CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:
21-306

STATISTICAL REVIEW(S)

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Translational Sciences
Office of Biostatistics

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/Serial Number:	21-306/000
Drug Name:	Butrans (buprenorphine) for transdermal administration
Indication:	Relief of moderate to severe pain in patients requiring continuous, around-the-clock opioid treatment for an extended period of time
Applicant:	Purdue Pharma
Date(s):	Stamp Date: 9/30/2009
	PDUFA Date: 3/30/2010
Review Priority:	Standard Resubmission
Biometrics Division:	DBII
Statistical Reviewer:	Jonathan Norton, Ph.D.
Concurring Reviewer:	Dionne Price, Ph.D.
Medical Division:	Division of Anesthesia, Analgesia, and Rheumatology Products
Clinical Team:	Robert Levin, M.D.
	Robert Shibuya, M.D.
Project Manager:	Matthew Sullivan

Keywords: Active control/superiority, clinical studies, data imputation, enrichment design, mixed models

Table of Contents

1 EXECUTIVE SUMMARY 3

1.1 CONCLUSIONS AND RECOMMENDATIONS 3
1.2 BRIEF OVERVIEW OF CLINICAL STUDIES 3
1.3 STATISTICAL ISSUES AND FINDINGS 4

2 INTRODUCTION 5

2.1 OVERVIEW 5
2.2 DATA SOURCES 6

3 STATISTICAL EVALUATION 7

3.1 EVALUATION OF EFFICACY 7
3.1.1 Study BUP3024 7
3.1.2 Study BUP3015 17
3.2 EVALUATION OF SAFETY 24

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS 25

4.1 GENDER, RACE AND AGE 25
4.1.1 Study BUP3024 25
4.1.2 Study BUP3015 26
4.2 OTHER SPECIAL/SUBGROUP POPULATIONS 27

5 SUMMARY AND CONCLUSIONS 28

5.1 STATISTICAL ISSUES AND COLLECTIVE EVIDENCE 28
5.2 CONCLUSIONS AND RECOMMENDATIONS 28
5.3 REVIEW OF THE PROPOSED LABEL 28

6 APPENDIX A: ADDITIONAL STATISTICAL DETAILS AND DISCUSSION 33

1 EXECUTIVE SUMMARY

1.1 Conclusions and Recommendations

The Applicant, Purdue Pharma LP, seeks to have Butrans, a transdermal buprenorphine patch, approved for management of moderate to severe pain. The two adequate and well-controlled studies featured in this submission support a finding of efficacy for this product.

1.2 Brief Overview of Clinical Studies

This NDA was originally submitted in November 2000. In August 2001, FDA sent a Not Approvable (NA) letter which cited 62 deficiencies. Deficiency #55 stated:

> You have not provided substantial evidence that the drug will have its intended clinical effect… Submit the results of additional adequate and well-controlled studies of appropriate duration and in relevant target populations to provide evidence of the effectiveness of the product and the durability of the treatment effect.

The Applicant had submitted two trials intended to serve as adequate and well-controlled studies, but the NA letter faulted the studies because their apparently positive results were an artifact of imputing good outcomes to patients who discontinued. Since that time this NDA has fallen within the purview of different divisions who have given advice to the Sponsor on multiple occasions.

In March 2007, the Division of Anesthesia, Analgesia, and Rheumatology Products (DAARP) stated that two appropriately designed studies in enriched populations could serve as adequate evidence of efficacy. However, the label would have to include the number of patients who failed to titrate to effect in each study. Successful studies would have to show efficacy at Week 12 of the double-blind period.

BUP3024 was a double-blind randomized withdrawal study to evaluate the efficacy of Butrans for the treatment of opioid-naïve subjects with chronic low back pain. It was conducted at 86 sites in the United States. The study enrolled 1027 patients (excluding screening failures), of whom 541 were randomized into the double-blind phase of the study.

Subjects in the study were titrated to either the 10 mcg/h dose (BTDS 10) or the 20 mcg/h dose (BTDS 20) of Butrans. They were then randomized to either the titrated dose or matching placebo. (BTDS 20 is a larger patch than BTDS 10.) Even after randomization, subjects assigned to BTDS 20 (or matching placebo) could be titrated down to BTDS 10 if they did not tolerate that dose. If the subject did not then respond to BTDS 10, then the investigator had the option to up-titrate back to BTDS 20. In contrast, subjects who started the double-blind period on BTDS 10 were not allowed to titrate during this phase of the study.

For the first six days of the Double-Blind Treatment Period, subjects were permitted to take immediate-release oxycodone 5 mg b.i.d. as rescue medication. For the remainder of the study, subjects were permitted to take acetaminophen or ibuprofen for this purpose.

The primary efficacy endpoint for this study was "average pain over the last 24 hours" at Week 12 of the Double-Blind Treatment Period, comparing titrated Butrans to placebo. This score was based on the 24 hours prior to the corresponding clinic visit, and the subject was told to abstain from rescue medication for three days prior to the visit. The secondary efficacy variables were the daily number of non-opioid rescue tablets and the Sleep Disturbance Subscale of the Medical Outcomes Study (MOS) Sleep Scale. The primary analysis population was randomized subjects who received a dose of double-blind study drug.

BUP3015 was a randomized, double-blind active comparator study in opioid-experienced subjects with moderate to severe low back pain. It was conducted in 75 sites in the United States. The study enrolled 1292 subjects, 662 of whom were randomized. Subjects were titrated to BTDS 20, then randomized to either BTDS 5, BTDS 20, or oxycodone 40 mg per day. Acetaminophen or ibuprofen was allowed as rescue medication.

The Applicant's primary efficacy variable was "average pain over the last 24 hours" at Weeks 4, 8, and 12, with the primary comparison between BTDS 5 and BTDS 20. Following the current thinking in the Division of Anesthesia, Analgesia, and Rheumatology Products (DAARP) on trials for chronic pain drugs, I used the pain at the Week 12 visit as the primary endpoint. The secondary efficacy variables were the number of tablets of rescue analgesia used in the double-blind phase, the Oswestry Disability Index, and the Sleep Disturbance Subscale of the MOS Sleep Scale. The full analysis population consisted of randomized subjects who received at least one dose of study medication.

1.3 Statistical Issues and Findings

The Applicant was granted a Special Protocol Agreement (SPA) for the primary efficacy analysis for BUP3024. This planned analysis showed the titrated Butrans dose to be statistically superior to placebo, decreasing the pain at double-blind Week 12 by .58 points on an 11-point scale. This magnitude of treatment effect is similar to what is found in other chronic pain trials. A number of sensitivity analyses were also conducted, including baseline observation carried forward (BOCF) imputation, and all either significantly favored Butrans or trended in the right direction. I was able to replicate the Applicant's findings, and did not identify any major issues that could undermine their validity.

For Study BUP3015, the Applicant's primary analysis was unacceptable for two reasons. First of all, they used average pain at Weeks 4, 8, and 12 as the primary endpoint. This goes against the current thinking in DAARP, which holds that a chronic pain drug must be effective at Week 12. The other problem with the Applicant's analysis is that they did not use any imputation for missing data. They provide an extensive rationale for this decision, which I address in the Appendix.

For my analysis of this trial, I focused on Week 12 and used BOCF imputation. This analysis showed BTDS 20 to be statistically superior to BTDS 5, reducing pain at Week 12 by .62 points. This is similar to the Applicant's finding of a .67 point reduction, also statistically significant.

The Applicant complicated interpretation of BUP3015 by halting the study after 74% of the planned sample size was enrolled. The stated rationale was "changing business conditions". This decision is troubling, as it raises the theoretical possibility of multiple looks at the data, which would invalidate the reported p-value. However, the Applicant states the trial was still unblinded when the decision was made to halt. Assuming that this was the case, the trial is acceptable to provide evidence of efficacy.

2 INTRODUCTION

2.1 Overview

The Applicant, Purdue Pharma LP, seeks to have a transdermal buprenorphine patch approved for management of moderate to severe pain. The patch, which has the proposed tradename Butrans, would be marketed in three dosage strengths: 5 mcg/h, 10 mcg/h, and 20 mcg/h. A larger dose is administered by using a larger patch. Herein I will refer to these products as BTDS 5, 10, or 20, respectively.

Buprenorphine is a Schedule III controlled substance that was originally approved in the US in 1981. Since that time, buprenorphine drugs have been approved for both pain and addiction indications, in either parental or sublingual dosage forms. This NDA was originally submitted in November 2000. In August 2001, FDA sent an NA letter which cited 62 deficiencies. Deficiency #55 stated:

> You have not provided substantial evidence that the drug will have its intended clinical effect… Submit the results of additional adequate and well-controlled studies of appropriate duration and in relevant target populations to provide evidence of the effectiveness of the product and the durability of the treatment effect.

The Applicant had submitted two trials intended to serve as adequate and well-controlled studies, but the NA letter faulted the studies because their apparently positive results were an artifact of imputing good outcomes to patients who discontinued. Also, Deficiency #59 faulted the studies for increasing the dose too quickly to properly assess efficacy, after as soon as three days.

In December 2004, the Division of Anesthetic, Critical Care, and Addiction Drug Products (DACCADP) met with the Applicant to discuss the statistical plan for a potential Phase 3 study. DACCADP stated that last observation carried forward (LOCF) imputation was not acceptable because it could introduce "bias". The Division suggested BOCF imputation as an alternative, and also requested a "responder" analysis in which all dropouts are treated as non-responders.

Due to a reorganization within FDA, this NDA fell within the purview of DAARP. In March 2007, DAARP held a teleconference with the Applicant to discuss a request for a Special Protocol Agreement for BUP3024. At this conference, DAARP stated that two appropriately designed studies in enriched populations could serve as adequate evidence of efficacy. However, the label would have to include the number of patients who failed to titrate to effect in each study. DAARP also stated that a successful trial would have to show efficacy at Week 12. The Division indicated that a "hybrid" LOCF/BOCF imputation would be acceptable, provided that the discontinuations were appropriately categorized.

In May 2007, DAARP issued a Special Protocol Agreement letter for BUP3024. In the letter, DAARP concurred with the following statement which the Applicant made in the briefing packet:

> Our understanding, based on the teleconference of March 20, 2007, is that sensitivity analyses are not required for approval, and if performed, would not be required to be statistically significant at the 5% level. We have, however, indicated some additional analyses which may be performed to assess the sensitivity of the primary efficacy analysis to the choice of missing data imputation and choice of primary time point. We do not expect all of these sensitivity analyses to reach statistical significance at the 5% level. However, we would look for a consistency of treatment effect (similar estimate of BTDS mean) across the primary and all sensitivity analyses.

The Applicant held a pre – Complete Response meeting with DAARP in September 2008. The format of the submission was discussed.

2.2 Data Sources

The electronic version of this NDA can be found at
\\CDSESUB1\EVSPROD\NDA021306\021306.enx

3 STATISTICAL EVALUATION

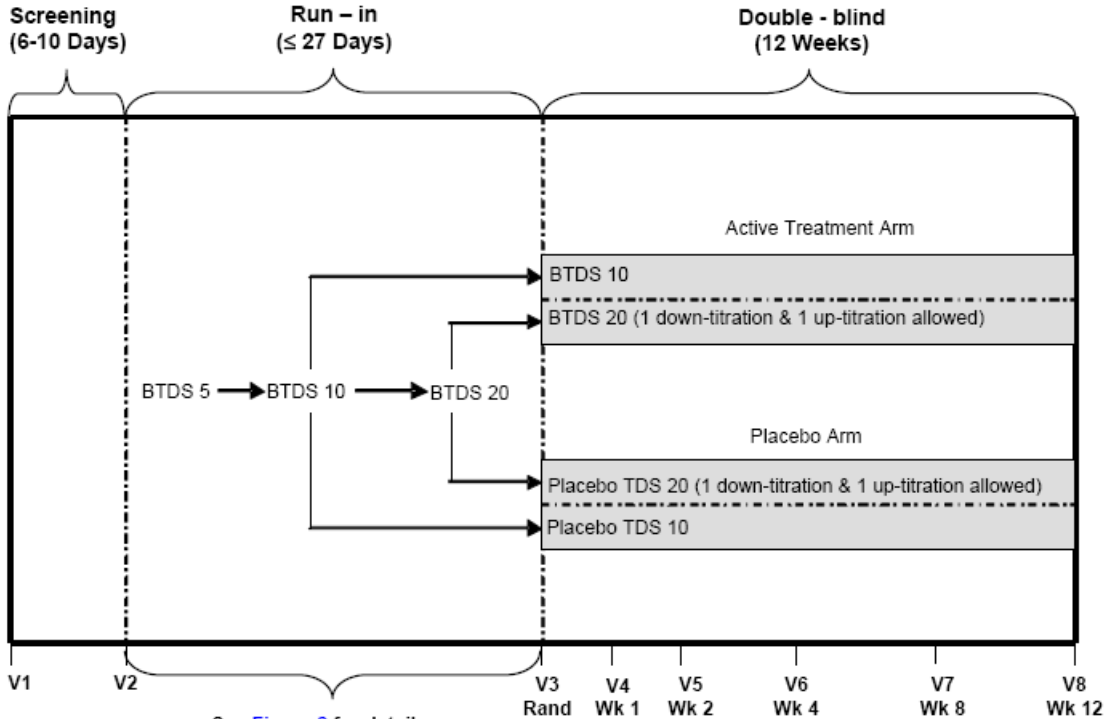
3.1 Evaluation of Efficacy

3.1.1 Study BUP3024

Study Design and Endpoints

BUP3024 was a double-blind randomized withdrawal study to evaluate the efficacy of Butrans for the treatment of opioid-naïve subjects with chronic low back pain. It was conducted at 86 sites in the United States. The study enrolled 1027 patients (excluding screening failures), of whom 541 were randomized into the double-blind phase of the study.

Figure 1: Overall Study Design (Source: Clinical Study Report, Figure 1)

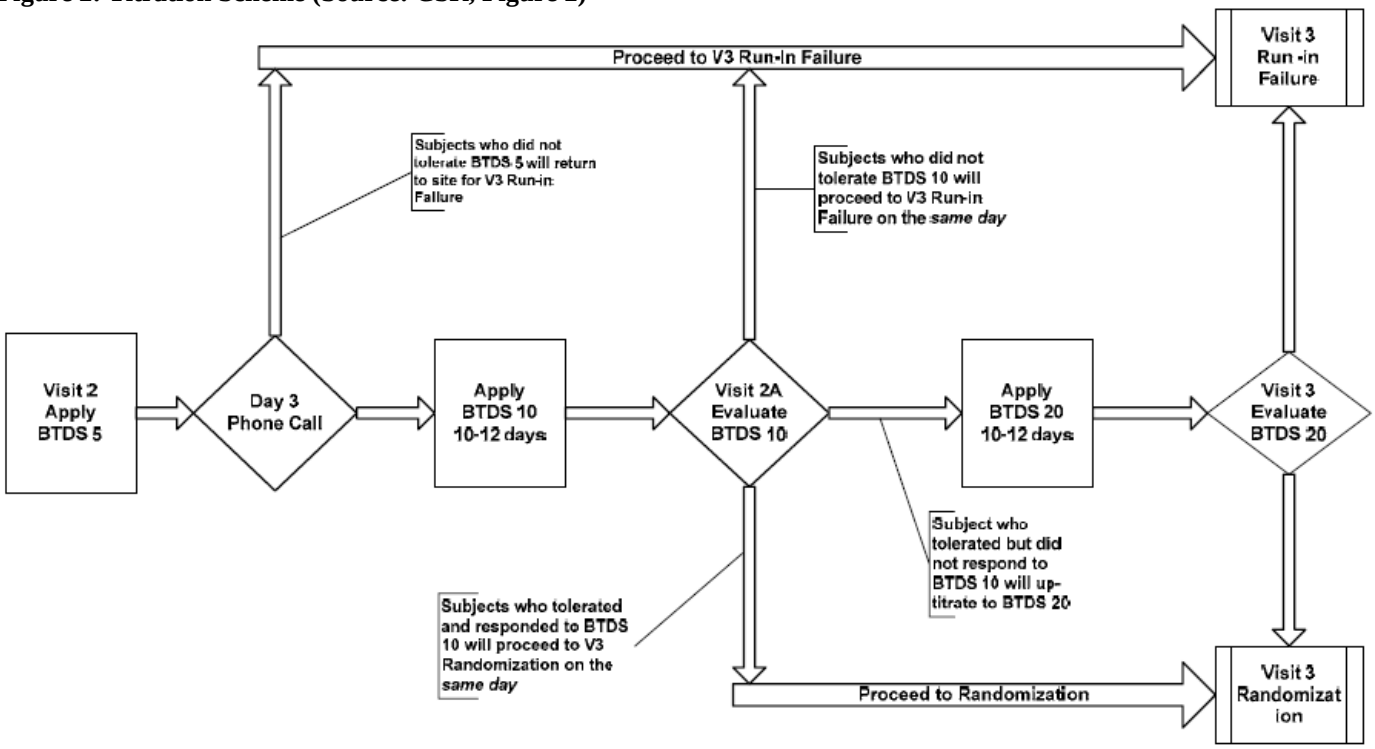


The overall design of the study is shown in Figure 1. During the Screening Period, subjects stopped all analgesics and "other medications used for chronic pain" and reported daily pain scores. Subjects entered the Run-In Period if and when they reported a pain score of at least 5, on an 11-point scale, for two consecutive days. The design of the Run-In Period is shown in Figure 2. The purpose of this period was to titrate each subject to either BTDS 10 or BTDS 20. Subjects were given BTDS 5 for three days. Those who tolerated this dose were switched BTDS 10; those who did not were discontinued from the study. Subjects stayed on BTDS 10 for 10 days, then were assessed for responsiveness and tolerance. Responsiveness was defined as a reduction of at least two points in daily pain for three consecutive days. Subjects who both tolerated and

responded to BTDS 10 were immediately randomized into double-blind treatment. Those who did not tolerate BTDS 10 were discontinued, and those who tolerated it but did not respond to it were switched to BTDS 20. After 10 days, those subjects who both tolerated and responded to BTDS 20 were randomized at that dose. Otherwise, they were discontinued.

Subjects began the Double-Blind Treatment Period on their titrated dose or matching placebo. If a subject assigned to BTDS 20 (or matching placebo) was not tolerating that dose then the investigator had the option to titrate down to BTDS 10. If the subject did not respond at that dose, then the investigator had the option to up-titrate back to BTDS 20. No further titrations were allowed. Those subjects who started this period on BTDS 10 were not allowed to titrate.

Figure 2: Titration Scheme (Source: CSR, Figure 2)



For the first six days of the Double-Blind Treatment Period, subjects were permitted to take immediate-release oxycodone 5 mg b.i.d. as rescue medication. For the remainder of the study, subjects were permitted to take acetaminophen (APAP) 500 mg every six hours. If APAP was contraindicated, then they could take ibuprofen 200 mg instead. The study medication was not tapered, as is commonly done to mitigate opioid withdrawal symptoms. Opioid rescue analgesia was allowed for six days, however.

The primary efficacy endpoint was “average pain over the last 24 hours” at Week 12 of the Double-Blind Treatment Period. This score was based on the 24 hours prior to the corresponding clinic visit, and the subject was told to abstain from rescue medication for three days prior to the visit. The secondary efficacy variables were the daily number of non-opioid rescue tablets and the Sleep Disturbance Subscale of the MOS Sleep Scale. The primary analysis population was randomized subjects who received a dose of double-blind study drug.

Patient Disposition, Demographic and Baseline Characteristics

Figure 3 and Table 1 show the disposition of subjects, with the reported reasons for discontinuation. They were provided by the Applicant, but I verified the contents. Note, per footnote *a*, that there were two additional randomized subjects who were excluded from Table 1. Subjects who were randomized to placebo were actually slightly more likely to complete the study and stay on the study treatment. This difference is not statistically significant, however.

Figure 3: Subject Disposition (Source: CSR, Figure 5)

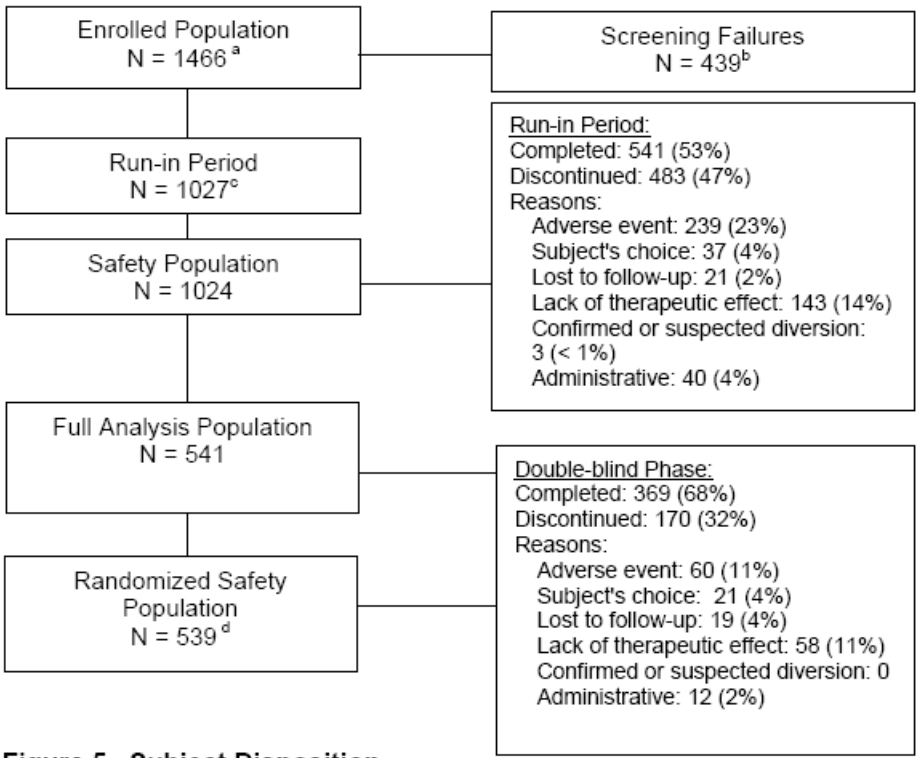


Table 1: Reasons for Discontinuation (Source: CSR, Table 5)

Category	Prerandomization Phase	Double-blind Phase[a] N = 539	
	Run-in Period[b] (n = 1024[c])	BTDS (n = 256)	Placebo (n = 283)
Completed period/phase on study drug, n (%)	541 (53)	170 (66)	199 (70)
Discontinued study drug – all cases,[d] n (%)	483 (47)	86 (34)	84 (30)
Drug discontinuation reason			
Adverse event	239 (23)	40 (16)	20 (7)
Subject's choice	37 (4)	10 (4)	11 (4)
Lost to follow-up	21 (2)	8 (3)	11 (4)
Lack of therapeutic effect	143 (14)	22 (9)	36 (13)
Confirmed or suspected diversion	3 (< 1)	0	0
Administrative	40 (4)	6 (2)	6 (2)
Discontinued study drug and study simultaneously, n (%)		70 (27)	71 (25)
Drug discontinuation reason			
Adverse event		32 (13)	18 (6)
Subject's choice		10 (4)	11 (4)
Lost to follow-up		8 (3)	11 (4)
Lack of therapeutic effect		15 (6)	25 (9)
Confirmed or suspected diversion		0	0
Administrative		5 (2)	6 (2)
Discontinued study drug and stayed in study, n (%)[e]		16 (6)	13 (5)
Drug discontinuation reason			
Adverse event		8 (3)	2 (< 1)
Subject's choice		0	0
Lack of therapeutic effect		7 (3)	11 (4)
Administrative		1 (< 1)	0

Cross-references: Table 14.1.1.1; Appendices 16.2.1.1 and 16.2.1.2.

[a] There were 2 subjects who were randomized but did not have safety data during double-blind treatment and thus were not included in the randomized safety population: subject 0043020 (randomized to placebo), discontinued due to subject's choice and subject 0018008 (randomized to BTDS), lost to follow-up.

[b] Does not include screen failures.

[c] Does not include subjects 0039009, 0019038, and 0067001 who entered the run-in period, but were not dosed and were therefore excluded from the safety population.

[d] Subjects could discontinue double-blind study drug but elect to complete all visits and assessments.

[e] Only the reason for study drug discontinuation is tabulated for subjects who discontinued study drug but continued study visits and discontinued from the study later.

Robert Levin, M.D., raised the question of whether BTDS 10 is an effective dose. One way to look at this question is to see whether subjects who where titrated to this dose managed to stay in the study. Table 2 shows that proportion of subjects who completed the study and stayed on drug by treatment arm and by the dose they were on when they began the double-blind treatment

period. Two features of the trial should be noted in order to properly interpret this table. First, the titrated starting dose simply determined the ***size of the patch*** at the beginning of the double-blind period. It is the treatment arm that determined whether or not the patch contained active drug or not. Secondly, subjects who began on a BTDS 20 patch (or matching placebo) were allowed to titrate down and back up once. On the other hand, subjects who began on a BTDS 10 patch were not allowed to titrate.

Table 2: Disposition by Starting Dose and Arm, Full Analysis Set (Source: Reviewer)

Starting Dose (Patch Size)	Completed?	BTDS	Placebo
10	Y (%)	84 (70%)	106 (77%)
	N (%)	36 (30%)	31 (23%)
20	Y (%)	86 (63%)	93 (63%)
	N (%)	51 (37%)	54 (37%)

Notes: The starting dose determined the size of the patch, which contained either BTDS or placebo.

As the table shows, subjects who started on active BTDS 10 actually trended towards being more likely to complete the study than those who started on the higher dose. This is noteworthy in light of the fact that those who started on BTDS 20 had the option of titrating down, while those on BTDS 10 could not titrate further. This finding must be interpreted carefully, however, as the subjects who were titrated to the BTDS 10 dose had less pain at screening. Taking completion status as a surrogate for analgesic effectiveness, we can state that the comparative completion rates show no evidence that the BTDS 10 is less effective when administered to the proper subject population. Results for the primary endpoint (see Results and Conclusions) also suggest that BTDS 10 is effective in the appropriate population.

Protocol Deviations

Of the 541 subjects in the full analysis set, 148 subjects (27%) had at least one protocol deviation, major or minor. The proportion of deviations was balanced across treatment groups. Further, 98 subjects (18%) had at least one major protocol deviation, as deemed by the Applicant. These were also balanced across treatment groups.

The subjects in the full analysis set had 167 deviations. Table 3 classifies the deviations according to category and whether the Applicant deemed them major. If a subject had multiple deviations of the same type (e.g., same deviation on multiple days), then they are counted as a single deviation. The concomitant medication incidents mostly involved rescue medication being used excessively or within 30 hours of a study visit; the Applicant addressed this issue using a sensitivity analysis. The 25 subjects with non-qualifying pain scores should not have been randomized; the Applicant excluded these subjects from their per protocol set.

Table 3: Protocol Deviations by Type and Severity (Source: Reviewer)

Type of Deviation	# Major	# Minor
Concomitant Medication	54	56
Non-Qualifying Pain Scores	25	0
Medical History	1	15
Run-In Qualification	9	0
Other	3	4

Statistical Methods

The primary efficacy analysis used a mixed effect generalized linear model with repeated measures. Subject was included in the model as a random effect, with treatment, screening pain, and pre-randomization pain as fixed effects. The treatment effect was allowed to vary over time, but the primary comparison was at Week 12. Missing pain scores due to early discontinuation were handled using a "hybrid" method: BOCF for subjects who discontinued study medication due to an adverse event, LOCF otherwise. The baseline pain in this study is the relatively high pain score recorded at screening, not the lower score reported prior to randomization. Patients who stopped taking study medication but stayed on study, whom the Applicant refers to as ***retained dropouts***, were treated as discontinuations for this analysis. There was no imputation for intermittent missing scores. The Applicant also used a number of alternative imputation methods which are described in the next section. Additional details of the primary efficacy analysis are given in Appendix A.

The analysis methods for the secondary efficacy variables were as follows. The number of analgesic rescue tablets was analyzed using an analysis of covariance (ANCOVA) model with terms for treatment, screening pain, and pre-randomization pain. Sleep disturbance was analyzed using a mixed effects model similar to that in the primary efficacy analysis. The estimate of the treatment effect was based on Weeks 4, 8, and 12, however. Since there were two secondary endpoints, the Applicant adjusted for multiplicity using Holm's method. The endpoint with the lower p-value was considered significant if $p < .025$. If that endpoint was significant, then the other endpoint was tested at the .05 level.

Results and Conclusions

Table 4 shows the Applicant's results for the primary endpoint. I was able to reproduce it from the submitted files. The primary analysis showed that BTDS was statistically superior to placebo, achieving an estimated effect of .58 points on an 11-point NRS. The confidence interval was fairly wide, however, showing that the effect could be between .14 and 1.02.

