CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

64160

STATISTICAL REVIEW(S)

4/26/99

Statistical Review: ANDA 64-160, Clindamycin Phosphate Gel USP, 1%, Altana Inc.

Material reviewed:
1. Red-Jacketed Volumes 3.1 through 3.7 of ANDA 64-160.
2. October 13, 1999 Medical Officer Review by
   Mary M. Fanning, M.D., Ph.D., Associate Director
   for Medical Affairs, Office of Generic Drugs.
3. Data files provided by the Sponsor on disk (ZIP Drive cartridge)

The issues in this review involve the sponsor's clinical bioequivalence study comparing their product, Clindamycin Phosphate Gel USP 1% from Altana, Inc., to the reference listed drug product, Cleocin T Gel 1%(Pharmacia & Upjohn), and to a vehicle control. The active products are for the treatment of Acne Vulgaris.

Based on discussions with the Medical Officer (Dr. Fanning) this statistical review is limited to The primary efficacy variables. These are:

1. Percent Change From Baseline (PCB) at the final visit for three lesion counts: inflammatory lesions, noninflammatory lesions, and total lesions.

2. Physician's Global Assessment at the final visit.

Study Design

The study was a parallel group design carried out at 8 clinical sites. The three treatments studied were

treatment 1. Altana's Clindamycin (Test product)

treatment 2. Cleocin T (Pharmacia & Upjohn, Reference product)

treatment 3. vehicle control

Within a site, subjects were assigned to treatments 1, 2, or 3 in approximately a 2:2:1 ratio.

The protocol called for patients to have 5 visits for assessment, nominally at 0, 21, 42, 63, and 84 days, the first visit (day 0) being the baseline visit.

Analysis Subsets

Two subsets of participating patients were specified for analysis of the primary efficacy variables: The Modified Intent-To-Treat (MITT) population consisted of patients who met all entry criteria and had at least one post-baseline visit. The Per Protocol population consisted of patients who met all entry criteria and completed all visits required by the protocol or were discontinued early due to adverse events or lack of efficacy. Missing visit data were supplied by the last observation carried forward (LOCF) method, except that the baseline visit was never carried forward.

The numbers of patients available for analysis were as follows

lesion count PCB

	MITT			Per Protocol		
	treatment			treatment		
site	1	2	3	1	2	3
1 (Davis)	22	24	12	21	23	12
2 (Kempers)	22	23	12	21	19	9
3 (Kaplan)	24	24	12	18	22	11
4 (Leyden)	31	32	16	29	30	12
5 (Lucky)	22	23	11	20	20	10
6 (Reyes)	24	24	12	23	19	12
7 (Rich)	28	29	15	25	22	10
8 (Savin)	16	16	8	12	15	8
total	189	195	98	169	170	84

physician's global assessment

	MITT			Per Protocol		
	treatment			treatment		
site	1	2	3	1	2	3
1 (Davis)	23	24	12	21	23	12
2 (Kempers)	22	23	12	21	19	9
3 (Kaplan)	24	24	12	18	22	11
4 (Leyden)	31	32	16	29	30	12
5 (Lucky)	22	23	11	20	20	10
6 (Reyes)	24	24	12	23	19	12
7 (Rich)	27	24	15	24	21	10
8 (Savin)	16	16	8	12	15	8
total	189	190	98	168	169	84

Lesion Count PCB

The sponsor has chosen to express the lesion count data as percent change from baseline (PCB), i.e.

100*(count - baseline count)/baseline count

In the case of noninflammatory lesions, for which the baseline count was zero in some cases, the sponsor adopted the convention of treating the baseline count as 1 if it was zero, for purposes of calculating PCB.

