CENTER FOR DRUG EVALUATION AND RESEARCH

APPROVAL PACKAGE FOR:

APPLICATION NUMBER

NDA 21-044

Statistical Review(s)

25 Page(s) Withheld

✓ § 552(b)(4) Trade Secret / Confidential

______ § 552(b)(5) Deliberative Process

______ § 552(b)(5) Draft Labeling

Statistical Review and Evaluation

CLINICAL STUDIES

NDA 21-044 (resubmission of approvable NDA)
Name of drug: Palladone (hydromorphone HCl) extended-release tablets
Applicant: Purdue
Indication: pain
Documents reviewed: CD-ROM dated 30 March 2001
Project manager: Judit Milstein
Medical officer: Michael Sevka, M.D.
Dates: user fee goal 30 September 2001

Reviewer: Thomas Permutt

BACKGROUND

NDA 21-044 concerns an extended-release formulation of hydromorphone, an opioid analgesic marketed in oral and parenteral dosage forms under NDAs 19-034, 19-891 and 19-892 as Dilaudid, as well as under ANDAs. This NDA was found to be "approvable" 29 December 1999. The NDA reported active-controlled studies against immediate-release hydromorphone as well as a placebo-controlled, rescue-sparing study in post-operative pain. The placebo-controlled study was believed by the sponsor to have a statistically significant result, but on review FDA disagreed. The active-controlled studies alone were considered by FDA not to provide substantial evidence of efficacy. Accordingly, the action letter conditioned the approvability on a further, successful study demonstrating efficacy in chronic pain. The present amendment reports the results of a placebo-controlled study of four weeks' duration (in addition to a titration period of up to two weeks) in osteoarthritis.

DESIGN AND PLANNED ANALYSIS

On meeting criteria for entry, patients were treated with Dilaudid (immediate-release, oral hydromorphone) for up to 14 days with titration "according to a dosing schedule determined by the Investigator." Patients achieving satisfactory results during this period with a stable dose between 8 mg and 14 mg were eligible for the double-blind phase. One hundred sixty patients at 19 sites were randomized in equal numbers to treatment with Palladone 12 mg q.d. or placebo.

Daily ratings of average pain over the preceding 24 hours were collected by an automatic telephone system. A five-point categorical scale was used (no pain, mild pain, moderate pain, moderately severe pain, and severe pain), and numerical scores of 0 to 4 were assigned to these categories. The primary analysis concerned the ratings on days 12, 13, 26 and 27. The scores for days 12 and 13 were to be averaged, and the scores for days 26 and 27 were to be averaged. The two resulting scores, one for week 2 and one for week 4, were then considered as repeated measures in testing for a treatment effect. This amounts essentially to a further averaging of the scores for each patient, so that the primary outcome would be the average of the scores on the four days in question.

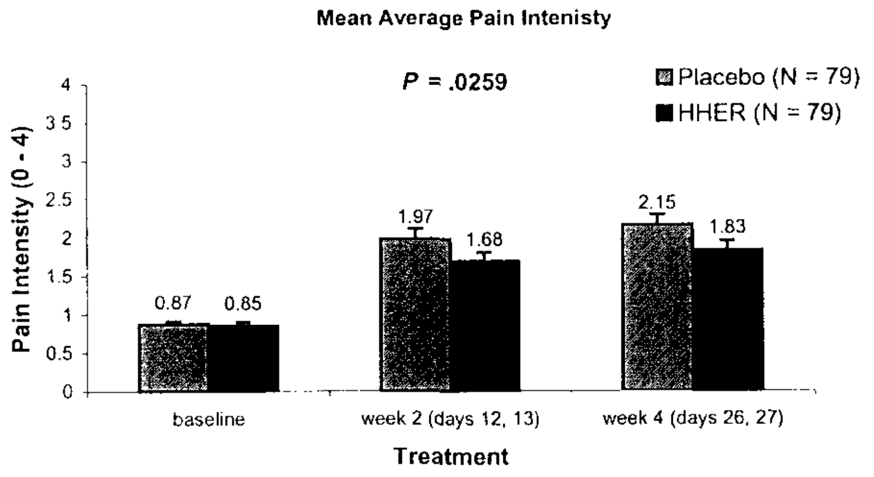
According to the protocol, missing values were to be treated differently depending on the reason. Missing scores followed by later, valid observations were to be interpolated. Missing values from patients dropping out were to be imputed by the last observation carried forward; but if the patient dropped out for lack of efficacy, the worst observation would be carried forward rather than the last.

The study report does not give the results of this planned analysis, nor does it say whether it was carried out, nor why it was or was not carried out. Instead it substitutes an analysis with last observation carried forward for all dropouts. Dropouts for lack of efficacy were still handled somewhat differently than others, however: the observations carried forward were from an exit interview if the patient dropped out for lack of efficacy, but from the usual telephone report if he or she dropped out for other reasons.

APPLICANT'S RESULTS

FIGURE 11.1.1.

Mean Average Pain Intensity by Treatment Group
Cross-reference: Figure 14.2.2; Appendix 16.2.5.2.



Note: *P* value for comparing 2 treatments was obtained from a repeated measures analysis of covariance with terms of treatment, center, and baseline pain intensity (mean Average Pain Intensity during the 2 consecutive days before randomization).
Note: Raised bars represent standard deviation.
0 = no pain, 1 = mild pain, 2 = pain, 3 = moderate pain, 4 = moderately severe pain

The figure above is copied from the electronic submission. The footnote is incorrect: the bars represent standard errors of the means. The standard deviations at weeks 2 and 4 were about 1 point on the pain intensity scale. Also, the verbal descriptions on the categorical pain scale are given incorrectly. The unusual term "mean average" is meant to indicate that the raw data were ratings of *average* pain over the last 24 hours, not "pain right now," and that the figure shows means of these.

The applicant characterizes the effect as "modest" but statistically significant. The submission goes on to discuss two secondary measures which, the applicant believes, may more clearly show the benefits of the drug. These are a subject global assessment, and the time to discontinuation for lack of efficacy. The results are shown below.

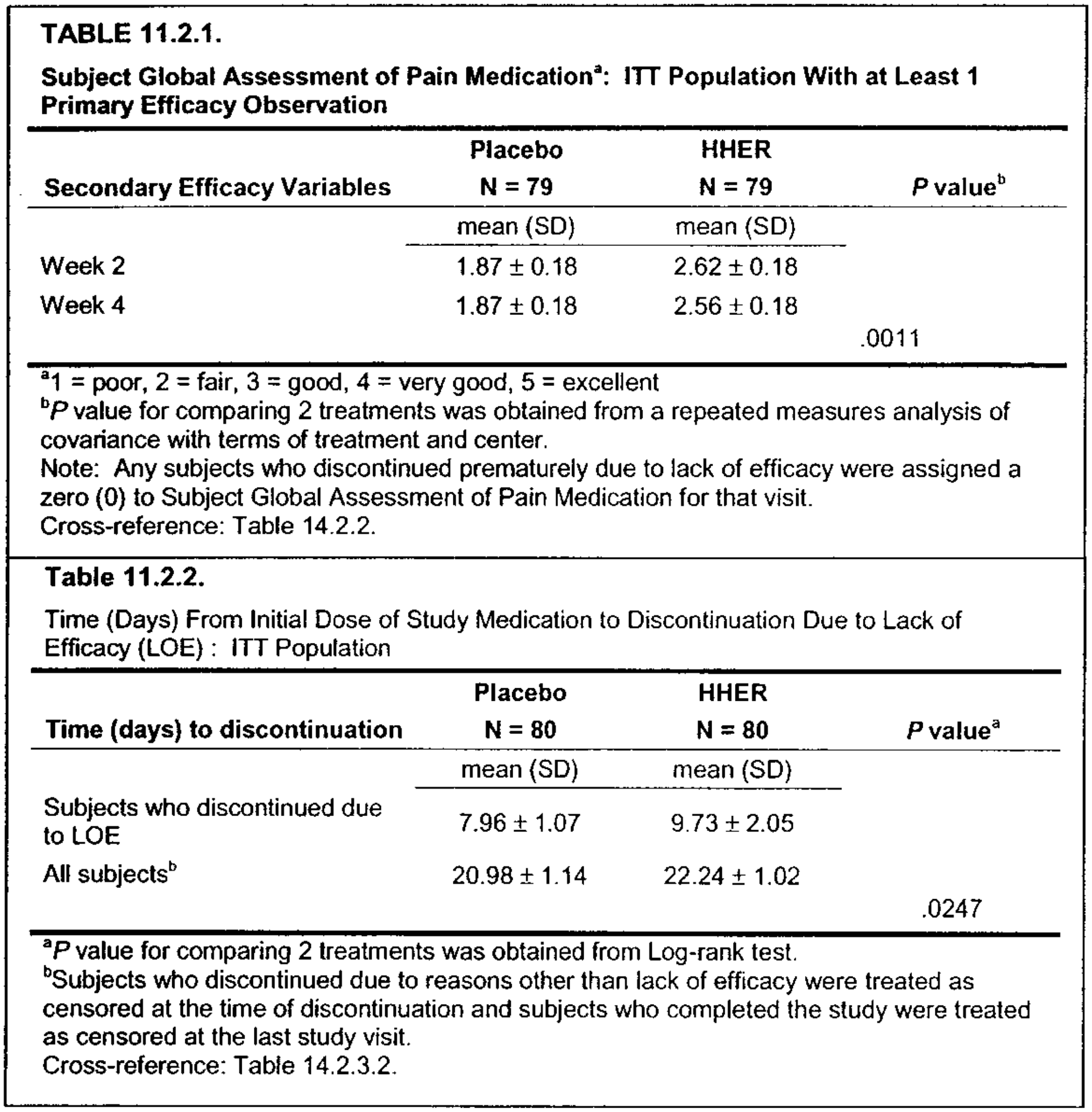


TABLE 11.2.1.

Subject Global Assessment of Pain Medication[a]: ITT Population With at Least 1 Primary Efficacy Observation

Secondary Efficacy Variables	Placebo N = 79	HHER N = 79	*P* value[b]
	mean (SD)	mean (SD)	
Week 2	1.87 ± 0.18	2.62 ± 0.18	
Week 4	1.87 ± 0.18	2.56 ± 0.18	
			.0011

[a] 1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent
[b] *P* value for comparing 2 treatments was obtained from a repeated measures analysis of covariance with terms of treatment and center.
Note: Any subjects who discontinued prematurely due to lack of efficacy were assigned a zero (0) to Subject Global Assessment of Pain Medication for that visit.
Cross-reference: Table 14.2.2.

Table 11.2.2.

Time (Days) From Initial Dose of Study Medication to Discontinuation Due to Lack of Efficacy (LOE) : ITT Population

Time (days) to discontinuation	Placebo N = 80	HHER N = 80	*P* value[a]
	mean (SD)	mean (SD)	
Subjects who discontinued due to LOE	7.96 ± 1.07	9.73 ± 2.05	
All subjects[b]	20.98 ± 1.14	22.24 ± 1.02	
			.0247

[a] *P* value for comparing 2 treatments was obtained from Log-rank test.
[b] Subjects who discontinued due to reasons other than lack of efficacy were treated as censored at the time of discontinuation and subjects who completed the study were treated as censored at the last study visit.
Cross-reference: Table 14.2.3.2.

The time to discontinuation requires some explanation. The figures 7.96 and 9.73 appear to be the average times of discontinuation (for lack of efficacy) for those patients who did in fact discontinue. In other words, subjects who discontinued for lack of efficacy in the active group did so slightly later than those in the placebo group; but this does not take into account that subjects in the active group were less likely altogether to discontinue for lack of efficacy. It is the figures 20.98 and 22.24 that are meant to reflect this, and on which the significance test is apparently based. The method of analysis is meant to estimate the mean time at which patients would have dropped out for lack of efficacy, assuming that all patients would eventually have done so if observed long enough. Patients who completed the study without dropping out, *as well as patients who dropped out for other reasons,* were supposed to have been "censored." That is, the time of their hypothetical, eventual dropping out for lack of efficacy could not be observed, but it could be estimated assuming the risk was the same as for patients being observed.

COMMENTS

LAST OBSERVATION CARRIED FORWARD

I reviewed the protocol for this trial. I commented then:

> It is anticipated that as many as half the patients may drop out before the end of the treatment period. Care has been taken to minimize the resulting bias. Patients dropping out for lack of efficacy would have their worst score carried forward. Patients dropping out for other reasons would have their last score carried forward.

I still am troubled by the number of dropouts. I still believe the reported analysis minimizes bias in a certain sense. I believe, however, that further exploration of the data is needed to clarify what sense that is.

		Placebo mean pain (N)	Hydromorphone mean pain (N)
Observed data:			
	day 12	1.56 (52)	1.47 (60)
	day 13	1.47 (55)	1.47 (53)
	day 26	1.50 (38)	1.41 (39)
	day 27	1.46 (39)	1.50 (42)
Imputed data[1]:			
	day 12	2.81 (27)	2.21 (19)
	day 13	3.08 (24)	2.23 (26)
	day 26	2.78 (41)	2.23 (40)
	day 27	2.78 (40)	2.22 (37)
All data:			
	day 12	1.99 (79)	1.65 (79)
	day 13	1.96 (79)	1.72 (79)
	day 26	2.16 (79)	1.82 (79)
	day 27	2.13 (79)	1.84 (79)

The observed data were very similar between the two treatment groups for all four days, but the imputed data were different. The "modest" difference seen in the applicant's figure can essentially all be attributed to the imputation of higher scores to placebo patients than to placebo patients when actual data were missing. (There is a slight additional effect for day 12, in that more placebo patients had imputed scores, and the imputed scores were higher than the observed scores.)

This difference does represent a real *effect* of the drug. Placebo patients had worse scores carried forward because they had more pain at the time they dropped out. Hydromorphone patients were more likely to drop out because of side effects, and they had less pain when they dropped out.

On the other hand, the difference does not seem to represent a real *benefit* of the drug in this population. Patients who dropped out, whether for lack of efficacy or for side effects, may be viewed as having been unsuccessfully treated. Approximately equal numbers in the hydromorphone and placebo groups failed in this way. The remaining patients, on average, fared no better on hydromorphone than on placebo. Overall, then, the placebo group seems

[1] CFFLAG > 0 in data set A_DAY

to have been very nearly as well off as the hydromorphone group: equal numbers of dropouts and equal pain for completers.

OTHER COMMENTS ON PRIMARY ANALYSIS

DEVIATIONS FROM PROTOCOL

Last vs. worst observation carried forward

As noted above, the protocol specified that the worst, rather than the last, observation should be carried forward for patients who dropped out for lack of efficacy. This analysis is not reported. However, the analysis in the protocol could only have been more favorable to the active drug. There being more dropouts for lack of efficacy in the placebo group, the imputed scores in the placebo group would have been worse than in the analysis actually performed, while the scores in the active group would have changed less. Thus the difference between the two treatments would have been more. The change in methods should have been explained in the study report, but it seems reasonable and conservative.

Treatment-by-center interaction

The protocol specified that the treatment effect was to be tested for significance in a repeated-measures analysis of covariance with effects for baseline, treatment, center, and treatment-by-center interaction. The reported analysis did not contain interaction terms. Some of the centers had few patients, and one center with only two patients had both assigned to the same treatment. The model specified in the protocol, with treatment-by-center interactions, therefore has too many parameters to be estimated at all, so that something else must be done. Dropping the interaction terms was reasonable.

It is not, however, the only reasonable approach. The center effects might also have been dropped. In this case, the p-value would change from the reported 0.03 to 0.08 by my calculation. I do not say this would have been preferable. Rather, I point out that the statistical significance of the primary result is sensitive to certain post-hoc choices of methods, and no justification is given for the choices made in the study report.

REPEATED MEASURES

The primary test of statistical significance compared the two groups on two derived measurements for each patient: the week 2 mean (days 12 and 13) and the week 4 mean (days 26 and 27). The placebo-controlled study in the original NDA used a somewhat similar method, and there was a critical difference of opinion between FDA and the applicant on the adequacy of that analysis. I therefore wish to make clear here that my comments on the earlier study do not apply to this one. The crucial difference is the introduction here of a random subject term into the repeated-measures analysis, as was discussed in my review of the original NDA. As applied in this study, the repeated-measures analysis appears to me to be appropriate.

COMMENTS ON SECONDARY ANALYSES

GLOBAL SATISFACTION

I observed above that on the whole the placebo group was as well off as the hydromorphone group, at least with respect to the primary endpoint. There is some evidence that the patients felt otherwise. The applicant's table 11.2.1, reproduced above on p. 3, says that placebo patients rated their medication about fair, on average, whereas hydromorphone patients rated it between fair and good.

Here again, however, the analysis was so conducted as to assign an artificially bad score (zero) to patients who dropped out for lack of efficacy. In contrast, patients who dropped out for other reasons had their last score carried forward, and it could not be worse than 1 (poor). Like the primary analysis, therefore, this may suggest only that patients perceived a *difference* between treatments (being more likely to drop out *for lack of efficacy* in the placebo group, though not more likely to drop out), without indicating that they were truly any better off.

TIME TO DROPOUT FOR LACK OF EFFICACY

In this other secondary analysis, patients who dropped out for lack of efficacy were again analyzed differently from those who dropped out for other reasons. At best, then, this analysis would show (again) that patients in the placebo group were more likely to give this reason for dropping out than patients in the hydromorphone group, without indicating that they were any better off.

I cannot see that it shows even that, however. Data from patients who dropped out for other reasons were considered to be censored. As I indicated above, the original interpretation of censoring in survival analysis was that patients continued to be at risk of the defining event (originally death; here, dropping out for lack of efficacy) but could not have their whole survival time observed, usually because they were still alive at the time of reporting. The present case is not similar. Patients who dropped out for adverse events were not at any risk of later dropping out for lack of efficacy. It is not that the time of dropping out for lack of efficacy was not observed (unlike the time of death in classical survival analysis): rather, this event did not occur, and never would have occurred. The patients in question had an event of another kind, which precluded them from having the defining event. There may be some sense in which this is analogous to censoring, but I cannot see it.

In any case, even with this questionable analysis, the difference in "survival" amounts only to a day: 21 days in the placebo group and 22 days in the hydromorphone group. It is hard to see this as representing a meaningful effect.

CONCLUSIONS

The analysis called primary in the study report deviated in at least two aspects from what was planned in the protocol. The last, rather than the worst, observation was carried forward for dropouts for lack of efficacy. Also, treatment-by-center interaction terms were dropped. The report did not explain these deviations, and it should have. It might have pointed out that the actual analysis was less favorable to the test drug than the planned one with respect to dropouts, and that the planned analysis was impossible with respect to interactions because one center had only one treatment. As the primary results were not overwhelmingly significant statistically, concerns about post-hoc choice of methods cannot easily be dismissed. On balance, however, I believe the primary result should be considered statistically significant.

That statistically significant result amounted to this: patients in the hydromorphone group who dropped out had less pain at the time of discontinuation than did patients in the placebo group who dropped out. Approximately equal proportions of patients dropped out on the two treatments. Furthermore, patients who did not drop out did about as well on placebo as on hydromorphone. If dropouts are considered failures, therefore, the placebo group had as many and as good successes as the hydromorphone group.

The primary analysis therefore does not seem to indicate any benefit of hydromorphone over placebo in this patient population. Neither do the two secondary endpoints discussed in the study report.

On the other hand, there was indeed a difference between hydromorphone and placebo with respect to the claimed effect of the drug, which is the relief of pain. In a narrow sense, this might be taken as satisfying the requirement in the action letter: "You must perform at least one adequate and well-controlled study in the setting of chronic pain, with multiple dosing, that demonstrates superiority over placebo or another control in order to establish the efficacy of your product."

There do not appear to be any statistical issues with regard to safety, which is discussed in the medical officer's review.

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

/s/

Thomas Permutt
9/24/01 10:22:03 AM
BIOMETRICS

S. Edward Nevius
9/24/01 01:05:16 PM
BIOMETRICS
Concur with review.

3/12/02

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

Statistical Review and Evaluation

CLINICAL STUDIES

NDA:	21-044/AZ (12 March 2002)
Name of drug:	Palladone (hydromorphone HCl extended-release) capsules
Applicant:	Purdue
Indication:	pain
Documents reviewed:	electronic submission:
	\\CDSESUB1\N21044\N_000\2002-03-12;
	amendment 15 April 2002 (1 volume)
Project manager:	Sara Shepherd
Clinical reviewer:	Michael Sevka, M.D.
Dates:	Received 3/13/02; user fee (6 months) 9/13/02
Statistical reviewer:	Thomas Permutt
Statistics team leader:	Thomas Permutt
Biometrics division director:	S. Edward Nevius, Ph.D.

Keywords: NDA review, clinical studies

1 BACKGROUND

Palladone is a new, extended-release formulation of hydromorphone, a pre-1938 opioid analgesic marketed in oral, rectal and parenteral forms under several NDAs and ANDAs. NDA 21-044 was found to be "approvable" 29 December 1999, but the conditions for approval including supplying additional evidence of efficacy. A response 30 March 2001 was found not approvable 4 October 2001. That response reported an additional trial of efficacy, but evidence of efficacy was still considered deficient. This further response reports another trial. There are also new stability data which will be the subject of a separate statistical review.

2 DATA ANALYZED AND SOURCES

This review concerns study HMP-3006, a multicenter, randomized, double-blind, placebo-controlled, parallel-group trial in chronic pain. The design and analysis of this trial were novel. Studies reported in the original application compared Palladone to immediate-release hydromorphone. The results were similar but left questions about assay sensitivity. The first resubmission reported a placebo-controlled study, but the interpretation was clouded by the large number of dropouts. The present study sought to turn to advantage the propensity of patients to drop out of placebo-controlled studies in chronic pain. The primary endpoint was a composite called Emergence of Inadequate Analgesia:

> Criteria for the Emergence of Inadequate Analgesia were: (1) a rating of 1 or 2 on the Subject Global Assessment of Pain Medication 5-point categorical scale, where, in response to the question, "How would you rate your medicine for pain?" the subject rated the medicine 1 = poor, 2 = fair, 3 = good, 4 = very good, or 5 = excellent; or (2) the subject rated pain on average as moderate to severe; or (3) the subject took more than 2 doses per week of a short-acting analgesic for acute pain; or (4) the subject discontinued double-blind study medication due to lack of efficacy.

Thus, differential dropouts for lack of efficacy would be considered evidence in themselves of efficacy of the drug. The primary statistical analysis was a log-rank test of time to Emergence of Inadequate Analgesia.

3 PRIMARY RESULTS

Two hundred twenty-one patients were randomized. The table below, copied from the electronic submission (study report, p. 44), summarizes their disposition.

Subject Disposition: ITT Population

	Treatment Groups		
Category	**Placebo**	**HHER 12 mg**	**Overall Total**
	n (%)	n (%)	n (%)
Randomized	111 (100.0)	110 (100.0)	221(100.0)
Completed	109 (98.2)	103 (93.6)	212 (95.9)
28 days (ie, End of the Study)	23 (20.7)	63 (57.3)	86 (38.9)
Emergence of Inadequate Analgesia[a, b, c]	86 (77.5)	40 (36.4)	126 (57.0)
Discontinued	5 (4.5)	9 (8.2)	14 (6.3)
Reason for discontinuation:			
Adverse Event[b]	4 (3.6)	7 (6.4)	11 (5.0)
Death	0	0	0
Lost to Follow-up	1 (0.9)	0	1 (0.5)
Protocol Violation	0	0	0
Other[c, d]	0	2 (1.8)	2 (0.9)

[a] These subjects are identified as "Discontinued due to ineffective treatment" in the CRFs but are considered complete because they met the study endpoint of Emergence of Inadequate Analgesia.

[b] Four subjects (Subjects 1892/12083 [HHER], 2149/8177 [placebo], 2154/14107 [placebo], and 2165/25248 [placebo]) were categorized by the investigator both as discontinuing due to an adverse event and as meeting the study endpoint of Emergence of Inadequate Analgesia. These subjects are counted in both categories.

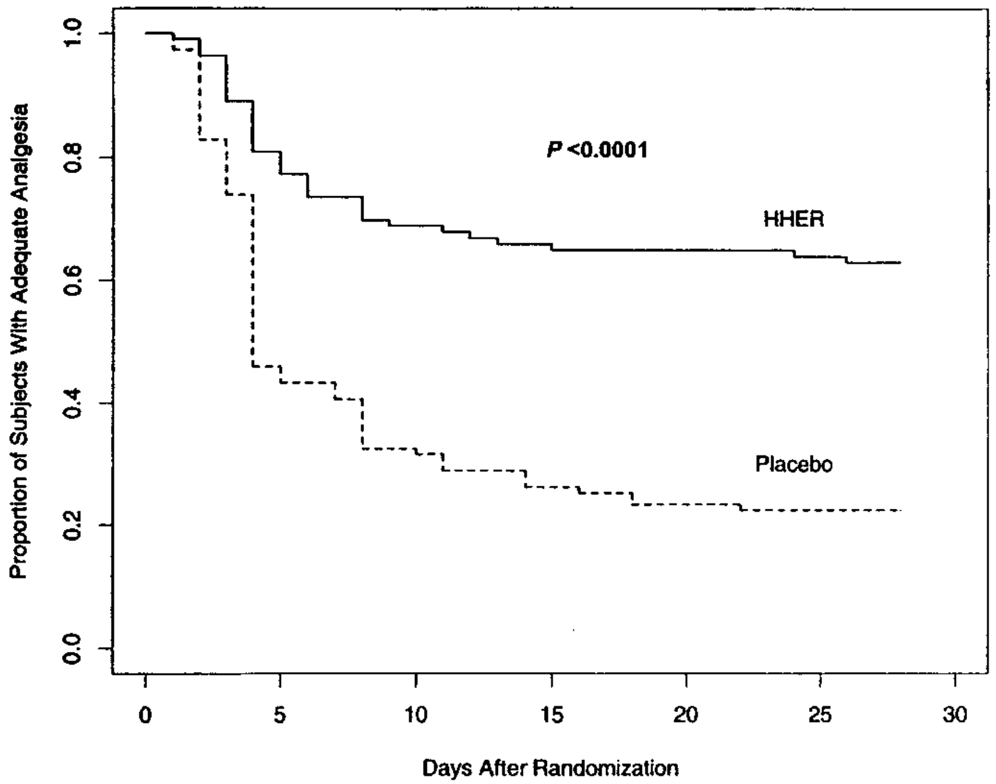
[c] One subject (Subject 2162/32229 [HHER]) was categorized by the investigator both as discontinuing due to withdrawal of consent and as meeting the study endpoint of Emergence of Inadequate Analgesia. This subject is counted in both categories.

[d] One subject (Subject 2164/23103 [HHER]) was non-compliant with treatment and pill count.

Cross-reference: Table 14.1.1.

More than three quarters of placebo patients experienced Emergence of Inadequate Analgesia, compared to only about a third of the Palladone (HHER, hydrocodone hydrochloride extended-release) group. A Kaplan-Meier plot of the time to this endpoint is given below (p. 51). The p-value is for the log-rank test. Most of the placebo patients had reached the endpoint within five days, while most of the Palladone patients never reached it.

Appears This Way
On Original



4 OTHER ANALYSES

Because of the design of the study, comparisons of pain scores between groups are not meaningful. At any given time, the groups being compared would include only those patients for whom analgesia was adequate, since on Emergence of Inadequate Analgesia their participation ended.

Two patients on placebo and seven patients on hydromorphone discontinued treatment but were not considered to have had Emergence of Inadequate Analgesia. In the active group, six of these seven patients withdrew because of adverse events. A conservative approach would be to consider these patients as unsuccessfully treated, the same as patients who discontinued because of inadequate analgesia. The number of such patients is so small that they could not substantially change either the overall appearance of the survival curves nor the statistical significance of the difference between them, regardless of when they dropped out. In other words, even with all dropouts conservatively counted as failures, there would still be a dramatic and statistically significant difference between treatments.

The applicant also includes some analyses to address the question of withdrawal. Hydromorphone is an addictive drug, and all the patients in the trial were taking hydromorphone at the beginning of the placebo-controlled phase. It is therefore possible that a difference between groups would result not from a beneficial effect of hydromorphone but rather from a deleterious effect of its discontinuation in the placebo group.

A preplanned analysis compared the means for the two treatment groups of a score that rated each of six withdrawal symptoms on a scale from zero to three and then summed the six ratings. The mean of this score at the time of each patient's discontinuation or completion was 0.5 in each group. However, nine patients, all in the placebo group, were observed to have symptoms consistent with withdrawal (p. 56).

Subjects with Adverse Events Consistent with Withdrawal Signs and Symptoms: ITT Population

Investigator/ Subject No.	Symptoms	Treatment	Completion Status[d]
Definite opioid withdrawal signs or symptoms[a]			
1820/9007	Shaking, sweating, anxiety, diarrhea, myalgia	Placebo	EIA[e]
1820/9216	Diarrhea, runny nose, chills	Placebo	EIA
Probable opioid withdrawal signs or symptoms[b]			
2149/8177	Diarrhea, nausea, abdominal cramps	Placebo	EIA
2151/26060	Nausea, vomiting, fever, chills	Placebo	EIA
2169/27092	Diarrhea, vomiting, sweating	Placebo	EIA
Possible opioid withdrawal signs or symptoms[c]			
1820/9132	"Cold" symptoms	Placebo	28 days
1892/12050	Stomach cramping	Placebo	EIA
1892/12293	Cold sweats, nervousness, dyspnea	Placebo	EIA
2168/20123	Stomach cramps, cold sweats	Placebo	EIA

[a] Diagnosed as opioid withdrawal by the investigator and reported as an adverse event.
[b] 2 – 3 opioid withdrawal signs or symptoms.
[c] 1– 2 opioid withdrawal signs or symptoms.
[d] Subjects completed either when they reached the endpoint of Emergence of Inadequate Analgesia, or at the completion of the 28-day double-blind phase. [e] Emergence of Inadequate Analgesia.
Cross-reference: Table 14.3.2.2, Appendices 16.2.1 and 16.2.7.1.

A placebo-controlled study in patients habituated to an addictive drug cannot definitely distinguish a beneficial effect of continuing the drug from a deleterious effect of stopping it. Nevertheless, there were only eight patients who had some withdrawal symptoms and were counted as having had Emergence of Inadequate Analgesia. Even if all these placebo patients were very conservatively considered never to have reached the endpoint (i.e., they were assumed to have been *successfully* treated), the results would not have been importantly affected.

5 DEMOGRAPHIC SUBGROUPS

An amendment 15 April 2002 reports analysis of the primary efficacy variable by sex, race and age. I compiled the table below from several tables in that submission. Palladone appeared to be effective in all the subgroups, with no indication of differential effectiveness, though the numbers of nonwhite and elderly patients were small.

Median days to emergence of inadequate analgesia
by demographic group

	hydromorphone		placebo	
	N	median	N	median
male	38	>28	35	7
female	72	>28	76	4
white	98	>28	96	4
nonwhite	12	>28	15	8
under 65	93	>28	93	4
over 65	17	>28	18	4

6 LABELING

The study reviewed here is described in the proposed label as follows:

This language fairly represents the results of the study.

In addition, the proposed label describes another study that I reviewed previously (NDA 21-044/AZ/30 March 2001, my review 24 September 2001):

[ ]

My earlier review concluded, "The primary analysis therefore does not seem to indicate any benefit of hydromorphone over placebo in this patient population. Neither do the two secondary endpoints discussed in the study report." Accordingly, I do not think it is useful to describe this trial in the label.

7 CONCLUSIONS AND RECOMMENDATIONS

Earlier, active-controlled studies of Palladone, previously reviewed, showed comparable effects on pain to immediate-release hydromorphone, but lacked internal evidence of sensitivity. The present placebo-controlled study did not use pain directly as a primary measure of effect. However, it clearly demonstrated an effect of hydromorphone on the primary measure, which was continuation without intolerable pain. Taken together, these studies demonstrate that Palladone has an analgesic effect, and they also provide evidence of the magnitude and duration of the effect.

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

/s/

Thomas Permutt
7/16/02 09:38:41 AM
BIOMETRICS

24 Page(s) Withheld

✓ § 552(b)(4) Trade Secret / Confidential

______ § 552(b)(5) Deliberative Process

______ § 552(b)(5) Draft Labeling

JUL 26 1999

Statistical Review and Evaluation

NDA 21-044
Name of drug: Palladone — (hydromorphone hydrochloride) controlled-release capsules
Applicant: Purdue Pharma
Indication: pain
Documents reviewed: volumes 1, 2, 105–144;
electronic copies of same; electronic data and programs
Project manager: Nancy Chamberlin
Medical officer: Monte Scheinbaum, Ph.D., M.D.
Dates: received 29 December 1998; user-fee goal (10 months) 29 October 1999

Reviewer: Thomas Permutt

INTRODUCTION

Hydromorphone is a pre-1938 opiate analgesic derived from morphine. It is marketed in oral, injectable and rectal formulations by several manufacturers. The subject application is for a new, controlled-release, oral formulation, intended to provide once-a-day dosing for chronic pain.

Evidence concerning efficacy comes from three studies. Two active-controlled, multiple-dose, crossover trials studied the efficacy of Palladone in a clinical setting close to the one in which it is meant to be used. One single-dose, three-arm, placebo-controlled (with rescue) trial purports to lend additional evidence of the efficacy both of the test drug and the active comparator.

ACTIVE-CONTROLLED TRIALS

Two identically-designed, multicenter, crossover trials (801 and 802) compared Palladone to an immediate-release (IR) preparation of the same active ingredient, hydromorphone hydrochloride. The studies enrolled "patients who required opioid analgesics for treatment of chronic cancer-related pain or chronic non-cancer pain"; about three fourths of the patients had cancer. At enrollment, patients were switched from morphine or other opioids to Palladone q.d. at a dose of 1 mg for each 8 mg/day of oral morphine, or for an amount of other opioids considered equivalent to 8 mg/day of oral morphine. The dose of Palladone was titrated over the course of 4 to 21 days "to achieve stable pain control." Patients were then randomized to either sequence of Palladone q.d. and IR hydromorphone q.i.d. (double dummy), at the same daily dose as at the end of the titration period. They took one treatment for 3 to 7 days and then crossed over to the other treatment for another 3 to 7 days. Neither the study reports nor the protocols are very clear on how the duration of these two double-blind periods was determined. Patients recorded their pain on a visual analog scale before each dose. They were asked to report both pain "right now" and average pain since the last dose.

PRIMARY ANALYSIS

The protocol specified both the primary measure of outcome and the statistical analysis in detail. The average (since the last dose) pain was to be averaged over two days of the treatment period (the last two days of treatment before a scheduled "PK/PD day" with frequent pain measurements and blood drawing). Again, it is not clear how the two days were chosen, but they seem to have been scheduled in advance. The two periods for each patient were then compared in an analysis of variance with terms for treatment, period, sequence and patient (nested in sequence). The protocols also provided for a center main effect and a center-by-treatment interaction, but the reports suggest that analysis without them is preferable because the patients were spread rather sparsely over centers (approximately 20 centers in each study). I repeated the principal analyses using the model specified in the protocols, and the results were substantially identical to those reported by the sponsor. Confidence intervals for the difference in pain scores between the two treatments were calculated. It was suggested a priori that a difference of two, on a scale of 0 to 10, might be considered clinically insignificant.

Results were calculated both for an "evaluable" population (67 and 91 patients in the two studies), who supplied the necessary measures, and an intent-to-treat population (104 and 113 patients), with last observation carried forward. The results of the two analyses were very similar. In a crossover study, even the evaluable population includes the same patients on both treatments, so that I see no clear reason to prefer the imputed, intent-to-treat results. I will therefore focus, as the sponsor does, on the efficacy population.

TABLE 8.10.4E.
Ninety Percent Confidence Interval Analysis of the Mean Average Pain Intensity Over the Last 2 Days Before Each PK/PD Day of the Double-Blind Periods

	Mean[a] (SE) Average Pain Intensity		Difference	90% CI of the Difference	
Study	HHCR	HHIR	(HHCR – HHIR)	Lower Bound	Upper Bound
HD95-0801 (N=67)	2.48 (0.07)	2.42 (0.07)	0.06	-0.11	0.23
HD95-0802 (N=91)	2.59 (0.08)	2.58 (0.08)	0.01	-0.17	0.19
Combined (N=158)	2.54 (0.05)	2.51 (0.05)	0.03	-0.09	0.16

Data for efficacy population from Periods 1 and 2 combined.

Cross-reference: Table 8.10.7.8.A

[a] Least squares mean.

The average scores were essentially identical for the two treatments in each study. The confidence bounds are so narrow, on the scale of 0 to 10, that it is hard to see how the difference could possibly be meaningful. The entire 90% confidence intervals lie well within the range of –2 to 2 suggested a priori. Ninety-five percent confidence intervals would still do so. Thus, with respect to the primary endpoint, the outcomes with Palladone were on average very similar to those with the active comparator.

RESCUE MEDICATION

Patients in these studies were allowed additional IR hydromorphone tablets as needed, in both treatment periods. The effectiveness of the study drugs, therefore, cannot be measured by the reported pain alone, unless the use of rescue medication was similar for the two treatments.

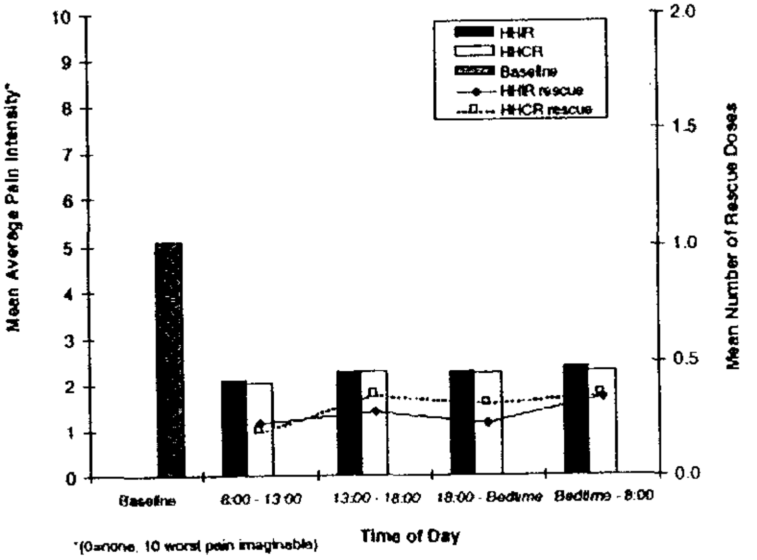
In fact, over the two-day periods that provided the primary pain data, patients averaged 1.3 rescue doses per day, on either treatment. Thus, the controlled- and immediate-release formulations were comparable with respect to use of rescue medication as well as reported pain.

Dr. Scheinbaum has conducted a more detailed analysis of rescue medication and pain combined. For each patient, he described one treatment as superior to the other if it had substantially less use of rescue or substantially less pain or both, provided that it did not have less rescue and more pain nor more rescue and less pain. By his criteria, in study 801 the outcome was superior on Palladone for 30 of 88 patients, and on IR hydromorphone for 34 patients. In study 802, the outcome was superior on Palladone in 34 of 106 patients, and on IR hydromorphone for 39 patients. A binomial significance test can be done conditional on the number for which one or the other treatment was superior, ignoring those patients for whom neither treatment was substantially better; this is analogous to McNemar's test. The difference between the treatments is not statistically significant, even if the studies are pooled.

TIME COURSE

The time course of action of sustained-release drugs is important. A recommendation of once-a-day dosing for an analgesic drug amounts to a marketing claim, and should therefore be based on evidence that the drug is effective for 24 hours. The figure, from the integrated summary of effectiveness, shows the time course of pain and of use of rescue medication, for the two studies pooled. There were no notable differences in either pain or rescue at any time of day. Thus, daily dosing with Palladone was similar to q.i.d. dosing with IR hydromorphone with respect to the time course of action.

Mean of Average Pain Intensity by Time of Day



DIFFERENCE-DETECTING ABILITY

With respect to the primary efficacy endpoint, Palladone and IR hydromorphone were indistinguishable in these studies. The same was true for the critical though "secondary" endpoint of rescue medication, as well as for a variety of other secondary measures of pain. As usual, the question remains of whether these active-controlled studies failed to distinguish two effective or two ineffective treatments. There are several possible approaches to answering this question.

The protocol incorporated an attempt at internal validation of the sensitivity of the study. Patients were to be given a lower dose (50% of the usual dose) on a random day. Poorer outcomes on that day would be evidence of a dose-response and so necessarily also of a drug effect. Based on logistical and ethical considerations, however, this experiment was moved in a protocol amendment from the main part of the study to an extension phase, and was actually carried out only in a handful of patients, with results that are not interpretable statistically.

The most persuasive indication that both treatments were effective is the condition of the patients at baseline. All had chronic pain. Most had cancer. They were converted to an average daily dose of about 20 mg of hydromorphone, at a ratio of 1:8 oral morphine equivalent; that is, on entry they were taking morphine or other opioids equivalent to about 160 mg/day of oral morphine. They reported an average baseline pain of about 5 on the scale of 0 to 10, even while on other opioids, and, after upward titration in most cases, achieved average scores of about 2 on hydromorphone. It is reasonable on historical grounds to suppose that these patients would have had scores higher than 2 if untreated; it would then follow that both hydromorphone treatments must have been effective.

There are limitations to this approach, of course. Patients were enrolled in the study because they required treatment for pain; although they had chronic conditions not likely to remit in most cases, some regression might be expected. Furthermore, the titration period served to enrich the population with patients who either responded well to Palladone or had spontaneous improvement: of 344 patients enrolled, only 219 were randomized, the rest having discontinued during the titration phase. In any case, while it might be reasonable to suppose that the study patients would have had more pain if untreated, it would be very difficult to say how much more.

In light of these limitations, it has been usual to view trials like these as giving substantial evidence that the test drug is effective, but not as justifying a claim of therapeutic equivalence. I believe that is also the correct interpretation of these trials.

DEMOGRAPHICS

The combined population in these two studies was 52 percent female and 10 percent nonwhite (mostly black). Thirty-seven percent were over 65 and 11 percent were over 75. Separate analyses were conducted by age, race and sex for the two studies pooled. No notable differences were seen between the treatments in any subgroup.

PLACEBO-CONTROLLED TRIAL

A third study (505) compared Palladone, IR hydromorphone and placebo in patients with post-operative pain. The application does not propose to label Palladone for this use. Consequently, both the sponsor and Dr. Scheinbaum consider this trial to be of secondary importance. By its design, however, the study had the potential to give evidence of efficacy both of Palladone and of the comparator in the active-controlled studies.

Patients were given fentanyl, a synthetic opioid analgesic, intravenously by a patient-controlled analgesia (PCA) device. They were then given one of three oral treatments, in a randomized, double-dummy, parallel-group design. One group had 24 mg of Palladone; another group had 6 mg of IR hydromorphone; and the third group had placebo. Patients were encouraged to use the PCA device to control their pain. It was expected that they would achieve similar levels of pain, and that there would be a difference between treatments in the amount of PCA fentanyl used.

The protocol specified both pain and fentanyl consumption as primary endpoints. It was vague as to the precise analysis to be performed, particularly as regards the time periods over which these outcomes were to be measured. The study report notes:

> Before the blind was broken, the statistical analysis plan, presented in the protocol (Appendix 16.1.1), was elaborated, clarified, and modified. This development of the statistical analysis plan was documented by memoranda copied to Sponsor files. The resulting statistical analysis plan is presented here. Any changes to the analysis after blind was broken is described in Section 9.8.

In fact there is little explanation of what the final methods were or why they were chosen. The statistical programs and the data were submitted electronically, however. I replicated the main analyses, so that I know what was done. The retrospective choice of methods without good documentation might raise concerns about multiplicity, if the results of these analyses had been more positive than they were.

The sponsor reports pain intensity in the following table:

TABLE 11.1.2
Study HD96-0505
Current Pain Intensity Over Time by Treatment

Time of Pain Intensity Postdose	Current Pain Intensity	ITT Population		
		HHCR* (N = 44)	HHIR* (N = 44)	Placebo* (N = 44)
0 (Baseline)†	Mean	5.68	5.55	5.55
	Range	5 – 9	5 – 8	4 – 8
24 Hours	Mean	1.40	1.72	1.83
	Range	0 – 5	0 – 5	0 – 5
Overall‡	Mean	2.48	2.76	2.69
	Range	0.56 – 4.72	1.00 – 4.91	1.03 – 5.09

* All patients received PCA fentanyl as rescue medication.
†Pain intensity after the PCA was discontinued and the patient first reported "moderate" (5-6) to "severe" (7-10) pain on the NRS.
‡Overall: Mean pain intensity is mean over hours of the mean over patients (by hour). Pain intensity was assessed using an NRS (0 = no pain to 10 = pain as bad as you can imagine) at baseline and at 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 20, and 24 hours.
(Cross-references: Table 14.2.1.B; Appendices 16.2.6.1 and 16.1.9; tabular data for the rescue medication population are presented in Table 14.2.1.A.)

No significance levels are reported, nor even any standard deviations. Palladone was numerically best, but the sponsor concluded, "No clinically meaningful differences were observed between ITT treatment groups; similar results were seen in the rescue medication population."

The report focuses on putative differences between groups in the other primary variable, the amount of fentanyl used. (Each injection was 25 μg.) As indicated in the footnote, the sponsor claims a statistically significant difference between each of the hydromorphone

TABLE 11.1.1.
Study HD96-0505
Total Amount of Fentanyl Rescue Medication Over 24 Hours and Mean Number of Rescue Injections by Time Interval

	ITT Population		
	HHCR (N = 44)	HHIR (N = 44)	Placebo (N = 44)
Total Amount of Fentanyl Rescue Medication (μg) Over 24 Hours			
Mean*	1004.0	985.8	1186.9
Range	125 – 2225	50 – 2625	175 – 3600
Mean Number of Rescue Injections by Time Interval (Hours)			
0 – 3	7.75	7.55	8.36
3 – 6	5.91	5.64	7.49
6 – 12	10.14	9.73	12.35
12 – 24	16.74	16.91	20.19
0 – 6	13.66	13.18	15.68
6 – 12	10.14	9.73	12.35
12 – 18	8.44	8.40	10.40
18 – 24	8.50	8.51	9.79

* HHCR was significantly different from placebo (p = 0.0086), and HHIR was significantly different from placebo (p = 0.0029). There was no significant difference between the HHCR and HHIR treatment groups (p = 0.7126).
(Cross-references: Table 14.2.2.B; Appendices 16.2.6.2 and 16.1.9; corresponding tabular data for the rescue medication population are presented in Table 14.2.2.A.)

treatments and the placebo. I believe this analysis is erroneous, and the difference is not statistically significant. The problem being technical, detailed discussion is deferred to the appendix. Here I simply note that the standard errors of the mean fentanyl consumption in the three groups were 89, 96 and 121 μg. The two-sample t-test, which is appropriate, therefore gives p-values of 0.23 (two-sided) for the comparison of Palladone to placebo and 0.20 for IR hydromorphone to placebo.

Even taking the numbers at face value without regard to their statistical significance, they seem to me rather to call into question than to support the efficacy of the controlled-release formulation. In all time periods the use of fentanyl was about the same or less in the IR group than in the Palladone group. This is a comparison of 24 mg of a formulation meant to be released over 24 hours to 6 mg of a formulation largely gone from the body after 12 hours; and even 18 to 24 hours after administration, the lower, IR dose group did as well as the higher, controlled-release dose group. If Palladone was better than placebo, it was no better than IR hydromorphone even after IR hydromorphone was gone.

In my opinion, this study was unsuccessful in demonstrating the efficacy of Palladone and of IR hydromorphone relative to placebo.

CONCLUSIONS AND RECOMMENDATIONS

Hydromorphone is an old drug and a derivative of an ancient drug, opium. Palladone is a new drug because it is a controlled-release dosage form. It was appropriately compared in clinical trials to an immediate-release preparation of hydromorphone given more frequently. It provided similar relief of pain and similar use of rescue medication, with a similar time course, in a population that would have been expected to experience substantially more pain if they were not being treated with an effective drug. Safety is discussed in the medical officer's review, no statistical issues having arisen with regard to safety.

Thomas Permutt 7/26/99
Thomas Permutt, Ph.D.
Mathematical Statistician (Team Leader)

Concur: S. Edward Nevius, Ph.D. 7/11/99
Director, Division of Biometrics II

Statistical Review and Evaluation Appendix

NDA 21-044
Name of drug: Palladone — (hydromorphone hydrochloride) controlled-release capsules
Applicant: Purdue Pharma
Indication: pain
Documents reviewed: volumes 1, 2, 105-144;
electronic copies of same; electronic data and programs
Project manager: Nancy Chamberlin
Medical officer: Monte Scheinbaum, Ph.D., M.D.
Dates: received 29 December 1998; user-fee goal (10 months) 29 October 1999

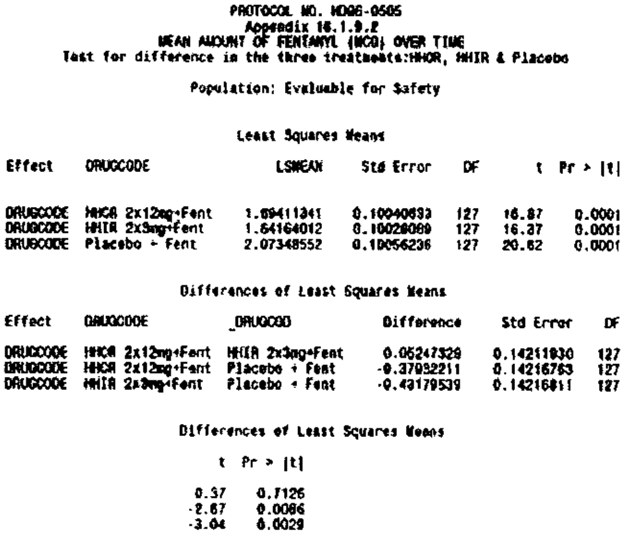
Reviewer: Thomas Permutt

The main efficacy result for study 505 was tested for significance by generalized least squares in a mixed-effects, repeated-measures analysis of variance. This analysis is not especially advantageous for testing treatment effects in a parallel-group study because the treatment effect is a between-subjects factor. It can produce correct results, however, if correctly applied. I think it has been incorrectly applied in this case; the standard errors of the estimated treatment effects have been grossly underestimated; and the significance levels have therefore been dramatically overstated.

The following SAS instructions were submitted by the sponsor. I ran them, and they do give the results reported by the sponsor.

```
proc mixed data=all;
class pno drugcode hour;
model injects=drugcode hour initial stabtime
  drugcode*hour/htype=1;
repeated/ type=ar(1) sub=pno(drugcode);
lsmeans drugcode/pdiff;
```

The relevant part of the output, as submitted by the sponsor, is as follows:



```
                          PROTOCOL NO. HD96-0505
                           Appendix 16.1.9.2
                   MEAN AMOUNT OF FENTANYL (MCG) OVER TIME
        Test for difference in the three treatments:HHOR, HHIR & Placebo

                     Population: Evaluable for Safety

                           Least Squares Means

Effect    DRUGCODE                LSMEAN    Std Error    DF        t  Pr > |t|

DRUGCODE  HHCR 2x12mg+Fent    1.69411341   0.10040693   127    16.87    0.0001
DRUGCODE  HHIR 2x3mg+Fent     1.64164012   0.10028089   127    16.37    0.0001
DRUGCODE  Placebo + Fent      2.07348552   0.10056236   127    20.62    0.0001

                    Differences of Least Squares Means

Effect    DRUGCODE          _DRUGCOD           Difference    Std Error    DF

DRUGCODE  HHCR 2x12mg+Fent  HHIR 2x3mg+Fent    0.05247329   0.14211930   127
DRUGCODE  HHCR 2x12mg+Fent  Placebo + Fent    -0.37932211   0.14216763   127
DRUGCODE  HHIR 2x3mg+Fent   Placebo + Fent    -0.43179539   0.14216811   127

                    Differences of Least Squares Means

                            t  Pr > |t|

                         0.37    0.7126
                        -2.67    0.0086
                        -3.04    0.0029
```

The title is incorrect; the results are in terms of injections per hour. The total amount of fentanyl in micrograms is 600 times this: 25 µg per dose times 24 hours. For the sake of comparison to the t-test discussed in the body of my review, I have recalculated the same analysis converting number of injections per hour to total fentanyl in 24 hours. The results, shown below, should be compared both to those above and to the simple means and standard errors for the three groups: 1004 ± 89 µg for Palladone, 986 ± 96 µg for IR hydromorphone, and 1187 ± 121 µg for placebo.

```
                        Least Squares Means

     Effect     DRUGCODE           LSMEAN      Std Error     DF        t   Pr > |t|

     DRUGCODE   806        1016.4686554    60.25472622    127    16.87     0.0001
     DRUGCODE   871         984.98393374   60.17946002    127    16.37     0.0001
     DRUGCODE   2085       1244.0642072    60.34836606    127    20.61     0.0001


                       Differences of Least Squares Means

Effect     DRUGCODE    _DRUGCOD      Difference       Std Error      DF          t    Pr > |t|

DRUGCODE   806         871          31.48472170     85.28704756     127      0.37      0.7126
DRUGCODE   806         2085       -227.5955517      85.31605346     127     -2.67      0.0086
DRUGCODE   871         2085       -259.0802734      85.31634726     127     -3.04      0.0029
```

The p-values are of course identical to the sponsor's: all that has changed is the units of measurement. The least-squares means are slightly different from the simple means, for two reasons. One is the presence of covariates (`initial` and `stabtime`) in the model, whose means are slightly different for the different treatment groups. The least-squares means are adjusted for these differences. The other is that the least-squares means use weighted rather than simple averages of the 24 hourly observations for each patient. In particular, in the first-order autoregressive, or AR(1), model specified by the sponsor, the first and last hours are weighted more heavily than each hour in between.

The main difference between this analysis and the simple one, however, is in the standard errors. The standard errors estimated by the generalized least-squares procedures are only about half the simple ones. Again, there are two reasons. The covariates explain some of the variability of the observations and so reduce the unexplained variance, on which the estimated standard error is based. This is good, but the effect is minor. The larger effect comes from the use of the estimated correlation matrix of the 24 observations within each patient. The AR(1) model imposes a structure on these correlations: the *i*th and *j*th observation for each patient are assumed to have correlation $\rho^{|i-j|}$; ρ was estimated to be 0.61. In this model, then, adjacent observations are estimated to be fairly highly correlated; but the correlation goes exponentially to zero as the observations become more separated in time. The first hour and the last would be estimated to have correlation 0.61^{23}, an extremely small number. In fact, the correlation between the first observation and the last for a given patient was 0.43[*]. Thus, the standard errors were calculated based on the assumption that observations separated by a few hours, even in the same patient, were approximately independent, so that there would be several essentially independent observations for each patient. In fact, *all* the observations for a given patient were substantially correlated. This was reasonably to be expected: many aspects of a given patient's condition may remain relatively constant over 24 hours, as compared to differences between patients.

The t-test allows for the correlation appropriately by treating each patient's average as a single observation. So does a repeated-measures analysis in which the subject effect is used as the error effect. While there is little advantage to the mixed-model analysis in estimating between-subject effects, it can also be applied, provided the correlation structure is realistic. The following code, for example, allows an additive subject effect in addition to the autoregressive structure. This produces a covariance structure in which the correlations decline exponentially, but to a nonzero floor. When allowed to estimate the correlations in this model, rather than forcing them quickly to zero, `proc mixed` produced estimates decreasing from 0.62 for adjacent observations to 0.46 for observations five hours apart, then remaining at 0.46.

[*] This is the simple correlation. Strictly speaking, what is wanted is the correlation of the residuals from the model fit. This may be expected to be somewhat less, as some correlation in the raw values is induced by the treatment effect itself: if patients on one treatment have all high scores and patients on another treatment have all low scores, then measurements on the same patient will be more alike than measurements on different patients on average because different patients may be from different groups. The effect is slight, because the fraction of variance explained by treatment in this case is modest. A formal estimate of the residual correlation is presented below, from the mixed model procedure without the exponential decline imposed by the autoregressive specification. It turns out to be 0.46 for observations separated by 5 hours or more.

```
proc mixed data=all1;
class pno drugcode hour;
model fentanyl=drugcode hour initial stabtime
   drugcode*hour/htype=1;
random intercept/ sub=pno vcorr;
repeated hour/ type=ar(1) sub=pno rcorr;
lsmeans drugcode/pdiff;
```

It produces the following results. Note that the (pooled) standard errors are similar to the simple ones, and that the differences between treatments are all now nonsignificant. (The closest to significance is between the active control and the placebo, not the test drug and the placebo.) This is the same model (same covariates) as the sponsor's; only the correlation structure is different.

```
                                Least Squares Means

   Effect      DRUGCODE          LSMEAN       Std Error      DF         t   Pr > |t|

   DRUGCODE   806         1034.1214728   107.32712347   2905      9.64     0.0001
   DRUGCODE   871         990.13890488   107.31033022   2905      9.23     0.0001
   DRUGCODE   2085        1279.4780283   107.43588873   2905     11.91     0.0001

                         Differences of Least Squares Means

Effect      DRUGCODE   _DRUGCOD     Difference       Std Error      DF          t   Pr > |t|

DRUGCODE   806         871          43.98256789   151.87524832   2905      0.29     0.7721
DRUGCODE   806         2085        -245.3565555   152.01386859   2905     -1.61     0.1066
DRUGCODE   871         2085        -289.3391234   151.97679264   2905     -1.90     0.0570
```

While I have used, in the body of my review, a simple alternative analysis (the t-test) because it is easy to understand as well as appropriate, I wish to make clear that this is not a question of a naïve vs. a more sophisticated analysis, nor of two equally correct analyses leading to different conclusions. It is, in my view, a case of the sophisticated analysis being carried out incorrectly. The generalized least-squares standard errors in this case are not alternative, valid estimates of the standard errors of the estimated treatment effects. Rather, they are incorrect estimates based on an assumption about correlation that is both implausible a priori and contradicted by the data.