Table 4: Primary Endpoint, BOCF/LOCF Imputation (Source: CSR, Table 9)

Weeks/Visits	BTDS (n = 257)	Placebo (n = 284)
Screening[a] (Visit 2)		
n	257	284
Mean (SD)	7.24 (1.263)	7.17 (1.223)
Prerandomization[b] (Visit 3)		
n	257	284
Mean (SD)	2.57 (1.283)	2.56 (1.207)
Double-blind Week 12 (Visit 8)		
n	257	283[d]
Mean (SD)	3.83 (2.738)	4.38 (2.690)
Repeated Measures Analysis/Least Squares Means (SE) at Week 12		
LS mean (SE)	3.81 (0.166)	4.39 (0.152)
Treatment Comparison at Week 12		
Difference in LS means from placebo	−0.58 (0.225)	
P value vs placebo[c]	.0104	
95% CI for difference from placebo	(−1.02, −0.14)	

Cross-references: Table 14.2.1.1; Appendices 16.2.6.1, 16.2.6.2, 16.2.6.4, and 16.1.9.2.1.
LS = least squares, SE = standard error
[a] Screening mean pain was the mean of the diary "average pain over the last 24 hours" scores during the 2 consecutive days in the screening period that qualified the subject to enter the run-in period.
[b] Prerandomization mean pain was the mean of the diary "average pain over the last 24 hours" scores during the days of the run-in period that qualified the subject for randomization to double-blind treatment.
[c] Statistics are based on a mixed effect general linear model with subject as a random component, treatment and time as fixed components, and the screening mean pain and prerandomization mean pain as fixed covariates.
[d] Does not include 1 subject who completed the study but the site did not collect the pain score.
Note: Number of subjects with data varies between visits because missing pain scores were only imputed for subjects who discontinued double-blind treatment early. Intermittent missing pain scores were not imputed.
Note: Based on the selection algorithm in the SAP, an unstructured covariance matrix was used.
Note: Pain scale is 11 points (0 = no pain to 10 = pain as bad as you can imagine).

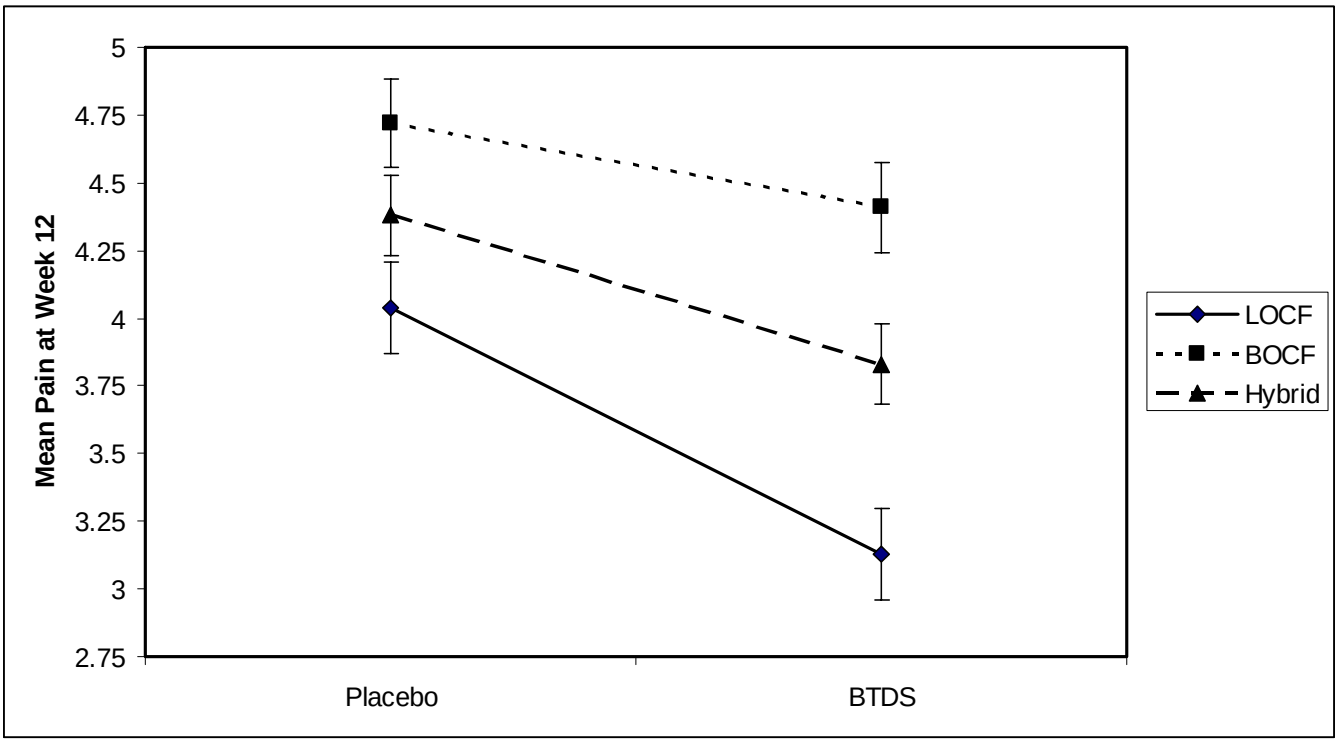
Table 5 shows the results of a number of sensitivity analysis conducted by the Applicant. I was able to reproduce the results from the submitted data. The first analysis in the table used the pain scores from Weeks 4, 8, and 12. The results represent an average treatment effect over time, rather then the Week 12 landmark effect estimated in the primary analysis. The Retained Dropout analysis included pain scores from subjects who discontinued study treatment but continued to provide data. There were 29 subjects who were in this category. This analysis used only observed pain scores. The Valid Pain Score Substitution analysis excluded the observed pain scores from subjects who took supplemental (rescue) or prohibited analgesics within a day of the study visit, using linear interpolation to impute a value. The Per-Protocol Analysis excluded subjects who were randomized without the required pain scores. The results from the various imputation methods were fairly consistent, with all but BOCF showing a significant treatment effect. The results from BOCF trended in the right direction.

Table 5: Applicant's Sensitivity Analyses (Source: CSR, Table 10)

Type of Analysis	BTDS vs Placebo		
	Difference from Placebo	*P* value	95% CI for Difference from Placebo
Hybrid Weeks 4, 8, 12	−0.62	.0016	−1.01, −0.24
BOCF Week 12	−0.34	.1502	−0.79, 0.12
LOCF Week 12	−0.93	< .0001	−1.33, −0.52
Retained Dropout ITT Week 12	−0.75	.0007	−1.18, −0.31
Valid Pain Score Substitution Week 12	−0.59	.0095	−1.03, −0.14
Per-Protocol Week 12 (hybrid imputation)	−0.52	.0260	−0.98, −0.06

From the treatment effects alone, one may underestimate the impact of the imputation method on the analysis. To remedy this, Figure 4 shows the mean pain in each treatment for three of the most common methods of imputation. We do not consider LOCF imputation appropriate for chronic pain trials, but I include it in the figure since it is related to the hybrid imputation method. It is clear from the figure that the change in estimated pain scores from using different imputation methods is substantially larger than the analgesic effect. The error bars show one standard error. As a caveat, one should resist the temptation to use the error bars to determine whether the LOCF and BOCF results are "significantly different". The two means are not based on independent samples.

Figure 4: Treatment Effect by Imputation (Source: Reviewer)



Note: Error bars show one standard error.

I noted earlier that subjects who began the double-blind period on the active BTDS 10 patch had a similar (but numerically higher) probability of completing the study to those who began that period on the higher dose. The two patch sizes also showed similar treatment effects, i.e., similar differences in least-squares means between the active patch of that size and the matching placebo. The estimated treatment effect was .50 (standard error = .30) for the BTDS 10 patch, compared to .69 (SE = .33) for BTDS 20. A review of the discussion around Table 2 may be helpful to interpret these results properly.

Responder Analysis

The Applicant performed two versions of a ***responder analysis***, also known as a cumulative responder analysis. In this analysis, one computes and plots the proportion of subjects who show at least *x%* improvement from baseline, where *x* is set to different cutpoints. (It is similar to a cumulative distribution function.) In the version of interest, a *0%* improvement is imputed to a subject who discontinues study medication. Table 6 and Figure 5 show the Applicant's results. My own calculations yielded quite similar results. The only differences are that I count all 284 subjects in the placebo arm in the denominators, and I also include an additional non-responder in this arm (for a total of 112).

It is common to put particular emphasis on the 30% cut point. Using this criterion, the proportions of responders in the two treatment arms are not significantly different at the conventional .05 level (chi-square test, p = .10).

Table 6: Applicant's Responder Analysis (Source: CSR, Table 14.2.3.1)

	BTDS (N=257)	Placebo (N=284)
Number of subjects	257	283
Percent reduction from baseline, n (%)		
0%	103 (40)	111 (39)
>0%	154 (60)	172 (61)
>=10%	153 (60)	165 (58)
>=20%	148 (58)	158 (56)
>=30%	137 (53)	131 (46)
>=40%	124 (48)	121 (43)
>=50%	113 (44)	101 (36)
>=60%	94 (37)	79 (28)
>=70%	75 (29)	62 (22)
>=80%	58 (23)	44 (16)
>=90%	32 (12)	22 (8)
=100%	32 (12)	22 (8)

Note: The "0%" row also includes subjects who discontinue study drug or show an increase in pain from baseline.

Figure 5: Applicant's Responder Analysis (Source: CSR, Figure 14.2.11.9)

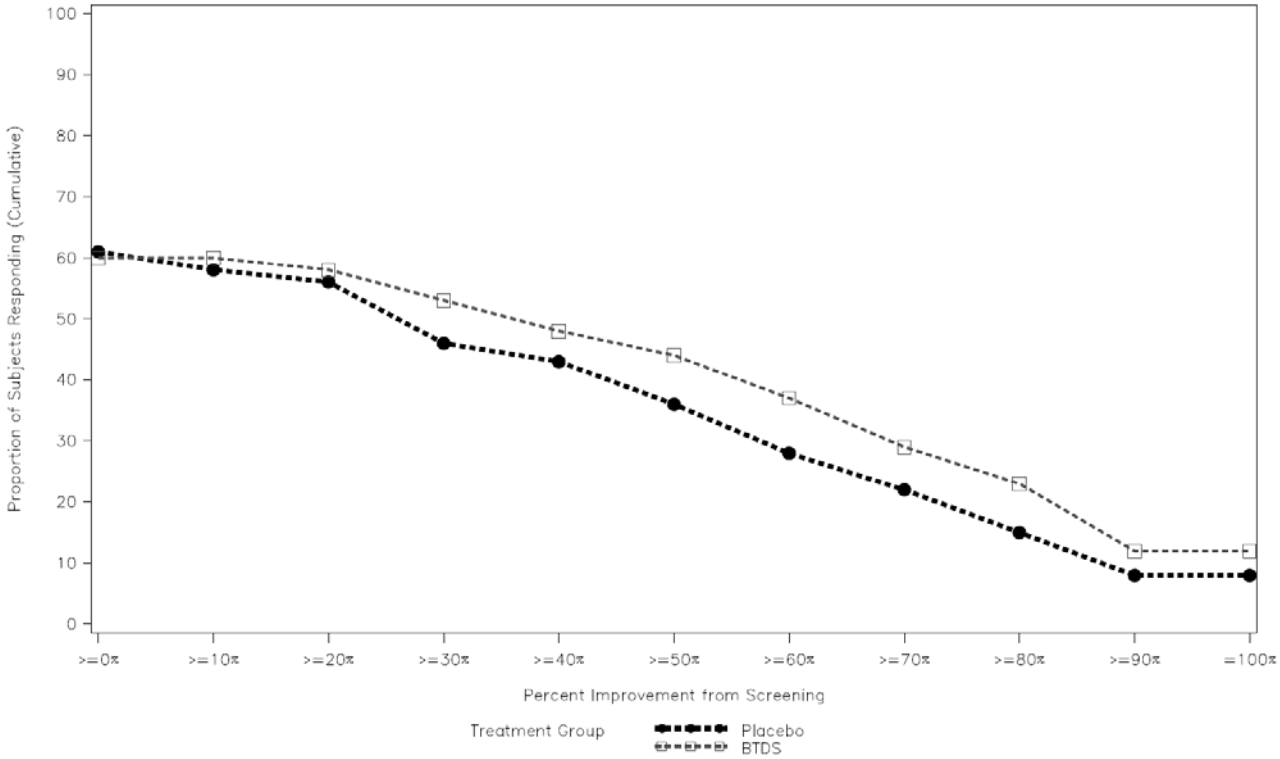


Table 7 shows the Applicant's results for the secondary endpoints, which I was able to reproduce. Although the SPA-approved analysis for sleep disturbance is based on Weeks 4, 8, and 12, it is more consistent with current practice to estimate the treatment effect at Week 12. Using this approach, the effect of BTDS is still significant (p = .035) with a similar point estimate (-3.78). The placebo group trended toward using more supplemental analgesia, but the difference was not statistically significant.

Table 7: Secondary Efficacy Endpoints (Source: CSR, Table 11)

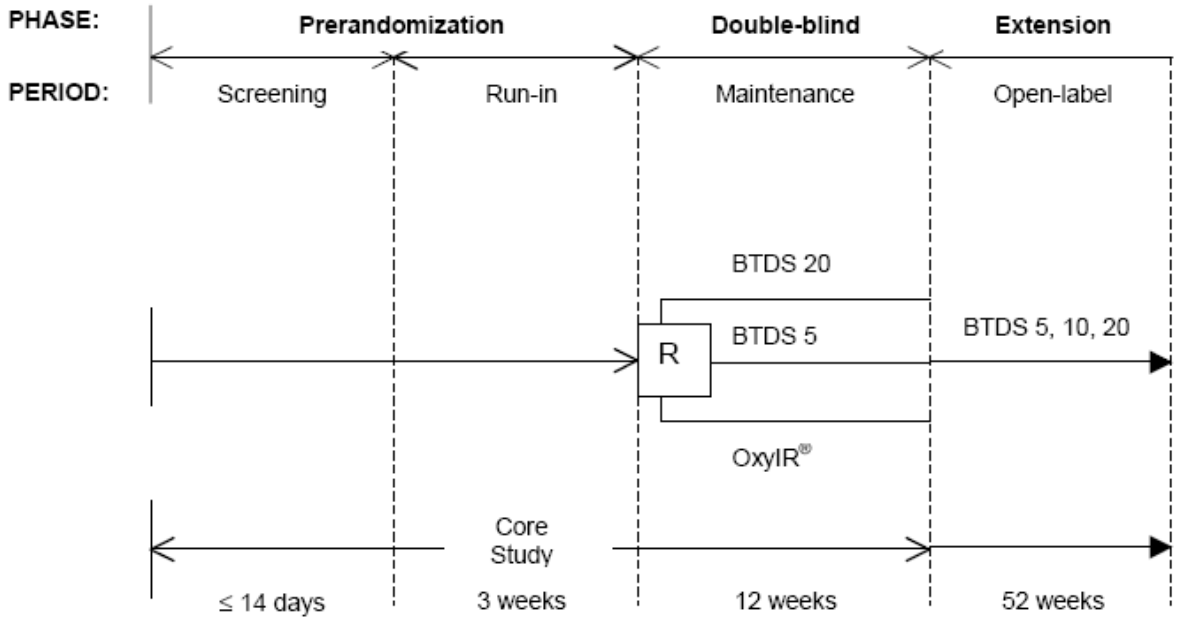
Treatments/Measurements	Treatment Difference (95% CI)	*P* Value	α-Level	Decision
Sleep disturbance subscale of the MOS-Sleep Scale at weeks 4, 8, and 12	−4.40 (−7.55, −1.25)	.0062	.025	Significant
Mean daily number of tablets of nonopioid supplemental analgesic during weeks 2 to 12 of double-blind treatment	−0.124 (−0.296, 0.048)	.1586	.05	NS

3.1.2 Study BUP3015

Study Design and Endpoints

BUP3015 was a randomized, double-blind active comparator study of Butrans in opioid-experienced subjects with moderate to severe low back pain. It was conducted in 75 sites in the United States. The study enrolled 1292 subjects, 662 of whom were randomized. The design of the study is shown in Figure 6.

Figure 6: Study Design for BUP3015 (Source: CSR, Figure 1)



During the Screening phase, subjects were assessed for compliance with study procedures and tapered from opioid medication. In order to continue to the Run-In phase, subjects had to have pain scores of at least 5 for two consecutive days with SOWS scores ≤ 23. In the open-label Run-In period, subjects began on BTDS 10 and were titrated to BTDS 20. Subjects qualified for the Double-Blind treatment phase when they tolerated BTDS 20 for 7 days and:

- had acceptable electrocardiogram (ECG) results
- had an "average pain over the last 24 hours" score of no more than 4 for at least 6 of the 7 days
- limited rescue medication to no more than 800 mg/day ibuprofen or 2000 mg/day APAP for at least 6 of 7 days
- used no non-study opioid medication

Qualifying subjects were randomized to BTDS 20, BTDS 5, or oxycodone 40 mg per day. A double-dummy design was used to maintain the blind during this phase. Subjects visited the clinic at Weeks 1, 2, 4, 8, 12, and reported the average and worst pain scores over the previous 24 hours. They were told to abstain from supplemental analgesia for 48 hours before the visit,

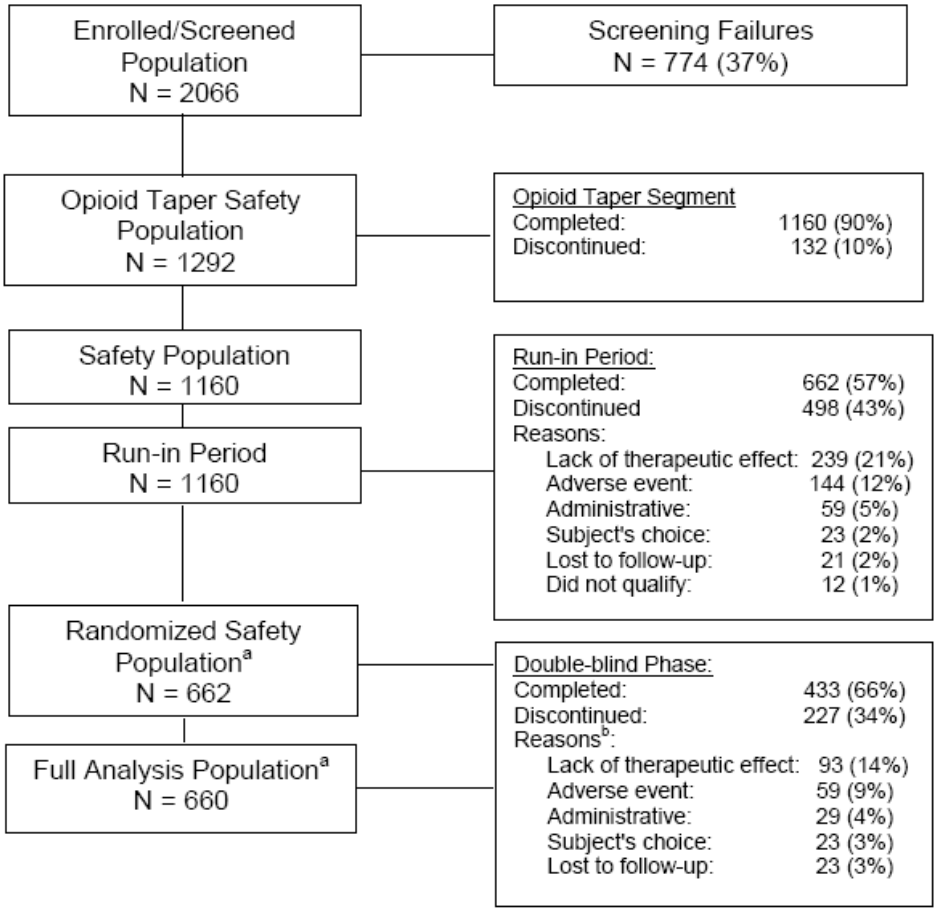
and the visit was rescheduled if they did not comply. At other times, subjects were allowed to take up to 3200 mg/day of ibuprofen, or 4000 mg/day APAP if they could not tolerate ibuprofen.

The Applicant's primary efficacy variable was "average pain over the last 24 hours" at Weeks 4, 8, and 12, with the primary comparison between BTDS 5 and BTDS 20. Following DAARP's current thinking on trials for chronic pain drugs, and consistent with the primary analysis for BUP3024, I used the pain at the Week 12 visit as the primary endpoint. The secondary efficacy variables were the number of tablets of rescue analgesia used in the double-blind phase, the Oswestry Disability Index, and the Sleep Disturbance Subscale of the MOS Sleep Scale. The full analysis population consisted of randomized subjects who received at least one dose of study medication.

Patient Disposition, Demographic and Baseline Characteristics

Figure 7 and Table 8 show the disposition of subjects and their reasons for discontinuation. They were provided by the Applicant, but I verified the contents.

Figure 7: Subject Disposition (Source: CSR, Figure 2)



Cross-references: Table 14.1.1.1 and 14.1.2.

[a] 662 subjects were randomized but 2 did not receive study drug.

[b]There were 3 subjects with a data discrepancy related to reason for discontinuation. For subject 28006, the reason for discontinuation was due to adverse event in the double-blind phase rather than loss of therapeutic effect (see narrative in section 14.3.3). For subject 47057, the reason for discontinuation was due to adverse event in the double-blind phase rather than administrative (see narrative in section 14.3.3). The reason for discontinuation was not clarified until after database lock and therefore the database, Table 14.1.1.1, and Appendix 16.2.1.3 were not updated in the CSR. For subject 71006, the study drug discontinuation CRF was received but not entered in the database and therefore this subject is not listed as a study drug discontinuation in Table 14.1.1.1 and Appendix 16.2.1.3.

Table 8: Reasons for Discontinuation (Source: CSR, Table 5)

Category	Prerandomization Phase Run-in Period[a] BTDS 10/20 (N = 1160)	Double-blind Phase Maintenance Period BTDS 5 (N = 221)	BTDS 20 (N = 219)	OxylR® (N = 220)
Completed Period/Phase on Study Drug, n (%)	662 (57)	128 (58)	146 (67)	159 (72)
Discontinued Study Drug - All Cases[b, c], n (%)	498 (43)	93 (42)	73 (33)	61 (28)
Adverse event	144 (12)	14 (6)	29 (13)	16 (7)
Lost to follow-up	21 (2)	7 (3)	6 (3)	10 (5)
Subject's choice	23 (2)	11 (5)	7 (3)	5 (2)
Administrative	59 (5)	9 (4)	6 (3)	14 (6)
Lack of therapeutic effect	239 (21)	52 (24)	25 (11)	16 (7)
Did not qualify	12 (1)	0	0	0
Discontinued Study Drug And Completed Double-blind Phase, n (%)[b]		20 (9)	9 (4)	10 (5)
Adverse event		0	3 (1)	1 (< 1)
Lost to follow-up		0	0	0
Subject's choice		1 (< 1)	0	1 (<1)
Administrative		0	0	0
Lack of therapeutic effect		19 (9)	6 (3)	8 (4)
Discontinued Study Drug And Did Not Complete Double-blind Phase, n (%)[b]		73 (33)	64 (29)	51 (23)
Adverse event		14 (6)	26 (12)	15 (7)
Lost to follow-up		7 (3)	6 (3)	10 (5)
Subject's choice		10 (5)	7 (3)	4 (2)
Administrative		9 (4)	6 (3)	14 (6)
Lack of therapeutic effect		33 (15)	19 (9)	8 (4)

Cross-references: Table 14.1.1.1, Appendices 16.2.1.2, 16.2.1.3.
[a] Opioid taper failures who did not qualify for the run-in period are not included.
[b] Subjects randomized to the double-blind phase who discontinued study drug were eligible to complete the double-blind phase off study drug. For the double-blind phase, "all cases" includes both subjects who discontinued study drug and either completed or did not complete the phase.
[c] There were 3 subjects with a data discrepancy related to reason for discontinuation. For subject 28006, the reason for discontinuation was due to adverse event in the double-blind phase rather than loss of therapeutic effect (see narrative in section 14.3.3). For subject 47057, the reason for discontinuation was due to adverse event in the double-blind phase rather than administrative (see narrative in section 14.3.3). For subject 71006, the study drug discontinuation CRF was received but not entered in the database and therefore this subject is not listed as a study drug discontinuation in Table 14.1.1.1 and Appendix 16.2.1.3.

Protocol Deviations

The Applicant reported 44 protocol deviations in the randomized patient population, involving 43 patients. The deviations were evenly balanced across the three treatment groups. Of the 44 deviations, 27 were screening deviations that preceded randomization. The deviations that

occurred after randomization were as follows: 13 subjects took prohibited analgesics less than 48 hours before a study visit, and 4 subjects in the oxycodone arm took less than 75% of the expected dose. The Applicant conducted a sensitivity analysis which excluded the subjects who took opioid analgesia before a study visit, and the results for the primary comparison were virtually unchanged.

In addition to these protocol deviations reported by the Applicant, there were 83 subjects who took non-study opioid medication during the double-blind period. The breakdown by treatment group was: 39 subjects (18%) in BTDS 5, 25 (11%) in BTDS 20, 19 (9%) in OxyIR.

The Applicant originally planned to have 891 randomized subjects, but terminated enrollment after 662 subjects were randomized. No statistical adjustment was made for early stopping. According to the Clinical Study Report, the early termination was due to "administrative reasons unrelated to safety and efficacy." In response to a request for clarification, the Applicant stated the following:

> In mid-2005, changing business conditions were the basis for discontinuation of further screening of additional subjects in Clinical Study BUP3015. Subjects active in the open-label run-in period of the study, who were not randomized, were discontinued. While no further screening or randomization was undertaken, all randomized subjects active within the clinical study followed all protocol-specified procedures until completion of the double-blind treatment phase. At the time of study completion, 662 subjects, representing roughly 74% of the planned sample size, had been randomized into the double-blind period. All study procedures, including data collection, site monitoring, data clarification, query resolution and database lock were performed according to standard operating procedures, and under blinded conditions. The study remained blinded until after the data base was locked and the statistical analysis plan was finalized.

Statistical Methodologies

The Applicant's primary efficacy analysis was a mixed effect linear model for the 'average pain over the last 24 hours' score at Weeks 1, 2, 4, 8, and 12. Fixed effects in the model included treatment, time (categorical effect), screening mean pain, and pre-randomization mean pain; subject was a random effect. The primary comparison was between BTDS 20 and BTDS 5. No missing data were imputed in the primary analysis. This analysis is unacceptable for two related reasons. First of all, it is DAARP's current thinking that chronic pain medications should be effective at Week 12 of the study. The rationale for this requirement is that a medication that became ineffective after only three months would clearly not be suitable for a chronic condition. The other problem is that the Applicant is only using the available data. They provided an extensive rationale for this decision which I address in Appendix A.

For my efficacy analysis, I took the primary endpoint to be the average pain at Week 12 and used an ANCOVA method. The terms in the model included treatment, screening pain, and randomization pain. I used BOCF imputation, treating retained dropouts as failures. See Appendix A for further discussion of these methods.

Results and Conclusions

Table 9 shows the Applicant's descriptive statistics and analysis for their primary endpoint, using only available data. In other words, there was no imputation for missing data, including intermittent missing data. I was able to verify their results.

Table 9: Applicant's Primary Endpoint – Average Pain at Weeks 4, 8, 12, No Imputation (Source: CSR, Table 9)

Visits/Weeks	BTDS 5 (N = 221)	BTDS 20 (N = 219)	OxyIR® (N = 220)
Screening[a]			
n	221	219	220
Mean (SE)	6.36 (0.075)	6.46 (0.084)	6.46 (0.079)
Median	6.0	6.5	6.0
Min, Max	4, 10	1, 10	3, 10
Prerandomization[b]			
n	221	219	219
Mean (SE)	2.84 (0.075)	2.91 (0.075)	2.74 (0.074)
Median	3.0	3.1	2.9
Min, Max	0, 7	0, 6	0, 5
Week 4			
n	154	176	184
Mean (SE)	3.79 (0.149)	3.40 (0.128)	3.14 (0.125)
Median	4.0	3.0	3.0
Min, Max	0, 8	0, 8	0, 7
Week 8			
n	138	164	173
Mean (SE)	3.83 (0.162)	3.35 (0.140)	3.24 (0.145)
Median	4.0	3.0	3.0
Min, Max	0, 9	0, 8	0, 10
Week 12			
n	127	142	154
Mean (SE)	4.02 (0.179)	3.35 (0.139)	3.26 (0.152)
Median	4.0	3.0	3.0
Min, Max	0, 9	0, 9	0, 8
Overall Statistics[c]			
Difference from BTDS 5 over weeks 4, 8, and 12		-0.67 (0.163)	-0.75 (0.161)
P value vs. BTDS 5		< .001	< .001
95% CI for difference from BTDS 5		(-0.99, -0.35)	(-1.07, -0.44)

Cross-references: Table 14.2.1, Appendices 16.2.6.1, 16.2.6.2.

[a]Screening mean pain is defined as the mean of diary "average pain over the last 24 hours" scores collected during the two consecutive days of the opioid taper segment that qualified the subject to enter into the run- in period.

[b] Prerandomization mean pain is defined as the mean of the diary "average pain over the last 24 hours" scores during the last 7 days of the run- in period.

[c]Statistics are based on a repeated measures, mixed effects linear model with fixed effects for treatment, time, screening mean pain, and prerandomization mean pain, and a random effect for subject.

Note: Number of subjects with data varies between visits because missing pain scores are only imputed for subjects who discontinue double-blind treatment early. Intermittent missing pain scores are not imputed.

Note: Based on the selection algorithm in the SAP, an unstructured covariance matrix was used.

Note: Pain scale is 11 points (0 = no pain to 10 = pain as bad as you can imagine).

As I noted in the previous section, we do not consider the Applicant's analysis of the available data from Weeks 4, 8, and 12 to be an appropriate primary analysis. For my own analysis I fit an ANCOVA model for pain at Week 12, using BOCF imputation. (The Applicant also reported results using a version of BOCF, but they did not impute failure for retained dropouts as I did.) This analysis also showed BTDS 20 to be statistically superior to BTDS 5 ($p < .01$). The estimated treatment effect (difference in least-squares means) was .62, with a standard error of .19. This is similar to the treatment effect of .67 found using the Applicant's analysis. The BTDS 5 treatment arm showed a (least-squares) mean pain at Week 12 of 4.96, and the mean pain for the BTDS 20 arm was 4.33.

As with Study BUP3024, the Applicant conducted two different versions of a cumulative responder analysis. Table 10 shows their results for the more "conservative" version which imputes zero improvement to all subjects who discontinue study medication. My own analysis yielded slightly different results; however, the percent responding did not differ by more than 1% in any cell. The Applicant also submitted Figure 8, which is cross-referenced to Table 10. They do not appear to quite match, however. Like the Applicant, I found that the BTDS 20 arm was significantly more likely to show 30% improvement than the BTDS 5 arm.

Table 10: Applicant's Responder Analysis (Source: Integrated Summary of Efficacy, Table 9.1)

	OxyIR (N = 220)	BTDS 5 (N = 221)	BTDS 20 (N = 219)
Subjects With Percent Improvement From Baseline [n (%)]			
0%	72 (32.7)	119 (53.8)	84 (38.4)
>0%	148 (67.3)	102 (46.2)	135 (61.6)
≥10%	139 (63.2)	98 (44.3)	129 (58.9)
≥20%	132 (60.0)	87 (39.4)	122 (55.7)
≥30%	112 (50.9)	73 (33.0)	107 (48.9)
≥40%	100 (45.5)	60 (27.1)	96 (43.8)
≥50%	78 (35.5)	46 (20.8)	77 (35.2)
≥60%	61 (27.7)	33 (14.9)	57 (26.0)
≥70%	39 (17.7)	14 (6.3)	28 (12.8)
≥80%	28 (12.7)	13 (5.9)	16 (7.3)
≥90%	12 (5.5)	6 (2.7)	3 (1.4)
≥100%	12 (5.5)	6 (2.7)	3 (1.4)

Note: The "0%" row also includes subjects who discontinue study medication or show an increase in pain from baseline.

Figure 8: Applicant's Responder Analysis (Source: ISE, Figure 5)

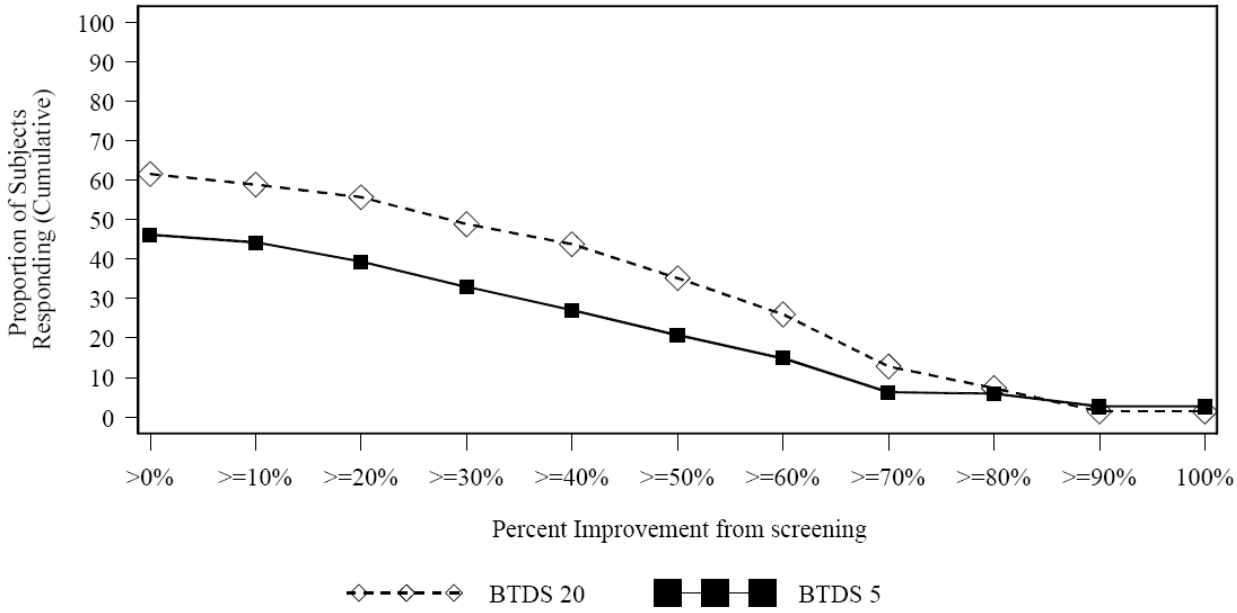


Table 11 shows the Applicant's results for secondary endpoints. I confirmed their significant findings for MOS Sleep Scale and supplemental analgesia, but did not look at the ODI outcome. As would be expected, the lower dose group used more supplemental analgesia.

Table 11: Secondary Efficacy Endpoints (Source: CSR, Table 12)

Treatments/Measurements	Treatment Difference (95% CI)	*P* Value	α-Level	Decision
BTDS 20 vs BTDS 5				
MOS Sleep Scale - Sleep Disturbance Subscale	−6.23 (−9.64, −2.82)	< .001	.0167	Significant
Supplemental analgesic medication	−0.5 (−0.90, −0.13)	.006	.025	Significant
ODI Score	−1.72 (−3.55, 0.11)	.065	.05	NS

The Applicant's decision to halt the trial after 74% enrollment is somewhat troubling in light of the fact that it was entirely volitional; there was no external agent who forced the Applicant to halt the trial. However, the Applicant states that the data were not unblinded before the trial was halted. Assuming that this is true, the resulting p-value should be valid.

3.2 Evaluation of Safety

The safety profile of Butrans was reviewed by Robert Levin, M.D.

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

4.1 Gender, Race and Age

4.1.1 Study BUP3024

Table 12 shows the descriptive statistics for the primary endpoint by sex. Hybrid imputation was used, following the SPA-approved primary analysis. There is clearly no evidence of an overall sex difference or treatment-by-sex interaction. I created all of the tables in this section.

Table 12: BUP3024, Pain at Week 12 by Gender, Hybrid Imputation

Sex	Treatment	N Obs	Mean	Std Dev
F	BTDS	134	3.8	2.8
	Placebo	164	4.4	2.8
M	BTDS	123	3.8	2.7
	Placebo	120	4.3	2.5

Table 13 shows the primary endpoint by race. I tested for an interaction between race and treatment, focusing on black and white subjects since the small numbers in other categories precluded meaningful inference. The interaction was not significant, but black subjects did report less pain overall ($p < .001$).

Table 13: BUP3024, Pain at Week 12 by Race, Hybrid Imputation

Race	Treatment	N Obs	Mean	Std Dev
Asian	BTDS	12	2.2	1.5
	Placebo	20	2.5	1.2
Black or African American	BTDS	58	3.5	3.0
	Placebo	62	3.8	2.8
Other	BTDS	3	4.0	2.6
	Placebo	5	4.8	2.8
White	BTDS	184	4.0	2.7
	Placebo	197	4.7	2.7

Table 14 shows the primary endpoint by age group. Using all three categories, there was neither an overall effect of age nor an interaction.

Table 14: BUP3024, Pain at Week 12 by Age, Hybrid Imputation

Age Group	Treatment	N Obs	Mean	Std Dev
18 - 39 years	BTDS	58	4.0	2.8
	Placebo	60	4.4	2.7
40 - 64 years	BTDS	170	3.8	2.8
	Placebo	186	4.5	2.7
65+ years	BTDS	29	3.8	2.3
	Placebo	38	3.8	2.5

4.1.2 Study BUP3015

Table 15 shows the results for the primary endpoint by sex. I used BOCF imputation, attributing zero improvement to all subjects who discontinued study medication prematurely. Focusing on the BTDS arms, there was not a significant effect of sex or nor an interaction with treatment.

Table 15: BUP3015, Pain at Week 12 by Sex, BOCF

Sex	Treatment	N Obs	Mean	Std Dev
Female	BTDS 5	101	4.8	2.2
	BTDS 20	113	4.6	2.2
	OxyIR	100	4.2	2.4
Male	BTDS 5	120	5.0	2.0
	BTDS 20	106	4.2	2.2
	OxyIR	120	4.1	2.1

Table 16 shows the primary endpoint by race. Since there were relatively few non-white subjects I did not conduct any statistical tests.

Table 16: BUP3015, Pain at Week 12 by Race, BOCF

Race	Treatment	N Obs	Mean	Std Dev
Black	BTDS 5	13	4.8	3.0
	BTDS 20	21	4.0	1.8
	OxyIR	14	3.7	2.7
White	BTDS 5	206	5.0	2.0
	BTDS 20	193	4.4	2.2
	OxyIR	201	4.2	2.2
Other	BTDS 5	2	3.0	2.8
	BTDS 20	5	4.6	2.6
	OxyIR	5	3.5	2.8

Table 17 shows the results by age group. Focusing on the BTDS arms, I found a significant overall effect of age ($p = .02$), but no interaction between age and treatment. Older subjects reported less pain.

Table 17: BUP3015, Pain at Week 12 by Age, BOCF

Age Group	Treatment	N Obs	Mean	Std Dev
21-39 years	BTDS 5	44	5.6	1.7
	BTDS 20	44	4.8	2.2
	OxyIR	46	4.9	2.3
40-64 years	BTDS 5	143	4.9	2.1
	BTDS 20	148	4.3	2.2
	OxyIR	151	4.0	2.1
65+ years	BTDS 5	34	4.3	2.2
	BTDS 20	27	4.4	2.1
	OxyIR	23	3.8	2.8

4.2 Other Special/Subgroup Populations

I did not investigate any other subgroups. All clinical sites in the studies of primary interest were in the United States.

5 SUMMARY AND CONCLUSIONS

5.1 Statistical Issues and Collective Evidence

BUP3024 was conducted under a Special Protocol Agreement, and was successful under the planned primary efficacy analysis. The estimated treatment effect of .58 (95% C.I. = [.14, 1.02]) is modest, but similar to what is found in other chronic pain trials for approved drugs.

As is common in chronic pain trials, a substantial number of subjects either discontinued study medication or discontinued the study altogether. Under the Special Protocol Agreement for BUP3024, pain scores were primarily imputed using a "hybrid" BOCF/LOCF method. The Applicant conducted a number of other imputations as planned sensitivity analyses, and all but BOCF showed a significant result. DAARP agreed with the Applicant that "we would look for a consistency of treatment effect (similar estimate of BTDS mean) across the primary and all sensitivity analyses". BOCF imputation showed a treatment effect of .34, compared to .58 using the primary imputation. It is not clear what constitutes "similar" in this context. Moreover, the BOCF and hybrid estimates are largely computed from the same data, so it does not make sense to compare them statistically. However, my subjective assessment is that the two estimates are "similar" enough that the study succeeded under the Agreement.

The Applicant's primary statistical analysis for BUP3015 was not appropriate according to the standards established by DAARP for chronic pain trials. However, my preferred analysis, a Week 12 landmark analysis with BOCF imputation, also supported a finding of efficacy. These two analyses produced comparable estimates of the treatment effect.

The Applicant complicated interpretation of BUP3015 by halting the study after 74% of the planned sample size was enrolled. The stated rationale was "changing business conditions". This decision is troubling, as it raises the theoretical possibility of multiple looks at the data, which would invalidate the reported p-value. However, the Applicant states the trial was still unblinded when the decision was made to halt. Assuming that this was the case, the trial is acceptable to provide evidence of efficacy.

5.2 Conclusions and Recommendations

The Applicant has submitted two adequate and well-controlled studies which support a finding of efficacy for Butrans.

5.3 Review of the Proposed Label

Selections from the Applicant's proposed label language are shown in *italics*, and my comments are shown in regular type. Note that any references to figures use different numbering than the rest of the report.

(b) (4)

(b) (4)

I did not find any errors in the previous paragraph.

(b) (4)

I confirmed this finding.

(b) (4)

Using the preferred method which imputes no improvement to subjects who drop out, 53% of BTDS subjects and 46% of placebo subjects showed 30% improvement. This difference is not statistically significant. Moreover, recent practice has been to show the entire responder graph and not focus on particular cutpoints.

(b) (4)

(b) (4)

Figure 3 should be replaced with one that which imputes zero improvement to all subjects who discontinue study medication.

(b) (4)

This statement is correct, but it is based on an average over Weeks 4, 8, and 12.

(b) (4)

I did not find any errors in the previous paragraph.

(b) (4)

The statement about oxycodone should be omitted because it is not relevant for showing the effectiveness of Butrans. I agree with the statements about Butrans.

(b) (4)

I substituted the percentages from a more "conservative" analysis that imputes failure to all withdrawals. The treatment effect remains significant (and is in fact larger).

(b) (4)

This statement is not relevant in the Butrans label.

The proportion of patients with various degrees of improvement from screening to study endpoint is shown in Figure 4 below.

(b) (4)

This figure should be replaced by one in which zero improvement is imputed for all subjects who discontinue study medication early.

(b) (4)

It is correct that BTDS 20 was superior to BTDS 5, averaging over Weeks 4, 8, and 12. According to the Applicant's own results, (b) (4) (See Table 12 in clinical study report.)

6 APPENDIX A: Additional Statistical Details and Discussion

BUP3024

The primary efficacy analysis was a mixed effect generalized linear with repeated measures. Subject was included in the model as a random effect, with treatment, screening pain, and pre-randomization pain as fixed effects. The following covariance structures were considered for the repeated measures analysis: unstructured, compound symmetry, compound symmetry with heterogeneous variances, autoregressive order 1, and autoregressive order 1 with heterogeneous variances. The unstructured covariance matrix was chosen because it yielded the lowest (best) Akaike Information Criterion among the models that converged.

A technical point for both studies: Sometimes SAS reported that a covariance structure converged but the Hessian matrix was not positive definite. In this case, it is difficult to tell whether the best covariance structure has been selected from among the candidates. However, I made sure that my major conclusions did not depend on the choice of covariance structure.

BUP3015

The statistical analysis plan for this study is accompanied by an extensive report on simulations intended to estimate the bias induced by different methods of handling missing data. The simulations cover a number of different missing data patterns, including missing-completely-at-random (MCAR) and various not-missing-at-random scenarios. They conclude from these simulations that using the available data results in less bias than LOCF or BOCF methods.

Unfortunately, the simulations are of limited use from a regulatory standpoint. The bias parameter in the simulations is based on what might be called the ***counterfactual analgesic effect***, i.e., what the effect of the analgesic ***would have been*** if all of the subjects completed the study. While the counterfactual analgesic effect may be of theoretical interest, DAARP has taken the position that completion status itself is one of the key outcomes in an analgesic trial. The argument for this position is that the entire purpose of an analgesic is to make the patient feel better, and that the subject who drops out of a trial has expressed a preference for not taking it (barring the possibility that the subject withdrew for a non-medical reason). DAARP has therefore argued that an imputation should be "conservative" in the specific sense that subjects who drop out are assigned poor pain scores, fully acknowledging that this imputed score is not meant to estimate the counterfactual pain score. BOCF is one way to implement this "conservative" strategy. While BOCF imputation is subject to various criticisms, I do not see any alternative which is clearly superior for analgesic trials. The Agency has let a contract with the National Academy of Sciences to study the issue of missing data.

Application Type/Number	Submission Type/Number	Submitter Name	Product Name
NDA-21306	ORIG-1	PURDUE PHARMA LP	BuTrans (buprenorphine transdermal system)

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

/s/

JONATHAN D NORTON
03/03/2010

DIONNE L PRICE
03/04/2010
Concur.

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

STATISTICAL REVIEW AND EVALUATION

CARCINOGENICITY STUDY

NDA Number:	21,306 / Serial 0000
Drug Name:	BuTrans™ (Buprenorphine Transdermal System)
Indication:	Relief of moderate to severe pain in patients requiring around-the-clock treatment for an extended period of time
Applicant:	Purdue Pharma L. P. Purdue Research Center 444 Saw Mill River Road Ardsley, New York 10502
Testing Facility:	(b) (4)
Date:	Submitted 25 September 2009 To reviewer 6 October 2009
Review Priority:	Standard
Biometrics Division:	Division 6
Statistical Reviewer:	Steve Thomson
Concurring Reviewer:	Team Leader: Karl Lin, Ph. D.
Medical Division:	Anesthesia and Analgeisia Products
Toxicologist:	Gary Bond, Ph.D.
Project Manager:	Matthew Sullivan
Keywords:	Carcinogenicity, Cox regresson, Kaplan-Meier product limit, Survival analysis, Trend test

Table of Contents

1. EXECUTIVE SUMMARY 3

1.1. CONCLUSIONS AND RECOMMENDATIONS 3
1.2. BRIEF OVERVIEW OF THE STUDIES 7
1.3. STATISTICAL ISSUES AND FINDINGS 7
1.3.1. Statistical Issues 7
1.3.2. Statistical Findings 11

2. INTRODUCTION 11

2.1. OVERVIEW 11
2.2. DATA SOURCES 12

3. STATISTICAL EVALUATION 12

3.1. EVALUATION OF EFFICACY 12
3.2. EVALUATION OF SAFETY 12
3.2.1. PROTOCOL BUP-N-004: BUPRENORPHINE BASE: A TWO YEAR DERMAL (SKIN PAINTING) CARCINOGENICITY STUDY IN RATS, ... 12
3.2.2. PROTOCOL BUP-P-012: BUPRENORPHINE: 26-WEEK DERMAL CARCINOGENICITY STUDY IN HEMIZYGOUS TG.AC MICE 18

4. FINDINGS IN SPECIAL/SUBGROUP POPULATIONS 23

5. SUMMARY AND CONCLUSIONS 23

5.1. STATISTICAL ISSUES AND COLLECTIVE EVIDENCE 23
5.2. CONCLUSIONS AND RECOMMENDATIONS 23

APPENDICES: 24

APPENDIX 1. SURVIVAL ANALYSIS 24
APPENDIX 2. FDA POLY-K TUMORIGENICITY ANALYSIS 28
APPENDIX 3. ANALYSIS OF PAPILLOMAS IN TG.AC MICE 36
APPENDIX 4. REFERENCES 48

1. EXECUTIVE SUMMARY

The Sponsor's reports indicate that the original objective of the rat study was to "evaluate the potential oncogenicity of the test article, buprenorphine base, when administered once daily to rats by the dermal (skin-painting) route for at least 104 weeks. However, the terminal necropsy in this study occurred during Week 100 after at least 99 weeks of daily treatment in order to assure adequate animals numbers for evaluation." (page 16 of report) The reported purpose of the mouse study was to assess whether or not daily dermal administration of the buprenorphine base for 26-weeks increased the incidence of skin tumors in hemizygous Tg.AC transgenic mice.

1.1. Conclusions and Recommendations

This submission summarizes the results of a two year rat study and a 26 week Tg.AC mouse study to assess the carcinogenic potential of the dermally applied Buprenorphine Transdermal System. The reported design of the rat study is summarized in the following table, for each gender:

Table 1. Design of Rat Study

Group	Interval	# animals	BuTrans Dosage (mg/kg/day)	Concentration (mg/mL) [1]	Volume (mL/kg/day)
1. Untreated	NA	65	0	0	8
2. Vehicle	Daily	65	0	0	8
3. Low	Daily	65	20	10	8
4. Medium	Daily	65	60	30	8
5. High	Daily	65	200	100/90[1]	2/2.22

[1]From days 1 through 177, the dosage concentration was 100 mg/mL and the dosage volume 2 mL/kg for Group 5. From day 178 through the end of study, the dosage concentration was 90 mg/mL with dosage volume 2.22.

The design of the Tg.AC mouse study can be summarized as follows:

Table 2. Design of Tg.AC Mouse Study

Group	# animals		Nominal Dosage (mg/kg/day)
	Males	Females	
1. Vehicle Control	25	25	Acetone vehicle
2. Positive Control	20	20	1.25µ TPA x3/week
3. Low	30	25	18.75 BuTrans mg/kg/day
4. Medium	30	25	37.5 BuTrans mg/kg/day
5. Medium-High	30	25	150 BuTrans mg/kg/day
6. High	30	25	600 BuTrans mg/kg/day

In addition both studies included toxicokinetic animals in the vehicle and BuTrans groups.

Table 3. Statistical Significances of Tests of Homogeneity and Trend in Survival

		Males		Females	
		Log rank	Wilcoxon	Log rank	Wilcoxon
Rats	Homogeneity over Groups 1 & 2	0.0928	0.1245	0.7659	0.9772
	Homogeneity over Groups 1-5	0.0110	0.0030	0.3065	0.4044
	Homogeneity over Groups 2-5	0.0495	0.0156	0.2861	0.4923
	Trend over Groups 2-5	0.5074	0.4121	0.6537	0.5401
Mice	Homogeneity over Groups 1- 6	0.3031	0.2688	0.5976	0.5597
	Homogeneity over Groups 1, 3-6	0.1956	0.1791	0.5179	0.4816
	Trend over Groups 1, 3- 6	0.0185	0.0179	0.3234	0.3076

From figure A.1.1 in Appendix 1, in male rats it seems that the low dose group has the highest mortality, with the vehicle, medium, and high dose groups generally intertwined with the next highest mortality, while the untreated group has the lowest mortality. However, the difference in survival between the vehicle and untreated dose groups was not quite statistically significant at the usual significance levels (Log rank $p = 0.0928$, Wilcoxon $p=0.1245$). On the other hand, at the usual 0.05 significance level, the hypothesis of homogeneity over all five dose groups would be rejected (Log rank $p = 0.0110$, Wilcoxon $p=0.0030$), as would be the more restricted hypothesis of homogeneity over the vehicle and the three actual dose groups (Log rank $p = 0.0495$, Wilcoxon $p=0.0156$). However, there is no evidence for a dose related trend (both $p \geq 0.4121$), consistent with the general observation that the low dose group had the highest mortality, while the vehicle, medium, and high dose groups were generally intertwined. In female rats, there was no strong evidence of heterogeneity or trend among any of the treatment groups (all $p \geq 0.2861$). This seems to be consistent with the observation, from figure A.1.2, that the dose group survival curves in female rats are generally crossing and somewhat intertwined.

From figure A.1.3 in the appendix, it is apparent that in male mice there is no strong evidence of heterogeneity in survival over the six treatment groups, including the TPA active control group (12-O- tetradecanoylphorbol 13 acetate in 0.1 mL acetone dosed three times weekly). That is, there is no strong evidence against homogeneity in survival (Log rank $p = 0.3031$, Wilcoxon $p = 0.2688$), nor among the five treatment groups deleting the TPA group (Log rank $p = 0.1956$, Wilcoxon $p = 0.1791$). Although mortality was low in this 27 week study, in male mice the high dose group had the highest mortality while the vehicle group had the lowest mortality, with the other groups rather intertwined. This is consistent with the statistically significant rejecting the hypothesis of test of no trend over the groups with vehicle (Log rank $p = 0.0185$, Wilcoxon $p = 0.0179$). In female mice there was no strong evidence of heterogeneity or trend among the treatment groups (all $p \geq 0.3076$). This also seems to be consistent with the observation, from figure A.1.4 below, that the survival curves of the different dose groups are generally intertwined.

In rats, the significance levels of the tests of tumorigenicity in the FDA analysis are based on poly-k tests applied to the data sets provided by the Sponsor. The poly-k test modifies the original Cochran-Armitage test of dose related trend in an event to adjust for differences in mortality (please see Bailer & Portier, 1988, Bieler & Williams, 1993). One problem with typical frequentist tumorgenicity analyses is that for each tumor-organ-gender-study combination there is one test of significance for each comparison of an actual treatment group to controls plus a test of overall trend. This implies a large number of tests, necessitating a multiplicity adjustment. For a single species, two gender study the usual so-called Haseman-Lin-Rahman rules adjust for the multiplicity of tests of tumorigenicity are that for a roughly 0.10 (10%) overall false positive error rate, one should claim statistical significance for tests of trend and the comparisons of the high dose group to controls if the nominal statistical significance is 0.05 (5%) level for rare tumors (with a historical control incidence less than 1%) and 0.01 (1%) for common tumors. These are discussed in Section 1.3.1.4 below. Since there are usually relatively few different types of tumors that develop in the short study period in genetically modified rodents, typically no adjustment for the multiplicity of tests is applied. Table 4 below, displays the tumor incidence in rats and the results of tests of no differences between treatments that had at least one test that achieved a nominal 0.05 level of significance.

Table 4. Potentially Statistically Significant Neoplasms in Rats

Organ / Tumor	Incidence Un-trt	Veh	Low	Med	Hi	Results of tests Untrt/ Veh	trend	Low/ Veh	Med/ Veh	High/ Veh
Male Rats										
PANCREAS										
B-ADENOMA, ISLET CELL	1	7	0	0	2	0.0209	0.8422	0.9831	0.9941	0.9106
TESTIS										
B-INTERSTITIAL CELL TUMOR	1	0	0	5	6	0.4643	0.0051	.	0.0332	0.0127
Female Rats										
ADRENAL, MEDULLA										
B-PHEOCHROMOCYTOMA	3	2	10	9	7	0.4895	0.2762	0.0119	0.0183	0.0633
Pheochromocytoma B+M	3	2	10	9	8	0.4895	0.1954	0.0119	0.0183	0.0379
CERVIX										
B-POLYP, ENDOMETRIAL STROMAL	0	0	1	1	3	.	0.0397	0.4824	0.4762	0.1079
UTERUS										
B-POLYP, ENDOMETRIAL,STROMAL	2	0	4	0	4	0.7416	0.0793	0.0500	.	0.0474
Uro-Genital										
Endo. Stromal Sarcoma/Polyp	3	0	6	2	10	0.8708	0.0019	0.0103	0.2297	0.0005
Endometrial Stromal Polyp	3	0	5	1	9	0.8708	0.0017	0.0228	0.4762	0.0010

Applying the Haseman-Lin-Rahman rules for a single study described above, and using the incidence in the no treatment group to decide if a tumor is rare or not, we would conclude, in male rats, that the test of trend in benign interstitial cell tumor was statistically significant ($p = 0.0051 < 0.01$). In the overall uro-genital system in female rats, tests of trend in endometrial stromal polyps alone and these tumors pooled with sarcomas were both statistically significant ($p = 0.0017$ and $p = 0.0019$, respectively, both < 0.01), as were the corresponding pairwise tests of the high dose group versus the vehicle ($p = 0.001$ and $p = 0.0005$, respectively, both < 0.01). Also in female rats the test of trend in endometrial stromal polyps in the cervix was statistically significant ($p = 0.0397 < 0.05$). When using the untreated group to define whether or not a tumor is classified as rare, following the Haseman-Lin-Rahman rules for tests of trend and

comparisons with the high dose group above, no other tests achieved statistical significance, though several were close. Other comparisons are based on the low or medium dose group and comparisons between the vehicle and untreated dose groups, plus the implications of using the vehicle group instead of the untreated group to classify whether a tumor is rare or not, are discussed in Section 3.2.1.2 and in Appendix 3 below.

In Tg.AC mice the primary analysis is usually based on profiles of papilloma counts. Dunson, et al (2000), recommend an analysis of the actual increase in maximum papilloma count. Two slightly different models are used. Because of difficulties in estimating the parameters in the original Dunson, et al model, a restricted Dunson model is specified. This assumes that there is no particular individual animal effect. The restricted Dunson model for active papillomas alone seems to show no particular evidence of a dose effect on papilloma onset (Males: p = 0.1551, Females p = 0.4995) or increase in maximum papilloma count (Males: p = 0.4065, Females: p = 0.5352). When dealing with pooled active and latent papillomas in male mice there is some questionable evidence of an effect on onset (p = 0.0556), but not in female mice (p = 0.2635). However, even with pooled active and latent papillomas there is no strong evidence of an effect on increase in papillomas after the first (Males: p = 0.1363, Females: p = 0.5507). An alternative model is proposed that incorporates a single animal effect on both papilloma onset and increase. Adding this parameter makes the effect of onset in male mice statistically insignificant (p = 0.1616). Detailed description of this model, plus the results and summary tables of actual papilloma counts are presented in Appendix 3.

As a secondary analysis we might model the number of Tg.AC mice with tumors using a poly-k methodology. Those organ tumor combinations that had a statistically significant result (i.e., a significance level at or below 0.05) are displayed in Table 5 below. In female mice there is a clear trend in papillomas of the stomach (p = 0.0142). The pairwise differences between the high dose and vehicle in female mice was also statistically significant (p = 0.0451). Perhaps partially due to the considerably higher mortality in the high dose group in males, the test of trend is very nonsignificant (p = 0.9998). However, the pairwise tests between the vehicle and the low, medium, and medium high dose groups are all statistically significant (p=0.0260, 0.0144, and 0.0095, respectively).

Table 5. Potentially Statistically Significant Neoplasms in Tg.AC Mice

Organ / Tumor	Incidence Veh	Low	Med	Med Hi	Hi	Results of tests trend	Low/ Veh	Med/ Veh	Medhi/ Veh	High/ Veh
Male Mice										
stomach										
papilloma	2	8	12	9	0	0.9998	0.0260	0.0144	0.0095	0.8167
Female Mice										
stomach										
papilloma	0	2	0	1	4	0.0142	0.3033	.	0.4500	0.0451

1.2. Brief Overview of the Studies

This submission had a standard rat study:

Protocol BUP-N-004: Buprenorphine Base: A Two Year Dermal (Skin Painting) Carcinogenicity Study in Rats,

and a Tg.AC mice study:

Protocol BUP-P-012: Buprenorphine: 26-Week Dermal Carcinogenicity Study in Hemizygous Tg.AC Mice

1.3. Statistical Issues and Findings

1.3.1. Statistical Issues

In this section, several issues, typical of statistical analyses of these studies, are considered. These issues include details of the survival analyses, tests on tumorigenicity, multiplicity of tests on neoplasms, and the validity of the designs.

1.3.1.1. Survival Analysis:

The survival analyses presented here are based on both the log rank test and the Wilcoxon test comparing survival curves. The log rank tests tends to puts higher weight on later events, while the Wilcoxon test tends to weight events more equally, and thus is more sensitive to earlier differences in survival. The log rank test is most powerful when the survival curves track each other, and thus the proportional hazard assumption seems to be true. Both tests were used to test both homogeneity of survival among the treatment groups and the effect of dose on trend in survival. Appendix 1 reviews the specific animal survival analyses in more detail. The results of the Sponsor's analyses are summarized in Sections 3.2.1.1 and 3.2.2.1.

1.3.1.2. Multiplicity of Tests on Survival:

Using either the logrank or the Wilcoxon test, there are eight tests of survival in each rat by gender combination, and six in each mouse by gender combination. If we were to assume the tests are independent across comparisons, which clearly half of them are not, and assume that there is absolutely no difference in survival, the probability of at least one statistically significant result in mice, at the usual 0.05 level, is about 0.2649, and in rats is about 0.3366. Such is the possible price paid for the multiplicity of hypothesis tests.

1.3.1.3. Tests on Neoplasms:

Appendix 2 presents the results from the FDA poly-k analysis on tumor incidence on rats, as well as the Tg.AC mice. In mice this should be considered as a secondary analysis. The poly-k test is a modification of the original Cochran-Armitage test of trend in response to dose, adjusted for differences in mortality (please see Bailer & Portier, 1988, Bieler & Williams, 1993). For rats the Sponsor provided the results of so-called Peto tests of

carcinogenicity, here applied to pairwise differences with the vehicle control. These tests require accurate specification of the cause of death, which is often difficult. It was noted in the report of the Society of Toxicological Pathology "town hall" meeting in June 2001 that the poly-k modification of the Cochran-Armitage tests of trend has been recommended over the corresponding Peto tests.

For Tg.AC mice the preferred method of analysis is to model the profiles of papilloma counts over time. Dunson, et al (2000) proposed an analysis based on the increase in maximum papilloma count, with separate effects for tumor onset and tumor increase. A fairly detailed description of this model, plus the results and summary tables of actual papilloma counts are presented in Appendix 3. Because of difficulties in estimating the parameters in the entire Dunson, et al model, a restricted Dunson model is used. The only parameter for an individual animal effect in the original Dunson model is for tumor onset, not increase. The restricted model assumes this parameter is zero. Results from an alternative model that incorporates a single animal effect on both both papilloma onset and increase are also presented in the aforementioned appendix. Although details do not seem to be provided, the Sponsor's analysis of papillomas is said to be based on Fisher Exact tests and ANOVAs. A secondary analysis is to model the counts of various animals with tumors, similar to that used with rats.

1.3.1.4. Multiplicity of Tests on Neoplasms:

Testing the various neoplasms necessitates a number of statistical tests, which in turn necessitates an adjustment in experiment-wise Type I error (i.e., the probability of rejecting a true null hypothesis). Based on his extensive experience with such carcinogenicity analyses in standard laboratory rodents, for pairwise tests between the high dose group and controls in two species, Haseman (1983) claimed that for a roughly 0.10 (10%) overall false positive error rate, rare tumors should be tested at a 0.05 (5%) level, and common tumors (with a historical control incidence greater than 1%) at a 0.01 level. Similarly, Lin and Rahman (1998) showed that tests of trend should be tested at a 0.025 level for rare tumors and 0.005 for common tumors. However, for a single species study both the test of trend and the test comparing the high dose to the appropriate controls should both be tested at the 0.05 level for rare tumors and 0.01 for common tumors. This approach is intended to balance both Type I error and Type II error (i.e., the error of concluding there is no evidence of a relation to tumorgenicity when there actually is such a relation).

Since there are usually only a few tumor types in transgenic mouse studies, with a small number of animals, over a short period of time, even the incidence analysis similar to that used for rats is generally tested at a 0.05 significance level. That is, no correction for multiplicity is applied. Further, with Tg.AC mice, interest is usually focused on papillomas so no particular adjustment for multiplicity is applied to the Tg.AC mouse results in this report. With the Dunson model and the varietions used here there are four key parameters tested (onset and in increase in both genders), so one might wish to apply a Bonferroni correction (i.e., for an overall 0.05 level, a test should achieve a 0.0125 level), or one of the less conservative step down or step up methods. However, that is not done here.

Thus, to summarize, the multiplicity adjustment used for the rat study corresponds to a single species study, while no adjustment is applied to the Tg.AC mouse results.

1.3.1.5. Validity of the Designs:

When determining the validity of designs there are two key points:

1) adequate drug exposure,
2) tumor challenge to the tested animals.

1) is related to whether or not sufficient animals survived long enough to be at risk of forming late-developing tumors and 2) is related to the Maximum Tolerated Dose (MTD), designed to achieve the greatest likelihood of tumorigenicity.

Lin and Ali (2006), quoting work by Haseman, have suggested that in standard laboratory rodent species, a survival rate of about 25 animals, out of 50 or more animals, between weeks 80-90 of a two-year study may be considered a sufficient number of survivors as well as one measure of adequate exposure. Note that as a percentage of animals that survived to week 91, this criterion is met in rats in all dose groups in both genders, except for the low dose group in males (Please see Table 9 on page 14 and Table 10 on page 15).

The mean weight values in the following tables were taken from the Sponsor's rat and mice reports (Table 17.6, Rat study pages 154-165, Table 17.8, Rat study pages 174-175,Table 6, Mice study pages 60-65):

Table 6. Mean Weights for Mice and Rats

Rats Dose Group	Males				Females			
	Week		Change from baseline	% change relative to vehicle	Week		Change from baseline	% change relative to vehicle
	1	100			1	100		
No Treat	201	849	649	120.6%	164	502	341	129.7%
Vehicle	199	734	538		162	423	263	
Low	210	565	357	66.4%	168	390	224	85.2%
Medium	210	572	357	66.4%	168	401	234	90.0%
High	211	566	347	64.5%	169	395	230	87.4%

Mice Dose Group	Males				Females			
	Days		Change from baseline	% change relative to vehicle	Days		Change from baseline	% change relative to vehicle
	1	169			1	169		
Vehicle	25.5	32.4	6.9		21.0	24.8	3.8	
TPA	25.9	32.3	6.5		20.4	24.4	4.0	
Low	25.0	31.6	6.6	95%	20.9	24.6	3.7	96%
Medium	23.9	32.3	8.2	118%	20.7	24.5	3.7	97%
Med-High	24.8	31.6	6.8	98%	20.9	23.6	2.7	71%
High	24.3	31.1	6.7	100%	20.4	25.5	5.1	132%

Chu, Ceuto, and Ward (1981), citing earlier work by Sontag *et al* (1976) recommend that the MTD "is taken as 'the highest dose that causes no more than a 10% weight decrement as compared to the appropriate control groups, and does not produce mortality, clinical signs of toxicity, or pathologic lesions (other than those that may be related to a neoplastic response) that would be predicted to shorten the animal's natural life span' " From Table 6, above, in male rats there is a clear decrement in the BuTrans groups compared to vehicle. However, it is clear this is not linearly related to dose and is apparently associated with the mere presence of BuTrans. In female rats the weight decrement is close to 10% for all three BuTrans groups. For mice the weight decrements and exceedances are rather erratically distributed.

In male and female rats, the Sponsor summarizes overall food consumption as follows: "during most weeks in which food consumption was measured, the mean food consumption of the acetone-treated group was significantly different (numerically higher) than that of the untreated control group. Additionally, in males and females, during most weeks in which food consumption was measured, mean values for the groups administered buprenorphine were significantly different (numerically lower) than those of the acetone control group. Thus, the lower body weight and body weight change values for the buprenorphine groups in relation to the vehicle control group correlate with the lower food consumption values for the buprenorphine groups in relation to the vehicle control. The treatment-related response for increased food consumption in the acetone control group appeared muted or absent in buprenorphine-treated groups." (page 41 of report)

In mice the Sponsor states that "there was no evidence of consistent dose or test article-related changes in food consumption. Total food consumption (from Day 1 to terminal sacrifice) in the females was statistically significantly decreased in Groups 4, 5, and 6 compared to the vehicle control. The decreases in female total food consumption were comparable in all dose levels (2.0 to 9.0% lower than the vehicle control) and did not occur in a dose-related manner, but may have been associated with treatment. There were no statistically significant differences in total food consumption in the test article-treated males when compared to the vehicle control group (Group 1)." (page 30 of report)

Whether this has any effect on the MTD is a decision requiring the expertise of the toxicologist.

Again from 2) above, excess mortality not associated with any tumor or sacrifice in the higher dose groups might suggest that the MTD was exceeded. If dosing is close to the MTD one would expect slightly higher mortality due to toxicity, but not so much that it largely reduces the number of animals exposed to the drug. In male rats there is some evidence of higher mortality in the high dose group than in the untreated group, but not in the vehicle group. In female rats and both genders in mice mortality in the high dose group does not seem to differ from the control groups. A related way to assess whether or not the MTD was achieved is to measure mortality not associated with any identified tumor. Table 7., below,

indicates that the number of animals in each dose group that died of a natural death or moribund sacrifice, but did not show any tumors (i.e., the "Event"):

Table 7. Natural Death with No Identified Tumor

Rats		1.Untreated 0 mg/kg	2. Vehicle 0 mg/kg	3. Low 20 mg/kg	4.Medium 60 mg/kg	5. High 200 mg/kg
Male	Event	6	12	24	18	20
	No event	59	53	41	47	45
Female	Event	0	1	7	8	5
	No event	65	64	58	57	60

Mice		1.Vehicle	2. TPA	3. Low	4.Medium	5. Med-Hi	6. High
Male	Event	0	2	2	0	1	6
	No event	25	17	28	30	29	24
Female	Event	0	4	0	0	5	3
	No event	25	15	25	25	20	22

To compare the incidence of deaths without tumors (i.e., the "Event" above) we can specify the usual survival tests where animals that die with a tumor or are sacrificed are considered as censored. The remaining animals are those that die a natural death prior to developing a tumor. One indication that the MTD is achieved would be dose related excess toxicity, resulting in a dose related increase in these deaths, particularly in the high dose group. Comparing the high dose to the untreated group in rats there are statistically significant differences in both genders (Male rats: log rank p = 0.0026, Wilcoxon p = 0.0086, Female rats: log rank p = 0.0206, Wilcoxon p = 0.0284). However, results are not so clear when comparing the high dose group to the vehicle (Male rats: log rank p = 0.1394, Wilcoxon p = 0.1378, Female rats: log rank p = 0.0836, Wilcoxon p = 0.1385). In mice results are rather more equivocal (Male mice: log rank p = 0.0193, Wilcoxon p = 0.0195, Female mice: log rank p = 0.0770, Wilcoxon p = 0.0771). Although this is a decision for the toxicologist, this may be evidence that the MTD was achieved in rats and male mice, but perhaps not in female mice. Note, however, the Sponsor suggests that in each species the high dose is the highest feasible dose with the current vehicle.

1.3.2. Statistical Findings

Please see Section 1.1 above.

2. INTRODUCTION

2.1. Overview

According to the Sponsor the objective of these studies was to assess the tumorigenicity of buprenorphine base when administered dermally once daily to standard laboratory rats and to

Tg.AC mice. The rat study was terminated after 99 weeks of treatment and the Tg.AC mouse study after 26 weeks of treatment.

2.2. Data Sources

A number of SAS transport files, one set for rats and another set for mice were provided by the Sponsor and placed in the CDER electronic data room (edr). These were labeled [name].xpt and each contained a SAS data set labeled [name].sas7bdat.

Rats:	cpchem	food	mortal	tumor
	cphem	macro	signsf	weights
	cpurine	micro	signsm	
Mice:	macro	mortal	tumor	
	micro	organwt	signs	

The Tg.AC mice tumor data sets included 5-10 animals with different, irregularly spaced times of death. The Sponsor classified these as being part of an interim sacrifice, and thus the times of death are treated as censored in any statistical tests. Presumably these animals should be considered as part of a moribund sacrifice. However, since they all occur in the TPA group, which is generally ignored, this classification was left unchanged.

As noted above, the primary analysis for Tg.AC mice is based on weekly papilloma counts, not included in the original data sets noted above. Later, upon request of the agency, the Sponsor submitted such a data set: papiloma.sas7bdat.

3. STATISTICAL EVALUATION

3.1. Evaluation of Efficacy

NA

3.2. Evaluation of Safety

3.2.1. Protocol BUP-N-004: Buprenorphine Base: A Two Year Dermal (Skin Painting) Carcinogenicity Study in Rats,

STUDY DURATION: 100 Weeks (99 weeks of treatment)
STARTING DOSING DATE: 9 June 2004
FINAL NECROPSY DATE: 4 May 2006
RAT STRAIN: Charles River Crl:CD(SD)IGS BR Rats
ROUTE: Dermal (skin painting)

The basic design of the rat study has five dose groups, summarized in Table 8 below:

Table 8. Design of Rat Study

Group	BuTrans Dosage (mg/kg/day)	Interval	Concen-tration (mg/mL)	Volume (mL/kg/ day)	# animals Main Study	# animals Toxico-kinetic
1. Untreated	0	NA	0	8	65	0
2. Vehicle	0	Daily	0	8	65	5[1]
3. Low	20	Daily	10	8	65	12[2]
4. Medium	60	Daily	30	8	65	12[2]
5. High[3]	200	Daily	100/90	2/2.22	65	12[2]

[1] Five animals/sex/time point were bled for toxicokinetic sampling.
[2] Three animals/sex/time point were bled for toxicokinetic sampling, and three animals/sex/group served as extras to be used as needed based on mortality.
[3]From days 1 through 177, the dosage concentration was 100 mg/mL and the dosage volume 2 mL/kg for Group 5. From day 178 through the end of study, the dosage concentration was 90 mg/mL with dosage volume 2.22.

According to the Sponsor the "dermal route of administration was selected because this is the intended route of human exposure. Dermal application in acetone was chosen because previous work on file with the sponsor indicated that higher exposure could be achieved by this route than via application of the buprenorphine transdermal system being evaluated in the clinic." (page 27 of report) "The maximum solubility of buprenorphine base in acetone is 100 mg/mL. The maximum-feasible dose (MFD) of approximately 200 mg/kg (using a maximum volume of 2 mL/kg) did not produce any dose-related pharmacotoxic signs or apparent dose-limiting toxicity in a 3-month dose range-finding study; therefore, the dose levels for this study are up to an MFD of approximately 200 mg/kg/day." (page 27 of report) Other doses were based on this MFD.

Animals were housed individually, with food and water available *ad libitum*.

The Sponsor states that "Animals were prepared for treatment by close clipping of the hair with an electric clipper prior to the first dose administration and as needed thereafter. Control and treatment room-specific clippers, clipper blades, and any other instrumentation were used. This clipped area extended from the scapular region to the hips and halfway down each flank. The test article was spread over the clipped area. The body surface to which the test article was applied to each animal was approximately 24 to 80 cm^2. The application area was clipped throughout the study as necessary. The days the fur was clipped were recorded. Care was taken to minimize clipper abrasions." (page 27 of report)

3.2.1.1. Sponsor's Results and Conclusions

This section will present a summary of the Sponsor's analysis on survivability and tumorigenicity in rats.

Survival analysis:

The Sponsor indicates that although "no significant trend with the vehicle was noted, Group 3 (20 mg/kg/day) was significantly increased over Control 2 for both Cox-Tarone and Gehan-Breslow tests (p = 0.0461 and 0.0108, respectively). No significances were noted for Control 1 versus Control 2." Continuing "for females, although no significant trend with Control 2 was noted, Group 4 (60 mg/kg/day) was significantly increased over Control 2 for both Cox-Tarone and Gehan-Breslow tests (p = 0.0227 and 0.0493, respectively). No significances were noted for Control 1 versus Control 2." (page 4187 of report)

Tumorigenicity analysis:

The Sponsor states that they used pairwise Peto tests to assess carcinogenicity, but did not adjust for the multiplicity of tests. The Sponsor summarizes carcinogenic results as follows: "in males, for testis benign interstitial cell tumor, a significant positive trend (p = 0.0012) was noted in Groups 3 (20 mg/kg/day), 4 (60 mg/kg/day), and 5 (200 mg/kg/day) with corresponding increases in Groups 4 and 5 (p = 0.0247 and 0.0030, respectively) versus the vehicle control group (Control 2). No other significances were noted in males." Continuing "in females, for adrenal medullary benign pheochromocytoma and benign and malignant pheochromocytoma combined, several significant increases in the incidences in treated groups were noted. Groups 3 (20 mg/kg/day), 4 (60 mg/kg/day), and 5 (200 mg/kg/day) were significantly increased over Control 2 (benign only: p = 0.0145, 0.0072, and 0.0419, respectively; benign combined with malignant: p = 0.0145, 0.0072, and 0.0297, respectively), but without a corresponding significant trend with Control 2 (benign only: p = 0.0935; benign combined with malignant: p = 0.0618). Significant increases were also noted for multiple organs (uterus/cervix/vagina) for benign endometrial stromal polyp alone and combined with malignant endometrial stromal sarcoma. For both cases, Groups 3 and 5 were significantly increased over Control 2 (polyp alone: p = 0.0041 and 0.0004, respectively; polyp combined with sarcoma: p = 0.0018 and 0.0002, respectively), with corresponding significant positive trends (polyp alone: p = 0.0015; polyp combined with sarcoma: p = 0.0011). No other significances were noted in females.

The Sponsor concludes that "Treatment with the test compound did not show any significant mortality pattern in animals of either sex. In males alone, treatment-related increases in benign interstitial cell tumors of the testis noted in Groups 4 (60 mg/kg/day) and 5 (200 mg/kg/day) were associated with a significant positive trend. For this case, Group 3 (20 mg/kg/day) was unaffected, and the two control groups were similar in incidence. No other significant effects were noted for any other neoplastic lesions in males.

"In females, significantly increased incidences of adrenal medullary benign and benign and malignant combined pheochromocytoma were noted in all three treated groups. The positive trend was not significant in this case because the dose response was not strictly monotonically increasing. Only one incidence of malignant pheochromocytoma was observed in Group 5 (200 mg/kg/day). The only other significant neoplastic effect was found in the benign endometrial stromal polyp and polyp and malignant stromal sarcoma combined

incidences of the female genito-urinary tract. They were significantly elevated in the 20- and 200-mg/kg/day dose treated groups. The 60-mg/kg/day group, in this case, was not statistically elevated, causing the dose-response curve to be nonmonotonic. The biological meaning of this nonmonotonicity is not known." (pages 4178-4179 of report)

3.2.1.2. FDA Reviewer's Results

This section will present the Agency findings on survival and tumorigenicity in male and female rats.

Survival analysis:

The following tables (Table 9 for male rats, Table 10 for female rats) summarize the mortality results for the dose groups. The data were grouped for the specified time period, and present the number of deaths during the time interval over the number at risk at the beginning of the interval. The percentage cited is the percent that survived at the end of the interval. The Kaplan-Meier survival plots of in Appendix 1 provide a more detailed picture of mortlity losses.

Table 9. Summary of Male Rats Survival (dosed at mg/kg/day)

Period (Weeks)	Untreated 0	Vehicle 0	Low 20	Medium 60	High 200
1-52	6/65[1] 90.8%[2]	7/65 89.2%	13/65 82.3%	5/65 92.3%	7/65 89.9%
53-78	13/59 70.8%	20/58 69.2%	24/52 66.2%	17/60 66.2%	19/58 60.0%
79-91	13/46 50.8%	14/38 55.4%	13/28 35.4%	17/43 40.0%	15/39 36.9%
92-100	11/33 33.9%	10/24 35.4%	4/15 26.2%	5/26 32.3%	9/24 23.1%
Terminal 100-101	22	14	11	21	15

[1] number of deaths / number at risk
[2] overall per cent survival to end of period.

In these tables the terminal period only includes those animals were sacrificed. Animals that died of other causes during the terminal period are included in the preceding, but overlapping time period. Note that the Sponsor's data set codes one animal in the high dose group as having a terminal sacrifice at Week 51, which appears to be a data error. Since this was only one animal, it was left unchanged.

Table 10. Summary of Female Rats Survival (dosed at mg/kg/day)

Period (Weeks)	Untreated 0	Vehicle 0	Low 20	Medium 60	High 200
1-52	1/65[1] 98.5%[2]	7/65 89.2%	5/65 92.3%	8/65 87.7%	7/65 89.2%
53-78	18/64 70.8%	13/58 69.2%	17/60 66.2%	10/57 62.3%	18/58 61.5%
79-91	19/46 41.5%	9/45 55.4%	20/43 35.4%	25/47 33.9%	16/40 36.9%
92-101	6/27 32.3%	13/36 35.4%	6/23 26.2%	10/22 18.5%	5/24 29.2%
Terminal 100-101	21	23	17	12	19

[1] number of deaths / number at risk
[2] per cent survival to end of period.

Table 11 below provides the significance levels of the tests of homogeneity and trend over dose groups as proposed in Section 1.3.1.1, above. Again, for reference, note that the

Table 11. Statistical Significances of Tests of Homogeneity and Trend in Survival

	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Rats Homogeneity over Groups 1 & 2	0.0928	0.1245	0.7659	0.9772
Homogeneity over Groups 1-5	0.0110	0.0030	0.3065	0.4044
Homogeneity over Groups 2-5	0.0495	0.0156	0.2861	0.4923
Trend over Groups 2-5	0.5074	0.4121	0.6537	0.5401

From figure A.1.1 in Appendix 1, in male rats it seems that the low dose group has the highest mortality, with the vehicle, medium, and low dose groups generally intertwined with the next highest mortality, while the untreated group has the lowest mortality. However, the difference in survival between the vehicle and untreated dose groups was not quite statistically significant at the usual significance levels (Log rank $p = 0.0928$, Wilcoxon $p=0.1245$). At the usual 0.05 significance level, the hypothesis of homogeneity over all five dose groups would be rejected (Log rank $p = 0.0110$, Wilcoxon $p=0.0030$), as would be the more restricted hypothesis of homogeneity over the vehicle and the three actual dose groups (Log rank $p = 0.0495$, Wilcoxon $p=0.0156$). However, there is no evidence for a dose related trend (both $p \geq 0.4121$), consistent with the general observation that the low dose group had the highest mortality, while the vehicle, medium, and low dose groups were generally intertwined. In female rats, there was no strong evidence of heterogeneity or trend among any of the treatment groups (all $p \geq 0.2861$). This seems consistent with the observation, from Figure A.1.2, that the dose groups in female rats are generally intertwined.

Tumorigenicity analysis:

As discussed in Section 1.3.1.4, for common tumors, the Haseman-Lin-Rahman rules for adjusting for the multiplicity of tests of tumorigenicity for single species, two gender, two year studies are that for a roughly 0.10 (10%) overall false positive error rate, one should claim statistical significance for either the test of trend or the comparisons of the high dose group to controls if the nominal statistical significance is 0.01 (1%). For tumors classified as rare, the corresponding level is 0.05 (5%).

Table 12, below, for male and female rats, displays the incidence of neoplasms and the corresponding p-values for all organ tumor combinations where there was at least p-value at 0.05 or less.

Even if we were to accept this somewhat higher type I error rate we would conclude that in male rats the incidence of islet cell benign adenoma in the vehicle group is significantly higher than in the untreated control group ($p = 0.0209 < 0.05$).

Table 12. Potentially Statistically Significant Neoplasms in Rats

Organ / Tumor	Incidence Un-trt	Veh	Low	Med	Hi	Results of tests Untrt/ Veh	trend	Low/ Veh	Med/ Veh	High/ Veh
Male Rats										
PANCREAS										
B-ADENOMA, ISLET CELL	1	7	0	0	2	0.0209	0.8422	0.9831	0.9941	0.9106
TESTIS										
B-INTERSTITIAL CELL TUMOR	1	0	0	5	6	0.4643	0.0051	.	0.0332	0.0127
Female Rats										
ADRENAL, MEDULLA										
B-PHEOCHROMOCYTOMA	3	2	10	9	7	0.4895	0.2762	0.0119	0.0183	0.0633
Pheochromocytoma B+M	3	2	10	9	8	0.4895	0.1954	0.0119	0.0183	0.0379
CERVIX										
B-POLYP, ENDOMETRIAL STROMAL	0	0	1	1	3	.	0.0397	0.4824	0.4762	0.1079
UTERUS										
B-POLYP, ENDOMETRIAL,STROMAL	2	0	4	0	4	0.7416	0.0793	0.0500	.	0.0474
Uro-Genital										
Endo. Stromal Sarcoma/Polyp	3	0	6	2	10	0.8708	0.0019	0.0103	0.2297	0.0005
Endometrial Stromal Polyp	3	0	5	1	9	0.8708	0.0017	0.0228	0.4762	0.0010

Applying the Haseman-Lin-Rahman rules for a single study described above, and using the incidence in the no treatment group to decide if a tumor is rare or not, we would conclude, in male rats, that the test of trend in benign interstitial cell tumor was statistically significant ($p = 0.0051 < 0.01$). The corresponding test comparing the high dose to vehicle was close to significance ($p = 0.0127 \approx 0.01$). If we use the vehicle group to determine whether or not a tumor is rare, we would have the same results in male rats, with the additional conclusion that the pairwise test between the high dose group and the medium dose group was statistically significant ($p = 0.0127$ and $0.0332 < 0.05$ respectively). Again using the untreated group to determine rarity, among female rats, the tests of trend in endometrial stromal polyps in the uro-genital system alone and with these tumors pooled with sarcomas were both statistically significant ($p = 0.0017$ and $p = 0.0019$, respectively, both < 0.01), as were the corresponding

pairwise tests of the high dose group versus the vehicle (p = 0.001 and p = 0.0005, respectively, both < 0.01). Again, using the incidence in the no treatment group, the test of trend in benign endometrial stromal polyps of the cervix of female rats was statistically significant (p = 0.0397 < 0.05). Strictly speaking, this adjustment for multiplicity applies to tests of trend and comparisons of the high dose to vehicle. Applying the adjustment to tests comparing the low, medium, and untreated dose group to vehicle will be anticonservative.

Despite its nominal statistical significamce, adjusting for multiplicity the test between the high dose and vehicle in pooled benign and malignant pheochromocytoma in the adrenal medulla of female rats was not statistically significant (p = 0.0379 > 0.01), nor were the pairwise tests of the medium and low dose groups with vehicle (p = 0.0379 and p = 0.0183, respectively, both > 0.01). Also in female rats the test of trend in endometrial stromal polyps in the cervix was not statistically significant (p = 0.0397 > 0.01). Using the incidence in the vehicle group to define the rarity of tumors in female rats, results would the same, except that the pairwise difference between the low dose group and the vehicle groups in endometrial stromal polyps in the uro-genital system alone and with these tumors pooled with sarcomas would both be statistically significant (p = 0.0228 and 0.0103, respectively, both < 0.05). Following the Haseman-Lin-Rahman rules above, no other tests achived statistical significance, though several were close.