The distribution across patients of the three variables - inflammatory lesion PCB at the final visit (PCBINF), noninflammatory lesion PCB at the final visit (PCBNON), and total lesion PCB at the final visit (PCBTOT) - was markedly skewed, with a long "tail" of higher positive values. Such a highly skewed distribution has consequences for the choice of statistical methods. One consequence is that p-values derived from ANOVA/linear models type analyses, based on an assumption of a normal distribution, may be inaccurate. Another consequence, relevant to the determination of average bioequivalence, is that the concepts of "average", "central tendency", and "typical value" are not so well defined. It is generally accepted that for highly skewed distributions the median of the distribution is a more reasonable measure of location, or "average", than the mean (There is an excellent discussion of this in the classic text *How To Lie With Statistics* by Darrell Huff [W.W. Norton & Co., New York, 1954], in the chapter titled "The Well-Chosen Average".) We may see the discrepancy between mean and median for these lesion count PCB variables by calculating the sample means and medians for each treatment (ignoring site) for the MITT population:

PCBINF	mean	median
treatment 1	-38.03	-50.00
treatment 2	-40.47	-47.37
treatment 3	-24.71	-31.17
PCBNON		
treatment 1	-1.81	-20.24
treatment 2	11.08	-31.25
treatment 3	3.45	-18.35
PCBTOT		
treatment 1	-29.50	-37.50
treatment 2	-31.35	-38.46
treatment 3	-18.92	-21.69

The discrepancy between mean and median in this data is most obvious for the PCBNON variable, for which the mean is positive (indicating a worsening of the condition over the course of the study) for treatments 2 and 3. The means for PCBNON would seem to indicate that treatment 2 (the Reference product) did worse than treatment 3, the vehicle control. The shifting of the mean toward higher values (either positive or less negative) relative to the median is apparent for all three variables.

The statistical method that the sponsor has chosen to deal with the skewness in this data is the Rank Transformation Method. This method, which is attributed to the statistician W. J. Conover, consists of the following:

1. Take the dataset and rank the response variable (in this case PCBINF, PCBNON, or PCBTOT) values from lowest to highest, without regard to any classification variables such as treatment or site, with tied observations receiving averaged ranks.

2. Replace the actual values in the dataset with their ranks.

3. Use the ANOVA/linear models method of your choice to analyze these rank data.

If the statistical model used in the analysis is a simple one, for example only including treatment as an explanatory variable, ignoring site, then this Rank Transformation Method corresponds at least approximately to standard nonparametric rank tests such as the Wilcoxon Rank Sum test (if the ranks are analyzed two treatments at a time) or the Kruskal-Wallis test (if all three treatments are analyzed at once). However, the Rank Transformation Method permits us to use a more complicated statistical model.

In using the Rank Transformation Method to analyze the lesion count PCB data, the sponsor is actually making inferences about the means of the ranks, which is closer to making inferences about the medians of the actual response variables.

Lesion Count PCB - choice of statistical model

The sponsor has chosen to include treatment (TRT), clinical site (SITE), and the interaction between TRT and SITE (TRT*SITE) in their statistical model. They claim that TRT*SITE was statistically significant for lesion count PCB at some time points. Based on my own analyses using the Rank Transformation, this must have been for some of the secondary efficacy variables (lesion count PCB at earlier visits), since TRT*SITE is not statistically significant for any of the three primary lesion count PCB variables PCBINF, PCBNON, and PCBTOT. On the other hand, SITE as a factor is highly statistically significant for PCBINF, PCBNON, and PCBTOT.

I will report analyses both with and without TRT*SITE, with TRT and SITE as factors in both cases.

Lesion Count PCB - do the active treatments beat vehicle control?

The p-values I have obtained for the three pairwise comparisons of treatments, using the Rank Transformation, are as follows:

MITT dataset p-values

	treatment	with	without
	comparison	TRT*SITE	TRT*SITE
PCBINF	1 vs. 2	0.7426	0.9271
	1 vs. 3	0.0018	0.0018
	2 vs. 3	0.0040	0.0022
PCBNON	1 vs. 2	0.5000	0.4309
	1 vs. 3	0.0444	0.0626
	2 vs. 3	0.0102	0.0119
PCBTOT	1 vs. 2	0.8065	0.6832
	1 vs. 3	0.0016	0.0027
	2 vs. 3	0.0007	0.0008

