment of Androgenetic Alopecia in Males

3.1.1 Study Design and Endpoints

This was a single trial comparing 5% minoxidil foam to vehicle in 352 male subjects with androgenetic alopecia having vertex pattern IIIv, IV, and V on the Norwood-Hamilton Scale with no known sensitivity to minoxidil. Subjects were randomized 1:1 to treatment in 14 centers. Treatment was applied for 16 weeks.

Efficacy in the treatment of androgenetic alopecia was compared between treatments based on (1) mean change in visualized hair count in a 1 cm^2 target region on the scalp (2) a subject's assessment of treatment benefit and (3) a panel review by three experts of treatment

benefit. The target region was within a circle, approximately 1.9 cm in diameter, positioned at the vertex portion of the scalp, on the leading anterior edge, will be identified as the site for hair clipping. At the screening visit and at visits on weeks 8, 12 and 16 "global" photographs were for review by the expert panel. Additional Polaroid photographs were taken at screening and Week 16 to aid the subject in assessing treatment benefit.

Once the hair was clipped at the baseline visit, a tattoo, approximately the size of a period (.), was placed in the center of the circle. This was used a guide for the photographs used in hair counts. The panel review consisted of an independent photographic assessment of the global photographs taken of the vertex area of the scalp (screening visit and weeks 8, 12 and 16) by three physicians.

The other primary efficacy endpoint was a Subject Rating of Treatment Benefit (assessed at Week 16 only) using the following scale:

Hair Loss Condition Rating (HLrate)

Score	Description	Score	Description
3	Significantly improved	-1	Slightly worse
2	Moderately improved	-2	Moderately worse
1	Slightly improved	-3	Significantly worse
0	No change		

3.1.2 Statistical Methodology

The protocol specified that, provided the data were normally distributed, both the change from baseline in hair count and the subject hair loss condition rating were to be analyzed with an analysis of variance (ANOVA) model with factors for treatment and center and covariates age and duration of hair loss:

response = age + duration of hair loss + treatment + pooled center + treatment by center.

If the data were not normal the protocol specified that a Wilcoxon test was to be used to compare median scores of the treatment groups. Change from baseline data was approximately normal (and did pass a test for normality), but the subject rating was highly skewed and did not. The stratified Wilcoxon test used by the Sponsor in their report, i.e., a Van Elteren test stratified on center, seems more appropriate than the simple Wilcoxon specified in the protocol. However, even with skewed data here, cell means seem to be roughly normally distributed and ANOVA would still be appropriate. But, since the hair loss condition rating is ordinal, one might, in principle, prefer a Cochran-Mantel-Haenszel (CMH) test of treatment differences. Results from the ANOVA, the van Elteren test, and CMH tests on dichotomized responses are all provided in the FDA reviewer's analysis.

The following table shows the number of subjects that were recruited into the study in each center:

Table 1. Subject Recruitment per Center

	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011
5% Foam	9	13	24	12	3	13	15	15	3	23	8
Placebo	9	13	22	12	3	13	14	15	3	22	6

	1012	1013	1014	Total
5% Foam	12	15	15	180
Placebo	12	13	15	172

Following the pooling algorithm in the statistical analysis plan, the Sponsor pooled centers 1005, 1009, and 1011 into one center, labeled 9001, while centers 1001 and 1004 were pooled into another center labeled 9002. This was designed to guarantee at least 10 subjects per arm per center. (However, note that the Statistical Analysis Plan is not dated and may be post-hoc). While this reviewer would have been satisfied with merely pooling centers 1005 and 1009, for convenience, the Sponsor's pooling of centers was used in the Agency analysis. (Appendix 1 includes an analysis of the primary endpoints with just these two small centers pooled).

3.1.3 Patient Disposition, Demographics and Baseline Characteristics.

The following table displays the final disposition of patients entering the trial:

Table 2. Patient Disposition

	5% Topical Foam	Placebo
Number Patients Enrolled	180	172
Number Patients Completed	164	151
Number Patients Discontinued	16	21
Reason for Discontinuation N (%)		
Adverse Event	3	2
Protocol Violation	0	1
Withdrew Consent	8	14
Loss to Follow-up	5	4

Thus, few subjects report withdrawing because of adverse events.

Patient demographic and baseline characteristics are summarized below:

Table 3. Subject Demographics and Baseline Hair Loss Characteristics

	5% Topical Foam	Placebo
Age in Years		
Mean (Std Dev)	40.1 (6.3)	38.3 (7.3)
Range (Min-Max)	20 – 49	21 – 49
Race N (%)		
White	151 (83.9%)	154 (89.5%)
Black	7 (3.9%)	5 (2.9%)
Hispanic	17 (9.4%)	7 (4.1%)
Asian/Pacific Islander	3 (1.7%)	3 (1.7%)
American Indian/Alaskan	2 (1.1%)	2 (1.2%)
Other	0	1 (0.6%)
Duration of Hair Loss		
Mean (Std Dev)	115.4 (77.0)	105.9 (67.0)
Range (Min-Max)	12-336	5-312
Pattern of Hair Loss N (%)		
Type IIIv	77 (42.8%)	63 (36.6%)
Type IV	53 (29.4%)	64 (37.2%)
Type V	50 (27.8%)	45 (26.2%)

Note that groups seem to be relatively balanced in terms of demographics and these hair loss characteristics. However, the maximum age of patients is 49. At the EOP2 meeting the Sponsor was requested to include older patients.

3.1.4 Reviewer Results and Conclusions

Recall there were two primary endpoints, each evaluated at Week 16 or time of early termination, 1) the mean change from baseline in hair counts, and 2) the Subject self assessment of the Hair Loss Condition Rating. Results are given for each endpoint in turn, followed by assessment of the secondary endpoint, the score from the expert panel.

3.1.4.1 Change from Baseline in Hair Count

Mean changes from baseline in hair count are given below:

Table 4. Week 16 (LOCF) Mean Change from Baseline in Hair Count

Treatment	N	Mean (Std Dev)	LS Mean (Std Error)
5% Foam	180	19.4 (22.3)	19.5 (1.6)
Placebo	172	4.3 (18.8)	3.9 (1.6)

Note that 142 (78.9%) of the 5% Foam group had an increase in hair counts versus 97 (56.4%) in the placebo group.

Least squares means are provided because the protocol specified tests that are based on differences in these means. The tests of treatment differences, with or without interaction were

highly statistically significant ($p < 0.0001$). The errors for the LS means in the table above are much smaller than the standard deviations since they are standard errors of the LS mean linear estimator, while the standard deviation is used to indicate variability around the mean, and is not adjusted for sample size.

The following table displays the actual baseline mean hair counts and the mean changes in hair counts over the study at Weeks 8, 12, and 16. This table is based on the group of subjects with data at the observed time, plus the intent-to-treat last-observation-carried-forward (ITT-LOCF) population at Week 16.

Table 5. Profiles Over Time of Mean Change from Baseline in Hair Count

			Week			
		Baseline	8	12	16	16 LOCF
5% Foam	N	180	157	166	167	180
	Mean	170.8	15.5	19.8	20.9	19.4
	Std Err	3.8	1.7	1.8	1.7	1.7
	Min, Max	79,329	-50,79	-50,87	-49,102	-49,102
Placebo	N	172	148	156	156	172
	Mean	168.9	5.2	5.0	4.7	4.3
	Std Err	3.7	1.6	1.4	1.6	1.4
	Min, Max	69,324	-55,93	-48,62	-47,61	-47,61

The following table shows summaries of the corresponding baseline mean hair count measures and the measures of mean changes in hair counts in the per protocol population.

Table 6. Profiles Over Time of Mean Change from Baseline in Hair Count (Per Protocol)

			Week		
		Baseline	8	12	16
5% Foam	N	159	149	158	159
	Mean	172.2	15.1	19.6	20.6
	Std Err	3.9	1.7	1.8	1.8
	Min	79,302	-50,79	-50,87	-49,102
Placebo	N	138	130	138	138
	Mean	168.7	4.0	4.9	4.3
	Std Err	4.2	1.5	1.5	1.6
	Min	71,324	-55,61	-48,62	-47,61

Again, in the per protocol population, the Week 16 tests of treatment differences, with or without interaction were highly statistically significant (both $p < 0.0001$). Note that at Week 16, 142 (78.9%) subjects in the 5% Foam group had an increase in hair counts versus 97 (56.4%) in the placebo group.

Thus, as required by the protocol for treatment success, there were highly statistically significant treatment differences in the Week 16, end of treatment, change from baseline in hair

count in both the ITT and the per protocol populations. Note that this does not address the issue of whether or not the actual hair count totals used are inflated by including vellus hairs. The regulatory history section (Section 2.1.2), above, includes a summary of the Sponsor's discussion of this issue. In addition, sensitivity of results to the inclusion of vellus hairs is partially addressed in Appendix 2.

3.1.4.2 Subject's Hair Loss Condition Rating

As discussed above, at the end of treatment or early termination patients are requested to rate the change in their hair loss. The following table displays the definitions of the different levels of that rating response and the corresponding frequencies of that particular level:

Table 7. Week 16/Early Termination Subject Hair Loss Condition Ratings (ITT Population)

Score	Description	5% Foam n (%)	Placebo n (%)
3	Significantly improved	39 (22.9%)	9 (5.6%)
2	Moderately improved	47 (27.7%)	28 (17.3%)
1	Slightly improved	41 (24.1%)	36 (22.2%)
0	No change	32 (18.8%)	56 (34.6%)
-1	Slightly worse	10 (5.9%)	25 (15.4%)
-2	Moderately worse	1 (0.6%)	8 (4.9%)
	Total	170	162

Treatment differences using a Van Elteren test were highly statistically significant ($p < 0.0001$). With the same ANOVA model as that used for mean change from baseline in hair count, treatment differences were also highly statistically significant ($p < 0.0001$).