3.2.2. Protocol BUP-P-012: Buprenorphine: 26-Week Dermal Carcinogenicity Study in Hemizygous Tg.AC Mice

STUDY DURATION: 26 Weeks
STARTING DOSING DATE: 12 February 2007
FINAL NECROPSY DATES: 13-17 August 2007
RAT STRAIN: FVB/N hemizygous Tg.AC mice
ROUTE: Daily dermal

According to the Sponsor "The purpose of this study was to assess whether daily dermal administration of the test article (buprenorphine base) for 26-weeks increased the incidence of skin tumors in hemizygous Tg.AC transgenic mice. Additionally, pathological evaluation of tissues was performed to identify other potential buprenorphine target organs of toxicity or tumorigenicity. This study consisted of a Main Study in Tg.AC mice and a concurrent Toxicokinetic (TK) Study in FVB/N mice. The Main Study consisted of one vehicle control group (Group 1), one positive control group (Group 2) and four test article treatment groups (Groups 3-6). The TK Study included a vehicle control group of 8 FVB/N mice/sex and four test article treatment groups. The positive control group was treated dermally with 1.25 µg TPA (12-O- tetradecanoylphorbol 13 acetate) in 0.1 mL acetone 3 times weekly. All other animals received daily dermal administrations of vehicle or test article in vehicle at a dosing volume of 6.7 µL/g/dose for 26 consecutive weeks. Test article-treated mice received one of four doses (18.75, 37.5, 150 or 600 mg/kg/day. In addition, blood samples were drawn from

FVB/N mice (TK Study) at various time points for plasma concentration analysis by the Sponsordesignated laboratory." (page 10 of report)

The design of the Tg.AC mouse study can be summarized as follows:

Table 13. Design of Tg.AC Mouse Study

Group	Nominal Dosage (mg/kg/day)	Main Study n		Toxicokinetic n	
		Males	Females	Males	Females
1. Vehicle Control	Acetone vehicle	25	25	8	8
2. Positive Control	1.25µ TPA x3/week	20	20	-	-
3. Low	18.75 BuTrans mg/kg/day	30	25	34	28
4. Medium	37.5 BuTrans mg/kg/day	30	25	34	28
5. Medium-High	150 BuTrans mg/kg/day	30	25	34	28
6. High	600 BuTrans mg/kg/day	30	25	34	28

The Sponsor also labeled these treatment groups as "Vehicle Control", "Positive Control", "Low", "Mid-Low", "Mid-High", and "High" dose groups, respectively.

Animals were housed individually, with food and water available *ad libitum*.

The Sponsor justifies dose as follows: "In the present study, the doses were selected based on data collected from a previous 28-day repeat dose dermal toxicity study performed in FVB/N mice . . . .The high dose was the maximum feasible dose that could be administered based on the limit of buprenorphine solubility in acetone and the maximum volume of dosing solution that could be administered to the skin site of administration without dripping of the solution off the animal." (page 17 of report)

"Skin masses were diagnosed in vivo and scored as a latent papilloma when they attained a 2 mm diameter and protruded from the surface. The date that these criteria were met was recorded and defined the latency period for that tumor. Upon remaining scorable for three consecutive weekly observations, skin papillomas were subsequently scored as actual papillomas. Once scored as an actual papilloma it was maintained in the count regardless of its fate (i.e., disappearance, death of the animal, conversion from papilloma to carcinoma). This procedure for scoring applied to both papillomas and carcinomas. In addition, the number of latent papillomas or carcinomas remained in the skin tumor count if it was present at the time any animal was removed from the study because of early death, moribund sacrifice or terminal sacrifice. Previously scored actual tumors which disappear were scored in the record as such. The skin lesion data differentiated papillomas from carcinomas. In vivo characterizations of papillomas or carcinomas was made on the basis of palpation, evidence of subcutaneous invasion, and ulceration." (page 20 of report)

Survival analysis:

The Sponsor notes that: "Mortality in the Main Study included 1/25, 3/20, 3/30, 3/30, 3/30 and 7/30 males and 2/25, 5/20, 3/25, 4/25, 6/25 and 5/25 Main Study females in Groups 1-6, respectively. Mortality in the End of Study TK animals in both sexes was similar in the test article treated groups to that observed in the Main Study (0/8, 4/22, 4/22, 8/22 and 10/22 males and 0/8, 6/16, 9/16, 10/16 and 16/16 females in Groups 1, 3, 4, 5 and 6, respectively). The increased mortality compared to the vehicle control was statistically significant in the End-of-Study TK females at the top two dose levels (150 and 600 mg/kg/day). However, there were no trends in cause of death that suggested relationship to treatment. The most common causality classification was "undetermined". There was an apparent slight (non-statistically significant) increase in the incidence of mortality in some Main Study treated groups compared to the controls and there was greater susceptibility to tumorigenicity in mice, mortality in FVB/N mice than Tg.AC and for females versus males, but the relationship of this increase to treatment is unclear and of dubious biological significance because of the small magnitude of the increase and the lack of any microscopic findings (neoplastic or otherwise) in Main Study animals that could account for the increased mortality in treated groups." (page 10 & 11 of report)

Tumorigenicity analysis:

"The number of mice that developed skin tumors at the skin site of application (SOA) per effective animal (animals surviving to scheduled or terminal sacrifice) was 2/25, 16/20, 2/30, 4/30, 2/30 and 3/29 for males and 2/25, 18/20, 3/25, 0/25, 0/25 and 1/23 for females (Groups 1-6, respectively). Statistical analysis revealed a statistically significant increase in the number of tumor bearing animals in the positive control group (Group 2) of both sexes and tumor multiplicity data indicated a statistically significantly increased mean tumor burden in the positive control group of both sexes when compared to the vehicle control group (Group 1). These findings demonstrated the sensitivity of the test system. No statistically significant difference in the number of tumor-bearing animals per group or tumor multiplicity was noted between the test article treatment groups and the vehicle control group (Group 1)." (page 10 of report). The Sponsor's statistical analysis of papillomas is said to be based on Fisher Exact tests and ANOVAs, although few details seem to be provided.

3.2.2.2. FDA Reviewer's Results

This section will present the Agency findings on survival and tumorigenicity in male and female mice.

Survival analysis:

The following tables (Table 14 for male mice, Table 15 for female mice) summarize the mortality results for the dose groups. The data were grouped for the specified time period, and present the number of deaths during the time interval over the number at risk at the beginning of the interval. The percentage cited is the percent survived at the end of the interval.

Table 14. Summary of Male Mice Survival

Period (Weeks)	Vehicle	TPA	Low	Med-Low	Med-High	High
0-12	0/25[1] 100%[2]	1/16 83.7%	1/30 96.7%	0/30 100%	3/30 90%	4/30 86.7%
13-18	0/25 100%	7/15 50.0%	1/29 93.3%	1/30 96.7%	0/27 90%	1/26 83.3%
19-24	1/24 96%	1/8 43.8%	1/28 90%	2/29 90%	0/27 90%	2/25 76.7%
Terminal (25-26)	25	10[3]	27	27	27	23

[1] number deaths / number at risk
[2] overall per cent survival to end of period.
[3] in this dose group, all these animals actually died before the final sacrifice period

In these tables the terminal period only includes those animals sacrificed in the terminal period. Animals that died of other causes during the terminal period are included in the preceding, but overlapping time period.

Table 15. Summary of Female Mice Survival

Period (Weeks)	Vehicle	TPA	Low	Med-Low	Med-High	High
0-12	0/25[1] 100%[2]	0/25 100%	0/25 100%	1/25 96%	2/25 92%	2/25 92%
13-18	1/25 96%	5/25 62.2%	0/25 100%	1/25 92%	3/25 80%	1/25 88%
19-24	1/24 92%	3/20 55.6%	3/25 88%	2/25 84%	1/25 76%	2/25 80%
Terminal 25-26	23	10	22	21	19	20

[1] number deaths / number at risk
[2] overall per cent survival to end of period.

Table 16 below provides the significance levels of tests of homogeneity and trend over the dose groups. Again. for reference, note that the Lixivaptan/Pluronic F-68 doses were, for groups 1-6, respectively 0/0, 0/1, 0/3.33, 0.5/1, 1/2, and 1.67/3.33. In both genders in mice, survival curves were fairly closely intertwined (see Figures A.1.3 and A.1.4 in Appendix 1), with no tests of differences or trend in survival being clearly statistically significant (for all 36 tests, $p \geq 0.0769$).

Table 16. Statistical Significances of Tests of Homogeneity and Trend in Survival

	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Mice Homogeneity over Groups 1-6	0.3031	0.2688	0.5976	0.5597
Homogeneity over Groups 1, 3-6	0.1956	0.1791	0.5179	0.4816
Trend over Groups 1, 3-6	0.0185	0.0179	0.3234	0.3076

Tumorigenicity analysis:

In Tg.AC mice the primary analysis is usually based on profiles of papilloma counts. Dunson, et al (2000), recommend an analysis of the actual increase in maximum papilloma count. A detailed description of this model, plus the results and summary tables of actual papilloma counts are presented in Appendix 3. Two models are presented. Because of difficulties in estimating the parameters in the original Dunson, et al model, a restricted Dunson model is specified. This assumes that there is no particular individual animal effect. The restricted Dunson model for active papillomas alone seems to show no particular evidence of a dose effect on papilloma onset (Males: p = 0.1551, Females p = 0.4995) or increase in maximum papilloma count (Males: p = 0.4065, Females: p = 0.5352). However, when dealing with pooled active and latent papillomas in male mice there is some questionable evidence of an effect on onset (p = 0.0556), but not in female mice (p = 0.2635). However, even with pooled active and latent papillomas there is no strong evidence of an effect on increase in papillomas after the first (Males: p = 0.1363, Females: p = 0.5507). An alternative model is proposed that incorporates a single animal effect on both both papilloma onset and increase. Adding this parameter makes the effect of onset in male mice statistically insignificant (p = 0.1616). Again, details are presented in Appendix 3.

As a secondary analysis we might model the number of mice with tumor using a poly-k methodology. Similar to the results in Table 4 above, those organ tumor combinations that had a potentially statistically significant result (i.e., a significance level at or below 0.05) are displayed in Table 17, below. In female mice there is a clear trend in papillomas of the stomach (p = 0.0142). The test of pairwise differences between the high dose and vehicle was also statistically significant (p = 0.0451). Perhaps partially due to the considerably higher mortality in the high dose group in males, the test of trend is very nonsignificant (p = 0.9998). However, the pairwise tests between the vehicle and the low, medium, and medium high dose groups are all statistically significant (p=0.0260, 0.0144, and 0.0095, respectively).

Table 17. Potentially Statistically Significant Neoplasms in Tg.AC Mice

Organ / Tumor	Incidence Veh	Low	Med	Med Hi	Hi	trend	Results of tests Low/ Veh	Med/ Veh	Medhi/ Veh	High/ Veh
Male Mice										
stomach										
papilloma	2	8	12	9	0	0.9998	0.0260	0.0144	0.0095	0.8167
Female Mice										
stomach										
papilloma	0	2	0	1	4	0.0142	0.3033	.	0.4500	0.0451

4. FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

NA

5. SUMMARY AND CONCLUSIONS

5.1. Statistical Issues and Collective Evidence

Please see Section 1.3 above.

5.2. Conclusions and Recommendations

Please see Section 1.1 above.

APPENDICES:

Appendix 1. Survival Analysis

Simple summary life tables in mortality are presented in the report (Tables 9, 10, 14, and 15), above. Kaplan-Meier estimated survival curves across dose groups for each gender in each study are displayed in Figures A.1.1-A.1.4 below. These plots include 95% confidence intervals around each survival curve (colored area around each curve). The plots are also supported by tests of homogeneity in survival over all five treatment groups, homogeneity over the four vehicle and BuTrans treatment groups, and finally, tests of trend in survival over the latter four groups. In addition there is a test of pairwise differences in survival between the untreated group and the vehicle group in rats. The statistical significances of the tests of differences in survival across treatment groups using the log rank and the so-called Wilcoxon test are given in Table A.1.1. below. One might note that the log rank tests places greater weight on later events, while the Wilcoxon test tends to weight weights them more equally, and thus places less weight on later events than does the log rank test.

Table A.1.1. Statistical Significances of Tests of Homogeneity and Trend in Survival

		Males		Females	
		Log rank	Wilcoxon	Log rank	Wilcoxon
Rats	Homogeneity over Groups 1 & 2	0.0928	0.1245	0.7659	0.9772
	Homogeneity over Groups 1-5	0.0110	0.0030	0.3065	0.4044
	Homogeneity over Groups 2-5	0.0495	0.0156	0.2861	0.4923
	Trend over Groups 2-5	0.5074	0.4121	0.6537	0.5401
Mice	Homogeneity over Groups 1-6	0.3031	0.2688	0.5976	0.5597
	Homogeneity over Groups 1, 3-6	0.1956	0.1791	0.5179	0.4816
	Trend over Groups 1, 3-6	0.0185	0.0179	0.3234	0.3076

From Figure A.1.1 below, in male rats it seems that the low dose group has the highest mortality, with the vehicle, medium, and low dose groups generally intertwined with the next highest mortality, while the untreated group has the lowest mortality. However, the difference in survival between the vehicle and untreated dose groups was not quite statistically significant (Log rank $p = 0.0928$, Wilcoxon $p=0.1245$). The hypothesis of homogeneity over all five dose groups was rejected (Log rank $p = 0.0110$, Wilcoxon $p=0.0030$), as was the hypothesis of homogeneity in the vehicle, and the three actual dose groups (Log rank $p = 0.0495$, Wilcoxon $p=0.0156$). However, there is no evidence for a dose related trend (both $p \geq 0.4121$), consistent with the general observation that the low dose group had the highest mortality, while the vehicle, medium, and low dose groups were generally intertwined. In female rats there was no strong evidence of heterogeneity or trend among the treatment groups (all $p \geq 0.2861$). This seems consistent with the observation, from Figure A.1.2, that the dose groups are generally intertwined.

Note that in rats there are eight tests in each statistic by species by gender combination. As discussed in Section 1.3.1.4. the number of these tests raises issues of multiplicity. Assuming different tests are independent and assuming there is absolutely no difference in survival, the a priori probability of at least one result in either the logrank or Wilcoxon tests exceeding the 0.05 significance level is about 0.3366. That does not resolve the issure of whether or not the statistically significant survival differences are real or an artifact, but does give some measure of the likelihood that it is an artifact.

Figure A.1.1 Kaplan-Meier Survival Curves for Male Rats

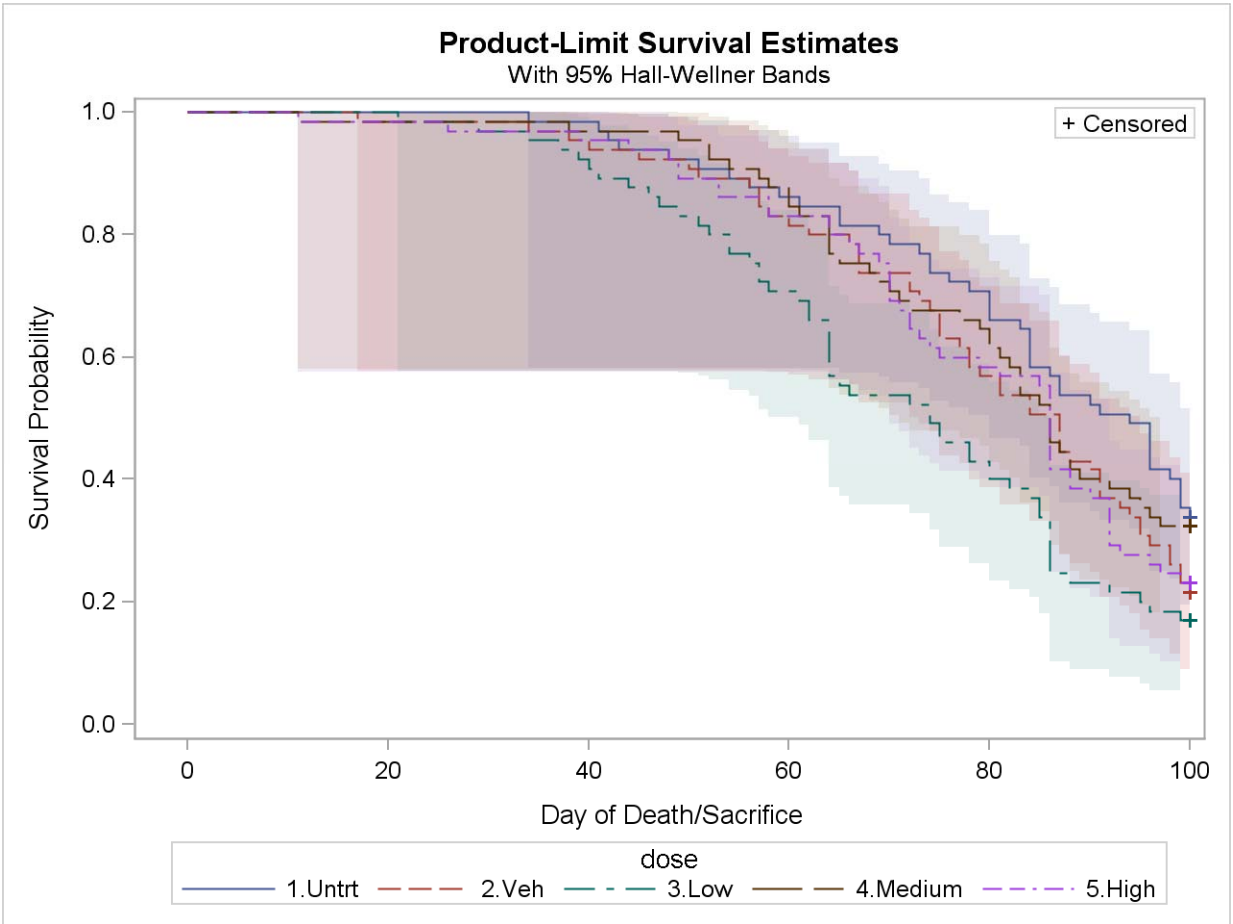
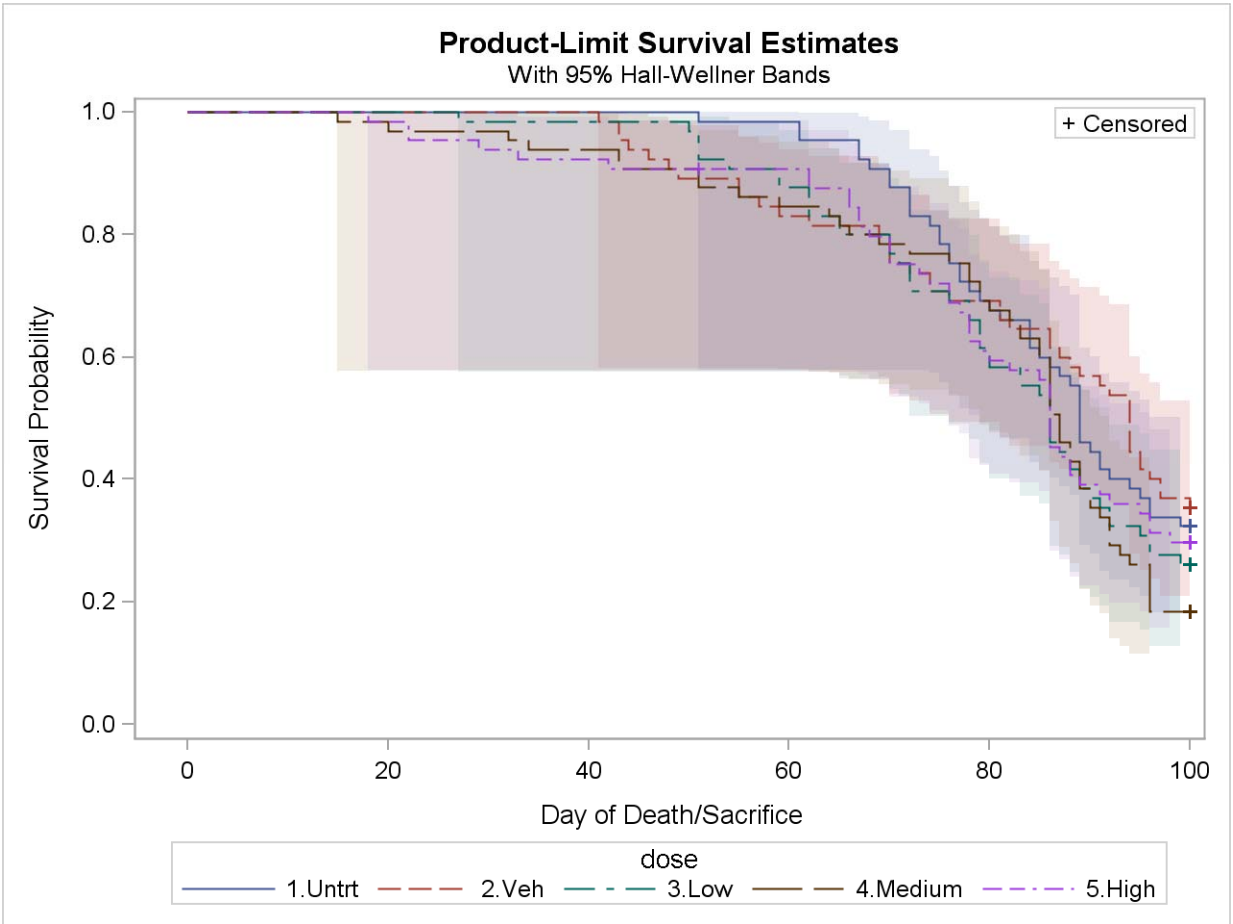


Figure A.1.2 Kaplan-Meier Survival Curves for Female Rats



From Figure A.1.3 below, in male mice there is no strong evidence of heterogeneity in survival over the six treatment groups, including the TPA active control group (Log rank p = 0.3031, Wilcoxon p=0.2688), nor among the five treatment groups deleting the TPA group (Log rank p = 0.1956, Wilcoxon p=0.1791). Although mortality was low in this 26 week study, the high dose group had the highest mortality while the vehicle group had the lowest mortality, with the other groups rather intertwined. This is consistent with the statistically significant trend (Log rank p = 0.0185, Wilcoxon p=0.0179). In female mice there was no strong evidence of heterogeneity or trend in survival among the treatment groups (all p ≥ 0.3076). This also seems consistent with the observation, from Figure A.1.4 below, that the dose groups are generally intertwined.

Figure A.1.3 Kaplan-Meier Survival Curves for Male Mice

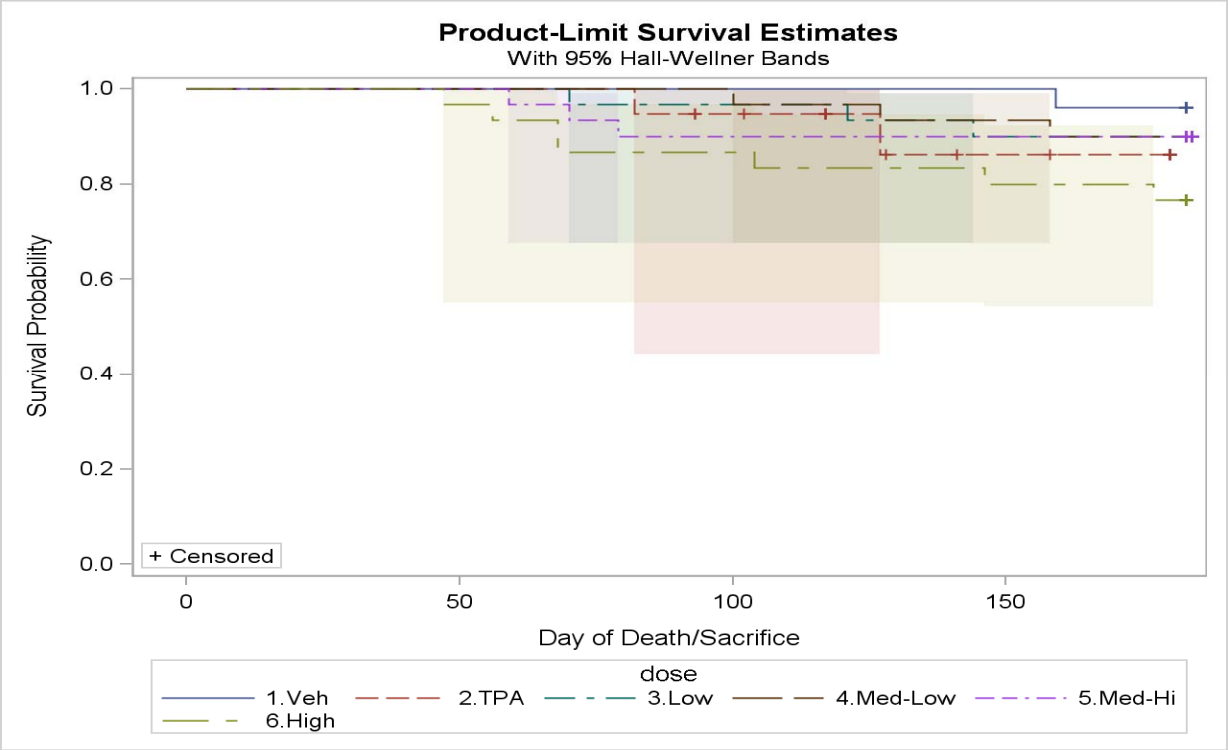
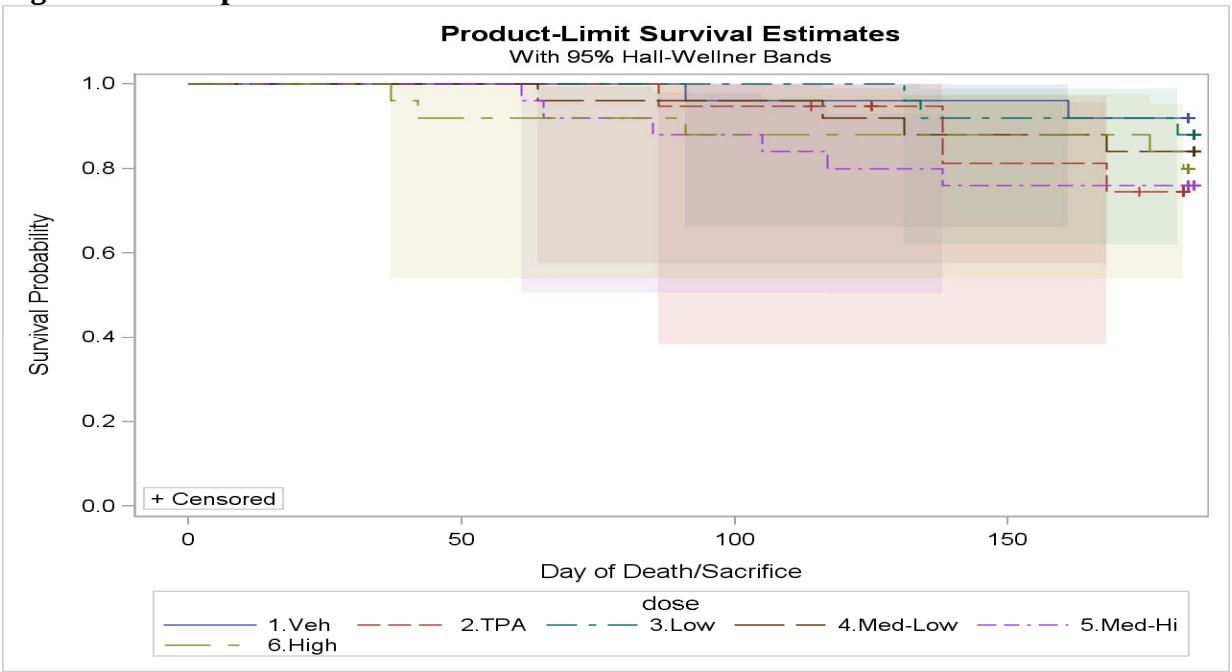


Figure A.1.4 Kaplan-Meier Survival Curves for Female Mice



Appendix 2. FDA Poly-k Tumorigenicity Analysis

Tables A.2.1 through A.2.6 below display, for each organ, the number of animals with one or more of the specified tumor, plus the statistical significance levels of the tests of no trend or no pairwise difference between the specified treatment groups. For Tg.AC mice this should be classified as a secondary analysis, with emphasis placed on the analysis of papillomas as in Appendix 3. Note that incidence data was not included for the TPA active control group in mice, so they are not displayed for that treatment group.

To be specific, in rats, the incidence in the untreated group is listed under the "Un-trt" heading, while the incidences in the vehicle, low (20 mg/kg/day), medium (60 mg/kg/day), and high (200 mg/kg/day) dose groups appear under the associated headings. In mice the incidences are similar, except that there is a medium-high dose group between the medium and high dose groups, plus there is no untreated group. That is, doses of BuTrans are 0, 18.75, 37.5, 150, and 600 mg/kg/day, respectively. In rats the significance levels of the poly-k test of no differences between the vehicle and no treatment group is listed next. For both species the column labeled "trend" denotes the significance level of the test of trend over the vehicle and specified dose groups. The significance levels of the pairwise tests between the vehicle and each of the low, medium, and high dose groups are follow, with an additional comparison to the medium-high group in mice.

The poly-k test, here with k=3, modifies the original Cochran-Armitage test to adjust for differences in mortality (please see Bailer & Portier, 1988, Bieler & Williams, 1993). The tests used here are small sample exact permutation tests of tumor incidence. These do assume all marginal totals are fixed, a debatable assumption. This assumption implies that in the pairwise tests when one dose group has no tumors of the specific type and the other does, there is only one permutation of this pattern. Since that means that the only permutation of the data is the one observed, that means that all possible permutations are as extreme as the pattern observed, and thus the significance level of the observed pattern can be logically expressed as 1.0. One could use the same sort of argument when there were no tumors of the specific type being analyzed in either cell of the 2x2 table corresponding to a pairwise comparison. Then an argument could be made that the p-value for this test should also be 1.0. However, largely for readability, in the tables below these p-values are considered as missing (i.e., corresponding to a null test), denoted by ".". Note that StatXact adjusts for the variance, which would be 0. Then the significance levels of the test statistics are based on the result of a division by 0, i.e., undefined, and hence StatXact codes these p-values as missing.

Up until recently, the Division has usually emphasized so-called Peto carcinogenicity tests, which require accurate specification of cause of death. This is the testing methodology used by the Sponsor. It was noted in the report of the Society of Toxicological Pathology "town hall" meeting in June 2001 that the poly-k modification of the Cochran-Armitage tests of trend has been recommended over such Peto tests.

To adjust for the multiplicity of tests involved in a typical Peto analysis, the so-called Haseman-Lin-Rahman rules discussed in Section 1.3.1.4 are usually applied. That is, when testing for trend over dose and the difference between the highest dose group with a control group, to control the overall Type I error rate to roughly 10% for a standard two species, two sex study, one compares the unadjusted significance level of the trend test to 0.005 for common tumors (incidence>1%) and 0.025 for rare tumors, and the pairwise test to 0.01 for common tumors and 0.05 for rare tumors. However, for a single species study both the test of trend and the test comparing the high dose to the appropriate controls should both be tested at the 0.05 level for rare tumors and 0.01 for common tumors. For transgenic mice, since there are only a few tumor types, with a small number of animals, over a short period of time, even the incidence analysis is generally tested at a 0.05 significance level. That is, there is no correction for multiplicity. In rats the no treatment group is used as the control while in mice the vehicle group is used.

Tables A.2.1 and A.2.2 in rats and mice, respectively, show the tumors that had at least one mortality adjusted test whose nominal statistical significance was at least 0.05. Tables A.2.3 and A.2.4 tediously display all test results for male and female rats, respectively, while Tables A.2.5 and A.2.6 present similar results in male and female mice. The p-values of the poly-k test are based on exact tests from StatXact as discussed above. As also noted above, the period '.' denotes tests of dose groups with no tumors in any group. Of course in mice the analysis of papilloma trajectories presented in Appendix 3 and summarized in the report should be considered as the primary analysis in that species.

In Table A.2.1, in male rats, following the adjustment for multiplicity in a single species study and using the incidence in the no treatment group to decide if a tumor is rare or not, we would conclude, in male rats, that the test of trend in benign interstitial cell tumor was statistically significant ($p = 0.0051 < 0.01$). The corresponding test comparing the high dose to vehicle was close to significance ($p = 0.0127 \approx 0.01$). If we use the vehicle group to determine whether or not a tumor is rare, we would have the same results in male rats, with the additional conclusion that the pairwise test between the high dose group and the medium dose group was statistically significant ($p = 0.0127$ and $0.0332 < 0.05$ respectively). Again using the untreated group to determine rarity, among female rats, the tests of trend in endometrial stromal polyps in the uro-genital system alone and with these tumors pooled with sarcomas were both statistically significant ($p = 0.0017$ and $p = 0.0019$, respectively, both < 0.01), as were the corresponding pairwise tests of the high dose group versus the vehicle ($p = 0.001$ and $p = 0.0005$, respectively, both < 0.01). Again, using the incidence in the no treatment group, the test of trend in benign endometrial stromal polyps of the cervix of female rats was statistically significant ($p = 0.0397 < 0.05$). Strictly speaking, this adjustment for multiplicity applies to tests of trend and comparisons of the high dose to vehicle. Applying the adjustment to tests comparing the low, medium, and untreated dose group to vehicle will be anticonservative.

Despite its nominal statistical significamce, adjusting for multiplicity the test between the high dose and vehicle in pooled benign and malignant pheochromocytoma in the adrenal

medulla of female rats was not statistically significant (p = 0.0379 > 0.01), nor were the pairwise tests of the medium and low dose groups with vehicle (p ≥ 0.0379 and p = 0.0183, respectively, both > 0.01). Also in female rats the test of trend in endometrial stromal polyps in the cervix was not statistically significant (p = 0.0397 > 0.01). Using the incidence in the vehicle group to define the rarity of tumors in female rats, results would the same, except that the pairwise difference between the low dose group and the vehicle groups in endometrial stromal polyps in the uro-genital system alone and with these tumors pooled with sarcomas would both be statistically significant (p = 0.0228 and 0.0103, respectively, both < 0.05). Following the Haseman-Lin-Rahman rules above, no other tests achived statistical significance, though several were close.

Again, the incidence analysis of tumors in Tg.AC mice is a secondary analysis. Those organ tumor combinations that had a potentially statistically significant result (i.e., a significance level at or below 0.05) are displayed in Table 5 below. For papillomas in the stomach, in male mice the pairwise test between the vehicle and the medium high dose group was statistically significant (p = 0.0095), as was the test of trend in the high dose group (p = 0.0142) and the test of pairwise differences between the vehicle and the high dose group (p = 0.0451). Other tests involved other dose groups.

Table A.2.1 Potentially Statistically Significant Neoplasms in Rats

Organ / Tumor	Incidence Un-trt	Veh	Low	Med	Hi	Results of tests Untrt/ Veh	trend	Low/ Veh	Med/ Veh	High/ Veh
Male Rats										
PANCREAS										
B-ADENOMA, ISLET CELL	1	7	0	0	2	0.0209	0.8422	0.9831	0.9941	0.9106
TESTIS										
B-INTERSTITIAL CELL TUMOR	1	0	0	5	6	0.4643	0.0051	.	0.0332	0.0127
Female Rats										
ADRENAL, MEDULLA										
B-PHEOCHROMOCYTOMA	3	2	10	9	7	0.4895	0.2762	0.0119	0.0183	0.0633
Pheochromocytoma B+M	3	2	10	9	8	0.4895	0.1954	0.0119	0.0183	0.0379
CERVIX										
B-POLYP, ENDOMETRIAL STROMAL	0	0	1	1	3	.	0.0397	0.4824	0.4762	0.1079
UTERUS										
B-POLYP, ENDOMETRIAL,STROMAL	2	0	4	0	4	0.7416	0.0793	0.0500	.	0.0474
Uro-Genital										
Endo. Stromal Sarcoma/Polyp	3	0	6	2	10	0.8708	0.0019	0.0103	0.2297	0.0005
Endometrial Stromal Polyp	3	0	5	1	9	0.8708	0.0017	0.0228	0.4762	0.0010

Again, since there are usually relatively few different types of tumors that develop in the short study period in genetically modified rodents, typically no adjustment for the multiplicity of tests is applied. In female mice there is a clear trend in papillomas of the stomach (p = 0.0142). The pairwise differences between the high dose and vehicle was also statistically significant (p = 0.0451). Perhaps partially due to the considerably higher mortality in the high dose group in males, the test of trend is very nonsignificant (p = 0.9998). However, the pairwise tests between the vehicle and the low, medium, and medium high dose groups are all statistically significant (p=0.0260, 0.0144, and 0.0095, respectively). In female mice, the

pairwise difference between the high dose and vehicle was also statistically significant (p=0.0451).

Table A.2.2 Potentially Statistically Significant Neoplasms in Mice

Organ / Tumor	Incidence Veh	Low	Med	Med Hi	Hi	Results of tests trend	Low/ Veh	Med/ Veh	Medhi/ Veh	High/ Veh
Male Mice										
stomach										
papilloma	2	8	12	9	0	0.9998	0.0260	0.0144	0.0095	0.8167
Female Mice										
stomach										
papilloma	0	2	0	1	4	0.0142	0.3033	.	0.4500	0.0451

Tables A.2.3 and A.2.4 display all test results for male and female rats, respectively, while tables A.2.5 and A.2.6 present similar results in male and female mice.

Table A.2.3 Incidence and Statistical Tests of Carcinogenicity in Male Rats

Organ / Tumor	Incidence Un-trt	Veh	Low	Med	Hi	Results of tests Untrt/ Veh	trend	Low/ Veh	Med/ Veh	High/ Veh
ADIPOSE TISSUE										
M-LIPOSARCOMA	0	0	0	0	1	.	0.2568	.	.	0.4935
ADRENAL, CORTEX										
B-ADENOMA	0	1	1	0	0	0.4643	0.8231	0.6932	0.5125	0.4935
ADRENAL, MEDULLA										
B-PHEOCHROMOCYTOMA	4	6	4	8	3	0.3256	0.8243	0.4478	0.3853	0.7203
M-MALIGNANT PHEOCHROMOCYTOMA	2	0	0	0	0	0.7160	.	.	.	.
Pheochromocytoma B+M	4	6	4	8	3	0.3256	0.8243	0.4478	0.3853	0.7203
AUDITORY SEB GL										
B-ADENOMA, SEBACEOUS-SQUAMOU	0	0	0	0	2	.	0.0672	.	.	0.2468
M-CARCINOMA, SEBACEOUS-SQUAM	0	1	1	0	0	0.4706	0.8204	0.6770	0.5062	0.4872
BRAIN										
M-ASTROCYTOMA	4	0	1	2	0	0.9126	0.6107	0.4348	0.2657	.
M-MENINGEAL SARCOMA	0	0	0	1	0	.	0.2568	.	0.5125	.
CAVITY, ORAL										
B-ODONTOMA	1	0	0	0	0	0.4588	.	.	.	.
CAVITY, THORACIC										
B-HIBERNOMA	0	1	0	0	0	0.4706	0.7315	0.4286	0.5062	0.4872
M-LIPOSARCOMA	0	0	1	2	0	.	0.6074	0.4429	0.2657	.
EPIDIDYMIS										
B-MESOTHELIOMA	1	0	0	0	1	0.4643	0.2617	.	.	0.5000
HARDERIAN GLAND										
B-ADENOMA	1	1	0	1	0	0.7227	0.6322	0.4286	0.2531	0.4872
HEMATO NEOPLASIA										
M-LYMPHOMA	0	0	0	1	0	.	0.2568	.	0.5125	
M-SARCOMA, HISTIOCYTIC	1	2	3	1	0	0.4554	0.9515	0.3790	0.5094	0.7403
JOINT, OTHER										
M-HEMANGIOSARCOMA	0	1	0	0	0	0.4706	0.7315	0.4286	0.5062	0.4872
KIDNEY										
M-CARCINOMA, TUBULAR CELL	1	0	0	0	0	0.4643	.	.	.	.
M-LIPOSARCOMA	1	0	0	1	0	0.4643	0.2568	.	0.5125	.
LIVER										
B-ADENOMA, HEPATOCELLULAR	5	1	0	1	1	0.8563	0.4239	0.4348	0.2595	0.7468
M-CARCINOMA, HEPATOCELLULAR	2	1	0	0	0	0.4554	0.7315	0.4286	0.5062	0.4872
LN, MESENTERIC										
B-HEMANGIOMA	0	2	1	0	0	0.2244	0.9359	0.3950	0.7531	0.7339
B-LYMPHANGIOMA	0	0	0	0	1	.	0.2617	.	.	0.5000

Table A.2.3 (cont.) Incidence and Statistical Tests of Carcinogenicity in Male Rats

Organ / Tumor	Incidence Un-trt	Veh	Low	Med	Hi	Results of tests Untrt/ Veh	trend	Low/ Veh	Med/ Veh	High/ Veh
LN, TRACHEOBRON										
M-LIPOSARCOMA	0	0	1	0	0	.	0.5302	0.4429	.	.
MAMMARY, MALE										
B-ADENOMA	1	1	0	0	0	0.7160	0.7365	0.4348	0.5125	0.4935
B-FIBROADENOMA	1	1	0	0	0	0.7160	0.7365	0.4348	0.5125	0.4935
MUSCLE, SKELETAL										
M-FIBROSARCOMA	1	0	0	0	0	0.4643	.	.	.	.
PANCREAS										
B-ADENOMA, ISLET CELL	1	7	0	0	2	0.0209	0.8422	0.9831	0.9941	0.9106
M-CARCINOMA, ISLET CELL	4	1	0	0	0	0.7722	0.7365	0.4348	0.5125	0.4935
M-HEMANGIOSARCOMA	0	1	0	0	0	0.4643	0.7365	0.4348	0.5125	0.4935
PARATHYROID										
B-ADENOMA	1	0	1	0	0	0.4643	0.5338	0.4348	.	.
PITUITARY										
B-ADENOMA	40	34	24	23	29	0.7284	0.6132	0.6861	0.9086	0.6693
SKIN										
B-FIBROMA	0	0	0	1	0	.	0.2568	.	0.5125	.
B-FIBROMA	0	1	0	0	0	0.4706	0.7315	0.4286	0.5062	0.4872
B-KERATOACANTHOMA	2	2	0	0	0	0.6360	0.9319	0.6841	0.7655	0.7468
B-MYXOMA	0	0	0	0	1	.	0.2568	.	.	0.4935
B-PAPILLOMA, SQUAMOUS CELL	1	0	0	2	0	0.4643	0.4463	.	0.2657	.
M-CARCINOMA, BASAL CELL	1	0	0	0	0	0.4643	.	.	.	.
M-CARCINOMA, SQUAMOUS CELL	1	2	0	0	0	0.4554	0.9293	0.6770	0.7593	0.7403
M-FIBROSARCOMA	2	1	1	1	0	0.4471	0.7461	0.6861	0.2531	0.4872
M-HEMANGIOSARCOMA	0	0	0	1	0	.	0.2568	.	0.5125	.
M-MYXOSARCOMA	1	0	0	0	0	0.4643	.	.	.	.
SKIN, TREATED										
B-KERATOACANTHOMA	0	1	0	0	0	0.4706	0.7315	0.4286	0.5062	0.4872
SPINAL CORD										
M-ASTROCYTOMA	0	0	1	0	0	.	0.5302	0.4429	.	.
SPLEEN										
B-HEMANGIOMA	0	0	0	0	1	.	0.2568	.	.	0.4935
M-HEMANGIOSARCOMA	1	0	0	0	0	0.4643	.	.	.	.
STOMACH, NONGL										
B-PAPILLOMA, SQUAMOUS CELL	0	0	0	0	1	.	0.2568	.	.	0.4935
Systemic										
Hemangioma	0	2	1	0	1	0.2244	0.6296	0.3950	0.7531	0.4712
Hemangioma/-sarcoma	1	4	1	1	1	0.1524	0.8456	0.7221	0.8204	0.7960
Hemangiosarcoma	1	2	0	1	0	0.4554	0.8395	0.6770	0.5094	0.7403
TESTIS										
B-INTERSTITIAL CELL TUMOR	1	0	0	5	6	0.4643	0.0051	.	0.0332	0.0127
THYMUS										
B-THYMOMA	1	0	1	1	0	0.4643	0.5207	0.4429	0.5125	.
THYROID										
B-ADENOMA, C-CELL	2	3	0	9	5	0.4436	0.1772	0.8195	0.0692	0.3559
B-ADENOMA, FOLLICULAR CELL	2	2	0	1	2	0.6461	0.3251	0.6770	0.5094	0.6827

Table A.2.4 Incidence and Statistical Tests of Carcinogenicity in Female Rats

Organ / Tumor	Incidence Un-trt	Veh	Low	Med	Hi	Results of tests Untrt/ Veh	trend	Low/ Veh	Med/ Veh	High/ Veh
ADIPOSE TISSUE										
B-LIPOMA	1	0	0	0	0	0.4889	.	.	.	.
ADRENAL, CORTEX										
B-ADENOMA	0	2	1	0	1	0.2472	0.6130	0.4554	0.7227	0.4647
M-CARCINOMA	0	1	1	0	0	0.5000	0.7938	0.7227	0.4706	0.4706
ADRENAL, MEDULLA										
B-PHEOCHROMOCYTOMA	3	2	10	9	7	0.4895	0.2762	0.0119	0.0183	0.0633
M-MALIGNANT PHEOCHROMOCYTOMA	0	0	0	0	1	.	0.2485	.	.	0.4824
Pheochromocytoma B+M	3	2	10	9	8	0.4895	0.1954	0.0119	0.0183	0.0379
AUDITORY SEB GL										
B-KERATOACANTHOMA	0	0	1	0	0	.	0.4878	0.4762	.	.
BRAIN										
M-ASTROCYTOMA	1	2	2	1	2	0.4917	0.4730	0.6461	0.4554	0.6559
CAVITY, THORACIC										
B-HIBERNOMA	0	1	0	0	0	0.5000	0.7273	0.4706	0.4706	0.4706
M-LIPOSARCOMA	0	0	2	1	3	.	0.0754	0.2356	0.4824	0.1079
CERVIX										
B-GRANULAR CELL TUMOR	0	1	0	0	0	0.4944	0.7317	0.4762	0.4762	0.4762
B-POLYP, ENDOMETRIAL STROMAL	0	0	1	1	3	.	0.0397	0.4824	0.4762	0.1079
CLITORAL GLAND										
B-KERATOACANTHOMA	2	0	0	0	0	0.7472	.	.	.	.
B-SQUAMOUS PAPILLOMA	0	1	0	0	0	0.5000	0.7273	0.4706	0.4706	0.4706
HARDERIAN GLAND										
B-ADENOMA	1	0	0	0	0	0.4944	.	.	.	.
HEART										
M-ENDOCARDIAL SCHWANNOMA	0	0	0	1	0	.	0.2439	.	0.4762	.
HEMATO NEOPLASIA										
M-LYMPHOMA	1	1	0	2	0	0.7416	0.6269	0.4762	0.4732	0.4762
M-SARCOMA, HISTIOCYTIC	1	2	1	1	3	0.4917	0.1948	0.4647	0.4647	0.4669
KIDNEY										
B-ADENOMA, NOS	0	0	1	0	0	.	0.4878	0.4762	.	.
B-ADENOMA, TUBULAR CELL	0	1	1	0	0	0.4944	0.7967	0.7350	0.4762	0.4762
M-NEPHROBLASTOMA	0	1	0	0	0	0.5000	0.7273	0.4706	0.4706	0.4706
LIVER										
B-ADENOMA, HEPATOCELLULAR	0	1	1	0	0	0.4944	0.7975	0.7286	0.4762	0.4762
MAMMARY, FEMALE										
B-ADENOMA	1	4	4	6	3	0.1804	0.6301	0.5895	0.3256	0.4474
B-FIBROADENOMA	32	25	17	24	13	0.7735	0.9599	0.8602	0.5392	0.9650
M-CARCINOMA	14	16	19	24	18	0.3889	0.3760	0.2572	0.0645	0.3028
M-HEMANGIOSARCOMA	0	0	0	0	1	.	0.2485	.	.	0.4824
M-SARCOMA	0	1	0	0	0	0.5000	0.7273	0.4706	0.4706	0.4706
MUSCLE, SKELETAL										
M-SARCOMA, NOS	0	0	0	1	0	.	0.2439	.	0.4762	.
OVARY										
B-GRANULOSA/THECA CELL TUMOR	0	0	0	0	1	.	0.2485	.	.	0.4824
PANCREAS										
B-ADENOMA, ISLET CELL	2	0	0	0	0	0.7416	.	.	.	.
PARATHYROID										
B-ADENOMA	0	1	0	0	0	0.4944	0.7317	0.4762	0.4762	0.4762
PINNA										
M-FIBROSARCOMA	0	0	0	1	0	.	0.2439	.	0.4762	.
PITUITARY										
B-ADENOMA	58	58	49	46	47	0.2703	0.8718	0.8717	0.9578	0.9270
SKIN										
B-KERATOACANTHOMA	0	1	0	0	0	0.5000	0.7273	0.4706	0.4706	0.4706
B-PAPILLOMA, SQUAMOUS CELL	1	0	0	0	1	0.4889	0.2439	.	.	0.4762
M-FIBROSARCOMA	0	0	0	0	1	.	0.2485	.	.	0.4824

Table A.2.4 (cont.) Incidence and Statistical Tests of Carcinogenicity in Female Rats

Organ / Tumor	Incidence Un-trt	Veh	Low	Med	Hi	Results of tests Untrt/ Veh	trend	Low/ Veh	Med/ Veh	High/ Veh
SKIN, TREATED										
B-KERATOACANTHOMA	0	1	0	0	0	0.5000	0.7273	0.4706	0.4706	0.4706
M-CARCINOMA, NOS	0	0	0	0	1	.	0.2439	.	.	0.4762
Systemic										
Hemangioma/-sarcoma	0	1	0	0	1	0.5000	0.4341	0.4706	0.4706	0.7291
THYMUS										
B-THYMOMA	0	1	1	0	0	0.5000	0.7930	0.7291	0.4706	0.4706
THYROID										
B-ADENOMA, C-CELL	9	9	6	7	8	0.5643	0.4109	0.6074	0.4883	0.3974
B-ADENOMA, FOLLICULAR CELL	2	0	0	0	0	0.7416	.	.	.	.
M-CARCINOMA, C-CELL	0	0	2	0	0	.	0.7361	0.2297	.	.
UTERUS										
B-HEMANGIOMA	0	1	0	0	0	0.5000	0.7273	0.4706	0.4706	0.4706
B-LEIOMYOMA	0	0	1	0	0	.	0.4848	0.4824	.	.
B-POLYP, ENDOMETRIAL, STROMA	2	0	4	0	4	0.7416	0.0793	0.0500	.	0.0474
M-SARCOMA, ENDOMETRIAL, STRO	0	0	1	1	1	.	0.2856	0.4762	0.4762	0.4824
Uro-Genital										
Endo0. Stromal Sarcoma/Polyp	3	0	6	2	10	0.8708	0.0019	0.0103	0.2297	0.0005
Endometrial Stromal Polyp	3	0	5	1	9	0.8708	0.0017	0.0228	0.4762	0.0010
VAGINA										
B-ENDOMETRIAL STROMAL POLYP	1	0	0	0	2	0.4889	0.0606	.	.	0.2297
M-FIBROSARCOMA	0	0	0	0	1	.	0.2485	.	.	0.4824

Table A.2.5 Incidence and Poly-k Statistical Tests of Carcinogenicity in Male Mice

Organ / Tumor	Incidence Veh	Low	Med	Med Hi	Hi	Results of tests trend	Low/ Veh	Med/ Veh	Medhi/ Veh	High/ Veh
eyelid										
papilloma	0	0	1	0	0	0.3881	.	0.6207	.	.
lip										
papilloma	1	1	0	1	0	0.7306	0.2609	0.6207	0.2609	0.5600
lungs with bronchi										
bronchiolo-alveolar adenoma	1	3	0	1	0	0.8992	0.3292	0.6207	0.2826	0.5600
multicentric										
leukemia	0	1	0	0	0	0.6471	0.5417	.	.	.
lymphoma	0	0	1	0	1	0.2311	.	0.6207	.	0.5600
perineum										
papilloma	3	4	6	6	5	0.3989	0.5555	0.5335	0.2468	0.4958
pituitary gland										
lymphoma	1	0	0	0	0	0.8358	0.5217	0.6207	0.5217	0.5600
salivary glands										
cystadenocarcinoma	1	0	0	0	0	0.8358	0.5217	0.6207	0.5217	0.5600
leiomyosarcoma	0	0	0	0	1	0.2090	.	.	.	0.5600
squamous cell carcinoma	1	0	0	0	0	0.8358	0.5217	0.6207	0.5217	0.5600
skin - non-soa										
papilloma	0	0	1	0	1	0.2311	.	0.6207	.	0.5600
skin - soa										
fibrosarcoma	1	0	0	0	0	0.8358	0.5217	0.6207	0.5217	0.5600
skin-soa										
papilloma	1	2	2	0	3	0.1961	0.5342	0.6839	0.5217	0.3957
stomach										
papilloma	2	8	12	9	0	0.9998	0.0260	0.0144	0.0095	0.8167
teeth										
chondroma	0	0	0	0	1	0.2090	.	.	.	0.5600
odontogenic tumor	3	4	9	8	5	0.5301	0.5555	0.2824	0.1021	0.4958

Table A.2.6 Incidence and Poly-k Statistical Tests of Carcinogenicity in Female Mice

Organ / Tumor	Incidence Veh	Low	Med	Med Hi	Hi	trend	Results of tests Low/ Veh	Med/ Veh	Medhi/ Veh	High/ Veh
ears										
papilloma	0	0	0	1	0	0.2200	.	.	0.4500	.
hindlimb										
papilloma	0	1	0	0	0	0.5000	0.5600	.	.	.
lungs with bronchi										
bronchiolo-alveolar adenoma	2	1	0	1	1	0.5423	0.5935	0.5417	0.4211	0.5000
multicentric										
leukemia	2	1	1	1	3	0.1487	0.5600	0.7304	0.4286	0.4551
lymphoma	0	0	2	0	0	0.6353	.	0.1103	.	.
ovaries										
hemangioma	1	0	0	0	0	0.7800	0.5600	0.3125	0.4500	0.5000
teratoma	0	0	1	0	0	0.4000	.	0.3125	.	.
perineum										
papilloma	0	2	1	1	1	0.4033	0.3033	0.3125	0.4500	0.5000
salivary glands										
papilloma	0	0	0	0	1	0.2200	.	.	.	0.5000
squamous cell carcinoma	2	0	0	1	0	0.7676	0.8167	0.5417	0.4211	0.7619
skin - non-soa										
papilloma	1	0	0	0	0	0.7800	0.5600	0.3125	0.4500	0.5000
skin - soa										
papilloma	2	2	0	0	0	0.9632	0.3957	0.5417	0.7105	0.7619
spleen										
leukemia	0	0	0	0	1	0.2200	.	.	.	0.5000
stomach										
papilloma	0	2	0	1	4	0.0142	0.3033	.	0.4500	0.0451
teeth										
chondroma	0	0	0	0	1	0.2200	.	.	.	0.5000
odontogenic tumor	6	8	3	9	3	0.9512	0.6075	0.3733	0.0937	0.8065
squamous cell carcinoma	0	0	1	0	0	0.4000	.	0.3125	.	.

Appendix 3. Analysis of Papillomas in Tg.AC Mice

Although apparently there are problems with this assay (e.g., please see Jacobsen-Kram et al, 2004), counts of papillomas in response to dermal administration of a test compound in Tg.AC genetically modified mice has been recommended as a feasible assessment of carcinogenicity. This particular study terminated during week 27. For each week in the study, Tables A.3.4-A.3.11 summarize treatment outcomes for each treatment group: the total number of animals alive during the specified week, the number of papillomas at various sites, the total number of tumors, and, for site of application papillomas, the mean number of tumors over all animals.

Dunson, et al (2000) proposed a model for the analysis of papilloma counts for Tg.AC mice. Since tumors can be absorbed, they suggest that dealing with the maximum papilloma counts over the course of the study at the relevant sites would be an appropriate endpoint. That is, for C_{ij} denoting the observed papilloma count at the specified site at the jth week for the ith animal. Then define $M_{ij} = \max\{C_{i1}, ..., C_{ij-1}\}$. The actual endpoint used in the analysis is defined as $Y_{ij} = M_{ij+1} - M_{ij}$, i.e., the increase in maximum papilloma count in animal i at week j+1. Observe that Y_{ij} will only be nonzero in those weeks when there is an actual increase in maximum papilloma count in the following week. As the values of the Y_{ij} increase, we could expect the proportion of values to decrease exponentially. Then, as suggested by Dunson, et al, this increase should be representable by a Poisson random variable, i.e., $Y_{ij} \sim \text{Poisson}(\mu_{ij})$. They propose to parameterize this mean as follows:

where $\mu_{ij} = \begin{cases} \exp(\beta_1 + (b_i + \gamma_1)t_j d_i) & if\ M_{ij} = 0 \\ \exp(\beta_2 + \gamma_2 d_i) & if\ M_{ij} > 0 \end{cases}$

with $b_i \sim$ Normal(0,σ^2), a random subject effect, t_j = j/26, and d_i dose level on a log scale (ignoring the TPA positive control group).

Then γ_1 represents the effect of increasing dose and time on the initial onset of papillomas, while b_i denotes the deviation of the ith subject on this initial deviation. In Dunson's model this is the only place where an individual effect is specified. Further, γ_2 defines the effect of increasing dose on the probability of an increase on maximum papilloma count. The beta parameters define the baseline effects on papilloma onset and increase. So interest is primarily focused on γ_1 and γ_2. Observe that $\exp(\beta_2)$ represents the spontaneous rate of increase in papilloma count in the control group while $1 - \exp(-\exp(\beta_1))$ denotes the probability of developing the first tumor in the control. One problem with this model is that it assumes that even within an individual, an increase in papilloma count at one point in time is stochastically independent of having a later increase, i.e., after tumor onset all animals within a dose group are assumed to be equally sensitive to dose. Further, within the control group there is no difference within and between animals in sensitivity to tumor onset and no differences in animals on tumor increase. These would seem to be questionable assumptions.

The usual frequentist approach to estimating the parameters in the Dunson model is clearly maximum likelihood. In general this model should be estimable, however note from Tables A.3.4-A.3.7 below that few animals have a first tumor. That suggests that there may be identification problems when estimating the parameters of the random effect b_i. That seems to be the case for this data set. For these estimability reasons the model that is actually fit is a restricted Dunson model with $b_i = 0$. The maximum likelihood estimates of the remaining parameters of the Dunson model are given in tables A.3.1 and A.3.2 below:

Table A.3.1 Estimated Parameters in Restricted Dunson Model for Increase in Maximum Count of Active Papillomas

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Lower	Upper
Male Mice						
beta1	-5.1896	0.3748	-13.85	<.0001	-5.9245	-4.4548
beta2	-4.1229	1.3859	-2.97	0.0030	-6.8401	-1.4057
gam1	-0.5827	0.4098	-1.42	0.1551	-1.3862	0.2207
gam2	-0.7965	0.9595	-0.83	0.4065	-2.6778	1.0847
	F Value	Pr > F				
gam1=gam2=0	1.36	0.2579				
Female Mice						
beta1	-6.0895	0.5878	-10.36	<0.0001	-7.2419	-4.9370
beta2	-2.6883	1.1695	-2.30	0.0216	-4.9813	-0.3953
gam1	-0.3849	0.5699	-0.68	0.4995	-1.5023	0.7326
gam2	-0.3760	0.6064	-0.62	0.5352	-1.5649	0.8129
	F Value	Pr > F				
gam1=gam2=0	0.42	0.6569				

The lower and upper limits of an asymptotic 95% confidence interval appear in the labeled columns.

Table A.3.2 Estimated Parameters in Restricted Dunson for Increase in Maximum Count of Active+ Latent Papillomas

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Lower	Upper
Male Mice						
beta1	-4.9786	0.3622	-13.74	<0.0001	-5.6887	-4.2684
beta2	-3.1229	0.8477	-3.68	0.0002	-4.7849	-1.4608
gam1	-0.9900	0.5171	-1.91	0.0556	-2.0039	0.02387
gam2	-1.3931	0.9350	-1.49	0.1363	-3.2264	0.4401
	F Value	Pr > F				
gam1=gam2=0	2.94	0.0529				
Female Mice						
beta1	-5.4983	0.4551	-12.08	<0.0001	-6.3906	-4.6060
beta2	-3.0408	0.8887	-3.42	0.0006	-4.7833	-1.2984
gam1	-0.5601	0.5009	-1.12	0.2635	-1.5421	0.4220
gam2	-0.3212	0.5382	-0.60	0.5507	-1.3766	0.7341
	F Value	Pr > F				
gam1=gam2=0	0.80	0.4479				

Again the primary interest is focused on the gamma parameters. The restricted Dunson model for active papillomas alone seems to show no evidence of a dose effect on onset (both p

≥ 0.1551) or increase (both p ≥ 0.4065). However, when dealing with pooled active and latent papillomas in male mice there is some evidence of an effect on onset (p = 0.0556), but not in female mice (p = 0.2635).

This reviewer prefers a model that allows within animal correlation in both papilloma onset and papilloma increase. One way to represent this a very similar model to that above, where $Y_{ij} \sim$ Poisson(μ_{ij}), with $\mu_{ij} = \begin{cases} \exp(\beta_1 + b_i + \gamma_1 t_j d_i) & if\ M_{ij} = 0 \\ \exp(\beta_2 + b_i + \gamma_2 d_i) & if\ M_{ij} > 0 \end{cases}$

Note that while this allows an individual animal effect, this animal effect does not differentiate between the effect on onset and the effect on papilloma increase. While it might be feasible in some circumstances to model both effects, for this particular data such a specification seems to lead to identification problems. Whether this is intrinsic to the model or dependent on the fact there are relatively few papillomas is a topic for further research.

Table A.3.4 Estimated Parameters in Alternative Model for the Increase in Maximum Counts of Papillomas

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Lower	Upper
Active Papillomas Only						
Males						
sigma	3.9211	4.1161	0.95	0.3424	-4.2147	12.0569
beta1	-6.6695	1.5006	-4.44	<.0001	-9.6356	-3.7034
beta2	-7.5003	2.8462	-2.64	0.0093	-13.1260	-1.8747
gam1	-0.4624	0.4960	-0.93	0.3528	-1.4427	0.5180
gam2	-1.0345	1.2940	-0.80	0.4253	-3.5921	1.5231
	F Value	Pr > F				
gam1=gam2=0	0.75	0.4761				
Females						
sigma	0.1429	3.1972	0.04	0.9644	-6.1853	6.4710
beta1	-6.1616	1.7188	-3.58	0.0005	-9.5635	-2.7597
beta2	-2.8458	3.6903	-0.77	0.4421	-10.1500	4.4583
gam1	-0.3826	0.5732	-0.67	0.5057	-1.5172	0.7520
gam2	-0.3856	0.6624	-0.58	0.5615	-1.6967	0.9254
	F Value	Pr > F				
gam1=gam2=0	0.40	0.6685				
Active + Latent Papillomas						
Male Mice						
sigma	5.6958	4.2660	1.34	0.1839	-2.7363	14.1280
beta1	-6.9095	1.4600	-4.73	<.0001	-9.7953	-4.0237
beta2	-7.2682	2.3827	-3.05	0.0027	-11.9777	-2.5587
gam1	-0.8638	0.6139	-1.41	0.1616	-2.0772	0.3496
gam2	-1.8214	1.3848	-1.32	0.1905	-4.5585	0.9157
	F Value	Pr > F				
gam1=gam2=0	1.82	0.1663				
Female Mice						

Model does not seem to be identified for female mice. Please see results for restricted Dunson model in Table A.3.2 above.

Absence of proof is not proof of absence, but the small asymptotic t-statistic for the variance of the individual female mice dose seems to suggest there is no evidence of an individual animal effect, i.e., $b_i = 0$. For male mice results are more ambiguous. While the tests of the hypothesis that the variance of the individual effect is 0 are not rejected (i.e., the hypothesis that $b_i = 0$ is not rejected) the t-statistics are not particularly small. In this model the evidence of an onset effect in pooled active and latent papillomas is not as close to being statistically significant (p = 0.1616). Other results seem to be consistent with those of the restricted Dunson model above.

The first four following tables A.3.4-A.3.7 display the number of animals still alive during the specified week ("n"), the total number of site of application ("SOA") tumors over all animals ("# tumors"), the number of all animals with SOA papillomas ("#anml w/tmr"), and finally the rounded mean number of tumors ("Tumor Mean").

Table A.3.4 Overall Trajectory of SOA Active Papillomas in Male Mice

Treat Group	Week 1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,	13,	14,	15,
Sex=Males															
Veh n	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25
# tumors	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
#anml w/tmr	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TPA n	20	20	20	20	20	20	20	20	20	20	20	20	19	19	17
# tumors	0	0	0	0	0	0	1	1	4	8	17	24	54	122	104
#anml w/tmr	0	0	0	0	0	0	1	1	4	6	6	8	11	14	13
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.4	0.9	1.2	2.8	6.4	6.1
Low n	30	30	30	30	30	30	30	30	30	30	29	29	29	29	29
# tumors	0	0	0	0	0	1	1	2	2	2	2	2	2	2	2
#anml w/tmr	0	0	0	0	0	1	1	2	2	2	2	2	2	2	2
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Med n	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30
# tumors	0	0	0	0	1	1	1	2	3	4	4	4	4	4	4
#anml w/tmr	0	0	0	0	1	1	1	2	3	4	4	4	4	4	4
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
MedHi n	30	30	30	30	30	30	30	30	30	29	28	28	27	27	27
# tumors	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
#anml w/tmr	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	30	30	30	30	30	30	30	29	28	28	26	26	26	26	26
# tumors	0	0	0	0	0	1	2	2	2	2	2	2	2	3	3
#anml w/tmr	0	0	0	0	0	1	2	2	2	2	2	2	2	3	3
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table A.3.4 (cont.) Overall Trajectory of SOA Active Papillomas in Male Mice

Treat Group	Week 16,	17,	18,	19,	20,	21,	22,	23,	24,	25,	26,	27,
Sex=Males												
Veh n	25	25	25	25	25	25	25	25	24	24	24	24
# tumors	1	1	1	1	1	1	2	2	2	2	2	2
#anml w/tmr	1	1	1	1	1	1	2	2	2	2	2	2
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1
TPA n	16	16	12	12	10	9	9	9	8	8	8	7
# tumors	109	168	96	111	68	65	65	70	58	61	57	54
#anml w/tmr	12	12	8	8	7	6	6	6	5	5	5	4
Tumor Mean	6.8	10.5	8.0	9.3	6.8	7.2	7.2	7.8	7.3	7.6	7.1	7.7
Low n	29	29	28	28	28	28	27	27	27	27	27	27
# tumors	2	2	2	2	2	2	2	2	2	3	3	3
#anml w/tmr	2	2	2	2	2	2	2	2	2	2	2	2
Tumor Mean	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Med n	29	29	29	28	28	28	28	28	27	27	27	27
# tumors	4	4	4	4	4	4	4	4	3	3	3	3
#anml w/tmr	4	4	4	4	4	4	4	4	3	3	3	3
Tumor Mean	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
MedHi n	27	27	27	27	27	27	27	27	27	27	27	27
# tumors	1	1	1	1	1	1	1	1	1	1	1	1
#anml w/tmr	1	1	1	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	26	26	26	26	26	26	25	25	25	25	25	25
# tumors	3	3	3	3	3	3	3	3	3	3	3	3
#anml w/tmr	3	3	3	3	3	3	3	3	3	3	3	3
Tumor Mean	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table A.3.5 Overall Trajectory of SOA Active Papillomas in Female Mice

Treat Group	Week 1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,	13,	14,	15,
Sex=Females															
Vehicle n	25	25	25	25	25	25	25	25	25	25	25	25	24	24	24
# tumors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TPA n	20	20	20	20	20	20	20	20	20	20	20	20	20	19	19
# tumors	0	0	0	0	0	0	0	0	0	4	9	13	30	42	73
#anml w/tmr	0	0	0	0	0	0	0	0	0	2	6	8	13	13	14
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.5	0.7	1.5	2.2	3.8
Low n	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25
# tumors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Med n	25	25	25	25	25	25	25	25	25	24	24	24	24	24	24
# tumors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MedHi n	25	25	25	25	25	25	25	25	25	24	23	23	22	22	21
# tumors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	25	25	25	25	25	25	24	23	23	23	23	23	22	22	22
# tumors	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table A.3.5 (cont.) Overall Trajectory of SOA Active Papillomas in Female Mice

Treat Group	Week 16,	17,	18,	19,	20,	21,	22,	23,	24,	25,	26,	27,
Sex=Females												
Veh n	24	24	24	24	24	24	24	24	23	23	23	23
# tumors	0	0	0	1	1	1	1	1	1	2	2	2
#anml w/tmr	0	0	0	1	1	1	1	1	1	2	2	2
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1
TPA n	19	19	18	15	14	12	12	12	12	11	11	9
# tumors	106	134	135	75	88	51	70	85	105	94	100	79
#anml w/tmr	14	14	14	12	11	9	9	9	9	9	9	7
Tumor Mean	5.6	7.1	7.5	5.0	6.3	4.3	5.8	7.1	8.8	8.5	9.1	8.8
Low n	25	25	25	25	24	23	23	23	23	23	23	22
# tumors	0	0	1	1	2	0	0	0	0	0	0	1
#anml w/tmr	0	0	1	1	1	0	0	0	0	0	0	1
Tumor Mean	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Med n	24	24	23	23	22	22	22	22	22	22	21	21
# tumors	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MedHi n	21	21	20	20	20	19	19	19	19	19	19	19
# tumors	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	22	22	22	22	22	22	22	22	22	22	21	20
# tumors	1	1	1	1	1	1	1	1	1	1	1	1
#anml w/tmr	1	1	1	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1

Table A.3.6 Overall Trajectory of SOA Latent + Active Papillomas in Male Mice

Treat Group	Week 1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,	13,	14,	15,
Veh n	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25
# tumors	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
#anml w/tmr	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TPA n	20	20	20	20	20	20	20	20	20	20	20	20	19	19	17
# tumors	0	0	0	0	1	1	4	8	19	27	58	146	169	201	232
#anml w/tmr	0	0	0	0	1	1	4	6	7	9	12	15	15	15	13
Tumor Mean	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.4	1.0	1.4	2.9	7.3	8.9	10.6	13.6
Low n	30	30	30	30	30	30	30	30	30	30	29	29	29	29	29
# tumors	0	0	0	1	1	2	2	2	2	2	2	2	2	2	2
#anml w/tmr	0	0	0	1	1	2	2	2	2	2	2	2	2	2	2
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Med n	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30
# tumors	0	0	1	1	1	2	3	4	4	4	4	4	4	4	4
#anml w/tmr	0	0	1	1	1	2	3	4	4	4	4	4	4	4	4
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
MedHi n	30	30	30	30	30	30	30	30	30	29	28	28	27	27	27
# tumors	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
#anml w/tmr	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	30	30	30	30	30	30	30	29	28	28	26	26	26	26	26
# tumors	0	0	0	1	2	2	2	2	2	2	2	3	3	3	3
#anml w/tmr	0	0	0	1	2	2	2	2	2	2	2	3	3	3	3
Tumor Mean	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table A.3.6 (cont.) Overall Trajectory of SOA Latent + Active Papillomas in Male Mice

Treat Group	Week 16,	17,	18,	19,	20,	21,	22,	23,	24,	25,	26,	27
Veh n	25	25	25	25	25	25	25	25	24	24	24	24
# tumors	1	1	1	1	2	2	2	2	2	2	2	3
#anml w/tmr	1	1	1	1	2	2	2	2	2	2	2	2
Tumor Mean	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
TPA n	16	16	12	12	10	9	9	9	8	8	8	7
# tumors	205	225	118	151	93	72	80	84	65	69	69	54
#anml w/tmr	12	12	9	9	7	6	6	6	5	5	5	4
Tumor Mean	12.8	14.1	9.8	12.6	9.3	8.0	8.9	9.3	8.1	8.6	8.6	7.7
Low n	29	29	28	28	28	28	27	27	27	27	27	27
# tumors	2	2	2	2	2	2	2	3	3	3	3	3
#anml w/tmr	2	2	2	2	2	2	2	2	2	2	2	2
Tumor Mean	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Med n	29	29	29	28	28	28	28	28	27	27	27	27
# tumors	4	4	4	4	4	4	4	4	3	3	3	3
#anml w/tmr	4	4	4	4	4	4	4	4	3	3	3	3
Tumor Mean	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
MedHi n	27	27	27	27	27	27	27	27	27	27	27	27
# tumors	1	1	1	1	1	1	1	1	1	1	1	1
#anml w/tmr	1	1	1	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	26	26	26	26	26	26	25	25	25	25	25	25
# tumors	3	3	3	3	3	3	3	3	3	3	3	3
#anml w/tmr	3	3	3	3	3	3	3	3	3	3	3	3
Tumor Mean	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table A.3.7 Overall Trajectory of SOA Active Latent + Papillomas in Female Mice

Treat Group	Week 1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,	13,	14,	15,
Veh n	25	25	25	25	25	25	25	25	25	25	25	25	24	24	24
# tumors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TPA n	20	20	20	20	20	20	20	20	20	20	20	20	20	19	19
# tumors	0	0	0	0	0	0	0	5	10	13	34	54	86	111	136
#anml w/tmr	0	0	0	0	0	0	0	2	6	8	13	14	15	14	14
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.7	1.7	2.7	4.3	5.8	7.2
Low n	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25
# tumors	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Med n	25	25	25	25	25	25	25	25	25	24	24	24	24	24	24
# tumors	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MedHi n	25	25	25	25	25	25	25	25	25	24	23	23	22	22	21
# tumors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	25	25	25	25	25	25	24	23	23	23	23	23	22	22	22
# tumors	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
#anml w/tmr	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table A.3.7 (cont.) Overall Trajectory of SOA Active Latent + Papillomas in Female Mice

Treat Group	Week 16,	17,	18,	19,	20,	21,	22,	23,	24,	25,	26,	27
Veh n	24	24	24	24	24	24	24	24	23	23	23	23
# tumors	0	1	1	2	1	2	1	2	2	2	2	2
#anml w/tmr	0	1	1	2	1	2	1	2	2	2	2	2
Tumor Mean	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.1	0.1	0.1
TPA n	19	19	18	15	14	12	12	12	12	11	11	9
# tumors	167	175	173	103	125	108	122	115	121	114	115	86
#anml w/tmr	16	17	15	12	12	9	11	10	10	9	9	7
Tumor Mean	8.8	9.2	9.6	6.9	8.9	9.0	10.2	9.6	10.1	10.4	10.5	9.6
Low n	25	25	25	25	24	23	23	23	23	23	23	22
# tumors	1	1	2	2	2	0	0	0	0	1	2	3
#anml w/tmr	1	1	1	1	1	0	0	0	0	1	1	2
Tumor Mean	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1
Med n	24	24	23	23	22	22	22	22	22	22	21	21
# tumors	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MedHi n	21	21	20	20	20	19	19	19	19	19	19	19
# tumors	0	0	0	0	0	0	0	0	0	0	0	0
#anml w/tmr	0	0	0	0	0	0	0	0	0	0	0	0
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High n	22	22	22	22	22	22	22	22	22	22	21	20
# tumors	1	1	1	1	1	1	1	1	1	1	1	1
#anml w/tmr	1	1	1	1	1	1	1	1	1	1	1	1
Tumor Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1

The following tables A.3.8-A.3.11 display the number of animals with tumors (denoted by “n”) at NSOA sites and the overall number of tumors in the treatment group. For each treatment group, the entry for “NSOA” gives the overall totals at all NSOA sites, while totals for several specific NSOA sites are listed below the NSOA entries.

Table A.3.8 Profile of NSOA Active Papilloma Incidence in Male Mice

Treat Group/ Organ	Week 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Vehicle																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	2	2	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	2	2	2
Anal n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1
Other n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1
TPA																											
NSOA n	0	0	0	1	1	1	1	1	1	1	3	3	3	3	3	6	5	6	6	4	5	4	4	3	3	3	3
# tumors	.	.	.	1	1	1	1	1	1	1	3	3	4	4	3	6	5	9	12	7	9	6	7	6	6	6	6
Other n	0	0	0	1	1	1	1	1	1	1	3	3	3	3	3	6	5	6	6	4	5	4	4	3	3	3	3
# tumors	.	.	.	1	1	1	1	1	1	1	3	3	4	4	3	6	5	9	12	7	9	6	7	6	6	6	6
Low																											
NSOA n	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
# tumors	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	3	3	3
Abdom n	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	2	2	2
Medium																											
NSOA n	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	4	6
# tumors	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	3	5	7
Eye n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1
Head n	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	5
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	3	5
Medium-High																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	3	3	3	5	6
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	3	3	3	5	8
Anal n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	3	3	3	5	5
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	3	3	3	5	7
High																											
NSOA n	0	0	0	0	0	0	0	1	2	3	3	3	3	3	3	3	4	4	4	4	6	6	7	7	8	9	9
# tumors	.	.	.	.	.	.	.	1	2	3	3	3	3	3	3	3	4	4	4	4	6	6	8	8	9	10	10
Abdom n	0	0	0	0	0	0	0	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
# tumors	.	.	.	.	.	.	.	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Ear n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1
Other n	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	3	3	4	4	4	5	5
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	3	3	5	5	5	6	

Note the clear evidence of an increasing trend in papillomas over increasing dose.

Table A.3.9 Profile of NSOA Active Papilloma Incidence in Female Mice

Treat Group/ Organ	Week 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Vehicle																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TPA																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	1	1	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1	1	3	3	.	.	.
Other n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	1	1	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1	1	3	3	.	.	.
Low																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	2	2	2	3
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	.	.	1	1	2	2	2	3
Ear n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1
Other n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	.	.	.	.	.	.	.	.
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	2
Medium																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Medium-High																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	2
Eye n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
Other n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1
High																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1

Unlike male mice there is no evidence of an increasing trend in active papillomas over increasing dose.

Table A.3.10 Profile of NSOA Active+Latent Papilloma Incidence in Male Mice

Treat Group/ Organ	Week 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Vehicle																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	2	2	2	2	2
Anal n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1
Other n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1
TPA																											
NSOA n	0	1	1	1	1	1	1	2	3	3	3	3	4	8	7	7	7	6	7	5	5	4	4	3	3	3	3
# tumors	.	1	1	1	1	1	1	2	3	3	4	4	5	10	8	10	13	12	14	9	10	7	7	6	6	6	6
Other n	0	1	1	1	1	1	1	2	3	3	3	3	4	8	7	7	7	6	7	5	5	4	4	3	3	3	3
# tumors	.	1	1	1	1	1	1	2	3	3	4	4	5	10	8	10	13	12	14	9	10	7	7	6	6	6	6
Low																											
NSOA n	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3	4
# tumors	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	3	3	3	4	5
Abdom n	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2	2	3	4
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	2	2	2	3	4
Medium																											
NSOA n	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	2	4	6	7	7
# tumors	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	3	5	7	9	8
Eye n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1
Head n	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3	5	6	6
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	3	5	7	6
Medium-High																											
NSOA n	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	3	3	3	5	6	7	7
# tumors	.	.	.	2	1	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	3	3	3	5	8	10	11
Anal n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1
Urogen n	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	3	3	3	5	5	7	7
# tumors	.	.	.	2	1	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	3	3	3	5	7	9	10
High																											
NSOA n	0	0	0	0	0	1	2	3	3	3	3	3	3	3	4	4	4	4	6	6	7	7	8	9	9	10	10
# tumors	.	.	.	.	.	1	2	3	3	3	3	3	3	3	4	4	4	4	6	6	8	8	9	10	10	11	11
Abdom n	0	0	0	0	0	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
# tumors	.	.	.	.	.	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Ear n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1
Other n	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	3	3	4	4	4	5	5	6	6
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	3	3	5	5	5	6	6	7	7

As with just active papillomas, in pooled active plus latent papillomas there is clear evidence of an increasing trend in papillomas over increasing dose.

Table A.3.11 Profile of NSOA Active+Latent Papilloma Incidence in Female Mice

Treat Group/ Organ	Week 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Vehicle																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	1
Abdom n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.
TPA																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	2	2	1	2	1	1	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	2	2	2	4	5	3	3	.	.	.
Other n	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	2	2	1	2	1	1	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	2	2	2	4	5	3	3	.	.	.
Low																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	2	2	2	3	4	4
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	2	2	2	3	4	4
Ear n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1
Other n	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	.	.	.	.	.	.	.	.
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	3	3
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	2	3	3
Medium																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2
Medium-High																											
NSOA n	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	2	2	3
# tumors	.	.	.	1	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	2	2	3
Eye n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1
Other n	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# tumors	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	2
High																											
NSOA n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1
Urogen n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
# tumors	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1

Again there is no evidence of an increasing trend in papilloma count over dose.

Appendix 4. References

Bailer, A. and Portier, C. (1988), "Effects of Treatment-Induced Mortality on Tests for Carcinogenicity in Small Samples", *Biometrics*, **44**, 4, 417-431.

Bieler, G.S., and Williams, R.L. (1993), "Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity", *Biometrics*, **49**, 4, 793-801.

Chu, K.C., Ceuto, C., and Ward, J.M. (1981), Factors in the Evaluation of 200 National Cancer Institute Carcinogen Bioassays, *Journal of Toxicology and Environmental Health,* 8, 251-280.

Dunson D.B., Haseman, J.K., van Birgelen, A.P.J.M., Stasiewicz, S., and Tennant, R.W. (2000) Statisrtical Analysis of Skin Tumor Data from Tg.AC Mouse Bioassays, *Toxicological Sciences,* **55**, 293-302.

Haseman, J. K. (1983), A Reexamination of False-positive Rates for Carcinogenicity Studies, *Fundamental and Applied Toxicology,* 3, 334-339.

Jacobson-Kram, D., Sistare, F.D., and Jacobs, A.C. (2004) Use of Transgenic Mice in Carcinogenicity Assessment, *Toxicologic Pathology,* **32 (Suppl. 1)**, 49-52.

Lin, K. K. and Ali, M.W. (2006), Statistical Review and Evaluation of Animal Tumorigenicity Studies, *Statistics in the Pharmaceutical Industry,Third Edition,* edited by C.R. Buncher and J.Y. Tsay, Marcel Dekker, Inc. New York.

Lin, K. K. and Rahman, M.A. (1998), Overall False Positive Rates in Tests for Linear Trend in Tumor Incidence in Animal Carcinogenicity Studies of New Drugs, *Journal of Biopharmaceutical Statistics*. 8(1), 1-15.

McConnell, E.E., Solleveld, H.A., Swenberg, J.A., and Boorman, G.A. (1986), Guidelines for Combining Neoplasms for Evaluation of Rodent Carcinogenesis Studies, *Journal of the National Cancer Institute*. 76, 283-289.

Peto, R., Pike, M.C., Day, N.E., Gray, R.G., Lee, P.N., Parrish, S., Peto, J., Richards, S., and Wahrendorf, J. (1980). Guidelines for sample sensitive significance tests for carcinogenic effects in long-term animal experiments, *IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans, supplement 2: Long term and Short term Screening Assays for Carcinogens: A Critical Appraisal*, International Agency for Research Against Cancer, 311-426.

Rahman, M.A. and Lin, K.K. (2008), A Comparison of False Positive Rates of Peto and Poly-3 Methods for Long Term Carcinogenicity Data Analysis Using Multiple Comparison

Adjustment Method Suggested by Lin and Rahman, *Journal of Biopharmaceutical Statistics*. 18, 949-958.

STP Peto Working Group (2002), Statistical Methods for Carcinogenicity Studies, *Toxicologic Pathology*. 30 (3), 403-414.

U.S. Department of Health and Human Services (2001), Guidance for Industry Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals (DRAFT GUIDANCE), Center for Drug Evaluation and Research, Food and Drug Administration.

Application Type/Number	Submission Type/Number	Submitter Name	Product Name
NDA-21306	ORIG-1	PURDUE PHARMA LP	BuTrans (buprenorphine transdermal system)

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

/s/

STEVEN F THOMSON
03/02/2010

KARL K LIN
03/03/2010
Concur with review

STATISTICAL REVIEW
Clinical Studies

NDA 21-306

Name of drug: Buprenorphine transdermal system
Applicant: Purdue Pharma L.P.
Indication: pain
Documents reviewed: Electronic documents BP96-0101.pdf, BP96-0102.pdf, BP96-0604.pdf, BP98-1201.pdf, BP99-0203.pdf, corresponding appendices, and ise.pdf

Dates: submitted 3 November 2000; 10-month user fee date 3 September 2001

Project manager: Sara Shepherd
Clinical reviewer: Gerald Dalpan, M.D.
Statistical Reviewer: Stella Grosser, Ph.D.

INTRODUCTION

Buprenorphine is an opioid analgesic, with partial μ-opioid agonist and κ-opioid antagonist properties. It has been available in the United States since 1982 in an injectable form and primarily used for the management of postoperative pain. Parenteral buprenorphine has a fairly short duration of action (6 to 8 hours), motivating the development of a transdermal system in hope of treating chronic pain.

BTDS (Norspan) is a transdermal system providing systemic delivery of buprenorphine continuously up to 7 days. The system is made in 3 dose sizes, namely 5, 10 and 20 mcg/h delivery rate, corresponding to an active surface area of 6.25, 12.5, and 25 cm^2.

EFFICACY STUDIES

The clinical program in this application includes five studies of the efficacy of BTDS. Two of these five, BP96-0101 and BP96-0102, were considered failed by the applicant. They are reviewed briefly below. The applicant claims the other threee—BP96-0604, BP990203, and BP98-1201—together demonstrate efficacy. This claim is open to question, as will be discussed in detail in this review.

Failed studies

Design

Studies BP96-0101 and BP96-0102 were forced titration studies. Patients initially received a 5-mg dose; those assigned to 10 or 20 mg had their dose increased to 10 mg at 6 days and further to 20 mg at 12 days if in the highest dose group. There was a placebo control as well as an oxycodone/acetaminophen (Oxy/APAP) active control arm in each study. BP96-0101 enrolled 270 osteoarthritis patients for 60 days. BP96-0102 enrolled 249 patients with low back pain, also for 60 days. Both studies had discontinuation rates of 40–50% in all treatment arms.

Outcome measures, analysis and results

The primary efficacy measures were pain on average and pain right now, as measured by the Brief Pain Inventory (BPI), an 11- point scale, where 10 is the greatest pain. Measurements were taken at baseline and at study days 9, 15, 30, 45 and 60. A linear model was fit to the data at each visit separately, with terms included for treatment, baseline pain, center, age and previous opioid use yes/no (study 96-0101) and treatment, baseline pain, center, age and gender (study 96-0102). A linear mixed effects model was also fit to all the measurements at days 9, 15, 30, 45 and 60 as part of a repeated measures analysis, using the same set of covariates as for the visit-by-visit analysis.
These covariates were chosen "as appropriate" from a list specified in protocol amendment 3 (study 96-0101) and amendment 1 (96-0102); the list was age, baseline pain, center, gender, previous opiod use, race and weight. I could not find a definition of "as appropriate." Missing data were imputed by carrying forward the last observation. All comparisons to placebo were non-significant at the 5% level. The results are shown in Tables 1 and 2 below.

Study BP96-0101

Change From Baseline, Least Square Means ± SEM(Days 9–60)

Efficacy variable	Placebo (N = 52)	Oxy/APAP (N = 55)	BTDS 5 (N = 55)	BTDS 10 (N = 54)	BTDS 20 (N = 54)	BTDS vs. Placebo
BPI (scale 0–10)						
Pain on the average	-0.7 ±0.3	-0.6 ± 0.3	-1.1 ± 0.3	-0.9 ± 0.3	-1.2 ± 0.3	ns
Pain right now	-0.2 ± 0.3	-0.9 ± 0.3	-0.9 ± 0.3	-0.8 ± 0.3	-1.0 ± 0.3	ns

Study BP96-0101

Change From Baseline, Least Square Means ± SEM(Days 9–60)

Efficacy variable	Placebo (N = 52)	Oxy/APAP (N = 55)	BTDS 5 (N = 55)	BTDS 10 (N = 54)	BTDS 20 (N = 54)	BTDS vs. Placebo
BPI (scale 0–10)						
Pain on the average	-1.2 ±0.2	-1.4 ± 0.2	-1.1 ± 0.2	-1.0 ± 0.2	-1.4 ± 0.2	ns
Pain right now	-1.1 ± 0.3	-1.2 ± 0.3	-0.9 ± 0.3	-1.2 ± 0.3	-1.5 ± 0.3	ns

Titration-to-effect studies

Study BP96-0604

Study BP96-0604 was a multicenter, randomized, double blind parallel group titration-to-effect trial, with both placebo and active (Oxy/APAP) control arms. One hundred thirty four patients with lower back pain were enrolled in this study (45 placebo, 43 Oxy/APAP, and 46 BTDS). A maintenance period of 63 days followed a 21-day titration period. Outcome information was collected at clinical visits on study days 7, 21, 30, 45, 60, 75 and 84.

Applicant's Analysis

Primary outcome variables
The primary outcome measures were pain on average since the previous visit and pain right now, both assessed on the BPI eleven-point scale; change from baseline was calculated to determine efficacy. The FDA criticized as unclear the analysis proposed in the original protocol. An amendment dated 11 March 1998 specified that

> …[A] repeated measures analysis will be performed to assess the effects due to treatment, center, and treatment by center interaction. Missing values will be extrapolated by the last observation carried forward (LOCF). Certain clinically important covariates (i.e., gender, age, race, weight, baseline pain, and previous opioid use) may be incorporated into the model or analyzed separately as a sub-population when appropriate. For the primary and the secondary comparisons of interest, pairwise t-tests will be used to test the null hypothesis of no treatment difference. A confirmatory analysis without extrapolating any missing values will also be performed using a general mixed model which constructs the likelihood for all of the data.

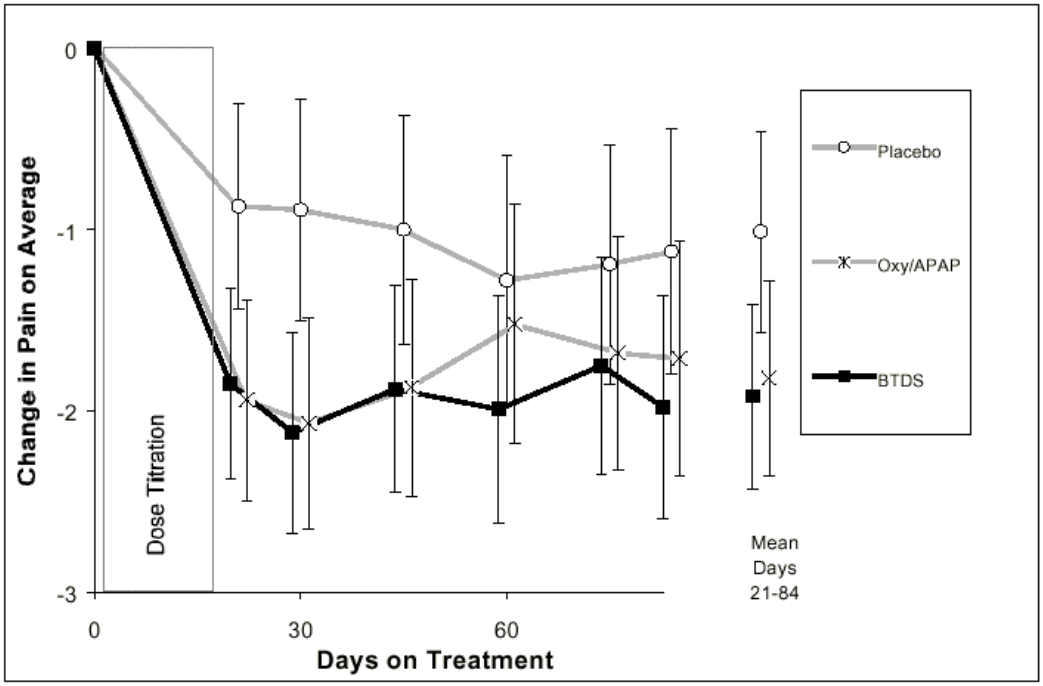
This plan is vague, since "repeated measures analysis" covers a wide variety of possible methods. What the applicant did was twofold:

A longitudinal mixed-effects linear model was fit to the data for all times, simultaneously. For subjects who discontinued, missing data were imputed by carrying forward the last observed outcome. Fixed-effect covariates were apparently chosen by backward elimination from the list given in the protocol amendment, keeping covariates significant at $p < 0.1$. (There is only a comment to this effect in a copy of the applicant's SAS program submitted in response to a request by the FDA.) This resulted in adjusting change in pain on average for center, baseline pain and prior opiod use, and change in pain right now for center, baseline pain, prior opiod use, and age. The within-patient error covariance matrix was assumed to be autoregressive; there was no subject random effect in the model. Confirmatory variants of this first analysis done by the applicant included using an unstructured covariance matrix and not imputing, i.e., not carrying forward the final observation.

In addition, for each of the primary outcome variables, a cross-sectional linear model was fit at each time. This model included the covariates of the respective longitudinal models, that is, center, baseline pain and prior opioid use for change in pain on average, and center, baseline pain, prior opiod use, and age for change in pain right now. This was done carrying the last observation forward as well as without imputing.

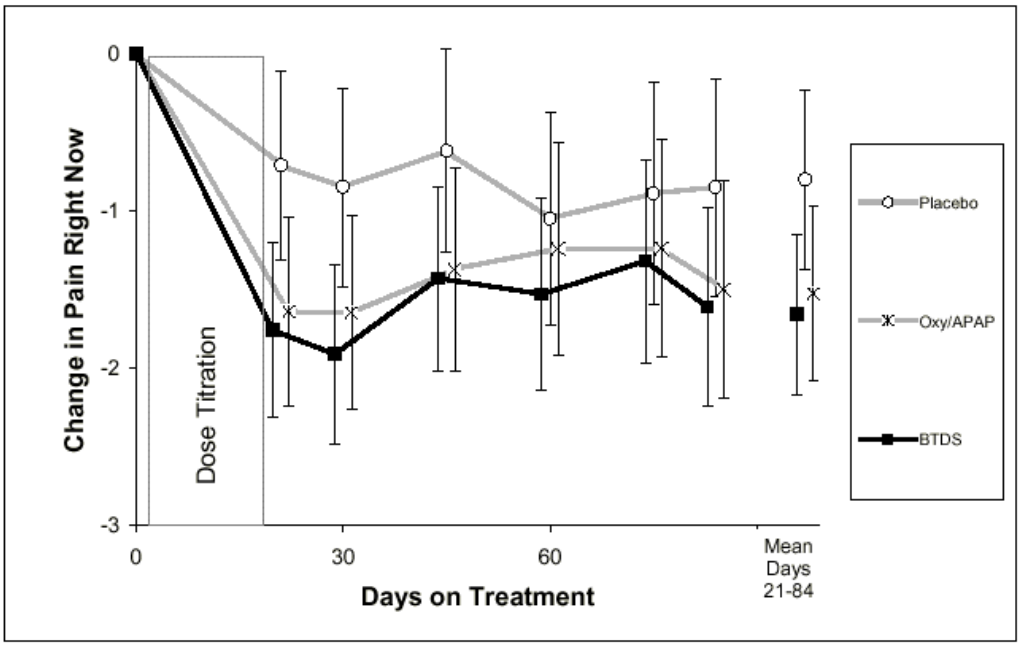
Applicant's results—primary efficacy analysis

The applicant's analysis, with the last observation carried forward, found statistically significant differences between placebo and BTDS in change in pain on average and pain right now for days 21-84, taken as a totality, as well as at days 21 and 30. Differences between arms, including those not statistically significant, were on the order of a point on the 11-point BPI scale. The results of the placebo – Oxy/APAP comparison follow much the same pattern. The results of the primary efficacy analysis, taken from the electronic submission, are shown below graphically and in tabular form:



		Day 21	Day 30	Day 45	Day 60	Day 75	Day 84	RM 21-84
Placebo	LS Mean ± SEM	-0.87 ± 0.38	-0.89 ± 0.41	-1.00 ± 0.42	-1.28 ± 0.46	-1.19 ± 0.44	-1.12 ± 0.45	-1.01 ± 0.37
N=45	N with data	34	23	22	18	18	18	45
Oxy/APAP	LS Mean ± SEM	**-1.94 ± 0.37**	**-2.07 ± 0.39**	-1.87 ± 0.40	-1.52 ± 0.44	-1.68 ± 0.43	-1.71 ± 0.43	-1.82 ± 0.36
N= 42	N with data	**31**	**32**	29	29	28	27	42
	Pairwise vs. Placebo	***P*=0.015**	***P*=0.013**	ns	ns	ns	ns	ns
BTDS	LS Mean ± SEM	**-1.85 ± 0.35**	**-2.12 ± 0.37**	-1.88 ± 0.38	-1.99 ± 0.42	-1.75 ± 0.40	-1.98 ± 0.41	**-1.92 ± 0.34**
N= 46	N with data	**33**	**29**	25	23	21	22	**46**
	Pairwise vs. Placebo	***P*=0.025**	***P*=0.0093**	ns	ns	ns	ns	***P*=0.035**

(Cross references: LS Means from Table 14.2.1.1.B2; Ns from Table 14.2.1.1.B3.)
Least squares (LS) means - corrected by SAS Proc Mixed for baseline pain, center and opioid experience
Bar indicates time of dose titration: all BTDS patients started with BTDS 5 and titrated dose on Day 7 and/or Day 14
N = Number of patients with data at that visit; N for LOCF = N for the treatment group and was consistent over time
Bars at each data point indicate ± 1.5 SEM
RM 21-84 values calculated via repeated measures analysis of all available data from Days 21-84 using SAS Proc Mixed
Pairwise vs. Placebo - results of comparison with placebo using SAS Proc Mixed
ns = difference not statistically significant (P> 0.05)
Bolding indicates statistically significant reesults



		Day 21	Day 30	Day 45	Day 60	Day 75	Day 84	RM 21-84
Placebo	LS Mean ± SEM	-0.71 ± .040	-0.85 ± 0.42	-0.62 ± 0.43	-1.05 ± 0.45	-0.89 ± 0.47	-0.85 ± 0.46	-0.80 ± 0.38
N=45	N with data	34	23	21	18	18	18	45
Oxy/APAP	LS Mean ± SEM	**-1.64 ± 0.40**	-1.65 ± 0.41	-1.37 ± 0.43	-1.24 ± 0.45	-1.24 ± 0.46	-1.50 ± 0.46	-1.53 ± 0.37
N= 42	N with data	**31**	32	29	29	28	27	42
	Pairwise vs. Placebo	***P*=0.049**	ns	ns	ns	ns	ns	ns
BTDS	LS Mean ± SEM	**-1.76 ± 0.37**	**-1.91 ± 0.38**	-1.43 ± 0.39	-1.53 ± 0.41	-1.32 ± 0.43	-1.61 ± 0.42	**-1.66 ± 0.34**
N= 46	N with data	**33**	**29**	25	23	21	22	**46**
	Pairwise vs. Placebo	***P*=0.022**	***P*=0.028**	ns	ns	ns	ns	***P*=0.045**

(Cross references: LS Means from Table 14.2.1.2.B2; Ns from Table 14.2.1.2.B3.)
Least squares (LS) means - corrected by SAS Proc Mixed for age category, baseline pain, center and opioid experience
Bar indicates time of dose titration: all BTDS patients started with BTDS 5 and titrated dose on Day 7 and/or Day 14
N = Number of patients with data at that visit; N for LOCF = N for the treatment group and was consistent over time
Bars at each data point indicate ± 1.5 SEM
RM 21-84 values calculated via repeated measures analysis of all available data from Days 21-84 using SAS Proc Mixed
Pairwise vs. Placebo - results of comparison with placebo using SAS Proc Mixed
ns = difference not statistically significant (P> 0.05)
Bolding indicates statistically significant results

Secondary outcome variables

The secondary outcome variables assessed were dropout due to lack of efficacy, MOS SF-36 (a quality of life survey), time to stable pain management, number of dose adjustments, and daily pain scores from telephone diary.

MOS SF-36 scores were similar across the three treatment arms. Time to stable pain management and the percentage achieving it were generally more favorable in the BTDS group and the Oxy/APAP group than in the placebo. Dose group titration and adjustments were comparable across the three treatment arms. Daily patient diary pain scores showed a slightly greater decrease in the BTDS and Oxy/APAP groups compared to the placebo treatment groups.

The proportion of patients discontinuing **due to lack of efficacy** was significantly higher in the placebo group than in either the BTDS or Oxy/APAP groups: 36% in the placebo, 2% in the Oxy/APAP, and 15% in the BTDS group. Essentially all of this discontinuation took place by study day 35 in all groups.

Other analyses

Of note in this trial are the large number of discontinuations for all reasons, combined. Forty-eight percent of the BTDS, 49% of the placebo, and 28% of the Oxy/APAP patients dropped out before the end of the study and in fact mostly before day 45. While it was not specified as an outcome variable, the proportion of patients dropping out **due to adverse events** was significantly higher in the BTDS and Oxy/APAP groups than in the placebo, namely 33%, 28% and 16%, respectively. This pattern is an opposite of that seen for discontinuation due to lack of efficacy.

The last observation in a patient who discontnued due to adverse events may be artificially high, in the sense that the dose was high enough to be efficacious perhaps only by being high enough to cause intolerable side effects. Because patients discontinued for different reasons in the placebo and the BTDS groups, and since the the dominant reason for discontinuation in the BTDS group was drug-related adverse events, the apparent advantage of BTDS in a LOCF analysis may be a artifact of differentially carrying forward artificially high pain relief values. Because of this, it is important to consider an alternative to the primary efficacy analysis done carrying forward the last observation. The confirmatory analysis provided by the applicant, of the observed data with no carry forward or other imputation, is one possibility.

Results of this analysis are shown in detail in the tables below. The longitudinal linear model used to analyze the all times together, the same model as used in the LOCF analysis, again yields a statistically significant difference between the placebo and BTDS groups in change in pain right now and pain on average, as well as between placebo and Oxy/APAP in change in pain right now. Again, the adjusted means in the placebo group are about a point worse than the means in the active treatment groups, though all groups appear to have a greater decrease in pain. The visit-by-visit means tell a different story.

For change in pain on average, the BTDS group is significantly better than the placebo group at days 21 and day 30. By day 60, and until the end of the study, there is almost no observable difference between these two groups. Change in pain right now shows a similar pattern, with a significant difference between the two groups at day 21, insignificant differences of the same order of about a point at days 30 and 45, and very little difference at the remaining visits.

The repeated-measures analysis compares weighted averages of the visit-by-visit values. The weights are different for the LOCF and non-LOCF analyses. In the LOCF analysis, there is a value (possibly imputed) for each patient at each timepoint, so that the timepoints are weighted equally. For the non-LOCF analysis, the early part of the study is weighted more heavily, because more data are available there, and this is where there is the largest difference between BTDS and placebo. In the later half of the study, that difference has evaporated, but the population under study has changed as well. Only about 50 percent of the patients completed the study, and it is likely to be a different 50 percent with the two treatments. Completers in the placebo group may be patients with tolerable pain, whereas completers in the BTDS group may be those who tolerate side effects, possibly because their pain is so severe as to make tolerating the side effects worthwhile.

Pain on Average (Change from Baseline), without LOCF

		Day 21	Day 30	Day 45	Day 60	Day 75	Day 84	RM 21-84
Placebo N=45	LS Mean ± SEM	-1.0 ± 0.5	-1.7 ± 0.5	-2.2 ± 0.6	-3.5 ± 0.7	-3.1 ± 0.7	-3.0 ± 0.7	-1.4 ± 0.5
	N with data	34	23	22	18	18	18	34
Oxy/APAP N=42	LS Mean ± SEM	-2.5 ± 0.4	-2.5 ± 0.4	-2.5 ± 0.5	-2.0 ± 0.6	-2.2 ± 0.6	-2.1 ± 0.6	2.3 ± 0.4
	N with data	31	32	29	29	28	27	32
	Pairwise vs. placebo	P=0.004	ns	ns	ns	ns	ns	ns
BTDS N=46	LS Mean ± SEM	-2.4 ± 0.4	-2.9 ± 0.5	-3.1 ± 0.6	-3.5 ± 0.7	-3.3 ± 0.7	-3.3 ± 0.6	-2.6 ± 0.4
	N with data	33	29	25	23	21	22	33
	Pairwise vs. placebo	P=0.008	P=0.046	ns	ns	ns	ns	P=0.02

Pain Right Now (Change from Baseline), without LOCF

		Day 21	Day 30	Day 45	Day 60	Day 75	Day 84	RM 21-84
Placebo N=45	LS Mean ± SEM	-0.9 ± 0.4	-1.5 ± 0.6	-1.2 ± 0.7	-2.8 ± 0.7	-2.3 ± 0.8	-2.4 ± 0.7	-1.1 ± 0.4
	N with data	34	23	22	18	18	18	34
Oxy/APAP N=42	LS Mean ± SEM	-2.4 ± 0.4	-2.0 ± 0.5	-2.1 ± 0.6	-1.7 ± 0.6	-1.8 ± 0.7	-2.0 ± 0.6	2.1 ± 0.4
	N with data	31	32	29	29	28	27	32
	Pairwise vs. placebo	P=0.005	ns	ns	ns	ns	ns	P=0.03
BTDS N=46	LS Mean ± SEM	-2.2 ± 0.4	-2.5 ± 0.5	-2.3 ± 0.6	-2.8 ± 0.6	-2.6 ± 0.7	-2.8 ± 0.7	-2.2 ± 0.4
	N with data	33	29	25	23	21	22	33
	Pairwise vs. placebo	P=0.008	ns	ns	ns	ns	ns	P=0.02

The choice of model may be post-hoc: I was unable to find documentation of the method used to select covariates from the protocol-specified list except as a comment in the SAS computer program that was sent in response to the Agency request after the NDA was filed. On the other hand, the algorithm for choosing covariates was reasonable, and my reanalysis using other p-values as cutoffs in the backward elimination procedure resulted in qualitatively the same contrast between BTDS and placebo. Similarly, there is no pre-specification or justification for the auto-regressive covariance structure, other than a statement that this structure had been chosen before the blind was broken. In fact, an auto-regressive structure theoretically does not describe data where half of the observations for half of the subjects have been carried forward—all the imputed values for an individual are exactly the same as his last observed value, and so the within-subject correlation is a constant of one rather than falling off exponentially with time. However, there is no qualitative difference in the primary results when other, more appropriate, covariance matrices, in particular an unstructured one, are used instead. The lack of a subject random effect also could be a problem, since the model implies that the correlation within an individual eventually fades to 0, and this is rarely the case. Again, adding a subject effect to the statistical model made little difference, probably because there were relatively few measurements per patient by design, and the estimate of the autoregressive parameter was close enough to 1 that it would take many measurement times before the correlation approached its asymptotic value.

Study BP99-0203

Study BP99-0203 enrolled 311 patients with osteoarthritis in the hip (n=140) and knee (n=171). There was a 21-day titration period, followed by a 7-day maintenance period.

Primary efficacy analysis

The primary endpoint was the patient-reported pain relief at the end of 28 days, scored on a five-point scale. Patients reporting good, very good or excellent pain relief were counted as successes; those reporting poor or fair were counted as failures. If a subject dropped out due to lack of efficacy, the applicant counted him as a failure. If a subject dropped out due to an adverse event, he was counted as a success or failure according to his pain-relief reported at the time of discontinuation.

The sponsor reported success rates of 44% (65/149) in the BTDS group, 32% (52/162) placebo. A chi-square test for treatment effect, using logistic regression with terms for treatment and center, gave an odds ratio of 1.7 with a 95% confidence interval of 1.04 to 2.7 and a p-value of 0.04. Other terms included initially and removed by backward elimination (P >0.1) were age, gender, weight, baseline average pain intensity, and joint site (knee or hip).

Analyzing hip and knee replacement sites separately gave odds ratios of 1.4 and 2.2 with corresponding confidence intervals 0.7 to 3.1 and 1.1 to 4.4, and p-values of 0.341 and 0.028, respectively.

These results are shown in the table below, taken from the applicant's electronic submission.

Treatment	BTDS			Placebo			Odds Ratio (OR)	
Measure	Success n/N (%)	Failure n/N (%)	Ratio	Success n/N (%)	Failure n/N (%)	Ratio	Observed OR	Adjusted OR [95% CI]
Knee	38/84 (45)	46/84 (55)	0.83	26/87 (30)	61/87 (70)	0.43	1.94	2.18 [1.099 to 4.403]
Hip	27/65 (42)	38/65 (58)	0.71	26/75 (35)	49/75 (65)	0.53	1.34	1.44 [0.681 to 3.073]
Combined	65/149 (44)	84/149 (56)	0.77	52/162 (32)	110/162 (68)	0.47	1.64	1.66 [1.035 to 2.693]

However, the medical reviewer, Dr. Dalpan, and I felt that a more appropriate definition of success and failure would classify as a failure someone who could not tolerate the treatment, regardless of reported pain relief at discontinuation. This is particularly true given the shortness of this trial for relief of chronic pain (28 days total with only one post-titration week, i.e., a single patch). Across the two pain sites combined, 45 patients discontinued due to drug-related adverse events: 31 BTDS, of which 15 had originally been classified as a success; and 14 in the placebo group, of whom 4 were successes. Reclassifying these gave new success rates of 34% and 30%, respectively. A chi-squared

test for association between success and treatment differences in proportions yields an non-significant p-value of 0.46. (I did not adjust for center in calculating this value. Recreating the adjustment in the precise way that the applicant had done it was not worth the trouble — it is unlikely that the value would change much if I had, since reanalyzing the original classification without adjusting for center gives a p-value of 0.036 that is nearly identical to the applicant's.)

Secondary endpoints

Secondary variables were change from baseline in average pain intensity at days 21 and 28, change from baseline of average diary pain scores during Days 21 to 28, patient satisfaction scores at Day 21 and Day 28, discontinuation due to lack of efficacy, investigator assessment of therapy, and dose level at end of titration. The BTDS group did slightly better than placebo on these measures, with significant differences in investigator assessment of therapy and dose level at the end of titration.

Study BP98-1201

Study BP98-1201 was an active-controlled, double-blind, parallel-group, multicenter, titration-to-effect study in chronic back pain patients. After a seven-day run-in period, patients were assigned to either BTDS or an active control of hydrocodone/acetaminophen tablets (HCD/APAP), and allowed to titrate over 21 days to one of three dose levels (one, two or three tablets of 2.5 mg hydrocodone/250 mg acetaminophen; 5, 10, or 20 mg BTDS). This titration period was followed by a 35-day maintenance period. Two hundred and sixty six patients were in the intent-to-treat population.

Highlights of the applicant's report are summarized below.

Patient disposition was similar across the two treatment arms, with approximately half the patients completing the study, 22% in the HCD/APAP vs. 24% in the BTDS group withdrawing due to drug-related adverse events and 14% HCD/APAP vs. 15% BTDS withdrawing due to lack of efficacy.

Primary efficacy outcome variables were average pain intensity (over the preceding 24 hours), measured on a 10-point scale and patient satisfaction, measured on a 5-point scale; both were assessed every 7 days. These measures were compared between treatment groups over days 21 to 56 using a repeated measures linear mixed-effects model with the last observation carried forward; least squares means were estimated from the final model. A 95% confidence interval was constructed for the difference in each outcome variable. The difference for average pain intensity was –0.08, with a 95% confidence interval of –0.6 to 0.4. The difference for patient satisfaction was 0.16, with a 95% confidence interval of –0.08 to 0.4. Both confidence intervals fell within the pre-specified limits for equivalence (upper limit less than 2 for average pain intensity and lower limit greater than –1 for patient satisfaction) as well as within non-inferiority

ranges, leading to a conclusion of non-inferiority and equivalence of BTDS relative to HCD/APAP.

Assay sensitivity was assessed by meta-analysis of HCD/APAP effect sizes in the literature. This analysis found a pooled effect size of 1.1, with a 95% confidence interval of 0.7 to 1.6 for the decrease in average pain of patients treated with HCD/APAP relative to placebo in well-controlled trials. Testing the null hypothesis that the difference between BTDS and HCD/APAP was greater than 1.1 gave a p-value of less than 0.001, leading to the conclusion that BTDS performed here within the range that HCD/APAP had performed in the prior placebo-controlled trials. However, it is not obvious that the studies cited were equivalent in design to the BP98-1201 and in particular whether the pain site studied was the back. Because these studies are not clearly comparable, and there is no internal placebo control, I do not review BP96-1201 in greater detail.

CONCLUSIONS

Based on the efficacy studies presented in this NDA, the evidence for efficacy of buprenorphine transdermal system is mixed. There is support for a conclusion of analgesic activity and short-term efficacy in the treatment of chronic back pain but very little for effectiveness in treating osteoarthritis.

Of the five studies, one is an active control study in back pain patients where BTDS and HCD/APAP apparently perform similarly. The effectiveness of HCD/APAP, the assay sensitivity, is not well-determined due to a meta-analysis of uncertain quality and lack of an internal placebo. The sponsor classified two as failed studies. The observed effect sizes are small, statistical significance is lacking and trends are not apparent in most cases (there is some dose-response for BTDS in study BP96-0102, forced titration for low-back pain).

Study BP96-0604 provides the strongest argument for efficacy. In this study in lower back pain, the beneficial effects of BTDS relative to a placebo control are seen in the first half of the 84-day trial and drive an estimate of overall pain relief leading to the applicant's conclusion that BTDS is effective relative to placebo. An important feature of this study is the large number of discontinuing patients in both the BTDS and placebo arms. There is more "lack of efficacy" in the placebo group, as judged by the reasons for dropping out; however, pain relief in the BTDS patients is balanced by unbearable toxicities and discomforts that lead to discontinuation. There is little difference at study's end between the groups, although both sets of completers continue to experience pain relief.

The success or failure of study BP98-0203 hinges on the definition of a binary outcome of an individual's success or failure. By the applicant's definition, there are significantly more successes in the BTDS group than in the placebo. By the reviewers' redefinition, there is little difference between the proportions. It is possible to view this single study

as two studies—one in hip pain, and one in knee pain, since the applicant powered the study for the two sites separately—and then even using the applicant's definition, one of these two studies failed.

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

/s/

Stella Grosser
8/20/01 03:43:03 PM
BIOMETRICS

Thomas Permutt
8/20/01 04:11:21 PM
BIOMETRICS
concur

S. Edward Nevius
8/21/01 10:58:20 AM
BIOMETRICS
Concur with review.