Per Protocol dataset p-values

	treatment	with	without
	comparison	TRT*SITE	TRT*SITE
PCBINF	1 vs. 2	0.7251	0.9765
	1 vs. 3	0.0007	0.0011
	2 vs. 3	0.0017	0.0010
PCBNON	1 vs. 2	0.2711	0.2359
	1 vs. 3	0.0489	0.0615
	2 vs. 3	0.0040	0.0046
PCBTOT	1 vs. 2	0.5867	0.4320
	1 vs. 3	0.0014	0.0025
	2 vs. 3	0.0003	0.0003

Note that for PCBNON, the comparison of 1 vs. 3 (Altana's Clindamycin vs. vehicle control) is statistically significant at the usual $\alpha=0.05$ level only for the model with TRT*SITE, both for the MITT dataset and for the Per Protocol dataset. In all other cases, both treatment 1 and treatment 2 are statistically significantly different from treatment 3.

Note that these p-values for the model with TRT*SITE differ somewhat from the p-values reported by the sponsor. I cannot discover a reason for this discrepancy in the material reviewed. The important thing is that the sponsor's p-values and my p-values are qualitatively similar. Both sets of p-values indicate that if TRT*SITE is included in the statistical model, both active treatments are statistically significantly different from vehicle control for PCBINF, PCBNON, and PCBTOT for both the MITT dataset and the Per Protocol dataset.

Lesion Count PCB - are Test and Reference equivalent?

As seen in the tables of p-values above, treatments 1 (Test product) and 2 (Reference product) were not statistically significantly different in any case. However, we have long known that just because treatments are not statistically significantly different, that does not necessarily mean that they are equivalent.

It is possible to obtain a 90% confidence interval for the difference between the average of treatment 1 and the average of treatment 2 ("average" as estimated by the Rank Transformation analyses) using the Rank Transformation method. This is accomplished by adding a constant, call it C, to all observations from one of the treatments. The Rank Transformation analysis is then run on this modified data and the p-value for 1 vs. 2 is noted. By varying the value of C, it is possible to identify all values of C for which the p-value of 1 vs. 2 is greater than or equal to 0.10. This set of C values constitutes a 90% confidence interval for the difference between the average of treatment 1 and the average of treatment 2.

The 90% confidence intervals I obtained using this method are as follows. I have limited the analyses to the statistical model with TRT*SITE, the model favored by the sponsor:

90% C.I. for the difference average(trt.1)-average(trt.2),
derived from Conover's Rank Transformation

	MITT	Per Protocol
PCBINF	-7.9 , 5.2	-8.4 , 5.5
PCBNON	-5.0 , 12.4	-2.9 , 15.4
PCBTOT	-5.2 , 7.2	-4.4 , 8.9

I have used a similar method to obtain a 90% confidence interval for the ratio of the average for treatment 1 over the average for treatment 2. All observations from the Reference product are MULTIPLIED by a constant, call it k. The Rank Transformation analysis is then run on this modified data and the p-value for 1 vs. 2 is noted. By varying the value of the multiplier k, it is possible to identify all values of k for which the p-value of 1 vs. 2 is greater than or equal to 0.10. This set of k values constitutes a 90% confidence interval for the ratio of the average of treatment 1 over the average of treatment 2. Note that multiplying the data values by a constant, rather than adding a constant, would be expected to change the spread among the values. For this reason, this method may be questionable if the values of k differ very much from 1.0.

The 90% confidence intervals I obtained using this method are as follows. Once again, I have limited the analyses to the statistical model with TRT*SITE:

90% C.I. for the ratio average(trt.1)/average(trt.2), derived
from Conover's Rank Transformation

	MITT	Per Protocol
PCBINF	91% , 116%	91% , 117%
PCBNON	69% , 114%	64% , 108%
PCBTOT	83% , 114%	80% , 112%

In the case of blood-level *in vivo* bioequivalence studies, it is standard to make the inference about the ratio of the averages, rather than the difference. However, in the case of lesion count PCB, the argument could possibly be made that the expression of the comparison into relative rather than absolute terms has already been made by expressing the lesion counts as percent changes. In the case of the 90% confidence intervals on the difference given above, the units for the confidence limits are the same as the units for the variables, namely percentage points of change from baseline.

If we compare the confidence limits for the ratio with the usual "goalposts" of 80% to 125% used in blood-level BE studies, we would have a problem with PCBNON, but PCBINF and PCBTOT would pass. Of course, "goalposts" other than 80% to 125% may be appropriate for clinical outcomes such as this. Looking at the confidence limits for the difference, the Per Protocol confidence interval for PCBNON indicates that the two averages may differ by as much as 15.4 percentage points.

Physician's Global Assessment at last visit

The six-point Physician's Global Assessment scale was defined as follows:

rating	definition
0	unchanged or worsened, compared to baseline
1	poor response, 1-24% improvement compared to baseline
2	fair response, 24-49% improvement compared to baseline
3	good response, 50-74% improvement compared to baseline
4	excellent response, 75-99% improvement compared to baseline
5	completely cleared, 100% improvement, defined as no papules, pustules, comedones, or nodulocystic lesions

The sponsor has analyzed the Physician's Global Assessment at the last visit (PHYGLOB) variable using analysis of variance methods. That is, they have analyzed it as a continuous variable. Note that they have analyzed the PHYGLOB ratings themselves, not the ranks as was the case with the lesion count PCB variables.

In this case, the sponsor has chosen to use a statistical model without the TRT*SITE statistical interaction, claiming that this interaction was not statistically significant at any visit. I will report analyses both with and without TRT*SITE in the model.

For the question "Did the active treatments beat vehicle control?", we may look at the p-values resulting from the ANOVA analysis.

PHYGLOB
MITT dataset
477 observations

	p-values	
	with TRT*SITE	without TRT*SITE
1 vs. 2	0.8557	0.9880
1 vs. 3	0.0091	0.0102
2 vs. 3	0.0137	0.0098

Per Protocol dataset
421 observations

	p-values	
	with TRT*SITE	without TRT*SITE
1 vs. 2	0.7965	0.9199
1 vs. 3	0.0017	0.0035
2 vs. 3	0.0033	0.0027

Using this analysis, the difference between either active treatment and the vehicle control is highly statistically significant.

For the question "Are the Test and Reference products equivalent?", we may calculate 90% confidence intervals for either the difference between the averages or the ratio of the averages. Confidence intervals for the ratio were calculated using Fieller's method. Let μT be the average PHYGLOB for treatment 1 and μR be the average PHYGLOB for treatment 2. In Fieller's method, the confidence interval consists of all values of k such that μT-k*μR was not statistically significantly different from zero at the $\alpha=0.10$ level of significance. The results are:

PHYGLOB

90% Confidence Intervals for the ratio $\mu T/\mu R$

	MITT (n=477)	Per Protocol (n=421)
with TRT*SITE	91.6% , 111.6%	91.7% , 112.6%
without TRT*SITE	90.7% , 110.0%	89.9% , 109.9%

90% Confidence Intervals for the difference $\mu T-\mu R$

	MITT (n=477)	Per Protocol (n=421)
with TRT*SITE	-0.1962 , 0.2449	-0.1984 , 0.2721
without TRT*SITE	-0.2175 , 0.2136	-0.2439 , 0.2158

The units of the confidence limits on the difference $\mu T-\mu R$ are the same as the units of the PHYGLOB rating scale.

If we compare the confidence limits on the ratio with the usual 80% to 125% "goalposts" used in blood-level BE studies, the products would be considered equivalent with respect to PHYGLOB. Of course, different "goalposts" may be appropriate for this variable.

If you prefer to compare the products in absolute terms, the confidence limits on the difference indicate that the average PHYGLOB rating for the two products may differ at most by a little more than a quarter of a rating point (0.2721 rating point, to be exact). Using the statistical model without TRT*SITE, as done by the sponsor, we would conclude that the averages differ at most by less than a quarter of a rating point (-0.2439 rating point, to be exact).

Alternate Equivalence Analyses of PHYGLOB

There may be at least two possible objections to analyzing the categorical variable PHYGLOB as a continuous variable, as has been done by the sponsor. The first is that the assumption of a normally distributed response is not strictly true. However, I do not feel that this is a serious objection for this dataset, which is not highly skewed. A second more serious objection is that this analysis implicitly assumes that the spacing between the six possible rating values is equal. That is, a rating of 1 differs from a rating of 0 by the same amount as, for example, a rating of 4 differs from a rating of 3. By assigning actual percentage improvement values to the definitions of each rating value the sponsor has improved the validity of this equal-spacing assumption, with the possible exception of the lowest and highest ratings.

If we wish to avoid this equal-spacing assumption, we might try using statistical methods appropriate to categorical data. If we ignore site, the PHYGLOB results for the two analysis subsets may be presented in tabular form, with the entries in the cells of the table being the number of subjects with the indicated PHYGLOB rating:

MITT dataset (n=477)

	PHYGLOB rating						
	0	1	2	3	4	5	totals
treatment 1	29	33	43	41	41	2	189
treatment 2	28	34	41	48	37	2	190
treatment 3	25	17	21	24	11	0	98

Per Protocol dataset (n=421)

	PHYGLOB rating						
	0	1	2	3	4	5	totals
treatment 1	28	24	38	38	39	1	168
treatment 2	24	29	34	45	35	2	169
treatment 3	20	16	19	20	9	0	84

Looking at these tables, it is evident that the numbers of subjects receiving a rating of 5 are rather sparse. Some statisticians would advocate combining the totals from the 4 and 5 ratings.

One possible approach to the analysis of this data would be to dichotomize the rating. With this 6-point scale, there are five ways to do that:

0	vs.	1 or higher
0 or 1	vs.	2 or higher
2 or lower	vs.	3 or higher
3 or lower	vs.	4 or 5
4 or lower	vs.	5

Let us define P(i|t) as the probability of obtaining a PHYGLOB rating less than or equal to i after receiving treatment t. For each of these five possible dichotomies, we may calculate a 90% confidence interval for the difference between P(i|1), the probability of seeing the lower rating using the Test product, and P(i|2), the probability of seeing the lower rating using the Reference product. For these calculations, I have pooled the observations from all eight sites. The confidence intervals were calculated using the Wald method with Yates continuity correction (as such, they do not use the information from treatment 3, the vehicle control).

90% confidence intervals for P(i|1)-P(i|2),
where P(i|1) = probability of a rating ≤ i using the Test product,
and P(i|2) = probability of a rating ≤ i using the Reference product.

i	dichotomy	MITT	Per Protocol
0	0 vs. 1 or higher	-6.0 , 7.2	-4.6 , 9.5
1	0 or 1 vs. 2 or higher	-8.3 , 8.6	-9.3 , 8.5
2	2 or lower vs. 3 or higher	-7.6 , 10.3	-7.5 , 11.6
3	3 or lower vs. 4 or 5	-9.7 , 5.3	-10.0 , 6.2
4	4 or lower vs. 5	-2.3 , 2.3	-1.7 , 2.9

The units of these confidence limits are probability units, expressed in percentage points.

If we compare these confidence limits to the usual "goalposts" of plus-or-minus 20 percentage points, as used for example in clinical BE studies of antiinfective products, the active treatments would be considered equivalent no matter which dichotomy we choose.

Another possible equivalence analysis approach for ordered categorical data such as the PHYGLOB data is the proportional odds ratio method as implemented in PROC LOGISTIC of SAS. Use of this method is based on an assumption of proportional odds ratios. That is, the assumption that the odds ratio

$$\frac{P(i|1)*(1-P(i|2))}{P(i|2)*(1-P(i|1))}$$

is constant for i = 0, 1, 2, 3, and 4. The LOGISTIC procedure in SAS provides a statistical test of this assumption, and the p-values for this test were 0.8454 for the MITT dataset and 0.8733 for the Per Protocol dataset. While the fact that these p-values are nowhere near statistically significant at the usual 0.05 (or even the 0.10) level of significance does not in itself establish that the proportional odds ratio assumption is true, it at least shows that there is nothing in the datasets themselves to indicate that the assumption is obviously false.

The 90% confidence intervals for the common odds ratio of treatment 1 compared to treatment 2 (using the Profile Likelihood method) using SAS PROC LOGISTIC are as follows

dataset	estimated odds ratio	90% confidence interval
MITT dataset	0.997	0.739 , 1.344
Per Protocol dataset	1.026	0.746 , 1.412

To our knowledge, there is currently no established method in OGD for setting equivalence limits when using the proportional odds method with ordered categorical endpoints. QMR has sometimes used equivalence limits of 3/7 (0.429) to 7/3 (2.333) to give an order of magnitude assessment of equivalence, acknowledging that these limits on the odds ratio may be too stringent. In the present case, the 90% confidence intervals for the common odds ratio fall comfortably within these 3/7-7/3 limits for both the MITT and the Per Protocol datasets.

In the proportional odds ratio analyses just described, I pooled the data from all eight sites, using treatment as the only explanatory variable in the logistic regression model. It is possible to include site as an explanatory variable as well, but when I did this the p-value from the statistical test of the proportional odds assumption was significant or near significant. When I ran the analysis anyway, including site as an explanatory variable, the results were qualitatively quite similar to those reported above. Also, the analyses reported above used the data from all three treatments, including the vehicle control. This differs from the analyses of the dichotomized rating, in which only treatments 1 and 2 were used. If treatments 1 and 2 data only are used in the proportional odds analysis, the results are almost identical to those reported above.

Summary

1. The extreme skewness of the lesion count percentage change from baseline (PCB) data led the sponsor to use Conover's Rank Transformation method to analyze the data. I have used this method in my analyses as well. Using this method, inferences concerning "average" equivalence are aimed more at the medians of the distributions rather than the means. This is reasonable, in my opinion, for this data.

2. p-values obtained from the Rank Transformation analysis indicate that the active treatments were statistically significantly different from the vehicle control for PCBINF and PCBTOT. This is also true of PCBNON if we use a statistical model including the TRT*SITE interaction factor. If we use the model without TRT*SITE, Cleocin-T (Reference treatment) is significantly different from vehicle for PCBNON but Altana's Clindamycin (Test treatment) is not, testing at the usual $\alpha=0.05$ level. The sponsor has chosen to use the model with TRT*SITE, indicating that this factor is statistically significant at some visits.

3. I am not aware of any established equivalence criteria, including "goalposts", for this type of clinical response variable. I have reported 90% confidence intervals for both the ratio and the difference between the averages for the Test and Reference treatments, to aid in assessing the equivalence of these treatments.

4. If the Physician's Global Assessment at the last visit (PHYGLOB) variable is analyzed as a continuous variable, as the sponsor has done, both active treatments are highly statistically significantly different from vehicle control. As with lesion count PCB, I am not aware of any established equivalence criteria, including "goalposts", for this

type of clinical response variable. I have reported 90% confidence intervals for both the ratio and the difference between the averages for the Test and Reference treatments, to aid in assessing the equivalence of these treatments with respect to PHYGLOB.

5. As an alternative equivalence analysis of PHYGLOB, I have looked at the five possible dichotomizations of the six-point PHYGLOB scale. 90% confidence intervals for the difference between the probabilities of achieving the lower values of the dichotomy, for Test and Reference, would lead to a conclusion of equivalence if the plus-or-minus 20 percentage points "goalposts", as used in clinical BE studies of antiinfectives, are applied. This is true of all five possible dichotomies.

   I have also reported results of a proportional odds ratio analysis, as implemented by SAS PROC LOGISTIC, to aid in assessing the equivalence of these products.

/S/

Donald J. Schuirmann
Expert Mathematical Statistician
Quantitative Methods & Research staff

/S/ 1/20/00

Concur: Stella Green Machado, Ph.D.
Director, Quantitative Methods & Research staff

cc:

Original ANDA 64-160

HFD-600	Mary M. Fanning
HFD-650	Dale P. Conner
HFD-615	Harvey A. Greenberg
HFD-650	Aida L. Sanchez
HFD-705	QMR Chron
HFD-705	Stella G. Machado
HFD-705	Donald J. Schuirmann