For reference, the subject hair loss condition rating score above was dichotomized using different cut-points as follows, with a final score of "1" denoting a "success":

Score3 = 1 if Score = 3, 0 otherwise
Score23 = 1 if Score $\geq$ 2, 0 otherwise
Score123 = 1 if Score $\geq$ 1, 0 otherwise

For the original ordinal subject hair loss condition rating score shown in Table 7 above, and for each of the dichotomized score variables score123, score23, and score3, treatment differences tested using CMH tests stratified on pooled center were highly statistically significant (all $p < 0.0001$). This is a completely post hoc analysis, but could have been requested and may be informative.

For some subjects the hair loss condition rating was not evaluated. An analysis of the sensitivity of these results to the missing subjects (10/treatment group) is provided in Appendix 3.

Table 8. Week 16 Results on Subject's Hair Loss Condition Rating (Per Protocol Population)

Score	Description	5% Foam n (%)	Placebo n (%)
3	Significantly improved	36 (22.9 %)	7 (5.1 %)
2	Moderately improved	45 (28.7 %)	21 (15.3 %)
1	Slightly improved	39 (24.8 %)	32 (23.4 %)
0	No change	28 (17.8 %)	50 (36.5 %)
-1	Slightly worse	8 (5.1 %)	21 (15.3 %)
-2	Moderately worse	1 (0.6 %)	6 (4.4 %)
Total		157	137

So, not only in the ITT population, but also in the PP population, using both the ANOVA model described above, and the Van Elteren tests on the hair loss condition rating, and the CMH analyses using the original ordered variables, plus each of the dichotomized score variables defined above, score 123, score23, and score3, treatment differences between the 5% foam and placebo were all highly statistically significant (all $p < 0.0001$).

Note, however, that despite the fact that the Change from Baseline in Hair Count and the Subject's Hair Loss Condition Rating are both intended as measures of hair growth, as discussed in Appendix 4, these measures have only a low association.

3.1.4.3 Secondary Endpoint:

The only secondary efficacy variable was based on an expert panel review of hair growth. Each of the three assessors compared global photographs of Baseline hair status to the Week 16 hair status, each scored from -3 to 3, with 1 to 3 denoting progressively increasing levels of benefit, "0" means no perceived benefit, and -1 to -3 progressively decreasing levels of benefit. The actual endpoint is defined as the median of the three expert panel scores.

Table 9. Week 16 Results on Expert Panel Review

Score	Description	5% Foam N (%)	Placebo N (%)
2	Positive	4 (2.7 %)	0
1		134 (91.2 %)	94 (57.7 %)
0	No benefit	8 (5.4 %)	55 (33.7 %)
-1	Negative	1 (0.7 %)	14 (8.6 %)
Total		163	147

The CMH test of mean differences in the Expert Panel Review score, stratified on site, was also highly statistically significant ($p < 0.0001$).

Note that the Review Team requested an analysis of the consistency of raters. The consistency can be described as generally slight to low fair. This analysis is given in Appendix 5.

3.1.5 Sponsor Results and Conclusions

The Sponsor states that "the primary analysis population was the intent-to-treat (ITT) population defined as all randomized subjects." However, for some analyses the Sponsor's ITT-LOCF population does not include all randomized subjects. The following table illustrates this:

Table 10. Sponsor Week 16 (ITT-LOCF) Mean Change from Baseline in Hair Count

Treatment	FDA Reviewer's Analysis		Sponsor's Analysis	
	N	Mean (Std Dev)	N	Mean (Std Dev)
5% Foam	180	19.4 (22.3)	167	20.9 (22.5)
Placebo	172	4.3 (18.8)	156	4.7 (19.7)

Other ITT analyses provided by the Sponsor do seem to include all subjects. Note that while the observations used can differ, the results provided by the Sponsor and the FDA reviewer's analyses are consistent.

However, for both the Subject Hair Loss Condition Rating and the Expert Panel Review , the FDA and Sponsor analyses are equivalent.

Table 11. Sponsor Week 16 (ITT-LOCF) Subject's Hair Loss Condition Rating

Treatment	N	Mean (Std Dev)
5% Foam	170	1.4 (1.23)
Placebo	162	0.5 (1.24)

The Sponsor's analyses dropped the duration of hair loss from the linear model and did not include the treatment by center interaction. However, efficacy conclusions based on the linear models were quite consistent with the FDA analysis. For both endpoints, in both the Sponsor's and the FDA reviewer's analyses treatment differences were highly statistically significant ($p < 0.0001$).

3.2 Evaluation of Safety

Please see the OTC Division Review.

4. FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

4.1 Gender, Race and Age

Note that all patients were male, so stratifying on gender would have been superfluous.

As discussed in the section on statistical issues (Section 1.3), with a single study submission it is particularly important that results be consistent across subgroups. Hence,

although the study was not powered to detect these differences, results of actual tests of differences in the primary endpoints are provided for race, age, and baseline hair loss category.

4.1.1 Stratification on Race:

Most patients were Caucasian, so they were grouped into only two subgroups. For the mean change in hair count we find the following:

Table 12. Week 16 (LOCF) Mean Change from Baseline in Hair Count by Race

	Caucasian		Other	
Treatment	N	Mean (Std Dev)	N	Mean (Std Dev)
5% Foam	151	19.9 (22.8)	29	16.4 (19.8)
Placebo	154	4.2 (19.0)	18	5.2 (17.5)

Specifying the same models for efficacy for each subgroup as the complete data model above, treatment differences in the Caucasian group were statistically highly significant ($p < 0.0001$) while in the small "Other" group differences were barely statistically significant ($p \leq 0.0475$). However, note that the ratio of observed means to standard deviations in the "Other" group is quite similar between the two race groups, and thus the higher statistical significance in the "Other" group is due to the much smaller sample size.

For the subject hair loss condition rating we get the following:

Table 13. Week 16 Results on Hair Loss Condition Rating by Race

		Caucasian		Other	
Score	Description	5% Foam n (%)	Placebo n (%)	5% Foam n (%)	Placebo n (%)
3	Significantly improved	32 (22.4%)	7 (4.8%)	7 (25.9%)	2 (11.8%)
2	Moderately improved	38 (26.6%)	23 (15.9%)	9 (33.3%)	5 (29.4%)
1	Slightly improved	37 (25.9%)	33 (22.8%)	4 (14.8%)	3 (17.7%)
0	No change	29 (20.3%)	51 (35.2%)	3 (11.1%)	5 (29.4%)
-1	Slightly worse	6 (4.2%)	24 (16.6%)	4 (14.8%)	1 (5.9%)
-2	Moderately worse	1 (0.7%)	7 (4.8%)	0	1 (5.9%)
Total N		143	145	27	17

Van Elteren tests of mean treatment differences in the hair Loss Condition Rating stratified on center were highly statistically significant in the Caucasian subgroup ($p < 0.0001$) while in the small "Other" group treatment differences were not statistically significant ($p \leq 0.3170$). The latter result is partially due to the small sample size. The corresponding ANOVA tests on Hair Loss Condition Rating as specified in the protocol, had similar results ($p < 0.0001$ and $p \leq 0.2109$, respectively).

4.1.2 Stratification on Age:

Defining subjects into two groups by age (less than 40 versus 40-49) we get the following:

Table 14. Week 16 (LOCF) Mean Change from Baseline in Hair Count by Age Group

	Age < 40		Age = 40-49	
Treatment	N	Mean (Std Dev)	N	Mean (Std Dev)
5% Foam	70	19.9 (24.2)	110	19.0 (21.1)
Placebo	88	5.0 (20.1)	84	3.6 (17.4)

The ANOVA tests of treatment differences in both age groups were statistically highly significant (both $p < 0.0001$).

Table 15. Week 16 Results on Hair Loss Condition Rating by Age Group

		Age < 40		Age = 40-49	
Score	Description	5% Foam n (%)	Placebo n (%)	5% Foam n (%)	Placebo n (%)
3	Significantly improved	12 (17.7%)	4 (6.3%)	27 (26.5%)	5 (6.3%)
2	Moderately improved	22 (32.4%)	13 (15.9%)	25 (24.5%)	15 (18.8%)
1	Slightly improved	18 (26.5%)	12 (14.6%)	23 (22.6%)	24 (30.0%)
0	No change	10 (14.7%)	30 (36.6%)	22 (21.6%)	26 (32.5%)
-1	Slightly worse	5 (7.4%)	18 (22.0%)	5 (4.9%)	7 (8.8%)
-2	Moderately worse	1 (1.5%)	5 (6.1%)	0	3 (3.8%)
Total N		68	82	102	80

CMH tests of treatment differences in hair loss condition rating for both age groups were statistically highly significant (both $p \leq 0.0001$). The corresponding ANOVA tests were also highly statistically significant (both $p < 0.0001$).

4.2 Other Special/Subgroup Populations

Stratification on Baseline Hair Loss Pattern:

At baseline patients were categorized by the Norwood-Hamilton classification of pattern of hair loss. Recall that the Sponsor was informed that it was important that treatment differences be apparent in each subgroup.

