CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

20-732

20-733

STATISTICAL REVIEW(S)

Dis. kit

Statistical Review and Evaluation
(Carcinogencity)

Date: APR 11 2000

NDA No: 20-733

Applicant: Reckitt & Colman Pharmaceuticals

Name of Drug: Suboxone

Document Reviewed: Protocol: Suboxone: 2 Year Oncogenicity Study in the Rat, undated.

Reviewing Pharmacologist: Anwar Goheer, Ph.D., HFD-170

Reviewing Statistician: Karl K. Lin, Ph.D., HFD-715

A. Introduction

Dr. Susmita Samantas of HFD-170 had requested the sponsor to submit this protocol of two year carcinogenicity study of Suboxone in the rat for his review. Dr. