

**CENTER FOR DRUG EVALUATION AND  
RESEARCH**

*APPLICATION NUMBER:*

**20-947**

**STATISTICAL REVIEW(S)**

## Comments on Tumorigenicity Analysis for N20-947

**NDA:** 20-947  
**Sponsor:** Nuvo Research, Inc.  
**Drug Identification:** Pennsaid® Topical Diclofenac Solution 1.5% w/w  
**Documents Reviewed:** Formal Consult dated 25 June 2009.  
E-mail dated 27 June 2009  
**Division:** Anesthesia, Analgesia, and Rheumatology Products  
**Toxicologist:** Adam Wasserman, Ph.D.  
**RPM:** Jessica Benjamin  
**Statistician:** Stella Machado, Ph.D.  
Steve Thomson  
**Type of Review:** Statistical Analysis of Sponsor's comments on  
tumorigenicity  
**Date:** 6/30/2009

### Introduction

The purpose of this consult was to evaluate the statistical arguments made by the Applicant relating to the incidence of lymphoma observed in the 26-week toxicity study in the rat relative to the background historical incidence. We were asked to comment in particular whether we agree with their statistical approach in which they conclude "Lymphomas occur spontaneously in the same strain of Sprague-Dawley rats as those in Nuvo Research's 6-month rat study, and the upper 95% confidence limits for male and female control groups projected from the ——— experience includes the observed incidence in the Nuvo study." In case of disagreement, we were asked to evaluate the data provided using appropriate methods and indicate the probability of observing these findings in the 26 week study, given the background incidence and our assessment as to whether the findings are likely to be unrelated to treatment.

b(4)

### Data on Lymphoma incidence, in same strain of rats.

Treated Groups from current study            2/200 (0/100 for M, 2/100 for F).  
Control data from current study:            0/50 (0/25 M and 0/25 F)  
Historical controls from 6-month studies:    3/1552 (1/767 for M, 2/785 for F)  
Historical controls from 2-year studies:    7/2548 (4/1284 for M, 3/1264 for F)  
by 26-30 weeks for males and 16-20 weeks for females

**The Sponsor's main points to support their contention that the current incidence rates are not different from historical control rates, and Reviewers' responses.**

**Main question:** Do we agree with "Lymphomas occur spontaneously in the same strain of Sprague-Dawley rats as those in Nuvo Research's 6-month rat study, and the upper

95% confidence limits for male and female control groups projected from the — experience includes the observed incidence in the Nuvo study.”?

b(4)

*Reviewer comment: the testing procedure followed by the Sponsor to make this statement is not clear, and it is difficult to agree/disagree with the assertion. The results from the current study have variability as well as those from the historical studies, and that does not seem to have been accounted for.*

The Sponsor quoted 2-year study historical control rates at — and rates in the literature. They said that though these data are from the 2-year studies, since the incidence gradually increases over the time-course of the studies, this establishes that lymphomas occur spontaneously in CD® Sprague-Dawley rats of approximately the same age as those in the 6-month current rat study.

b(4)

*Reviewer comment: the ideal control data are from the current 6-month study, but the number of rats (50) is small for drawing reasonable inference, especially when the incidence of the event is very small. The next best control data are from the 6-month historical studies. It is doubtful whether the 2-year study historical data can carry much weight, because of all the assumptions (e.g., changes in response over time, rate of change with age of tumor incidence, no effect of terminal sacrifice) that would have to be made. Further, the apparent lengths of time in study differ across gender in the two-year studies.*

*Reviewer comment: should the data be evaluated for the 2 sexes pooled or separate? This is for the toxicologist/pharmacologist to decide. Our analyses were done both ways.*

The Sponsor’s consultant was asked to assess whether the 2 cases of lymphoma in the Nuvo 6-month study would be expected just by chance. He concluded that that the finding of 2 lymphomas in 200 is not larger than would be expected to occur by chance.

The difference in lymphoma incidence between the treated rats (200) and control rats (50) in the current 26-week study is far from significant. (Fisher’s exact test: p-value = 0.65. The consultant acknowledged that 50 was a small number for drawing reasonable inference.

The difference in lymphoma incidence between the treated rats (200) in the current study and historical 6-month study controls is not significant at the 0.05 level of significance (Fisher’s exact test: p-value = 0.103 for both males and females; p=0.065 for females alone).

*Reviewer: we agree with these calculations.*

The consultant made the point that it is not appropriate to consider the 1552 rats as the whole population – that it is a sample, and subject to variability.

*Reviewer comment: we agree that allowing for variability in the historical control data would improve the results.*

*Reviewer comment: We pursued 3 different analysis approaches for evaluating whether the lymphoma rates in the current 6-month study and those in the historical studies could be consider similar or not. Results were obtained for males, females and both sexes pooled. The methods are: 1) to calculate the probability of observing the same or a more extreme number of lymphomas given a fixed historical rate, 2) Fisher's exact test, which allows for sampling variability, and 3) a Bayesian analysis.*

**RESULTS**

Method 1: The probabilities of observing the same or a more extreme number of rats with lymphomas, using the binomial distribution and assuming the historical control rates to be fixed are shown in Table 1.

**Table 1.**

Study	Lymphoma rates			Probability of $\geq$ cases		
	males	females	both	males (0)	females (2)	both (2)
Current study (trt)	0/100	2/100	2/200			
6-month.HCD	1/767	2/785	3/1552	1.0	0.0272	0.0578
2yr. HCD	4/1284	3/1264	7/2548	1.0	0.0239	0.1053
Pooled HCD	5/2501	5/2049	10/4100	1.0	0.0252	0.0863

It is preferable to use the 6-month control data rather than 2-year study control data because it may not be reasonable to assume comparability due to changes in response rates over time, and changes over time due to age; however, the 2-year data are included in the table for reference.

Results: compared with the 6-month historical controls, the chance of observing 2 or more lymphomas among 100 females is 0.0272. The chance of 2 or more lymphomas among 200 animals of both sexes is 0.0578. Since there were 0 lymphomas for males, not much can be said - the result is not significant.

Comment: the historical lymphoma rates were assumed fixed for the analysis in Table 1. The Sponsor's consultant pointed out that it would be better to account for variability in the historical control data, and we consider this a valid point. The p-values in the tables are likely somewhat underestimated, and thus the test may be somewhat anti-conservative.

Method 2. Using Fisher's exact test, one computes the probability that the difference in lymphoma rates between the treated animals and the 6-month historical controls is at least as extreme as the observed difference.

Table 2 shows the p-values from the Fisher's exact test comparing lymphoma incidence rates from the current 6-month study with the pooled current and 6-month historical controls. The appendix includes the p-values from the Fisher's exact test comparing lymphoma incidence rates from the current 6-month study.

**Table 2.**

Pooled Current Study and Historical Controls

Lymphomas	Females		Males		Both	
	Y	N	Y	N	Y	N
Control	2	808	1	791	3	1599
Treat	2	98	0	100	2	198
	p = 0.0619		p = 0.8879		p = 0.0978	

To summarize, these are tests of treatment differences conditional on the observed marginal totals. Again, with so few lymphoma responses the data still have to work very hard to show differences. The statistical significance of these differences in females are close to the traditional 0.05 level ( $p=0.0619$ ), with clearly no evidence of differences in males ( $p=0.8879$ ) and debatable results pooling males and females ( $p=0.0978$ ). The hypothesis test here is that, if we assume the rates in the pooled control group and the experimental group are the same, how likely would we be to get a difference in rates as extreme or more extreme than that which was actually observed. Note that rejecting the null hypothesis does not mean the proportions are close, only that the data do not provide evidence that they are not close.

Method 3. An alternative approach is a so-called Bayesian analysis. The exact tests above assess if the observed results are reasonable assuming that the treatments are the same. The Bayesian formulation assesses whether or not the incidence proportions are close given the observed responses. These approaches address slightly different questions. In a Bayesian analysis one attempts to quantify the initial lack of knowledge about the parameters, and express this initial uncertainty in probability distributions. Then the data are used to update this initial specification of uncertainty.

In this particular case we will examine the incidence rates in the historical control and in the current study treatment group and see what the data suggest about the distribution of the parameters, including inquiring if the data suggest these rates are close to each other. One measure of closeness in the rates is the absolute value of their difference. If we assume that we have no prior certainty about the proportion of animals with lymphoma, it might be reasonable to specify a uniform distribution to model the uncertainty about the parameter. Binomial likelihoods are then used to model the probability of tumor incidence with these parameters. Denote the lymphoma rate in the controls as  $p_0$  and in the current treatment group as  $p_1$ .

The Bayesian approach computes the distribution of any function of the parameters conditional upon the observed data. Here we model the difference in tumor rates,  $\text{diff} = p_1 - p_0$ , and the logarithm of the odds ratio of rates,  $\text{LOR} = \log(p_1/(1-p_1)/(p_0/(1-p_0)))$ . As measures of separation of the parameters, we also compute the posterior probability that the magnitude of the difference in rates is no more than 0.01, and no more than 0.02. These values are arbitrary, but should be informative when dealing with low incidence rates. The resulting posterior distribution for females is summarized in the following table:

**Table 3 Females**

Summary of posterior distribution

node	mean	sd	2.5%	median	97.5%	start	sample
LOR	2.076	0.8948	0.2965	2.076	3.872	2001	18000
$ p_1-p_0  < 0.01$	0.1671	0.3731	0.0	0.0	1.0	2001	18000
$ p_1-p_0  < 0.02$	0.4313	0.4953	0.0	0.0	1.0	2001	18000
diff	0.02556	0.01668	0.00183	0.02261	0.06523	2001	18000

Again, the distribution of the updated knowledge about the functions of the rates is summarized above. The 2.5%, 50%, and 97.5% percentiles of the corresponding probability distribution of the parameters are listed, and the interval between the 2.5% and the 97.5% percentile is a so called “95% credible interval”. There is roughly a 0.95 probability that the value of the function lies in this interval.

Thus, for example, the probability that in females the probabilities  $p_0$  and  $p_1$  are within 0.01 is estimated to be 0.1671, i.e., the estimated probability that the rates differ by at least 0.01 is  $0.8329 = (1-0.1671)$ . When discussing tumor rates in the range of about 0.01 to 0.03, is a 0.01 difference “close”? That is a decision requiring the expertise of the toxicologist. Note the 95% credible interval for the log odds ratio (0.2965 to 3.872) excludes the value zero, suggesting a difference in the odds. The difference in proportions of animals with lymphoma is also bounded away from zero, but the lower bound of the credible interval is close to zero.

Plots of the posterior distributions for the difference and log odds ratio, along with further results for males and for both genders pooled, as well as the results assuming a more concentrated prior, are all presented in Section 3 of the Appendix.

### Conclusion and Recommendations

Drawing inferences about differences in proportions is generally statistically challenging, especially when expected rates are very small, and the number of animals is not large. The methods we used are in common usage. We consider the Bayesian approach most illuminating.

It does appear that there is no evidence of treatment differences in lymphoma incidence in male animals. In female animals the results are more equivocal. Both the frequentist tests (Method 1 calculating the binomial probabilities using fixed historical rates; Method 2 using Fisher’s exact test) and the Bayesian analysis using the historical

controls provide suggestive, but not conclusive, evidence of treatment differences in lymphoma incidence in female animals.

Stella G. Machado  
Division Director, Biometrics VI

Steve Thomson  
Mathematical Statistician, Biometrics VI

## Appendix

### 1. Binomial Tests

Table 1 in the text shows, for males, females, and pooled genders, the probability of observing the same or a more extreme number of animals with lymphoma when administered Pennsaid.

As noted in the report, the Sponsor's analyst does criticize this approach and notes that it ignores the variation in the historical control. But if animal variances are of the same magnitude, whether one restricts attention to each gender separately or analyzes the pooled genders, the variance in study group should be about eight times the variance in the control group. So as a "quick and dirty" test this should be reasonable. Note the following analyses do adjust for differences in variances, either implicitly (in the exact tests) or explicitly (in the Bayesian tests).

### 2. Fisher's Exact Tests

In the Sponsor's discussion much ado is made of the results of the Fisher exact tests. The Fisher exact tests compute the proportion of permutations of the within table responses that are as "extreme" as the observed result, while holding all table marginal totals fixed (i.e. treatment group totals and response group totals). In a typical study animals are assigned to the treatment group, so holding treatment group totals fixed is very reasonable. Holding response group totals fixed seems to be more problematic. If subjects can be assumed to be randomly assigned to treatment and responses are independent across subjects, this may be a reasonable test. It does not assume anything about the probability distribution of the responses except that all permutations are equally likely. However, the latter is a very strong assumption.

The following tables display the results of the exact tests using the current within study control group. Note that conditional on the table marginal totals, the observed pattern is the most extreme pattern that could be observed, and hence with only two tumors, only in the larger treatment group, there is no way that an exact test would ever reject the hypothesis that treatments have no effect on responses.

#### Within Study Control

Lymphomas	Females		Males		Both	
	Y	N	Y	N	Y	N
Control	0	25	0	25	0	50
Treat	2	98	0	100	2	198
	p = 0.6387		p = NA (or 1.0)		p = 0.6394	

Since all permutations with fixed marginal totals in the male group would result in the same table, all tables are as "extreme" as the observed table and hence one could say that the significance of the observed table is 1.0. Again, the problem with the other two tables is that with only two positive responses almost all permutations will be as

extreme as the observed table, and hence, with the given marginal totals, no table can indicate treatment differences, i.e., the exact test will never reject the null hypothesis.

However, the historical group does provide a larger pool of animals for comparison. Using the historical controls provided in Dr. Wasserman's e-mail provides the following:

#### Historical Controls Only

Lymphomas	Females		Males		Both	
	Y	N	Y	N	Y	N
Control	2	783	1	766	3	1549
Treat	2	98	0	100	2	198
	p = 0.0652		p = 0.8847		p = 0.1028	

#### Pooled Current Study and Historical Controls

Lymphomas	Females		Males		Both	
	Y	N	Y	N	Y	N
Control	2	808	1	791	3	1599
Treat	2	98	0	100	2	198
	p = 0.0619		p = 0.8879		p = 0.0978	

To summarize again, these are tests of treatment differences conditional on the observed marginal totals. With so few lymphoma responses, the data still have to work very hard to show differences. The statistical significance of these differences in females are close to the traditional 0.05 level ( $p=0.0619$ ), with clearly no evidence of differences in males ( $p=0.8879$ ) and debatable results pooling males and females ( $p=0.0978$ ).

### 3. Bayesian Analysis

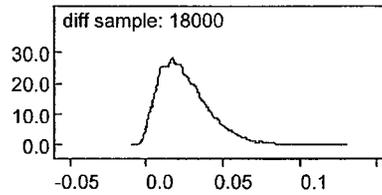
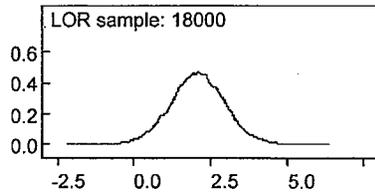
A Bayesian analysis does require an initial assessment of the uncertainty about the tumor incidence proportions. One simple approach is to assume that we have no certain knowledge about the tumor incidence, a description that could be interpreted as saying that, prior to collection of the data, any proportion between 0 and 1 is equally likely to hold. That is, the two proportions have a uniform distribution over the interval 0 to 1. We use binomial likelihoods to model the probability of tumor incidence within the historical controls (denoted as  $p_0$ ) and within the treatment group in the current study (denoted as  $p_1$ ). The Bayesian approach then computes the posterior distribution of any function of these parameters. Here we model the difference in tumor proportions,  $\text{diff} = p_1 - p_0$ , and the logarithm of the odds ratio,  $\text{LOR} = \log(p_1/(1-p_1) / (p_0/(1-p_0)))$ . As measures of separation of the parameters we also compute the posterior probability that the magnitude of the difference in probabilities is no more than 0.01, and is no more than 0.02.

First, we assume that the prior distribution for  $p_0$  and  $p_1$  is uniform, i.e., for example the prior probability that each proportion is  $< 0.1$  is the same as the probability it is  $> 0.9$ . The plots below estimate the posterior distributions of the  $\text{diff}$  and  $\text{LOR}$ . The

tables summarize the posterior distribution. It is of special interest to see if 0 is in the interval between the 2.5% percentile and the 97.5% percentile. Note that “ $|p_1 - p_0| < \text{value}$ ” gives the posterior probability that the two proportions are within the specified value of each other.

Uniform Prior:

Females

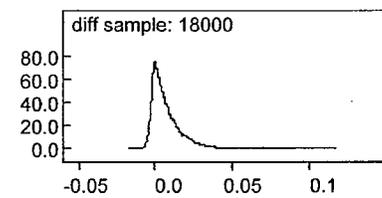
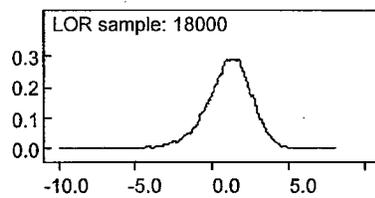


Summary of posterior

node	mean	sd	2.5%	median	97.5%	start	sample
LOR	2.076	0.8948	0.2965	2.076	3.872	2001	18000
$ p_1 - p_0  < 0.01$	0.1671	0.3731	0.0	0.0	1.0	2001	18000
$ p_1 - p_0  < 0.02$	0.4313	0.4953	0.0	0.0	1.0	2001	18000
diff	0.02556	0.01668	0.00183	0.02261	0.06523	2001	18000

Thus, for example, the probability that in females the probabilities  $p_0$  and  $p_1$  are within 0.01 is estimated to be 0.1671. It is up to the toxicologist to determine if that defines “close”. Note the 95% credible interval for the log odds ratio (0.2965 to 3.872) seems to be bounded away from zero, suggesting differences in the odds. The difference in proportions is also bounded away from zero, but is close to zero.

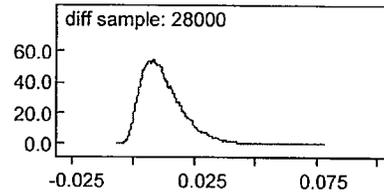
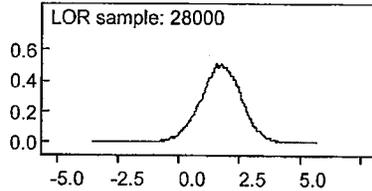
Males



node	mean	sd	2.5%	median	97.5%	start	sample
LOR	1.039	1.514	-2.372	1.153	3.709	2001	18000
$ p_1 - p_0  < 0.01$	0.7174	0.4503	0.0	1.0	1.0	2001	18000
$ p_1 - p_0  < 0.02$	0.898	0.3026	0.0	1.0	1.0	2001	18000
diff	0.00719	0.009819	-0.004027	0.004433	0.03289	2001	18000

Consistent with the tests in Methods 1 and 2, there is no evidence that these lymphoma incidence rates differ in males.

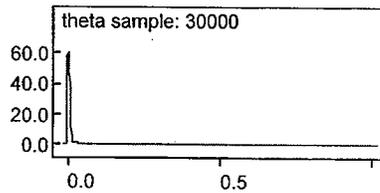
Both Genders



node	mean	sd	2.5%	median	97.5%	start	sample
LOR	1.724	0.8258	0.05314	1.739	3.309	2001	28000
p1-p0  < 0.01	0.4603	0.4984	0.0	0.0	1.0	2001	28000
p1-p0  < 0.02	0.833	0.3729	0.0	1.0	1.0	2001	28000
diff	0.01228	0.008561	1.712E-4	0.01075	0.0332	2001	28000

For the diff and log odds ratio, results from pooling genders are similar to those for females alone, since the 95% credible intervals for LOR and diff exclude zero. But they are much less extreme.

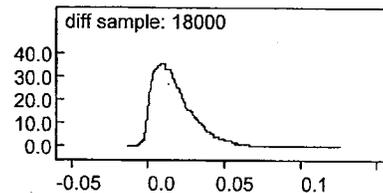
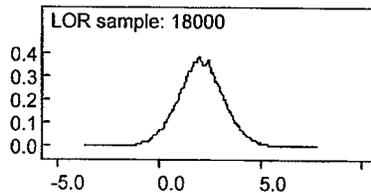
We also addressed the question of whether a uniform prior is appropriate. A prior that may reflect a belief that lymphomas are rare would be a Beta(0.03,0.97) distribution, with modal and mean value at 0.03. An estimated plot of this prior distribution is as follows:



This prior is meant to reflect that we would expect lymphoma incidence rates to be close to zero. For example, with this prior the probability of a lymphoma rate above 0.1 is only about 0.0711. Such a prior may better reflect knowledge about the low probability of lymphomas. Then, as with the uniform prior, we get:

Beta(0.03,0.97) prior:

Females

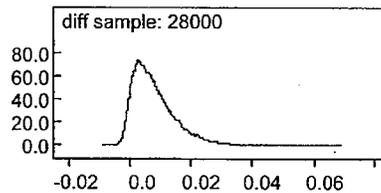
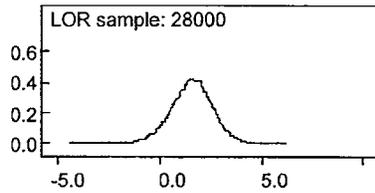


node	mean	sd	2.5%	median	97.5%	start	sample
LOR	2.072	1.129	-0.1675	2.063	4.312	2001	18000
p1-p0  < 0.01	0.3472	0.4761	0.0	0.0	1.0	2001	18000
p1-p0  < 0.02	0.6556	0.4752	0.0	1.0	1.0	2001	18000
diff	0.01749	0.01399	-5.357E-4	0.0145	0.05246	2001	18000

So now the 95% credible intervals for the log odds ratios and the differences in proportions do include zero, though only barely. Further, the estimated probability that the two proportions are more than 0.01 apart is 0.6528 (=1-0.3472), although the probability they are more than 0.02 apart is only 0.3444. However, with such low incidence rates a difference of 0.01 or 0.02 may be considered as a considerable difference.

Results for males will be even less extreme than with the uniform prior, with very little evidence of differences in the tumor proportions, and thus are not given here.

### Both Genders



node	mean	sd	2.5%	median	97.5%	start	sample
LOR	1.561	0.005194	-0.5208	1.591	3.49	2001	28000
p1-p0  < 0.01	0.6815	0.002835	0.0	1.0	1.0	2001	28000
p1-p0  < 0.02	0.9311	0.001426	0.0	1.0	1.0	2001	28000
diff	0.008131	3.867E-5	-0.001112	0.006603	0.02593	2001	28000

Again, the credible intervals for the log odds ratios and the differences in proportions do include zero, though only perhaps somewhat marginally. Further, the estimated probability that the two incidence rates are more than 0.01 apart is still 0.3185. Again, with such low rates, this may be a considerable difference.

These analyses were performed using WinBUGS 1.4.3. Also, for convenience the Bayesian analyses were done using only the current study incidence and the 6-month historical controls.

-----  
**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  
7/2/2009 03:00:33 PM  
BIOMETRICS

Stella Machado  
7/2/2009 03:03:37 PM  
BIOMETRICS



DEPARTMENT OF HEALTH AND HUMAN SERVICES  
FOOD AND DRUG ADMINISTRATION  
CENTER FOR DRUG EVALUATION AND RESEARCH  
OFFICE OF BIostatISTICS

## Statistical Review and Evaluation CLINICAL STUDIES

NDA: 20-947

Name of drug: Pennsaid Topical Solution (1.5% w/w diclofenac sodium)

Applicant: Dimethaid/Nuvo

Indication: osteoarthritis

Documents reviewed: electronic submission at  
\\CDSESUB1\N20947\N 000\2006-06-28 and  
\\CDSESUB1\N20947\N 000\2006-09-19 ;  
electronic data at  
\\CDSESUB1\N20947\N 000\2006-08-17 ;  
previous submission 7 August 2001;  
reviews of previous submission by Suktae Choi, Ph.D.  
(26 July 2002) and James Witter, M.D. (1 August 2002);  
minutes of meetings of 29 May 2002, 10 September 2002,  
19 November 2002, 29 August 2003, 4 November 2003

Medical division: Anesthesia, Analgesia and Rheumatology

Project manager: Paul Balcer

Clinical reviewer: Larissa Lapteva, M.D.

Review priority: standard (resubmission)

Dates: letter 28 June 2006; user fee goal (class 2 resubmission,  
6 months) 28 December 2006

Statistical reviewer: Thomas Permutt (acting division director)

Keywords: NDA review, clinical studies, combination drug

1 Executive Summary	3
2 Introduction	3
2.1 Overview	3
2.2 Data Sources	5
3 Statistical Evaluation	5
3.1 Evaluation of Efficacy	5
3.1.1 <i>Statistical Methods</i>	7
3.1.2 <i>Detailed Review of Individual Studies</i>	8
3.1.2.1 <i>Study 112</i>	8
3.1.2.2 <i>Study 109US</i>	11
3.1.3 <i>Statistical Reviewer's Findings</i>	12
3.2 Evaluation of Safety	13
4 Findings in Special/Subgroup Populations	14
5 Summary and Conclusions	14
5.1 Statistical Issues and Collective Evidence	14
5.2 Conclusions and Recommendations	15
6 Labeling	15

---

## 1 EXECUTIVE SUMMARY

---

Diclofenac in the test article can confidently be concluded to have a modest effect on signs and symptoms of osteoarthritis. Whether the effect is sufficient depends on the balance of benefit and risk, but there is little doubt that there is an effect.

There is no direct evidence of a benefit of DMSO in the test article. If DMSO is considered an active ingredient in a combination product, the combination should not be approved. The applicant's method of addressing the combination policy would render the policy useless in preventing the marketing of irrational mixtures: if a component contributed to the claimed effects, the combination would be acceptable; but if it did not, the combination would be claimed not to be a combination at all. If there are other reasons, however, to consider DMSO an excipient rather than an active ingredient, then the studies conducted are sufficient to demonstrate the efficacy of the single active ingredient diclofenac.

---

## 2 INTRODUCTION

---

### 2.1 OVERVIEW

Diclofenac is a nonsteroidal anti-inflammatory drug used orally in arthritis and other inflammatory conditions. This application concerns a topical solution of diclofenac in DMSO (dimethyl sulfoxide) to be applied as 40 drops four times a day to an osteoarthritic knee.

NDA 20-947 was found to be not approvable 7 August 2002. Two studies (109 and 109US: these were different studies despite the naming) were described as pivotal, but the division director, Lee Simon, M.D., noted the same shortcomings in both and found inadequate evidence of efficacy. His three principal concerns related to the treatment of both knees; to the exclusion of some patients from analyses, even those analyses labeled "intent-to-treat"; and to the possibility that DMSO in this preparation is an active ingredient.

Patients were selected who had osteoarthritis in at least one knee, and the worse afflicted knee was the subject of study measurements. They were also allowed to apply the drug to the other knee as needed. The reviewers of the original NDA opined that patients who treated the other knee could not meaningfully be compared to those who did not, and stratified analysis was inconclusive.

From 10 to 20 percent of patients were excluded from various "intent-to-treat" analyses. These comprised not only patients whose primary outcomes were unknown but also a larger number for which the data existed but were considered "invalid" because the final assessment was performed more than 48 h after the last application of drug. Per-protocol analyses were reported excluding even larger numbers, but no analysis was given for all treated patients or even all patients with complete data. Suktae Choi, Ph.D., the statistical reviewer, performed such analyses with various methods of imputation of missing or invalid

data. He found that the results varied so much with the method of imputation as to cast substantial doubt on the efficacy of the drug.

The action letter also noted that the preparation might be considered a fixed-ratio combination drug product with DMSO as an active ingredient. The applicant apparently considered DMSO an excipient and had not anticipated the possible relevance of the combination drug policy.

The Agency and the applicant held meetings or teleconferences 29 May 2002, 10 September 2002, 19 November 2002, 29 August 2003 (erroneously "2002" in some documentation) and 4 November 2003. The present submission alludes to putative agreements about the further course of development, but the applicant's characterizations differ from my reading of the minutes. In particular, the submission (Integrated Summary of Efficacy, p. 3) claims

It is understood from the guidance received from the Division that another Phase 3 trial was required to move towards NDA approval and that, as recommended by the Division, the PEN-03-112 study protocol design responds to all the remaining issues identified in the NA letter. It is further understood that this one study will complete the totality of the evidence for the basis of approval for PENNSAID® Topical Solution NDA 020-947 amendment.

In fact, the applicant sought but apparently did not get this agreement (minutes of teleconference 4 November 2003):

Question 2: Once finalized and approved, this clinical study should meet all outstanding issues relating to the efficacy and safety of the product, and will form the primary basis for the marketing approval of PENNSAID® Topical Solution, 1.5% w/w diclofenac sodium, NDA 20-947. Does the FDA agree, assuming that the study will be successful and will meet the study objectives?

Initial FDA Response:

Approval will be based on the results of the review of this study, along with data from the prior NDA submission.

Meeting comments:

Approval of the drug will be based on the totality of the new evidence in addition to the past evidence submitted to the Division.

Sponsor's pivotal study should show an improvement in the target knee.

There is also some discrepancy with respect to the combination drug issue. According to the ISE (pp. 14–15),

no statistically significant difference was observed between placebo (P) and vehicle-control (VC) in any of the clinical efficacy measures. These results show that the vehicle (i.e., 45.5% w/w DMSO) in PENNSAID does not have an independent effect in symptomatic relief of OA of the knee. This finding directly addresses the combination rule question in the manner requested by FDA.

The Agency did remark on several occasions that the question needed to be addressed, but I can find no indication that we requested it be addressed in this manner.

## 2.2 DATA SOURCES

The application now purports to rely on a new Study 112 along with the previously submitted Study 109US. These are the only 12-week studies, and studies of at least 12 weeks have been considered essential in osteoarthritis. The applicant considers the data from several shorter studies, including the previously reviewed study 109, to be “supportive.”

Electronic data were submitted with the application but were not in a convenient form for review. At the Agency’s request additional files were submitted 17 August 2006.

The written submission is in electronic form at \\CDSesub1\N20947\N\_000\2006-06-28 with an amendment at \\CDSesub1\N20947\N\_000\2006-09-19.

Electronic data are at \\CDSesub1\N20947\N\_000\2006-08-17.

I have also consulted the previous submission of 7 August 2001; the reviews of the previous submission by Suktae Choi, Ph.D. (26 July 2002) and James Witter, M.D. (1 August 2002); and the minutes of meetings of 29 May 2002, 10 September 2002, 19 November 2002, 29 August 2003 and 4 November 2003.

---

## 3 STATISTICAL EVALUATION

---

### 3.1 EVALUATION OF EFFICACY

The main results from the new Study 112 and the old Study 109US are summarized in the two tables below, copied from the Integrated Summary of Effectiveness.

**Table 5: Study PEN-03-112: Efficacy Results:**  
**Change in Score from Baseline**

Treatment Group <sup>1</sup>	Intent to Treat Population					Per Protocol Population		
	Pain	Physical Function	POHA	Stiffness	PGA	Pain	Physical Function	POHA
	N Mean (SD)	N Mean (SD)	N Mean (SD)	N Mean (SD)	N Mean (SD)	N Mean (SD)	N Mean (SD)	N Mean (SD)
Group 1 (PEN+OD)	151 -6.95 (4.76)	150 -18.69 (14.03)	148 -0.95 (1.21)	150 -2.30 (2.00)	150 -1.53 (1.27)	109 -6.86 (4.61)	110 -18.93 (13.98)	106 -1.02 (1.23)
Group 2 (PEN+OP)	154 -6.02 (4.54)	154 -15.75 (15.14)	154 -0.95 (1.30)	154 -1.93 (2.01)	154 -1.36 (1.19)	109 -6.36 (4.66)	109 -17.31 (15.47)	108 -0.98 (1.36)
Group 3 (VC+OP)	161 -4.7 (4.31)	161 -12.13 (14.58)	160 -0.65 (1.12)	161 -1.48 (2.07)	161 -1.07 (1.10)	116 -5.02 (4.36)	117 -12.87 (14.63)	114 -0.72 (1.14)
Group 4 (P+OP)	155 -4.74 (4.35)	153 -12.34 (14.72)	152 -0.37 (1.04)	153 -1.52 (2.05)	153 -1.10 (1.18)	113 -5.02 (4.33)	111 -13.34 (15.27)	106 -0.39 (1.10)
Group 5 (P+OD)	151 -6.43 (4.11)	151 -17.48 (14.33)	150 -0.88 (1.31)	151 -2.07 (2.02)	151 -1.42 (1.29)	113 -6.65 (4.09)	113 -17.83 (13.74)	113 -0.93 (1.31)
<b>Group Comparisons</b>	<b>P-values<sup>2</sup></b>							
PEN vs. P (2 vs. 4)	0.0150	0.0344	0.0000	0.1120	0.0165	0.0347	0.0522	0.0006
PEN vs. VC (2 vs. 3)	0.0094	0.0255	0.0158	0.0347	0.0181	0.0144	0.0144	0.0700
VC vs. P (3 vs. 4)	0.8855	0.9266	0.0376	0.6156	0.9481	0.7435	0.6272	0.0909
PEN vs. OD (2 vs. 5)	0.4290	0.3189	0.9565	0.5960	0.4392	0.5740	0.7453	0.8348
<sup>1</sup> PEN=PENNSAID, OD=Oral diclofenac, VC=Vehicle-control solution, P=Placebo solution, OP=Oral, POHA=Patient Overall Health Assessment, PGA=Patient Global Assessment <sup>2</sup> Statistical test used: ANCOVA with baseline score as covariate; N/A=not available Source: PEN-03-112 Study Report, Tables 14.2.16, 14.2.17, 14.2.18								

**Table 10: Study RA-CP-109-US: Efficacy Results:  
 Comparison: PENNSAID® vs. Vehicle-control**

Efficacy variables	Treatment Group <sup>1</sup>	Intent-to-Treat (previously submitted)			All Patients (re-analysis)		
		N	Mean (SD) change from baseline	p-value <sup>2</sup>	N	Mean (SD) change from baseline	p-value <sup>2</sup>
Pain	PEN	133	-6.4 (4.8)	0.0001	164	-5.9 (4.7)	0.0017
	VC	144	-4.3 (4.5)		162	-4.4 (4.4)	
Physical Function	PEN	132	-16.9 (15.7)	0.0003	164	-15.3 (15.2)	0.0024
	VC	144	-10.2 (14.1)		162	-10.3 (13.9)	
Patient Global Evaluation	PEN	131	-1.4 (1.2)	0.0004	164	-1.3 (1.2)	0.0052
	VC	144	-0.9 (1.2)		162	-1.0 (1.1)	
Stiffness	PEN	133	-2.0 (2.2)	0.0006	164	-1.8 (2.1)	0.0086
	VC	144	-1.2 (2.0)		162	-1.3 (2.0)	

<sup>1</sup>PEN = PENNSAID; VC = vehicle-control  
<sup>2</sup>Statistical test used: ANCOVA with baseline score as a covariate  
 Source: RA-CP-109-US Study Report, Tables 59, 60, 61, 62

The interpretation of the numerical scores is rather obscure. All assessments were on a categorical scale from 0 (best) to 4 (worst). The pain score, however, is the sum of 5 such assessments, the function score of 17, and the stiffness score of 2, while the overall health assessment and global assessment were answers to single questions. Thus, the range of the pain score was 0–20, of function 0–68, of stiffness 0–8, and of the two overall assessments 0–4. For each measure, therefore, there were improvements from baseline on the order of 1 point out of 4 in all treatment groups, and the differences between groups were generally small fractions of a point.

As is customary in trials in osteoarthritis, all of pain, function and global assessment were considered primary endpoints in the sense that a positive effect on each was essential. The overall health assessment, which was the patient’s answer to a differently worded question, was primary in preference to the other global assessment in Study 112. Stiffness was included as a potentially important secondary claim.

The applicant notes statistically significant differences in the primary variables between the test article and the vehicle control in Study 109US and between the test article (group 2) and placebo (group 4) in Study 112. For Study 109US the “all patients” re-analysis corresponds more closely to what is usually thought of as intent-to-treat analysis, and to the intent-to-treat analysis of study 112, than does the analysis labeled “intent-to-treat.”

### 3.1.1 STATISTICAL METHODS

For all variables, changes from baseline were compared pairwise between groups using analysis of covariance with the baseline value as a covariate. This method is appropriate.

Missing data were imputed by last observation carried forward (LOCF). In study 109US, there were no observations between the baseline and final observations, so that LOCF amounts to carrying forward the baseline (BOCF), or imputing zero change, which is

appropriately conservative. In study 112 LOCF is not an appropriate method, but BOCF analyses were also performed post hoc.

### 3.1.2 DETAILED REVIEW OF INDIVIDUAL STUDIES

#### 3.1.2.1 *Study 112*

Study 112, newly submitted, was a randomized, double-blind trial at 61 centers, 40 in Canada and 21 in the U.S. Seven hundred seventy-five patients were randomized in approximately equal numbers to five arms. For evaluation of efficacy, the critical arms were designated groups 2 and 3. Group 2 was treated with the test drug, a 1.5% solution of diclofenac sodium in a vehicle containing 45.5% DMSO. Group 3 was treated with the same vehicle without diclofenac. Another arm, group 4, used a control with most of the DMSO removed as well as the diclofenac; but because DMSO has a characteristic odor and taste, even when applied to the skin, 2.3% DMSO was used in this preparation to improve masking. This treatment is referred to in the submission as “placebo” while the treatment for group 3 is called “vehicle control.” Groups 1 and 5 were treated with oral diclofenac 100 mg sustained-release once daily in addition to the test article and the placebo lotion, respectively; groups 2–4 had oral dummies. The purpose of the oral treatments appears to have been to study the safety profile in the case of concomitant oral therapy, and they also furnish a frame of reference for the magnitude of the effects of the topical treatment.

Three measures were identified as primary in the protocol, in the sense that all three should show a significant effect: pain, overall health assessment and physical function. The protocol specified that the comparison of group 2 to group 4 was primary, for reasons that are not clear; there may have been some miscommunication with the Agency over what comparison was most important for approval.

The table below, copied from the study report, gives the results of the post-hoc analysis with baseline carried forward.

**Table 27: Baseline Observation Carried Forward Analysis of the Primary Efficacy Variables**

**ITT, Mean (SD) Change in Score**

Variable	Group <sup>1</sup>				
	1 PEN+OD N=151	2 PEN+OP N=154	3 VC+OP N=161	4 P+OP N=155	5 P+OD N=151
Pain	-6.51 (4.87)	-5.81 (4.53)	-4.42 (4.31)	-4.60 (4.33)	-6.11 (4.26)
Physical function	-17.79 (14.24)	-15.05 (15.04)	-11.41 (14.43)	-11.92 (14.70)	-16.59 (14.64)
Patient Overall Health Assessment	-0.91 (1.17)	-0.92 (1.30)	-0.61 (1.11)	-0.35 (1.01)	-0.84 (1.30)
<i>Tests of statistical significance:</i>					
Pain:	Group 2 vs. Group 4, p = 0.0233				
	Group 2 vs. Group 3, p = 0.0070				
Physical function:	Group 2 vs. Group 4, p = 0.0553				
	Group 2 vs. Group 3, p = 0.0261				
Patient Overall Health Assessment:	Group 2 vs. Group 4, p < 0.0000				
	Group 2 vs. Group 3, p = 0.0119				
(p-value from ANCOVA with baseline score as covariate)					
<sup>1</sup> PEN = PENNSAID; OD = oral diclofenac; P = topical placebo; VC = topical vehicle-control; OP = oral placebo					
Source: Table 14.2.16.4					

Except for the absurd “p<0.0000” (the correct p-value is indeed zero to four decimal places, but not less than zero) the results are generally similar to those with last observation carried forward. Note that the comparison of physical function between groups 2 and 4, ostensibly the primary comparison, changes from significant to nonsignificant at level 0.05.

In my opinion, notwithstanding the protocol, the crucial comparison is between the test article (group 2) and its vehicle without diclofenac. It is an accepted principle of drug testing to isolate so far as possible the effect of the putatively active drug substance, here diclofenac, from possible effects of other ingredients by leaving out only the active substance. If the concern were that DMSO also may be an active ingredient, this approach does nothing to allay it. Indeed, in studies of combination drugs, the combination product needs to be compared to formulations lacking *each* of the putatively active ingredients. Thus, the comparison of test article to vehicle would still be critical, along with a comparison, which was not performed, to a preparation without DMSO but with diclofenac. The only real value in the comparison to the vehicle without DMSO is to exclude the possibility of a *deleterious* effect of DMSO overcome by a larger benefit of diclofenac.

In any case, the results are similar. All the primary analyses but one show a statistically significant though very modest benefit of the test article. The one borderline test is for the overall health assessment. Arguably an “overall” assessment is not very meaningful anyway, in the case where only one knee is treated in a patient with osteoarthritis in both knees.

Again, the magnitude of the effects requires some attention to interpret correctly. Consider, for example, the average changes in pain score of -5.8 for the test article and -4.4 for the vehicle control. The difference of 1.4 represents a difference of 1.4/5 = 0.3 points on a

scale of 0–4 for each of five components of the WOMAC (Western Ontario–McMaster) pain scale.

The Agency had expressed some concern about the possible magnitude of the treatment effect in relation to the sample size: “The division considers the change in WOMAC score of at least 10% of scale from the baseline score to represent a minimal clinical importance difference for this protocol *regardless of the number of patients in each arm.*” (Emphasis in original.) “Ten percent of scale” in this case would be  $(0.1)(4)(5) = 2$ . The applicant correctly points out that the change from baseline in the test group was well above this. Of course, it was also so in the control groups: as is common in osteoarthritis trials, which enroll patients when their symptoms are relatively severe, the patients in all groups improved substantially. The treatment effect, however, is smaller than this. Although the wording of the Agency’s comment is awkward, I think we were referring to the size of the treatment effect. I cannot see why the amount of spontaneous improvement would be more important than the amount of improvement attributable to the test drug.

There is a third way of interpreting the 10% requirement. Instead of looking at the average change within a group or at the difference between these averages, we can ask how many individuals had a 10% improvement. The numbers and percentages by group are shown in the table below. I also tabulated these for alternative criteria of improvement, namely a change of at least 30% or 50% of the baseline score. In each case patients with missing data were considered to be nonresponders.

	Group 1	Group 2	Group 3	Group 4	Group 5
N	151	154	161	157	151
Improvement at least:					
10% of scale	121 (80%)	129 (84%)	115 (71%)	113 (72%)	126 (83%)
30% of baseline	110 (73%)	102 (66%)	86 (53%)	86(55%)	105 (70%)
50% of baseline	85 (56%)	73 (47%)	56 (35%)	68 (43%)	76 (50%)

I focus on groups 2 and 3, the test article and the vehicle, both without concomitant oral diclofenac. Most of the patients (71%) improved by more than two points (10% of scale) even in the control group 3. Improvement in an additional 13% (84% – 71%) of patients can be attributed to the diclofenac preparation in group 2. This gives a number needed to treat of 1/0.13 or about 8: eight patients had to be treated for each one with an attributable improvement.

It is also worth considering the ratio 13%/84% or its inverse, which is about 6. There is no standard term for this, but it might be called the number needed to continue treatment. If this were small, we might hope, after initially treating 8 patients, to identify the 1 who benefited and discontinue therapy for the other 7. In fact, though, we expect about 6 of 8 patients to improve with treatment. The 1 who truly benefited from the active drug will be indistinguishable from 5 others who would have improved regardless of treatment. Thus, 6 patients would have to continue treatment to assure the continued benefit to the 1 unidentifiable patient who benefited from the drug.

The table also shows similar calculations for responders defined as patients with improvement of 30% or 50% of their baseline score, rather than 10% of scale. The number needed to treat remains about 8 for each of these analyses, though the number needed to continue treatment goes down as the criterion for a response gets stricter.

### *3.1.2.2 Study 109US*

Study 109US was reported in the last cycle and reviewed in detail by Drs. Choi and Witter. The applicant has submitted a re-analysis including all patients treated, with missing data imputed by last observation carried forward. There being no intermediate observations in this study between the baseline and final observations, the last observation to be carried forward was in fact the baseline observation. Thus, dropouts had zero change from baseline imputed, which is appropriately conservative. This analysis corresponds to Dr. Choi's "method 1," and the applicant's analyses substantially agree with Dr. Choi's. There were modest but clearly statistically significant differences between the test article and the vehicle on all three primary measures of outcome.

As Dr. Choi noted, even with his more conservative method 2, the differences remained significant. In method 2, missing data were imputed in a "worst-case" way, with good scores imputed to vehicle patients with missing data but bad scores to active-drug patients with missing data.

Dr. Choi also noted that the significance was lost, and in fact the estimated treatment effects were even in the wrong direction, if worst-case imputations were also applied to data that were present but classified by the applicant as "invalid." The reasons for these "invalid" data were not clear in the original submission, nor for that matter in the resubmission, but have been clarified in a subsequent amendment 29 September 2006. The final scores were not counted because the patient questionnaire was completed more than 48 h after the last dose. For the pain scores, for example, 7 scores were actually missing, and have now been imputed by "LOCF," which, because there were no observations between the baseline and the final observation, amounts to BOCF. An additional 33 scores were excluded as invalid but have now been included as recorded.

I believe the most appropriate way to handle these observations is to include them in the analysis as they were observed. This is what the applicant has now done, and what Dr. Choi did as method 1. The actual observations are the best information we have on the patients' status at the end of the trial. The applicant's misplaced zeal for perfect data led them to exclude these patients, and then they confused the issue by labeling this restricted analysis as "intent-to-treat." Absent clear information about what was done, Dr. Choi appropriately did a very conservative analysis. Now that we know, I think the results of that additional analysis can be disregarded. The nearest thing possible to intent-to-treat analysis in the circumstances is Dr. Choi's method 1.

### 3.1.3 STATISTICAL REVIEWER'S FINDINGS

The new Study 112 showed a statistically significant effect of the diclofenac preparation compared to its vehicle on the primary measures of outcome as well as on the secondary outcome of stiffness. The benefit was small, perhaps unprecedentedly so for a drug to be approved for osteoarthritis. The study was large enough, however, to establish the existence and direction of this small effect with confidence.

After resolving the issues with excluded patients, I reach similar conclusions with respect to the previously submitted Study 109US. My view of the design of that study differs from that of Drs. Choi, Witter and Simon. Many patients with osteoarthritis of the knee have it in both knees. If such patients had been excluded, the study population would have been restricted in an artificial way and would not have corresponded well to the target population: even if the product were labeled for use on one knee only, it seems unlikely that it would be used that way. If, as in Study 112, they had been instructed to treat only one knee, again the trial conditions would have deviated from the likely clinical use of the drug, and global assessments would be rendered difficult to interpret. In contrast, the actual trial seems more realistic, and the interpretation is straightforward: treat as many knees as hurt, and this is the estimated effect. The reason for restricting the *assessment* to a single knee is statistical: the unit of randomization is the patient, not the knee. If knees are the unit of analysis, then the randomization is in clusters of one or two knees. There are various ways of analyzing cluster-randomized studies, but analyzing a single, prespecified unit from each cluster is a straightforward and entirely correct one. This is what was done.

In any case, we now have one one-knee and one two-knee study with similar results. There can be little doubt that there was an effect of diclofenac in the test article, but the effect was small. The Agency advised the sponsor in specific terms that the effect would need to be larger than this to be clinically meaningful. It is neither customary nor statistically important to prespecify the magnitude of effect that will be considered meaningful, in the sense that it is important to prespecify other aspects of analysis to avoid problems of multiplicity. I believe the magnitude of the benefit needed should be determined in light of the observed risk, therefore necessarily post hoc.

### 3.2 EVALUATION OF SAFETY

DMSO has been associated with the formation of cataracts in some animal species. In an extension of study 112 in which some patients were exposed for as long as a year, ocular examinations were performed at six months and at the end of the patient's participation as well as on entry. The proportions of patients having new cataracts or worsening of existing cataracts was compared to historical data for the incidence of cataracts in a general, elderly population. The tables below (from study report 112E, pp. 54–55) summarize the comparison.

The comparison is rather superficial. The patient characteristics in the studies have not been closely compared, and no attempt has been made to adjust for the length of follow-up. In particular, the crude incidence rates for study 112E at 1 year are compared to those reported by Leske et al. at 2 years.

The incidence of cataracts, however, does not appear alarming in historical context. Even multiplying the rates by two, which would be a crude but likely conservative adjustment for the shorter exposure, the rates on the whole do not exceed the historical rates.

**Table 24: Incidence of New Cataract – Comparison with Published Data**

Age Range	Incidence Rate in this Study		Rate as per Leske et al. <sup>1,2</sup>		Rate as per Taylor and Munoz <sup>3</sup>	
	Nuclear	Cortical	Nuclear	Cortical	Nuclear	Cortical
<65	1.6%	0.3%	3.4%	4.1%	N/A	N/A
≥65	0.5%	0	10.3%	9.5%	N/A	N/A
Total:	1.2%	0.2%	5.9%	6.5%	11–20%	4%

<sup>1</sup>Leske et al., 1996; rates at 2<sup>nd</sup> year of follow-up  
<sup>2</sup>Leske et al., 1997; rates at 2<sup>nd</sup> year of follow-up  
<sup>3</sup>Taylor and Munoz, 1991; rates at one year of follow-up

**Table 25: Rate for Progression of Cataract – Comparison with Published Data**

Age Range	Progression Rate in this Study		Rate as per Leske et al. <sup>1,2</sup>		Rate as per Taylor and Munoz <sup>3</sup>	
	Nuclear	Cortical	Nuclear	Cortical	Nuclear	Cortical
<65	0	5.1%	32.6%	10.4%	N/A	N/A
≥65	4.1%	5.5%	37.2%	8.3%	N/A	N/A
Total:	2.7%	5.4%	35.8%	8.9%	14–16%	18–21%

<sup>1</sup>Leske et al., 1996; rates at 2<sup>nd</sup> year of follow-up  
<sup>2</sup>Leske et al., 1997; rates at 2<sup>nd</sup> year of follow-up  
<sup>3</sup>Taylor and Munoz, 1991; rates at one year of follow-up

Other aspects of safety are discussed in Dr. Lapteva's review.

#### 4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

The tables below, copied from the electronic submission of 29 September 2006, break down the pain outcome in study 112 by age, race, sex and weight. There were no remarkable interactions of treatment with any of these categories.

Table 2.6.1.1  
 Subset Analyses of Mean Change in Pain (Final minus Baseline)  
 Intent To Treat Data Set

VARIABLE		STATISTIC	GROUP1 (N=151)	GROUP2 (N=154)	GROUP3 (N=161)	GROUP4 (N=155)	GROUP5 (N=151)
GENDER	MALE	N	50	50	71	65	56
		MEAN	-7.00	-6.38	-4.93	-5.08	-6.46
		S.D.	4.62	4.14	3.99	4.26	4.00
	FEMALE	N	101	104	90	90	95
		MEAN	-6.92	-5.85	-4.52	-4.49	-6.41
		S.D.	4.86	4.73	4.57	4.41	4.19
AGE	35-55	N	54	38	37	46	41
		MEAN	-6.96	-5.97	-5.00	-5.48	-7.49
		S.D.	4.65	5.06	4.58	4.70	3.66
	56-75	N	94	104	111	94	89
		MEAN	-6.65	-6.21	-4.74	-4.49	-6.15
		S.D.	4.61	4.33	4.24	4.32	4.29
	≥76	N	13	12	13	15	21
		MEAN	-8.77	-4.50	-3.54	-4.00	-5.57
		S.D.	6.03	4.70	4.31	3.16	3.94

GROUP1: PENNSAID + ORAL DICLOFENAC.      GROUP2: PENNSAID + ORAL PLACEBO.      GROUP3: VEHICLE-CONTROL SOLUTION + ORAL PLACEBO.  
 GROUP4: PLACEBO SOLUTION + ORAL PLACEBO.      GROUP5: PLACEBO SOLUTION + ORAL DICLOFENAC.

Table 2.6.1.2  
 Subset Analyses of Mean Change in Pain (Final minus Baseline)  
 Intent To Treat Data Set

VARIABLE		STATISTIC	GROUP1 (N=151)	GROUP2 (N=154)	GROUP3 (N=161)	GROUP4 (N=155)	GROUP5 (N=151)
RACE	WHITE	N	116	120	123	120	119
		MEAN	-7.09	-6.12	-4.94	-4.71	-6.62
		S.D.	4.83	4.45	4.35	4.39	4.00
	OTHER	N	35	34	38	35	32
		MEAN	-6.49	-5.68	-3.92	-4.83	-5.72
		S.D.	4.58	4.90	4.16	4.26	4.51
HEIGHT, KG	≤80	N	58	63	52	57	46
		MEAN	-6.71	-5.98	-4.63	-5.11	-6.00
		S.D.	4.90	4.13	4.52	4.47	3.83
	>80 AND <90	N	28	30	40	27	30
		MEAN	-6.79	-6.00	-4.55	-4.00	-6.50
		S.D.	4.02	5.47	4.07	3.31	4.95
	≥90	N	64	61	69	71	74
		MEAN	-7.27	-6.07	-4.84	-4.72	-6.70
		S.D.	5.01	4.53	4.34	4.60	3.96

GROUP1: PENNSAID + ORAL DICLOFENAC.      GROUP2: PENNSAID + ORAL PLACEBO.      GROUP3: VEHICLE-CONTROL SOLUTION + ORAL PLACEBO.  
 GROUP4: PLACEBO SOLUTION + ORAL PLACEBO.      GROUP5: PLACEBO SOLUTION + ORAL DICLOFENAC.

#### 5 SUMMARY AND CONCLUSIONS

##### 5.1 STATISTICAL ISSUES AND COLLECTIVE EVIDENCE

Studies 112 and 109US each show a modest effect of diclofenac in the test article on the signs and symptoms of osteoarthritis.

5.2 CONCLUSIONS AND RECOMMENDATIONS

Diclofenac in the test article can confidently be concluded to have a modest effect on signs and symptoms of osteoarthritis. Whether the effect is sufficient depends on the balance of benefit and risk, but there is little doubt that there is an effect.

There is no direct evidence of a benefit of DMSO in the test article. If DMSO is considered an active ingredient in a combination product, the combination should not be approved. The applicant's method of addressing the combination policy would render the policy useless in preventing the marketing of irrational mixtures: if a component contributed to the claimed effects, the combination would be acceptable; but if it did not, the combination would be claimed not to be a combination at all. If there are other reasons, however, to consider DMSO an excipient rather than an active ingredient, then the studies conducted are sufficient to demonstrate the efficacy of the single active ingredient diclofenac.

---

6 LABELING

---

The Clinical Studies section of the proposed labeling is as follows:

---

---

---

---

---

---

---

---

---

---

b(4)

This appears to be generally accurate and not overly promotional. I would, however, recommend some changes, as follows:

---

---

---

---

---

**b(4)**

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

/s/

-----  
Thomas Permutt  
11/8/2006 05:32:08 PM  
BIOMETRICS



DEPARTMENT OF HEALTH AND HUMAN SERVICES  
FOOD AND DRUG ADMINISTRATION  
CENTER FOR DRUG EVALUATION AND RESEARCH  
OFFICE OF BIostatISTICS

## Statistical Review and Evaluation CLINICAL STUDIES

NDA: 20-947

Name of drug: PENNSAID® Topical Lotion

Applicant: Dimethaid International Inc.

Indication: \_\_\_\_\_

**b(4)**

Documents reviewed: Sponsor's submitted hard copy: Vol. 1, Vol. 64 – 106  
Sponsor's additional submitted data in 6/28/02 by this reviewer's request

Project manager: Nancy Halonen

Clinical reviewer: James Witter, M.D.

Dates: Received 1/3/02; user fee (2 months) 3/3/02

Statistical reviewer: Suktae Choi, Ph.D.

Statistics team leader: Stan Lin, Ph.D.

Biometrics division director: Mo, Huque, Ph.D.

Keywords: NDA review, clinical studies, analysis of covariance

1 Executive Summary of Statistical Findings	3
1.1 Conclusions and Recommendations	3
1.2 Overview of Clinical Program and Studies Reviewed	3
1.3 Principal Findings	4
2 Statistical Review and Evaluation of Evidence	5
2.1 Introduction and Background	5
2.2 Data Analyzed and Sources	5
2.3 Statistical Evaluation of Evidence on Efficacy / Safety	5
2.3.1 <i>Sponsor's Results and Conclusions</i>	5
2.3.2 <i>Statistical Methodologies</i>	6
2.3.3 <i>Detailed Review of Individual Studies</i>	6
2.3.3.1 <i>Study 107-96</i>	6
2.3.3.2 <i>Study RA-CP-109</i>	6
2.3.3.3 <i>Study RA-CP-109US</i>	7
2.3.4 <i>Statistical Reviewer's Findings</i>	8
2.4 Conclusions and Recommendations	10
2.5 Appendix	11
2.5.1 <i>Tables</i>	11
2.5.2 <i>Figures</i>	14

---

## 1 EXECUTIVE SUMMARY OF STATISTICAL FINDINGS

---

### 1.1 CONCLUSIONS AND RECOMMENDATIONS

This NDA failed to show convincing evidence of efficacy of PENNSAID Topical Lotion versus a control lotion in the treatment of patients with symptoms of primary osteoarthritis (OA) of the knee. All the pivotal studies were not well controlled in many aspects.

### 1.2 OVERVIEW OF CLINICAL PROGRAM AND STUDIES REVIEWED

This NDA is for diclofenac topical lotion \_\_\_\_\_  
\_\_\_\_\_ To support the efficacy claim of this new drug the sponsor submitted 5 Phase III studies; one 4-week study (107-96), three 6-week studies (102-93-1, 108-97, RA-CP-109), and one 12-week study (RA-CP-109US). Among them, study 102-93-1 and study 108-97 were failed to show enough evidence of efficacy based on sponsor's analysis results as submitted. Sponsor insists three studies 107-96, RA-CP-109, and RA-CP-109US, as pivotal trials. However, the duration of study 107-96 (4 weeks) was too short to be considered as pivotal because agency requests at least two 12-week well-controlled success studies. Therefore, this review focused only on RA-CP-109 and RA-CP-109US.

b(4)

#### Study RA-CP-109

This study was a double-blind, randomized, 6-week, DMSO-controlled (vehicle), two-way parallel safety and efficacy study, initiated on November 29, 1999 and completed on August 21, 2000. Patients were randomly assigned to treat their osteoarthritic knee(s) with 40 drops (approximately 1 mL) per knee of PENNSAID, or DMSO solution, four times daily for 42 days. PENNSAID contained 1.5% w/w diclofenac sodium in the full carrier lotion which includes DMSO 45.5% w/w. Patients were allowed of same treatment to both knees if there was pain in both knees, and the more painful knee was included for subsequent analysis. There were three primary efficacy variables: the change from baseline to final assessment in WOMAC LK3.1 Index Pain Subscale score, Physical Function Subscale score, and Patient Global Assessment score. The change from baseline to final assessment in WOMAC Index LK3.1 Stiffness Dimension was a secondary efficacy endpoint. All the efficacy variables were observed only at baseline and at the end of study. ANCOVA using baseline score as a covariate was used for statistical comparison of efficacy endpoints between two treatment groups. Of the 216 randomized and treated patients (PENNSAID:107, DMSO:109), 155 patients (72%) completed 6-week period (PENNSAID:85, DMSO:70), and 128 patients (59%) were included in PP group (PENNSAID:73, DMSO:55). All primary and secondary efficacy endpoints show significant difference between treatment groups by sponsor's analysis.

#### RA-CP-109US

This study was identically designed to study RA-CP-109, except the duration, 12-week instead of 6-week. This study was initiated on December 19, 2000 and completed on May 18, 2001. Of the 326 randomized and treated patients (PENNSAID:164, DMSO:162), 225 patients (69%) completed 12-week period (PENNSAID:119, DMSO:106), and 171 patients

(52%) were included in PP group (PENNSAID:88, DMSO:85). All primary and secondary efficacy endpoints show significant difference between treatment groups by sponsor's analysis.

### 1.3 PRINCIPAL FINDINGS

1. Many randomized patients were not qualified for the objective of these studies (RA-CP-109: 15/216, RA-CP-109US: 49/326) because of many reasons (invalid baseline or final assessment, no radiology evidence of OA, patient does not have primary OA, etc). Sponsor excluded these patients in their primary analyses, and analyses based on this restricted population showed that PENNSAID treated group is significant better than DMSO treated group for all primary efficacy endpoints. However, all randomized and treated patients should be included in an ITT analyses, therefore this reviewer did sensitivity analyses by including these patients with different imputation methods. When very conservative method was used, all the primary efficacy endpoints showed reverse direction (DMSO group showed better efficacy than PENNSAID group).
2. These studies included patients with osteoarthritis of either one knee or both knees. In addition, both knees were allowed treated with study lotion (40 drops for each knee - approximately 1 mL - four times a day). Therefore, some patients were treated only one knee during the whole study period, some patients treated both knees during the whole study period, and others were treated both knees during some period. In other words, even in a same randomized group, subjects were treated differently. This is a study design flaw and reviewer's additional analyses by these subgroups showed not consistent results.
3. No efficacy variables were measurements were made between baseline and final assessments for both studies. Observations in early or middle of the stage are considered as very important secondary efficacy endpoints because they give information about process of the drug efficacy. For example, one is clueless as to when an effect might begin to show and whether any effect was maintained or diminishing at the end of study.
4. DMSO was used as a control drug. In addition, PENNSAID® also contained DMSO. However, DMSO was not considered as a placebo because of its potential efficacy. Moreover PENNSAID® also contains DMSO. Therefore, this drug may have to be considered as a combination drug.

---

## 2 STATISTICAL REVIEW AND EVALUATION OF EVIDENCE

---

### 2.1 INTRODUCTION AND BACKGROUND

This NDA is for diclofenac topical lotion for the treatment of symptoms and signs of osteoarthritis of the knee. To support the efficacy claim of this new drug the sponsor submitted 5 Phase III studies; one 4-week study (107-96), three 6-week studies (102-93-1, 108-97, RA-CP-109), and one 12-week study (RA-CP-109US). Among them, study 102-93-1 and study 108-97 were failed to show enough evidence of efficacy. Sponsor insists three studies 107-96, RA-CP-109, and RA-CP-109US, as pivotal trials. However, the duration of study 107-96 (4 weeks) was too short to be considered as pivotal because agency requests at least two 12-week well-controlled success studies. Therefore, this review focused only on RA-CP-109 and RA-CP-109US. Since RA-CP-109 is a 6-week study, this study should demonstrate “a very high level of efficacy” based on agreement with sponsor’s meeting in June 5, 2000. Following is quoted from the meeting minutes.

FDA responded that there would still be only a 4-week and a 6-week study for assessment of efficacy, and this was problematic in view of the current approach, which involved two 12-week efficacy studies for evaluation of topical products for OA. FDA noted that in light of the duration of ongoing development, a 5-week and a 12-week study could be adequate. The sponsor inquired whether a very high level of efficacy (e.g., a very low p-value) in the ongoing 6-week study would make another study unnecessary. FDA explained that if one can split a study in half and still have significance, that suggests robustness. However, this does not address the need for two separate studies of adequate duration.

There is no doubt that “a very high level of efficacy.” can be achieved only from a very well controlled study.

### 2.2 DATA ANALYZED AND SOURCES

This reviewer requested efficacy data for study RA-CP-109 and RA-CP-109US, and submitted by sponsor in 6/28/02. These data can be found from [\\Cds030\daaodps1\NDA 20-947 PENNSAID\data submission by boistat request](#)

### 2.3 STATISTICAL EVALUATION OF EVIDENCE ON EFFICACY / SAFETY

#### 2.3.1 SPONSOR'S RESULTS AND CONCLUSIONS

For both pivotal studies RA-CP-109 and RA-CP-109US, based on Table 1 in appendix, sponsor concluded as follow;

The efficacy of PENNSAID® was demonstrated by each of the three primary variables, WOMAC Osteoarthritis Index Pain subscale, WOMAC Physical Function subscale and Patient Global Assessment, and was further confirmed by the secondary variable, MOMAC stiffness subscale.

However, many randomized patients were not qualified for the objective of these studies (RA-CP-109: 15/216, RA-CP-109US: 49/326) because of many reasons (invalid baseline or final assessment, no radiology evidence of OA, patient does not have primary OA, etc). Sponsor excluded these patients in their primary analyses.

### 2.3.2 STATISTICAL METHODOLOGIES

Efficacy analysis was performed for intent-to-treat and per-protocol data sets. The change from baseline to final in WOMAC scores and Patient Global Assessment, with baseline as a covariate, was analyzed using ANCOVA to determine a difference between treatment groups.

### 2.3.3 DETAILED REVIEW OF INDIVIDUAL STUDIES

#### 2.3.3.1 *Study 107-96*

This study was a four-week, three way, double-blind, placebo-controlled, parallel safety and efficacy study. Three treatment groups were PENNSAID 40 drops (approximately 1 mL), control (contains just carrier with no diclofenac), and placebo (a token amount of DMSO). There was a substantial discrepancy between the number of patients randomized and treated (248) and the number included in the ITT group (170), because 77 subjects had no valid baseline assessment and 1 had no final assessment. The primary efficacy endpoint was the change from baseline of WOMAC pain. Contrast between least-square means using 2-way ANOVA revealed a significant difference between PENNSAID and control and between PENNSAID and placebo, supporting the efficacy of PENNSAID.

Note that this study was a 4-week study, which is too short to support OA indication. In addition, sponsor's efficacy analysis results are not reliable because more than 30% of the randomized and treated patients were excluded in their primary efficacy analyses. ITT population must include all the randomized and treated patients. There won't be more discussion about this study in this review.

#### 2.3.3.2 *Study RA-CP-109*

Following summary of the study is quoted from sponsor's submission;

##### Design

Protocol #RA-CP-109 was a double-blinded, randomized, 42-day, multi-centered, vehicle-controlled, two-way parallel clinical trial to confirm the safety and efficacy of PENNSAID® in the treatment of the osteoarthritic knee. The study was performed at seventeen centers in Canada. Every patient suffered from osteoarthritis of at least one knee, based on both standard radiological criteria (Altman Atlas), and clinical criteria of pain defined as at least a moderate flare of pain at baseline assessment, as compared with the screening visit assessment, after withdrawal of NSAID or other regularly-used analgesic. There were three primary efficacy variables: the change from baseline to final assessment in (i) WOMAC

LK3.1 Osteoarthritis Index pain subscale score, (ii) WOMAC LK3.1 Osteoarthritis Index physical function subscale score and (iii) Patient Global Assessment score. The secondary variable was the change from baseline to final assessment in WOMAC Osteoarthritis Index stiffness subscale score. Patients were randomly assigned to treat their osteoarthritic knee(s) with 40 drops (approximately 1 mL) per knee of PENNSAID<sup>®</sup>, or vehicle control solution, four times daily for 42 days. The clinical trial PENNSAID<sup>®</sup> formulation was the same as the proposed marketing formulation; the vehicle control contained just the carrier with no diclofenac. Approximately half of the patients suffered bilateral osteoarthritis pain and, for practical reasons, treated both knees. Rescue analgesia with acetaminophen, 325 mg up to four times a day, as needed, was permitted throughout the treatment phase of the study, except for the final week, week 6, leading up to the final WOMAC assessment on day 43.

#### **Efficacy Analysis Results**

Of the 411 screened patients, there were 195 screening failures resulting in 216 randomized and treated patients. The prime reason for screening failure was the lack of a flare of pain, following washout of prior stable analgesic therapy. During auditing of the study it appeared to the sponsor that among these 216 treated patients there were 22 randomized and treated patients who did not meet strictly-defined ICH criteria and should be eliminated from the ITT analysis group (invalid baseline assessment, inadequate washout of prior treatment, no radiographic evidence of primary OA, invalid final assessment, etc.). A detailed Statistical Analysis Plan, filed prior to blind-breaking, then defined two analysis groups: Intent to treat (ITT) - 194 patients, and Per Protocol (PP) - 128 patients. For each of the two data sets analyzed, descriptive statistical analysis revealed that the PENNSAID<sup>®</sup> group had the greatest improvement in pain score, physical function score and patient global assessment. ANCOVA, using baseline score as a covariate, revealed a significant difference between the two treatment groups ( $p < 0.05$ ). Analysis of the secondary variable demonstrated a similar advantage of PENNSAID<sup>®</sup>. For each data set analysed, the PENNSAID<sup>®</sup> group had the greatest improvement in stiffness score. As the WOMAC Index claims that each of its dimensions, (pain, stiffness, and physical function) is an independent assessment, it speaks to the robustness of the conclusion that the analysis of each of its dimensions confirms the efficacy of PENNSAID<sup>®</sup>.

Details of sponsor's analysis results are summarized in Table 1 of appendix, and number of patients who were excluded in sponsor's analysis by the reasons are summarized in Table 2 of appendix.

#### ***2.3.3.3 Study RA-CP-109US***

Following summary of the study is quoted from sponsor's submission;

##### **Design**

Protocol #RA-CP-109-US was a double-blinded, randomized, 12-week, multi-centered, vehicle-controlled, two-way parallel clinical trial to confirm the safety and efficacy of PENNSAID<sup>®</sup> in the treatment of the osteoarthritic knee. The study was performed at 43 centers in the USA. Every patient suffered from osteoarthritis of at least one knee, based on both standard radiological criteria, and clinical criteria of pain defined as at least a moderate flare of pain at baseline assessment, as compared with the screening visit assessment, after withdrawal of NSAID or other regularly-used analgesic. There were three primary efficacy variables: the change from baseline to final assessment in (i) WOMAC LK3.1 Osteoarthritis Index pain subscale score, (ii) WOMAC LK3.1 Osteoarthritis Index physical function subscale score and (iii) Patient Global Assessment score. The secondary variable was the change from baseline to final assessment in WOMAC Osteoarthritis Index stiffness subscale

score. Patients were randomly assigned to treat their osteoarthritic knee(s) with 40 drops (approximately 1 mL) of PENNSAID<sup>®</sup>, or vehicle control solution, four times daily for 84 days. The clinical trial PENNSAID<sup>®</sup> formulation was the same as the proposed marketing formulation; the vehicle control contained just the carrier with no diclofenac. Rescue analgesia with acetaminophen, 325 mg up to four times a day, as needed, was permitted throughout the treatment phase of the study, except for the final week, week 12, leading up to the final WOMAC assessment on day 85.

#### Efficacy Analysis Results

A detailed Statistical Analysis Plan defined two analysis groups: Intent to treat (ITT) - 277 patients, and Per Protocol (PP) - 171 patients. For each of the two data sets analyzed, descriptive statistical analysis revealed that the PENNSAID<sup>®</sup> group had the greatest improvement in pain score, physical function score and patient global assessment. ANCOVA, using baseline score as a covariate, revealed a significant advantage for PENNSAID<sup>®</sup> over the control-DMSO.

Details of sponsor's analysis results are summarized in Table 1 of appendix, and number of patients who were excluded in sponsor's analysis by the reasons are summarized in Table 2 of appendix.

#### 2.3.4 STATISTICAL REVIEWER'S FINDINGS

##### 1. Randomized and treated, but unqualified subjects

Many randomized patients were not qualified for the objective of these studies by the sponsor's judgement (RA-CP-109: 15/216, RA-CP-109US: 49/326), the reasons are summarized in Table 2 of appendix. The sponsor excluded these patients in their primary analyses, and analysis results based on this restricted population showed that PENNSAID treated group is significant better than DMSO treated group for all primary efficacy endpoints. However, all the randomized and treated patients must be included in ITT, therefore this reviewer did sensitivity analyses by including these patients with three different imputation methods. Three different imputations methods used are as follow;

Method 1. Preserve the values if both baseline and final scores are observed. In one of them are missing, impute 0.

Method 2. Preserve the values if both baseline and final scores are observed. If one of them are missing, impute by worst case scenario (defined below)

Method 3. Impute by worst case scenario for all the unqualified subjects.

##### Definition of "Worst case scenario"

For PENNSAID<sup>®</sup> treated group, impute the least improved value (maximum negative change from baseline among its treatment group), and for DMSO treated group, impute the most improved value (minimum negative change from baseline among its treatment group). The imputed values are summarized in Table 4 of appendix.

Analysis results are summarized in Table 3 of appendix. As shown, Method 1 preserved statistical significant difference results, while Method 2 lost statistical significant results with same direction and Method 3 lost even directions. For study RA-CP-109US, Method 3 showed significant results with reverse direction. It is arguable how conservative this method was. However, it is clear that these sensitivity analyses results with worst case scenario demonstrate that it is possible that sponsor's efficacy result may not hold.

2. Inconsistency of symptoms and treatment

These studies included patients with osteoarthritis of either one knee or both knees. In addition, both knees were allowed to be treated with study lotion (40 drops for each knee - approximately 1 mL – four times a day). Therefore, patients can be divided into following three categories by their symptoms and treatments for each randomized group;

Category 1: Feel pain and treated only one knee during the whole study period

Category 2: Feel pain and treated both knees during the whole study period

Category 3: Sometimes feel pain and treated one knee, and other times, both knees.

Numbers of patients in these three categories are summarized in Table 5 of appendix. These different categories may effect the efficacy outcomes very complicatedly in each patient. If a patient feel pain in one knee only, the patient will rely his/her body weight on the other healthy knee in ordinary life. However, if a patient feels pain in both knees, the patient will rely on the less painful knee, or distribute his/her body weight evenly, or may varies time to time. If the patient feels pain in the other knee time to time, it will be very complicate. Therefore the patient's improvement of OA are affected by categories. Moreover, this study allowed to be treated on their knees as the patient want, which made the problem too complicate to adjust. Basically, this study did not consider these complicate but possibly big imbalanced effect issues.

Additional analyses by these three categories were done and summarized in Table 5 and Figure 5 to 7 of appendix. Especially for Study RA-CP-109, Category 2 showed little difference between treated groups, which are the results of such complicate effects.

3. No interim efficacy measurement

No efficacy variables were measurements were made between baseline and final assessments for both studies. Observations in early or middle of the stage are considered as very important secondary efficacy endpoints because they give information about process of the drug efficacy. For example, one is clueless as to when an effect might begin to show and whether any effect was maintained or diminishing at the end of study.

4. DMSO-controlled

DMSO was used as a control treatment ingredient for both studies. However, DMSO is not considered as a placebo because of its potential efficacy. In addition, PENNSAID® also contained DMSO. Therefore, this drug may have to be considered as combination drug.

#### 2.4 CONCLUSIONS AND RECOMMENDATIONS

This NDA failed to show enough evidence of efficacy of PENNSAID Topical Lotion versus a control lotion in the treatment of patients with symptoms of primary osteoarthritis (OA) of the knee.

Among five submitted studies, this review concentrated on two pivotal studies, one 6-week (RA-CP-109) and one 12-week (RA-CP-109US). The agency agreed a 6-week and a 12-week study could be adequate if a very high level of efficacy was demonstrated. However, these two studies can not be considered as well-controlled study because of following reasons: First, many patients were randomized and treated but their efficacy data are not qualified. This reviewer's sensitivity analysis results did not preserve the sponsor's analysis results based on restricted population. Second, patients were allowed to treat both knees if they feel pain, so that, symptoms and treated amount of study drug were not standard. Third, only baseline and final efficacy variables were observed but no interim efficacy variables were observed. Forth, DMSO was used as a control ingredient.

"A very high level of efficacy" can be achieved only from a well-controlled study. However, sponsor failed to plan and to perform both pivotal studies as well controlled.

2.5 APPENDIX

2.5.1 TABLES

Table 1. Sponsor's primary efficacy analysis results (Sponsor's restricted ITT)

Primary Efficacy Endpoints	RA-CP-109			RA-CP-109US		
	LS Mean (Std) <sup>a</sup>		P-value <sup>b</sup>	LS Mean (Std) <sup>a</sup>		P-value <sup>b</sup>
	PENNSAID	DMSO		PENNSAID	DMSO	
WOMAC Pain	N=98 -5.6 (4.9)	N=96 -3.5 (4.2)	0.0035	N=133 -6.4 (4.8)	N=144 -4.3 (4.5)	0.0001
WOMAC Physical Functions	N=97 -14.3 (16.4)	N=97 -7.2 (13.1)	0.0005	N=132 -16.9 (15.7)	N=144 -10.2 (14.1)	0.0003
Patient's Global	N=96 -1.3 (1.3)	N=97 -0.7 (1.2)	0.0001	N=131 -1.4 (1.2)	N=144 -0.9 (1.2)	0.0004

a. Change from baseline  
b. baseline as covariate

Table 2. Disposition of patients with unqualified efficacy data (excluded from sponsor's ITT)

Study	Description	WOMAC Pain		WOMAC Physical Functions		Patient's Global	
		PENN	DMSO	PENN	DMSO	PENN	DMSO
RA-CP-109	Reasons						
	Invalid baseline assessment	1 (11%)	2 (15%)	3 (30%)	2 (17%)	3 (27%)	2 (17%)
	Invalid final assessment	7 (78%)	8 (62%)	5 (50%)	7 (58%)	6 (55%)	7 (58%)
	No radiological evidence of OA	1 (11%)	1 (8%)	1 (10%)	1 (8%)	1 (9%)	1 (8%)
	Patient does not have primary OA	0 (0%)	2 (15%)	1 (10%)	2 (17%)	1 (9%)	2 (17%)
	Total	9	13	10	12	11	12
	Data existence						
Both baseline and final	3 (33%)	9 (69%)	3 (30%)	10 (83%)	3 (27%)	9 (75%)	
Missing baseline or final	6 (67%)	4 (31%)	7 (70%)	2 (17%)	8 (73%)	3 (25%)	
RA-CP-109US	Reasons						
	Invalid baseline assessment	0 (0%)	0 (0%)	1 (3%)	0 (0%)	2 (6%)	0 (0%)
	Invalid final assessment	28 (90%)	12 (90%)	28 (88%)	12 (67%)	28 (85%)	12 (67%)
	Patient does not have primary OA	0 (0%)	1 (0%)	0 (0%)	1 (6%)	0 (0%)	1 (6%)
	Use of Prohibited Medication/Others	3 (10%)	5 (10%)	3 (9%)	5 (28%)	3 (9%)	5 (28%)
	Total	31	18	32	18	33	18
	Data existence						
Both baseline and final	25 (81%)	17 (81%)	24 (75%)	17 (94%)	24 (73%)	17 (94%)	
Missing baseline or final	6 (19%)	1 (19%)	8 (25%)	1 (6%)	9 (27%)	1 (6%)	

**Table 3. Reviewer's analysis results including all the randomized and treated subjects with different imputation methods for unqualified data; All Randomized and Treated (ITT)**

Primary Efficacy Endpoints	RA-CP-109			RA-CP-109US		
	LS Mean (Std Err) <sup>a</sup>		P-value <sup>b</sup>	LS Mean (Std Err) <sup>a</sup>		P-value <sup>b</sup>
	PENNSAID	DMSO		PENNSAID	DMSO	
Method 1 <sup>c</sup>						
WOMAC Pain	N=107 -5.21 (0.42)	N=109 -3.47 (0.42)	0.0040	N=164 -5.93 (0.34)	N=162 -4.41 (0.34)	0.0017
WOMAC Physical Functions	N=107 -13.11 (1.37)	N=109 -7.45 (1.35)	0.0036	N=164 -15.27 (1.09)	N=162 -10.43 (1.09)	0.0019
Patient's Global	N=107 -1.26 (0.11)	N=109 -0.70 (0.11)	0.0004	N=164 -1.32 (0.09)	N=162 -0.96 (0.09)	0.0036
Method 2 <sup>d</sup>						
WOMAC Pain	N=107 -4.9 (0.47)	N=109 -4.1 (0.47)	0.2578	N=164 -5.68 (0.37)	N=162 -4.53 (0.37)	0.0281
WOMAC Physical Functions	N=107 -11.2 (1.62)	N=109 -8.4 (1.59)	0.2155	N=164 -14.32 (1.18)	N=162 -10.82 (1.18)	0.0363
Patient's Global	N=107 -1.1 (0.13)	N=109 -0.8 (0.13)	0.0614	N=164 -1.24 (0.10)	N=162 -0.99 (0.10)	0.0619
Method 3 <sup>e</sup>						
WOMAC Pain	N=107 -4.57 (0.53)	N=109 -5.19 (0.53)	0.4097	N=164 -3.85 (0.51)	N=162 -6.05 (0.51)	0.0026
WOMAC Physical Functions	N=107 -10.02 (1.88)	N=109 -12.12 (1.85)	0.4274	N=164 -9.40 (1.61)	N=162 -16.03 (1.61)	0.0039
Patient's Global	N=107 -1.08 (0.14)	N=109 -1.08 (0.14)	0.9703	N=164 -0.77 (0.12)	N=162 -1.28 (0.12)	0.0035

- a. Change from baseline
- b. baseline as covariate
- c. Method 1: Preserve the values if both baseline and final scores are observed. If one of them are missing, impute 0.
- d. Method 2: Preserve the values if both baseline and final scores are observed. If one of them are missing, impute by worst case scenario
- e. Method 3: Impute by worst case scenario for all the unqualified subjects.

**Table 4. Summary of imputed values for worst case scenario**

	RA-CP-109			RA-CP-109		
	WOMAC Pain	WOMAC Physical Function	Patient Global	WOMAC Pain	WOMAC Physical Function	Patient Global
PENNSAID Group	6	40	2	7	22	2
DMAO Group	-17	-51	-4	-20	-62	-4

**Table 5. Subgroup analysis results of patient's dominated knee by three categories; All Randomized and treated (ITT), Method 1 for imputation of unqualified data**

Primary Efficacy Endpoints	RA-CP-109			RA-CP-109US		
	LS Mean (Std Err) <sup>a</sup>		P-value <sup>b</sup>	LS Mean (Std Err) <sup>a</sup>		P-value <sup>b</sup>
	PENNSAID	DMSO		PENNSAID	DMSO	
<b>Category 1<sup>c</sup></b>						
WOMAC Pain	N=23 -5.52 (4.54)	N=18 -3.44 (4.16)	0.0763	N=37 -6.11 (4.64)	N=33 -4.15 (4.27)	0.0873
WOMAC Physical Functions	N=23 -12.3 (12.98)	N=18 -6.67 (11.95)	0.1149	N=37 -14.78 (14.24)	N=33 -9.03 (13.59)	0.1174
Patient's Global	N=23 -1.17 (1.23)	N=18 -0.44 (1.25)	0.0070	N=37 -1.24 (0.93)	N=33 -0.73 (1.10)	0.0472
<b>Category 2<sup>d</sup></b>						
WOMAC Pain	N=63 -5.24 (5.34)	N=75 -3.15 (4.28)	0.0200	N=100 -5.92 (4.73)	N=110 -4.40 (4.68)	0.0182
WOMAC Physical Functions	N=63 -13.25 (18.08)	N=75 -6.68 (13.50)	0.0192	N=100 -16.13 (15.92)	N=110 -10.49 (14.75)	0.0081
Patient's Global	N=63 -1.25 (1.36)	N=75 -0.67 (1.15)	0.0035	N=100 -1.35 (1.28)	N=110 -1.00 (1.16)	0.0393
<b>Category 3<sup>e</sup></b>						
WOMAC Pain	N=19 -5.11 (4.34)	N=15 -5.2 (4.30)	0.9500	N=24 -6.50 (4.59)	N=19 -4.84 (3.48)	0.1222
WOMAC Physical Functions	N=19 -13.32 (14.00)	N=15 -12.53 (14.33)	0.8112	N=24 -14.54 (14.22)	N=19 -11.47 (9.48)	0.4178
Patient's Global	N=19 -1.21 (1.32)	N=15 -1.33 (0.98)	0.8635	N=24 -1.42 (1.10)	N=19 -1.11 (1.15)	0.3676

- a. Change from baseline of dominated knee selected by patient
- b. baseline as covariate
- c. Category 1: Feel pain and treated only one knee during the whole study period
- d. Category 2: Feel pain and treated both knees during the whole study period
- e. Category 3: Sometimes feel pain and treated one knee, and other times, both knees.

2.5.2 FIGURES

Figure 1. Mean  $\pm$  Standard Error of WOMAC Pain change from baseline by three imputation methods; All Randomized and Treated

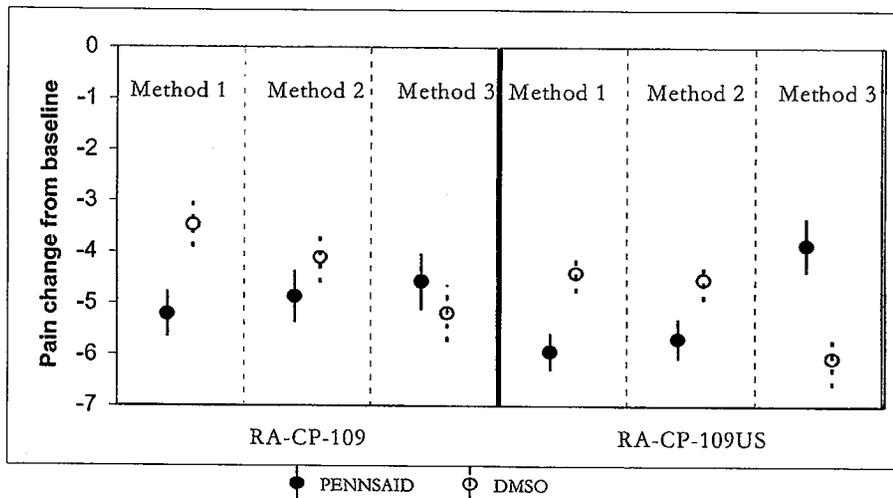


Figure 2. Mean  $\pm$  Standard Error of WOMAC Physical Function change from baseline by three imputation methods; All Randomized and Treated

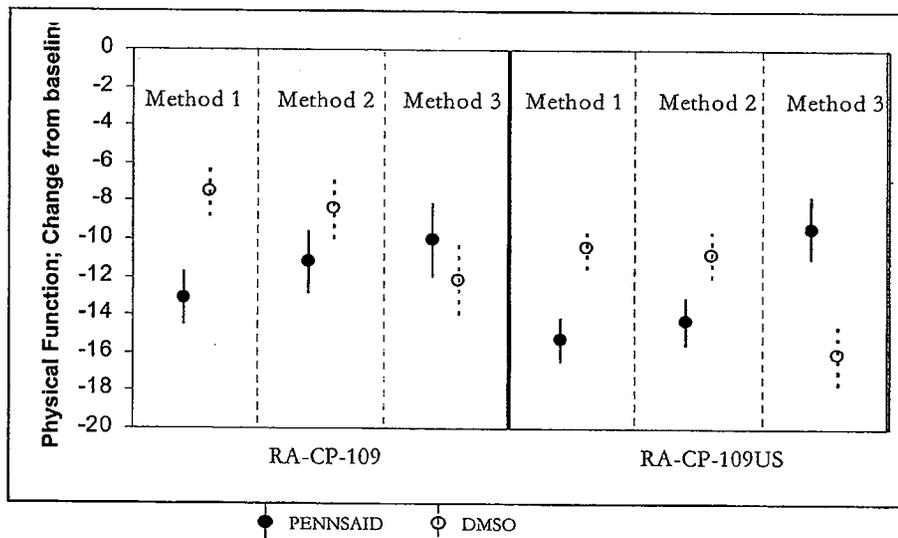


Figure 3. Mean  $\pm$  Standard Error of Patient Global change from baseline by three imputation methods; All Randomized and Treated

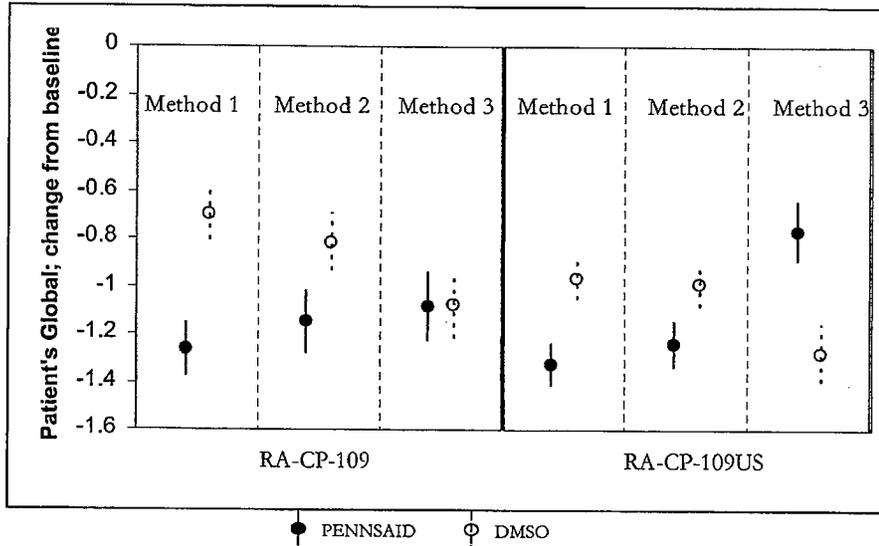


Figure 4. Mean  $\pm$  Standard Error of WOMAC Pain change from baseline by three categories; All Randomized and Treated

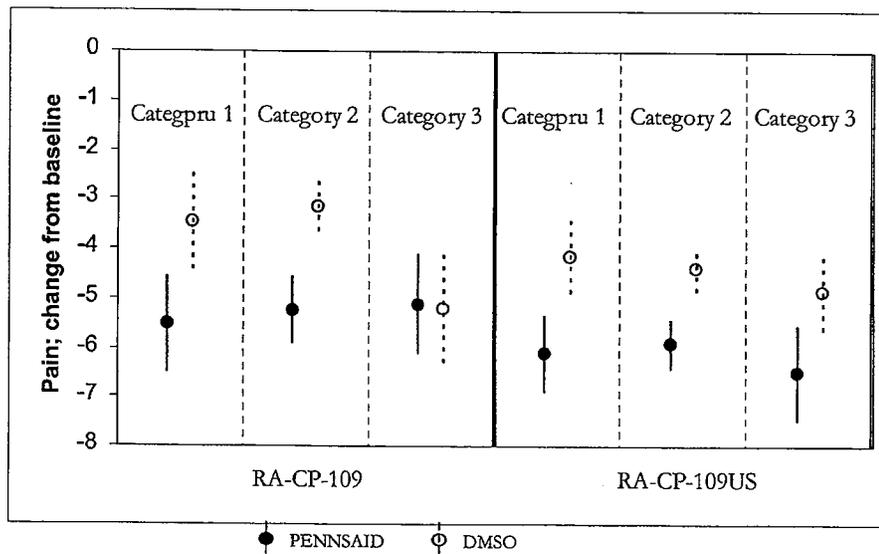


Figure 5. Mean  $\pm$  Standard Error of WOMAC Physical Function change from baseline by three categories; All Randomized and Treated

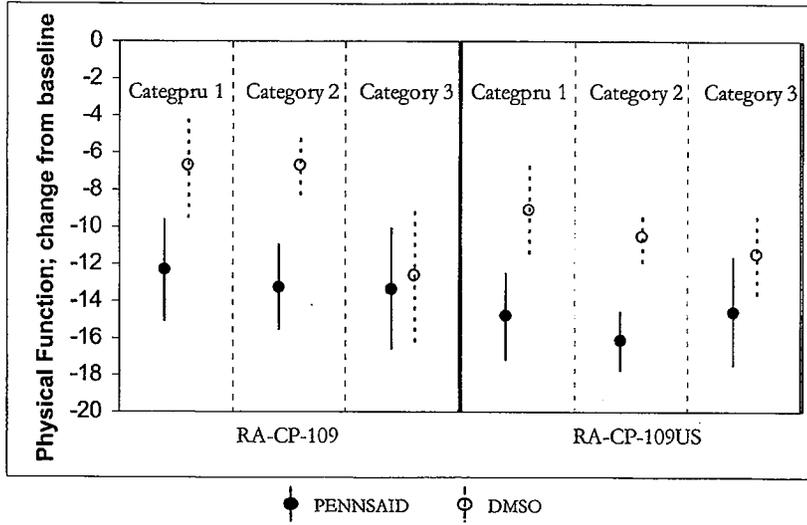
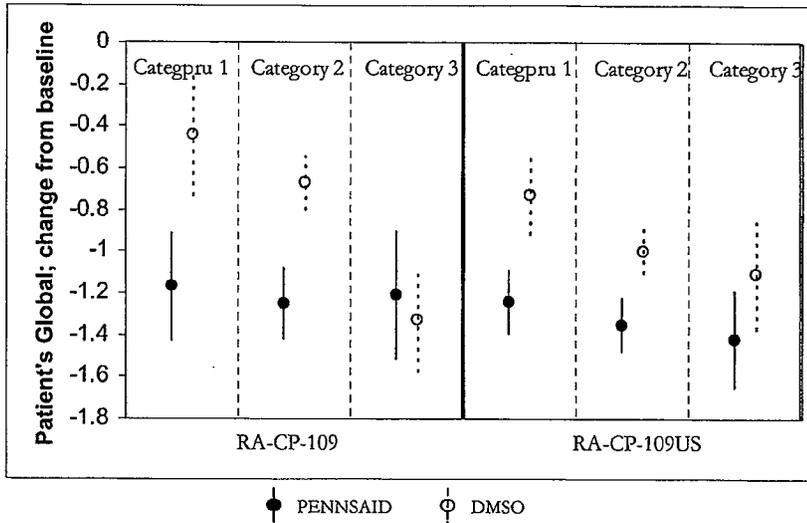


Figure 6. Mean  $\pm$  Standard Error of Patient Global change from baseline by three categories; All Randomized and Treated



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

/s/

-----  
Suktae Choi  
7/26/02 05:44:09 PM  
BIOMETRICS

Stan Lin  
7/26/02 05:51:47 PM  
UNKNOWN