Table 16. Week 16 (LOCF) Mean Change from Baseline in Hair Count by Baseline Hair Loss Pattern

Pattern:	IIIv		IV		V	
Treatment	N	Mean (Std Dev)	N	Mean (Std Dev)	N	Mean (Std Dev)
5% Foam	77	21.5 (19.2)	53	14.7 (25.1)	50	21.1 (23.3)
Placebo	63	7.3 (18.9)	64	4.0 (17.4)	45	0.5 (20.2)

ANOVA tests of treatment differences in the mean change from baseline in hair count were statistically significant for all three baseline hair loss patterns ($p \leq 0.0001$, $p \leq 0.0184$, and $p \leq 0.0020$, respectively).

Table 17. Week 16 Results on Hair Loss Condition Rating by Baseline Hair Loss Pattern

Pattern	IIIv		IV		V	
Score	5% Foam n (%)	Placebo n (%)	5% Foam n (%)	Placebo n (%)	5% Foam n (%)	Placebo n (%)
3	19 (25.7%)	3 (5.2%)	8 (16.7%)	4 (6.5%)	12 (25.0%)	2 (4.8%)
2	20 (27.0%)	12 (20.7%)	15 (31.3%)	11 (17.7%)	12 (25.0%)	5 (11.9%)
1	14 (18.9%)	7 (12.1%)	16 (33.3%)	17 (27.4%)	11 (22.9%)	12 (28.6%)
0	15 (20.3%)	25 (43.1%)	5 (10.4%)	13 (21.0%)	12 (25.0%)	18 (42.9%)
-1	6 (8.1%)	10 (17.2%)	3 (6.3%)	11 (17.7%)	1 (2.1%)	4 (9.5%)
-2	0	1 (1.7%)	1 (2.1%)	6 (9.7%)	0	1 (2.4%)
Total	74	58	48	62	48	42

Stratifying on center, Van Elteren tests of treatment differences in mean hair loss condition rating were statistically significant for all three baseline hair loss patterns ($p < 0.0001$, $p \leq 0.0026$, and $p \leq 0.0006$, respectively). The corresponding ANOVA tests of treatment differences were also statistically significant for all three baseline hair loss patterns ($p \leq 0.0001$, $p \leq 0.0110$, and $p \leq 0.0002$, respectively).

5. SUMMARY AND CONCLUSIONS

5.1 Statistical Issues and Collective Evidence

1. At the EOP 2 meeting on January 16, 2003, the Division encouraged the Sponsor to either conduct two placebo controlled Phase 3 trials, or to conduct a 3-arm Phase 3 trial including an arm with the currently labeled ointment. However, the Agency also stated that "a single . . . study in which the new product is compared to its vehicle might be acceptable in the study of androgenic alopecia on the vertex." With a single study, either 2-arm or 3-arm, the Sponsor was told they need very small significance levels for tests of the primary endpoints, plus consistency across centers and study subgroups. The Sponsor has submitted a single placebo-controlled, 2-arm study of androgenetic alopecia on the vertex. Whether or not this study satisfied those criteria and the choice of endpoints seem to be the key points in determining that the results in this trial are conclusive. The discussion with the Sponsor is summarized above in the regulatory history section 2.1.3 of this report.

2. At the EOP 2 meeting and later, the Division requested that only non-vellus hairs be included in the hair count totals. In response to an information request (received September 2, 2005), the Sponsor stated that "Pfizer believes that the magnification level of 5.7 fold yields visualization and thereby counting of non-vellus hairs (i.e. those ≥ 0.03 mm in diameter) and adequately filters out vellus and insignificant miniaturized non-vellus hairs (i.e. those < 0.03 mm in diameter). The Sponsor further contends that this magnification is consistent with previous

submissions. This argument is also summarized in the regulatory history section, 2.1.3, of this report, while the Sponsor's argument is included in Appendix 6.

3. Also at the EOP 2 meeting, the Sponsor was also reminded of the need to achieve very small significance levels for tests of the primary endpoints, plus consistency across centers and study subgroups. At that meeting the Division recommended that, since the condition was common in men aged over 49, "there be no upper age limit." Although arguably not primarily a statistical issue, one might argue that the Sponsor's restriction to subjects aged 49 or less implies an insufficient analysis in all relevant demographic subgroups.

4. Despite some limitations related to actual interpretation, in many circumstances p-values can be useful as rough measures of the weight of evidence. To show superiority the Agency generally recommends two-sided statistical tests with a significance level of 0.05 or smaller. But since the drug would not be rejected if it was considerably superior to its comparator, in effect, as evidence of efficacy, the Agency requests one-sided tests with statistical significance levels at or below 0.025 in two studies. Then with independent studies the probability of a type I error in both studies is $(0.025)^2 = 0.000625$. To achieve an equivalent level of significance with a single study, we could use a 0.000625 level in a single study. Note that the Sponsor did generally achieve this level of significance in the overall tests on the primary endpoints, so we would conclude that they achieved the "small" significance levels.

5. The protocol specifies that both the change from baseline in hair count and the subject hair loss condition rating are to be analyzed with an analysis of variance (ANOVA) model with factors for treatment and center and covariates age and duration of hair loss. The protocol also specified that if the residuals are not normal a Wilcoxon test is to be used to compare median scores of the treatment group. However, the actual test used by the Sponsor was a Wilcoxon test stratified on center, i.e., a Van Elteren test. Even with skewed data as here, cell means are roughly normally distributed and ANOVA would still be appropriate. However, since the hair loss condition rating is ordinal, one might, in principle, prefer a Cochran-Mantel-Haenszel (CMH) test of treatment differences. Using modified ridit scores in the CMH test leads to the Van Elteren test. Results from both statistical tests are reported here, and are always essentially equivalent.

6. Several centers only recruited a small number of patients into the study. Pooling of subjects for the analysis was specified in the analysis plan, but not in the Study protocol. Since the dating of the analysis plan was not clear, this pooling may be post hoc. However, this pooling (see 2.2 below) was deemed to be acceptable, and for convenience was followed in the Agency analysis. Note that endpoints, methods of analysis, etc. were specified in the protocol.

7. The Review team wanted to investigate the consistency of ratings across the three raters for the expert panel review of hair growth. In response, the Sponsor provided kappa statistics, including a pooled kappa across the three raters. As discussed in Appendix 5, there are possible problems with using kappa statistics to assess consistency, but as also shown in that appendix

both the kappa statistics and a corresponding latent trait model suggest a lack of consistency among the raters.

Statistical Findings

For both mean change from Baseline in visualized hair counts in the target region and the subject rating of treatment benefit in the differences in favor the 5% Foam were highly statistically significant ($p < 0.0001$). This was true both in the intent to treat (ITT) and the per protocol populations. However, as detailed in Appendix 4, the association between these two measures of hair growth was generally low.

Grouping ITT patients into two roughly equally sized groups of those aged less than 40 and those 40-49, for both primary endpoints treatment differences in each group were still highly statistically significant in favor of 5% minoxidil foam (all $p \leq 0.0001$). Most patients were Caucasian, and among them all differences were still highly statistically significant in favor of minoxidil (both $p \leq 0.0001$). Although actual success proportions among the non-Caucasian patients were similar to those among the Caucasian, there were only a relatively small number of patients in this subgroup (44) and treatment differences were only barely statistically significant for change from baseline in hair counts ($p \leq 0.0475$) and non-significant for the subject rating ($p \leq 0.2109$). For the Norwood Baseline Hair Loss Patterns IIIv, IV and V treatment differences in the change from baseline were all statistically significant ($p \leq 0.0001$, $p \leq 0.0184$, and $p \leq 0.0020$, respectively). Results for the subject rating were similar ($p \leq 0.0001$, $p \leq 0.0110$, and $p \leq 0.0002$, respectively). An expert panel review found similar overall results, although there was evidence that raters were not highly consistent in their evaluations (see Appendix 5).

5.2 Conclusions and Recommendations

The two primary endpoints in the study were 1) the change from baseline in hair count in a small area on the vertex of scalp described above, and 2) a subject rating of overall treatment benefit scored on a seven point ordinal scale. For both endpoints differences in favor of the 5% Foam were very highly statistically significant (both $p < 0.0001$). This was true both in the intent to treat (ITT) and the per protocol (PP) populations. However, despite the fact these are both measures of hair growth, as discussed in Appendix 4, these measures have only a low association.

There are two primary issues in this submission. The first issue is whether or not the primary endpoint includes counts of small vellus hairs. The Sponsor argues that their magnification and technology only counts a relatively small number of such hairs. The Sponsor's argument is summarized in Section 2.1.3 above and presented completely in Appendix 6. Note that the analysis supposedly confirming the claim that few vellus hairs are included seems to be based on a small study. However, the Sponsor contends that this endpoint is the same as that used in prior submissions for this and similar indications. This seems to be

confirmed by the Kaufman, K. D. et al, (1998): "Finasteride in the treatment of men with andorgenetic alopecia," *J. Am. Acad Dermatol* **39**: 578-89, paper, available on the internet.

The second issue is whether or not a single study is adequate to demonstrate efficacy. At the End of Phase 2 meeting held on January 16, 2003, the Division noted that "a single . . . study in which the new product is compared to its vehicle might be acceptable in the study of androgenic alopecia on the vertex." At that meeting the Sponsor was also reminded of the need to achieve very small significance levels for tests of the primary endpoints, plus consistency across centers and study subgroups. As discussed in Section 5.1 above, one could argue that a 0.0006 level of significance could be used to define a conclusively small p-value for the overall analysis. For the primary endpoints, overall tests of treatment differences achieved this level of significance. Results were also required to be consistent across demographic and hair loss subgroups. Most demographic and hair loss subgroups had statistically significant differences. However, patients were aged 49 or below. Since this condition is common in patients aged 50 and above, it is not clear if the exclusion of older patients satisfies the criterion of consistency across relevant study subgroups. Note that at the EOP 2 meeting, the Division recommended that such older patients be included in the study.

However, for the specified endpoints treatment differences were highly statistically significant. Further, treatment differences within the included subgroups were statistically significant.

APPENDICES:

Appendix 1. Sensitivity Analysis to Vellus Hairs in the Hair Count Totals

One problem with this submission is the possibility that very thin hairs (e.g., "peach fuzz" / vellus hairs) are included, and inflate the hair count totals. Before specifying models to assess sensitivity, note that 142 (78.9%) of the 5% minoxidil foam treatment group had an actual increase in hair counts in the target area versus 97 (56.4%) in the placebo group.

Since we are primarily interested in the effect of including vellus hairs on the increase in hair counts, it seems appropriate, at least as a sensitivity analysis, to assess the impact of assuming various proportions of the increase in hair counts correspond to thin, vellus hairs that should not be counted. One approach is to simply down-weight the increase in those cases that showed an actual increase in hair counts. Note that since we only down-weight cases that show an increase, 78.9% of the minoxidil group will be down-weighted versus only 56.4% in the placebo group. For this analysis counts that show a decease will be left unchanged. The following table provides the resulting means and standard deviation for each treatment group, and the corresponding significance levels of the test of treatment differences assuming the linear model specified earlier.

For the "Percent Reduced" column, the first proportion refers to the reduction in increase in the 5% minoxidil foam treatment group and the second proportion to the corresponding reduction in the placebo group to reflect this down-weighting.

Table A.1.1. Week 16 (LOCF) Mean Change from Baseline in Hair Count

Percent Reduced	Mean (Std Dev)		Significance Level
	5% Foam	Placebo	
10%, 05%	24.3 (16.0)	15.5 (13.0)	$p< 0.0001$
20%, 10%	21.6 (14.2)	14.7 (12.3)	$p< 0.0001$
30%, 15%	18.9 (12.5)	13.9 (11.6)	$p< 0.0001$
40%, 20%	16.2 (10.7)	13.1 (10.9)	$p\leq 0.0173$
50%, 25%	13.5 (8.9)	12.2 (10.2)	$p\leq 0.2284$
60%, 30%	10.8 (7.1)	11.4 (9.6)	$p\leq 0.7121$
10%, 10%	24.3 (16.0)	14.7 (12.3)	$p< 0.0001$
20%, 20%	21.6 (14.2)	13.1 (10.9)	$p< 0.0001$
30%, 30%	18.9 (12.5)	11.4 (9.6)	$p< 0.0001$
40%, 40%	16.2 (10.7)	9.8 (8.2)	$p< 0.0001$
50%, 50%	13.5 (8.9)	8.2 (6.8)	$p< 0.0001$
60%, 60%	10.8 (7.1)	6.5 (5.5)	$p< 0.0001$
70%, 70%	8.1 (5.3)	4.9 (4.1)	$p< 0.0001$
80%, 80%	5.4 (3.6)	3.3 (2.7)	$p< 0.0001$
90%, 90%	2.7 (1.8)	1.6 (1.3)	$p< 0.0001$

The first part of the table above assesses the impact of assuming that among patients showing an increase in hair counts, the proportion of that increase due to counting vellus hairs in the minoxidil group is twice that in the placebo group. The second part of the table above assumes that among those subjects showing an increase, the proportion of vellus hairs counted in the increase is the same in the two treatment groups. But note that since only patients with an increase are reduced, this still penalizes the 5% Foam group more than the placebo group.

Assume that the proportion of new vellus hairs in the 5% Foam group is twice the proportion of new vellus hairs in the placebo group. If we assume that as much as 40% of the increase in hair count in the foam treatment group is due to vellus hairs, and only 20% in the placebo group, then the adjusted difference between the 5% Foam and placebo will still be statistically significant, though barely ($p < 0.0173$). Further, if we assume that among those cases in both treatment groups that show an increase in hair count, 80% or 90% of that increase corresponds to vellus hairs that should not be included in the total, the adjusted difference between the 5% Foam and placebo treatment groups will still be statistically significant ($p < 0.0001$).

Note, however, the actual difference in hair counts depends upon the proportion of vellus hairs actually counted, and while possibly statistically significant, the difference in hair counts may not be clinically significant.

Appendix 2. Sensitivity Analysis to Centers

The goal of this analysis was to assess the impact of individual centers on the final results. Centers 1005 and 1009 had relatively few subjects and for this analysis were merged in a single new center 8001. Since this particular analysis is meant to investigate differences in centers it makes sense to use this less restrictive pooling rather than that specified by the Sponsor.

Table A.2.1 Week 16 Per Investigator Least Squares Means in Change From Baseline in Hair Counts and Subject Rating of Treatment Benefit

Investigator	Change From Baseline				Subject Rating of Treatment Benefit			
	5% Foam		Placebo		5% Foam		Placebo	
Number	N	LS Mean	N	LS Mean	N	LS Mean	N	LS Mean
1001	9	22.14	9	3.43	9	1.65	9	0.87
1002	13	19.28	13	7.18	11	2.15	13	1.23
1003	24	22.55	22	10.04	23	1.59	21	0.66
1004	12	5.78	12	-5.74	11	1.41	10	0.56
1006	13	28.16	13	9.03	12	0.88	12	0.50
1007	15	28.40	14	0.72	15	1.60	12	-0.37
1008	15	16.52	15	-2.77	14	1.13	15	0.25
1010	23	22.61	22	9.66	21	0.98	22	0.22
1011	8	16.96	6	-4.35	7	1.83	5	-0.11
1012	12	17.17	12	7.79	11	1.22	10	0.49
1013	15	18.03	13	1.04	15	1.59	13	1.26
1014	15	15.37	15	-0.13	15	1.52	14	0.21
8001	6	9.34	6	8.75	6	0.65	6	0.39

Least squares means are used because the protocol specified ANOVA tests are based on the differences in these means. For each center, both the least squares mean change from baseline in hair counts and the subject rating of treatment benefit, are better than the corresponding scores in the placebo group. Note that there is a wide variation among treatment means, however, except for the pooled center 8001, the within investigator treatment differences in mean scores are generally of roughly the same magnitude. (Note that the difference in center 8001 is unusually small). With additive center/investigator effects, this suggests that in general, while center differences are large, perhaps reflecting different patient populations, there are no particular center by treatment differences; i.e., treatment differences within centers are consistent. For both the change from baseline in hair count and the subject rating of treatment benefit, the ANOVA tests of investigator by treatment interaction were not statistically significant ($p \leq 0.8979$ and $p \leq 0.4315$, respectively).

To allow further assessment of the effect of possibly deleting a center, the following table displays the significance level of the tests of no treatment differences for both primary endpoints, the change from baseline in hair count and the subject hair loss condition rating. The rows labeled "None" correspond to no deleted centers. The other rows correspond to the tests of treatment effect when the indicated center is deleted. The model:

response = age + duration of hair loss + treatment + center + treatment by center,

was used for both endpoints. Because they are so small, the significance levels are given in exponential notation.

Table A.2.2 Week 16 Effect of Deleting a Center for Change From Baseline in Hair Counts and Subject Rating of Treatment Benefit

Investigator Deleted	Endpoint	Error Degrees of Freedom	F-ratio	Significance Level	5% Foam LSM	Placebo LSM
None	Change Hair	324	41.1385	5.0196E-10	18.6389	3.43336
None	Subj Rating	304	42.3342	3.1651E-10	1.3997	0.47482
1001	Change Hair	308	36.2886	4.84651E-9	18.3126	3.45940
1001	Subj Rating	288	39.8500	1.03408E-9	1.3771	0.44065
1002	Change Hair	300	39.8048	1.00506E-9	18.5495	3.17638
1002	Subj Rating	282	38.9080	1.62037E-9	1.3369	0.41085
1003	Change Hair	280	36.3364	5.22358E-9	18.1923	2.84567
1003	Subj Rating	262	37.3963	3.48682E-9	1.3878	0.46103
1004	Change Hair	302	38.6968	1.64771E-9	19.7497	4.18159
1004	Subj Rating	285	40.1126	9.3096E-10	1.4019	0.46850
1006	Change Hair	300	36.6277	4.25921E-9	17.7703	3.02759
1006	Subj Rating	282	44.0654	1.6252E-10	1.4446	0.47111
1007	Change Hair	297	34.3481	1.22379E-8	17.8450	3.62570
1007	Subj Rating	279	30.5933	7.32114E-8	1.3849	0.54682
1008	Change Hair	296	35.4987	7.23626E-9	18.8462	3.93042
1008	Subj Rating	277	39.2460	1.42243E-9	1.4211	0.49638
1010	Change Hair	281	37.5944	2.94262E-9	18.3233	2.82037
1010	Subj Rating	263	40.2866	9.556E-10	1.4342	0.50101
1011	Change Hair	312	38.0064	2.17862E-9	18.8008	4.10837
1011	Subj Rating	294	35.5496	7.11633E-9	1.3611	0.52136
1012	Change Hair	302	38.2924	1.98008E-9	18.8210	3.10961
1012	Subj Rating	285	40.1466	9.1686E-10	1.4113	0.47171
1013	Change Hair	298	38.1400	2.15319E-9	18.7464	3.55415
1013	Subj Rating	278	41.5368	5.0883E-10	1.3793	0.40275
1014	Change Hair	296	36.5121	4.55182E-9	18.9513	3.75894
1014	Subj Rating	277	34.9093	1.01109E-8	1.3945	0.49885
8001	Change Hair	314	49.1633	1.4485E-11	19.4050	3.01037
8001	Subj Rating	294	47.1099	3.9826E-11	1.4631	0.48315

Note that none of the significance levels are above 0.0000001, that is, one out of ten million. So even when deleting any single center, results on the primary endpoints remain extremely statistically significant.

As noted in the report, an alternative assessment is to dichotomize the Hair Loss Condition Rating and analyze the resultant variable using a Mantel-Haenszel test stratified on center. For each of the three dichotomized subject ratings of treatment benefit defined earlier in the report, namely where treatment success on the was defined as either a score of 3, a score of 2 or 3, or a score of 1 or higher, all tests of differences were highly statistically significant (all $p <$

0.0001) regardless of which single center was deleted. The size of the computed chi-square statistic indicates that the actual computed significance level would be considerably smaller.

Thus deleting any single center seems to have no effect on final efficacy conclusions, suggesting that no one center is problematical. Center 1007 does show the largest discrepancy between treatments, but it does appear to be consistent with the other centers.

Appendix 3. Sensitivity Analysis to Missing Data in the Subject Hair Loss Condition Rating.

The following table reiterates the Week 16 results on the subject hair loss condition rating. Note that this is one of the primary endpoints. Ten subjects in each treatment group were in the ITT-LOCF population, but not in the population of subjects with Week 16/early termination data. As a sensitivity analysis one can assign to the 10 Placebo patients missing from this ITT population the best observed score, i.e., a score of "3" , ("Significantly improved"), and to the 10 missing 5% Foam patients the worst observed score, i.e., "-2", ("moderately worse"). This is nearly the most extreme imputation possible against the Sponsor.

Table A.3.1. Frequency of Subject Hair Loss Condition Rating

Score	Description	5% Foam	Placebo
3	Significantly improved	39	9 + 10*
2	Moderately improved	47	28
1	Slightly improved	41	36
0	No change	32	56
-1	Slightly worse	10	25
-2	Moderately worse	1 + 10*	8
Total		180	172

* Use these values to match the total number of subjects in the ITT-LOCF population

Using this extreme imputation we still have statistically significant treatment differences in favor of the 5% Foam. The ANOVA model:

Hair Loss rate = age + duration of hair loss + treatment + pooled center + treatment by center,

specified in the protocol, showed statistically significant differences in favor of the 5% Foam over placebo ($p \leq 0.0002$). Using this imputation, the Hair Loss Condition Rating was dichotomized three ways so that treatment success was defined as either a score of 3, a score of 2 or 3, or a score of 1 or higher. The corresponding CMH tests comparing mean hair loss rating and stratified on the Sponsor pooled center were highly statistically significant (all $p \leq 0.0001$).

This seems to strongly suggest that results on the subject hair loss condition rating are not sensitive to these missing cases.

Appendix 4. Association between the Subject Self Assessment and Hair Count Measures

The Review team expressed interest in investigating the association between the hair count measures and the Subject Hair Loss Condition Rating. It should be noted that the hair count measure is based on a square centimeter at the vertex, while the Subject Hair Loss Condition Rating is a global measure. Still one would expect an association between these measures. The Subject Hair Loss Condition Rating is only assessed at Week 16 or at early termination, and hence is only compared to the corresponding count measure at nominal Week 16 in the ITT-LOCF population. For correlation type measures, since the Subject Hair Loss Condition Rating is measured on an ordinal scale, and the actual hair count and the change from baseline are measured on ratio scales (i.e., an interval scale with a fixed zero), nonparametric correlation measures like Spearman's and Kendall's are arguably most appropriate.

The following table displays these correlations and the corresponding significance levels for each treatment group and overall the data.

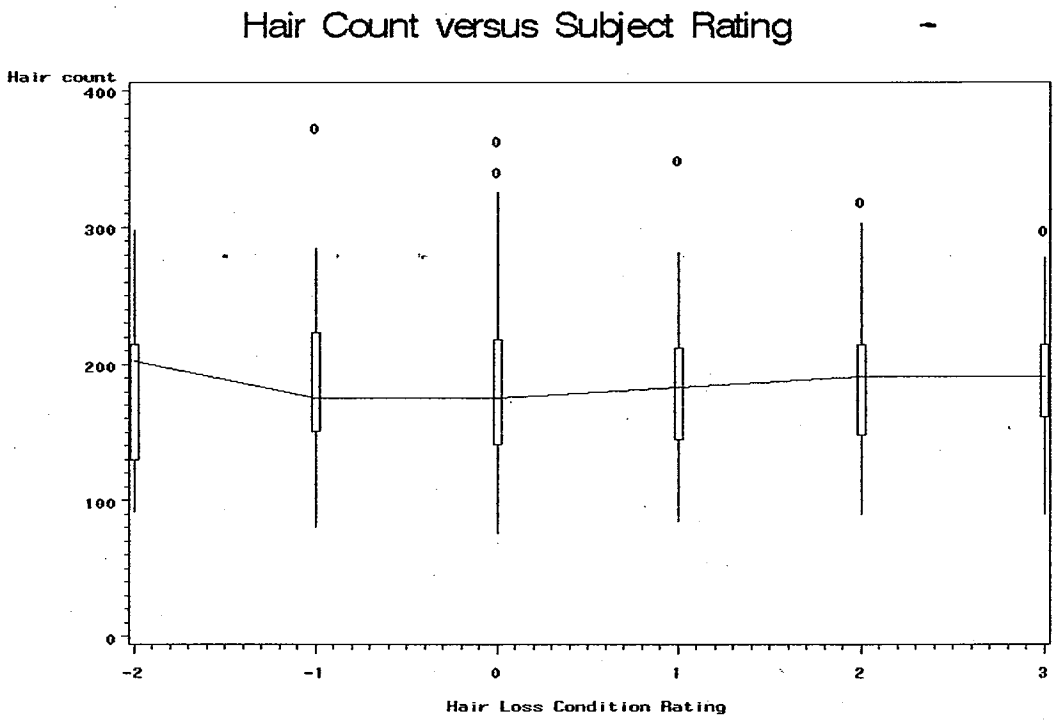
Table A.4.1 Correlations Between the two Hair Count Measures and Subject Hair Loss Condition Rating.

Variable	Statistic	5% Foam	Placebo	Overall
Hair Count vs. Hair Loss Condition Rating	Spearman rho	-0.0035	0.0033	0.0527
	p-value for test rho=0	0.9642	0.9669	0.3789
	Kendall tau	-0.0040	0.0004	0.0384
	p-value for test tau=0	0.9436	0.9940	0.3412
Change from baseline in hair count vs. Hair Loss Condition Rating	Spearman rho	0.1246	0.0773	0.2056
	p-value for test rho=0	0.1056	0.3280	0.0002
	Kendall tau	0.0879	0.0004	0.1518
	p-value for test tau=0	0.1250	0.3247	0.0002

The small, generally statistically non-significant results, suggest that there is no evidence that the Subject Hair Loss Condition Rating is linearly associated with the actual hair count, and only weak evidence that it is linearly related to the change from baseline in Hair Count. The statistically significant results for the correlational measures in the overall group versus the same measures within each treatment group seem to be at least partly due to the separation of these groups.

Graphically the weakness of these relations can also be displayed by looking at the box and whisker summaries of the distributions of the hair count measures at each level of the Subject Hair Loss Condition Rating. Recall that the box displays the data between the first and third quartiles, and the "whiskers" extending from the box display the distribution of data up to 1.5 times the interquartile range. Data points outside that range are indicated by a "0". The lines connect medians.

The following plot summarizes the distribution of the hair count measure for each level of the Subject Hair Loss Condition Rating.

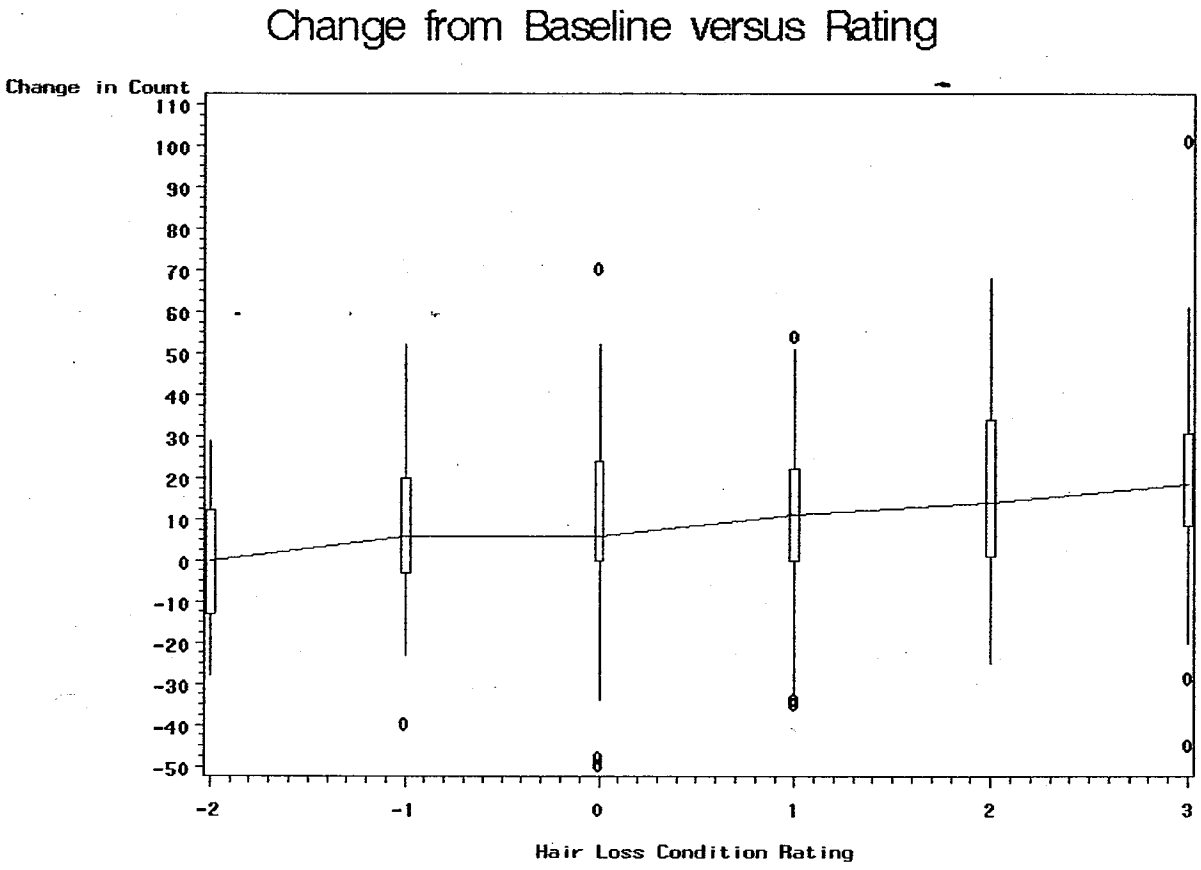


As indicated by the both Spearman's rho and Kendall's tau, there seems to be very little relation between the actual hair count and the level of the Subject Hair Loss Condition Rating. Note that variation in hair counts seems to be relatively consistent across levels of the Subject Hair Loss Condition Rating.

Note that since the Subject Hair Loss Condition Rating assesses change from baseline it probably is more appropriate to compare it to the change from baseline in the hair count in the target region as summarized below.

Appears This Way
On Original

The plot below summarizes the distribution of the change from baseline hair count measure for each level of the Subject Hair Loss Condition Rating.



Since the Subject Hair Loss Condition Rating is meant to measure improvement in hair, it is not surprising that there is some trend in greater increases over baseline in hair counts for each level of the rating. What might be surprising is how weak this trend actually is. It does appear that, at least for this population, the overall Subject Hair Loss Condition Rating is only weakly related to the actual hair count increase in the 1 cm^2 target area.

Although not displayed here, after adjusting for small centers, for both the hair counts and the change from baseline in hair counts, results are consistent across treatment groups.

Appendix 5. Frequentist Analysis of Consistency of Expert Panel Raters

The Review team inquired about the consistency across raters in the Expert Panel review. The following table displays the three 5x5 tables comparing the three expert panel raters, , and .

Table A.5.1 Cross-classification Among Raters

		Score						Score				
		-2	-1	0	1	2	-2	-1	0	1	2
Dr.	score										
	-2	.	.	.	.	.	.	.	.	.	.
	-1	1	2	4	.	.	1	2	4	.	.
	0	.	2	170	17	2	.	14	166	11	.
	1	.	.	41	41	5	.	2	62	21	2
	2	.	.	2	9	14	.	.	7	14	4
Dr.	score										
	-2						.	.	1	.	.
	-1						1	1	2	.	.
	0						.	16	185	16	.
	1						.	.	46	20	1
	2						.	1	5	10	5

The Sponsor provided the following table of marginal distributions and associated unweighted Kappa statistics.

Table A.5.2 Agreement Among Raters (from Sponsor)

	Placebo (N=172)			5% Foam (N=180)			Overall (N=352)		
-2=Moderately Worse	1		1				1		1
-1=Slightly Worse	11	7	4	7			18	7	4
0=No Change	127	118	124	112	73	93	239	191	217
1=Slightly Improved	7	22	17	39	65	50	46	87	67
2=Moderately Improved	1		1	5	25	20	6	25	21
Missing	25	25	25	17	17	17	42	42	42
Kappa (a)									
vs		0.2285			0.1504			0.2135	
vs		0.2258			0.2291			0.2507	
vs		0.2764			0.4814			0.4683	
vs vs (b)		0.24801			0.26418			0.30956	

(a) Kappa were calculated, using Proc freq, based on complete or square data in each treatment group.
(b) Kappa (among multiple raters) were calculated using macro MAGREE from SAS.

The Kappa's under the columns labeled "Placebo" and "5% Foam" are the restricted to those treatment groups. The Corresponding marginal relative frequency distributions are as follows:

Table A.5.3 Agreement Among Raters (Marginal Distribution) FDA Analysis
Score:

	-2	-1	0	1	2
Dr [redacted]	0.3%	5.8%	77.1%	14.8%	1.9%
Dr. [redacted]	0.0%	2.3%	61.6%	28.1%	8.1%
Dr. [redacted]	0.3%	1.3%	70.0%	21.6%	6.8%

Note that Dr. [redacted] is generally more conservative than either of the other two raters, while Dr. [redacted] tends to give the highest scores. Further, the marginal distributions seem to be unimodal and could potentially fit a latent trait model.

For the provided data, the FDA reviewer computed slightly different Kappa statistics:

Table A.5.4 Agreement Among Raters (Kappa Statistics) FDA Analysis

	5% Foam	Placebo	Overall
Dr. [redacted] vs Dr. [redacted]	0.1297	0.2411	0.2145
Dr. [redacted] vs Dr. [redacted]	0.2025	0.2258	0.2507
Dr. [redacted] vs Dr. [redacted]	0.4814	0.2846	0.4670
Drs. [redacted] vs [redacted] vs [redacted]	0.2642	0.2480	0.3096

Landis and Koch (1977) suggest the following rough guide to interpreting the computed values of Kappa statistics to assess the strength of agreement among raters:

Kappa ≤ 0	Poor	Kappa in 0.4-0.6	Moderate
Kappa in 0 - 0.2	Slight	0.6-0.8	Substantial
0.2-0.4	Fair	0.8-1.0	Almost perfect

Using this scale we conclude that consistency is generally only slightly within the fair category, with Dr. [redacted] giving generally lower scores.

Despite the fact that the marginal distribution computed by the FDA reviewer agrees with that computed by the Sponsor, the computed Kappas differ slightly. This may be due to the method of treating missing values. Since the response "missing" is generally not under the control of the rater, and is not due to a rating by that rater, it seems to make more sense to ignore unrated cases. That was done in the Agency analysis. A less likely, but possible technical source of the discrepancy is the fact that to compute the measures of agreement in SAS requires square tables. Computationally, this is done by providing small weights (e.g., 1.0E-20) to sufficient cells needed to make the contingency table square. An alternative would be to pool categories to make the resulting table square. Thus it is possible that applying different weights, or pooling cells, could also explain the discrepancies.

One possible problem with using kappa statistics to assess consistency is that one can argue that kappa statistics actually measure how far the observed ratings are from what one would expect if ratings were independent. That is, instead of measuring consistency of ratings, they measure lack of independence in the ratings. Since raters evaluate the same subjects we

expect lack of independence in the ratings, so one could argue that the distance from independence is not likely to be a reasonable measure of consistency in ratings.

An alternative approach is use a located latent class analysis (see Uebersax, 1993). For each of the 5% Foam, Placebo, and Overall pooled groups of subjects, the covariance matrices of the three observed ratings tend to have a single dominant eigenvalue. This is at least consistent with the notion that the ratings are largely unidimensional, i.e., that the ratings might be well summarized by a single latent trait. A located latent class model postulates that the estimated latent classes are determined by such a single latent trait.

Let π_{ijk} denote the probability that rater 1 classifies response in category i, rater 2 in category j, and rater 3 as category k. There seemed to be identifiability problems in the estimation, and since this was only a secondary analysis it was felt adequate to pool the -2,-1, and 0 responses, and code the resulting response categories as 1-3. So manifest categories 1 and 2, above, are denoted as 2 and 3 in the output below. For three latent classes, if we denote the probability of class c as λ_c, for c=1,2,3, and the probability of response pattern ijk in class c as $\pi_{ijk|c}$, we model π_{ijk} as follows:

$$\pi_{ijk} = \lambda_1 \pi_{ijk|1} + \lambda_2 \pi_{ijk|2} + \lambda_3 \pi_{ijk|3}$$

As is usually true of latent class analyses we assume conditional independence of raters (within a class). Then

$$\pi_{ijk|c} = \pi_{i|c1} \pi_{j|c2} \pi_{k|c3},$$

where the ci subscript refers to rater i in class c. Here both c,i = 1,2,3.

The model defines the probabilities of a subject in a latent class having a certain response by whether a specified threshold was exceeded. For some distribution $\psi_{cr}(\ .\ )$, the probability that a randomly observed member of the latent class c will have an apparent trait level, x_c, that exceed raters r threshold for trait level m is defined by:

$P(x_c) = 1 - \psi_{cr}(\tau_{m\,r})$, m=2,3 (so for latent classes 1 to 3 we have thresholds 2 and 3). .

Then $\pi_{1|cr} = \psi_{cr}(\tau_{2r})$

$\pi_{m|cr} = \psi_{cr}(\tau_{m+1,r}) - \psi_{cr}(\tau_{m\,r})$ for m = 2

$\pi_{3|cr} = 1 - \psi_{cr}(\tau_{3r})$

The distribution used here is the logistic approximation to the normal distribution.

For identification purposes we restrict attention to 3 classes. Note interest is not focused on these classes but rather on the hypothesized underlying latent trait.

This model above defines what Uebersax (1993) labels as the basic model. Note that identification restrictions are still required and discussed by Uebersax. He defines a model where the category widths $\tau_{m+1,r} - \tau_{mr}$ are the same across raters as the simple bias model. If we

further restrict thresholds to be equal across raters we have what he labels as the identical thresholds model.

Using Uebersax's LLCA program, we can specify these models with either equal or unequal error variances in the manifest variables. For the case with three latent classes, the following table displays the corresponding likelihood ratio chi-square statistics:

Table A.5.5 Likelihood Ratio Statistics for Three Latent Classes

Likelihood Ratio Test Statistics	Different Measurement Errors		Equal Measurement Errors	
	L.R. χ^2	Df	L.R. χ^2	Df
Basic Model	9.52	14	15.65	16
Simple Bias	11.60	16	16.84	18
Identical Threshholds	40.55	18	87.84	20

Note that, given the basic model, either assuming different or equal measurement errors, the likelihood ratio tests of the restriction of equal category differences (i.e. the further restriction of the simple bias model over the simple bias model) are not statistically significant. That is,

observed $\chi^2(2\ df) = 11.60 - 9.52 = 2.08$ ($p \leq 0.3535$) and observed $\chi^2(2\ df) = 16.84 - 15.65 = 1.19$ ($p \leq 0.5516$), respectively.

ictio

In both cases we would accept the hypothesis of equal category differences. However, with or without equal measurement errors, the further restriction implying identical thresholds is strongly rejected, i.e., $\chi^2(2\ df) = 40.55 - 11.60 = 28.95$ ($p < 0.0001$) and observed $\chi^2(2\ df) = 87.84 - 16.84 = 71$($p << 0.0001$), respectively

Within each of the first two models, the results testing the assumption of equal measurement error are equivocal. That is,

observed $\chi^2(2\ df) = 15.65 - 9.52 = 6.13$ ($p \leq 0.0467$) and observed $\chi^2(2\ df) = 16.84 - 11.60 = 5.24$ ($p \leq 0.0728$), respectively.

The AIC and BIC are minimized for the simple bias model specified with different measurement errors, and were used to determine the final model. A summary of the relevant LLCA output, slightly annotated, for this simple bias model with different measurement errors, is presented as follows:

```
3. PARAMETER ESTIMATES

   a. Latent class location parameters
      1     -1.500   (fixed for identification)
      2     -4.605   (estimated cut-point)
      3      1.500   (fixed for identification)

   b. Latent class probabilities
      1      0.211
```

2	0.718
3	0.071

c. Measurement error

Item/ rater	alpha	rho
1	0.789	0.828
2	0.493	0.678
3	0.670	0.781
Mean	0.651	0.762

d. Threshold parameters (for three classes)

Item/ rater	Manifest category	Threshold estimate
all	2	-2.165
all	3	1.248

Bias for item/rater 1: -1.061 (Dr. ———)
Bias for item/rater 2: 1.389 (Dr. ———)
Bias for item/rater 3: -0.328 (Dr. ———)

This confirms the notion that Dr. ——— scores are generally the lowest among the three raters, with Dr. ——— the highest.

4. OBSERVED/EXPECTED FREQUENCIES

Cell	Observed frequency	Expected frequency	Rating Pattern
1	170.	170.45	1 1 1
2	16.	17.61	1 1 2
3	1.	0.57	1 1 3
4	9.	6.89	1 2 1
5	1.	1.89	1 2 2
6	1.	0.14	1 2 3
7	35.	34.77	2 1 1
8	27.	26.55	2 1 2
9	2.	2.40	2 1 3
10	6.	5.22	2 2 1
11	13.	12.71	2 2 2
12	2.	2.28	2 2 3
13	1.	1.42	2 3 2
14	1.	0.68	2 3 3
15	1.	1.32	3 1 1
16	3.	3.61	3 1 2
17	3.	1.84	3 1 3
18	1.	0.61	3 2 1
19	6.	5.06	3 2 2
20	7.	7.47	3 2 3
21	4.	3.46	3 3 3

Total number of cases	310.
Total expected frequencies	306.9463

The generally close similarity between the observed frequencies and those expected under the model suggests strong model fit.

5. CONDITIONAL RATING PROBABILITIES

Item/ rater	Latent class	Manifest category	Conditional probability
1	1	1	0.0898
1	1	2	0.8160
1	1	3	0.0942
1	2	1	0.8642
1	2	2	0.1341
1	2	3	0.0016
1	3	1	0.0018
1	3	2	0.1447
1	3	3	0.8536
2	1	1	0.6474
2	1	2	0.3224
2	1	3	0.0302
2	2	1	0.9613
2	2	2	0.0364
2	2	3	0.0023
2	3	1	0.1292
2	3	2	0.5926
2	3	3	0.2782
3	1	1	0.2442
3	1	2	0.6960
3	1	3	0.0598
3	2	1	0.9172
3	2	2	0.0809
3	2	3	0.0019
3	3	1	0.0105
3	3	2	0.3301
3	3	3	0.6594

6. CLASSIFICATION/SCORING (only included for completeness)

Cell	Observed frequency	Assigned class	Modal proba-bility	Latent trait score	Rating Pattern
1	170.	2	0.9945	-4.5882	1 1 1
2	16.	2	0.8494	-4.1373	1 1 2
3	1.	2	0.5970	-3.3364	1 1 3
4	9.	2	0.9327	-4.3961	1 2 1
5	1.	1	0.6968	-2.4172	1 2 2
6	1.	1	0.8015	-1.4641	1 2 3
7	35.	2	0.7568	-3.8496	2 1 1
8	27.	1	0.9074	-1.7563	2 1 2

9	2.	1	0.8645	-1.2291	2 1 3
10	6.	1	0.8054	-2.0811	2 2 1
11	13.	1	0.9440	-1.3744	2 2 2
12	2.	3	0.5464	0.1366	2 2 3
13	1.	1	0.7903	-0.8948	2 3 2
14	1.	3	0.8580	1.0734	2 3 3
15	1.	1	0.7409	-2.1862	3 1 1
16	3.	1	0.7703	-0.8581	3 1 2
17	3.	3	0.8697	1.1082	3 1 3
18	1.	1	0.7902	-0.9894	3 2 1
19	6.	3	0.7262	0.6780	3 2 2
20	7.	3	0.9840	1.4521	3 2 3
21	4.	3	0.9968	1.4903	3 3 3

Percent correctly classified: 0.9271 (Thus the classes reproduce the observed assignment quite well.)

Percent assigned to:

Latent Class 1	0.1806
Latent Class 2	0.7452
Latent Class 3	0.0742

Classification/membership joint probabilities:

Latent Class	Assigned Latent Class 1	2	3
1	0.1596	0.0414	0.0103
2	0.0139	0.7037	0.0000
3	0.0072	0.0000	0.0638

So, at least for this model, the output section 4. shows the model has good fit. Output section 3.d shows that there are statistically significant differences between raters, with Dr. —— s scores generally the lowest among the three raters, and Dr ——'s the highest.

References:

Landis, J.R. and Koch G.G. (1977), "The measurement of observer agreement for categorical data," *Biometrics*, **33**, 159-174.

Uebersax, J.S. (1993), "Statistical modeling of expert ratings on medical treatment appropriateness," *Journal of the American Statistical Association*, **88**, 421-427.

Appendix 6. Sponsor Responses to Information Requests on Hair Measurement Techniques:

On August 22, 2005 the Sponsor sent the following responses to clinical/statistical questions sent on August 12, 2005:

---

1. What is the cutoff of diameter for hair measurement using dot mapping technique (smallest hair diameter measured)?

   The cutoff diameter is 0.03 mm for the dot mapping technique used.

2. Could you please provide a reanalysis of the data using a hair counting technique that excludes hair with a diameter of less than 0.03 mm and an additional analysis excluding hair of less than 0.05 mm diameter?

   Reanalysis of the data using diameter of hair as cut points is not possible using the photographs taken in the trial (MINO-9140-006). While the hair counting technique used permits the establishment of a lower threshold (0.03 mm), it does not permit direct measurement of individual hair diameter.

   There are newer hair counting technologies that permit concurrent measurement of individual hair diameter and target area hair count. These other hair counting techniques were not included in this trial as the trial was designed to use an established, validated and published technique that was accepted by the Agency for the approval of both Propecia and Minoxidil products.

3. Provide a graphic plot showing the distribution of hair diameters vs. counts for their given study population with a comparison between the different arms?

   Graphic plots showing distribution of hair diameters can not be provided since individual hair diameters can not be measured. See response to question 2 above.

4. Provide a graphic plot showing the distribution of hair diameters vs. counts for the study population with a comparison between the different arms.

   Graphic plots showing distribution of hair diameters can not be provided since individual hair diameters can not be measured. See response to question 2 above.

The Review team requested a further teleconference with the Sponsor and requested further analysis on these, as well as other, issues. On September 2, 2005, Tte Sponsor sent the following response:

1) What is the support for 0.03 mm (30 microns) as the appropriate diameter threshold for identifying non-vellus hairs?

Androgenetic alopecia (AGA) is characterized by follicular miniaturization and by affecting the hair cycle, resulting in a reduction in the length, diameter and overall hair density of normal, non-vellus (visualized) hairs. Minoxidil reverses miniaturization of non-vellus follicles, shifting vellus-like hairs in those follicles to non-vellus hairs. Vellus-like hairs are those hairs produced from a previously normal non-vellus producing follicle that has been miniaturized due to androgenetic alopecia and, without treatment, is no longer able to produce normal non-vellus hairs. Our clinical methodologies were designed to assess the regrowth of non-vellus hairs based on the definitions which follow.

Scalp hairs are defined as non-vellus, vellus-like, or vellus, based on the parameters of length, pigmentation and diameter as shown in Table 1.1. As with many classifications in the biological sciences, these distinctions are not absolute, and ranges, exceptions and transitions do exist.

Table 1.1: Definition of Non-Vellus, Vellus-like and Vellus Hairs

Parameter	Non-Vellus	Vellus-like	Vellus
Length	> 1 cm	< 1 cm	< 1 cm
Pigmentation	Present	Absent	Absent
Diameter	≥ 0.03mm	< 0.03mm	< 0.03mm

Hair follicles are defined as non-vellus, miniaturized non-vellus and vellus, based on parameters of inner root sheath/hair diameter ratio, location of stela in the dermal layer, and size and location of hair bulb as shown in Table 1.2.

Table 1.2: Definition of Non-Vellus, Miniaturized Non-Vellus and Vellus Follicles

Parameter	Non-Vellus	Miniaturized Non-Vellus	Vellus
Inner root sheath: hair diameter	Hair diameter > thickness of inner root sheath	Hair diameter may be< thickness of inner root sheath	Hair diameter < thickness of inner root sheath
Location of stela	Lower dermis and subcutaneous tissue	Lower dermis and subcutaneous tissue	Upper dermis
Size and location of hair bulb in dermis (depending on stage of hair cycle)	Lower dermis or subcutaneous tissue	Upper or mid levels of dermis	Upper levels of dermis

As noted in the Table 1.1 above, 0.03 mm is accepted as the upper limit diameter for vellus hairs. For example, Hordinsky et al states "Small hairs, with no pigment or medullary cavity, a diameter less than 0.03 mm, and a length of less than 1 cm, are classified as vellus (downy) hairs."[i] Whiting states that "Vellus hairs are inconspicuous and are 0.03 mm or less in diameter and often less than 1 cm in length

and lack melanin and medulla".[ii] Additional references [iii iv v] are provided as well as the photocopies of the specific sections of these references for your review.

Based on the definitions outlined above and the literature references provided, Pfizer believes 0.03 mm is the appropriate diameter threshold for identifying visualized non-vellus hairs.

2) How was the 5.7 fold final magnification determined to be the appropriate magnification at which only non-vellus hairs would be visualized, and thereby counted, in the hair counting methodology used?

Over the past twenty years, hair re-growth has been evaluated using a consistent approach of counting visualized hairs, first by direct observation and then with the use of photographs magnified to a sufficient extent to exclude the visualization, and thereby counting, of vellus hairs. The technology used in Pfizer clinical trial MINOB-9140-006 (006), which has been the basis of regulatory approvals, is this current technology of photographic magnification. Of course without a diameter measurement of each individual hair, insignificant numbers of vellus and miniaturized non-vellus hairs may be counted. The technology to measure the diameter of each individual hair to distinguish between non-vellus and vellus hairs is being explored for future use.

A publication by Kaufman et al [vi], reporting on a study to evaluate the efficacy of Finasteride, utilized the photographic magnification technique. Macro photographs were taken using a 35 mm film camera system with fixed settings and the resulting photographs were enlarged sufficiently to permit counting of non-vellus hairs. Table 2.1 compares the techniques used in the Pfizer 006 study to that of the technique in the Kaufman published study.

Table 2.1: Comparison of Finasteride Study Technique vs Minoxidil 006 Study Technique

Technique	Finasteride Study	Minoxidil 006 Study
Macro photographs	Yes	Yes
Preset 35mm film SLR camera system	Yes	Yes
Final magnification	5.7*	5.7

* [redacted] has confirmed that a final magnification of 5.7 fold was used for both Pfizer's 006 and Merck Finasteride studies. Final enlargement is calculated by taking the area of the enlarged device template, multiplying this by the un-enlarged area of the scalp to be counted, then dividing this by the un-enlarged area of the device template. For the Merck studies, these areas were rounded to one decimal place before calculation resulting in a reported final magnification of 5.84 fold (rounded to two decimal places). For the Pfizer studies, all numbers were rounded to four decimal places before calculation resulting in a reported magnification of 5.7 fold (rounded to one decimal place). The 5.7 fold final magnification is confirmed by the direct measurement of an internal standard which is in every macro photograph.

It is Pfizer's belief that this study was one of the pivotal studies used to approve Finasteride as noted in the Summary Basis of Approval (SBA). Pfizer also received assurances from [redacted] technical/medical imaging service provider), that the photographic equipment and hair counting technique used in the Pfizer trial was the same technique used in the Finasteride trial. Pfizer therefore initiated the 006 trial with the understanding that this methodology is currently accepted by Regulatory Authorities to support the approval of hair regrowth products.

Pfizer participated in exploratory research on hair count methodology in a study comparing hair counts from the technique used in the 006 study to that of the newer technology, which determines actual hair diameter measurements. Because this was an exploratory study, the subject numbers are limited; however, the results are important in supporting the magnification used in 006 and the accuracy of counting non-vellus hairs.

The results of this study were presented as a poster [vii] at the recent European Hair Research Society meeting in July, 2005. This study compared the number of non-vellus hairs ($\geq$ 0.03 mm) counted with the magnification technique to the actual number of hairs with a diameter of $\geq$ 0.03 mm as measured by the newer technology using digital techniques and image analysis. The results of this study showed the mean target area hair count using the 5.7 fold magnification technique was 169.1 and the number of hairs in the same target area with a measured diameter of $\geq$ 0.03 mm was 166.6. Results of this study lend support to 5.7 fold as the magnification level which yields visualization and counting of non-vellus hairs in the photographic magnification technique.

There is further support of the appropriateness of this magnification level when one compares the total number of non-vellus hairs counted in target areas from the leading anterior edge of the vertex area from patients with androgenetic alopecia in different studies. The non-vellus hair count from the Finasteride study was 175/cm^2 and the number of non-vellus hairs determined by histology from biopsies was 167/cm^2 (both of which were derived from the number of non-vellus hairs counted and then interpolated to the common target area of 1 cm^2). The baseline non-vellus counts for the Pfizer 006 study were 169/cm^2 and 171/cm^2 (for placebo and active group respectively). Table 2.2 summarizes these findings.

Table 2.2: Hair Methodology Baseline Comparison Chart

Methodology	Source	N	Avg Hair Counts Per cm^2
Photographic Magnification *	Kaufman et al. [6]	1553	175
Biopsy	Whiting [viii]	278	167
Photographic Magnification *	Pfizer 006 placebo active	172 180	169 171

*35mm traditional hair count by [redacted]

In summary, based on all of the above support, Pfizer believes that the magnification level of 5.7 fold yields visualization and thereby counting of non-vellus hairs (i.e. those $\geq$ 0.03 mm in diameter) and adequately filters out vellus and insignificant miniaturized non-vellus hairs (i.e. those < 0.03 mm in diameter).

References

[i] Hordinsky M, Sawaya M, Scher R: Atlas of Hair and Nails. Philadelphia, PA, Churchill Livingstone, 2000, p 10.

[ii] Whiting DA: The Structure of the Human Hair Follicle. Fairfield, NJ, Canfield Publishing, 2004, p 5.

[iii] Unger W, Shapiro J: Hair Transplantation. Marcel Dekker, Inc, 2004, p 31-32.

[iv] Shapiro J: Hair loss: Principles of Diagnosis and Management of Alopecia . Martin Dunitz, 2002.

[v] Olsen E (Editor): Disorders of Hair Growth Second Edition – McGraw-Hill Companies, Inc., 2003, p 7.

[vi] Kaufman K. et al. Finasteride in the treatment of men with androgenetic alopecia. JAAD 1998:39:578-589

[vii] Kohut B. et al. A Methodology Study Comparing Traditional 35mm Hair Counts to Automated Image Analysis Measurements, and Assessing Visualization Sensitivity of Hair Dyeing when Quantifying Hair Loss in Men and Women with Androgenetic Alopecia. Poster # 25. Presented at the European Hair Research Society Meeting Zurich July 2005.

[viii] Whiting DA: The Structure of the Human Hair Follicle. Fairfield, NJ, Canfield Publishing, 2004, p 26.

SIGNATURES/DISTRIBUTION LIST

Primary Statistical Reviewer: Steve Thomson
Date: 12 December 2005

Concurring Reviewer:

Statistical Team Leader: Mohamed Alosh, Ph.D., HFD-725

cc:
HFD-540/Phyllis Huene, M.D.
HFD-560/Daiva Shetty, M.D.
HFD-560/Tia Frazier.
HFD-540/Markham Luke, M.D., Ph.D.
HFD-725/Steve Thomson
HFD-725/ Mohamed Alosh, Ph.D.
HFD-725/Stan Lin, Ph.D.
HFD-725/Mohammed Huque, Ph.D.,
HFD-700/Robert O'Neill, Ph.D.
HFD-700/Lillian Patrician

c:\Data\Profiles\MyDocuments~\N21812S000RogainePfizer.doc

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
12/19/2005 08:35:05 AM
BIOMETRICS

Mohamed Alosh
12/19/2005 09:08:54 AM
BIOMETRICS
Concur with review

Appears This Way
On Original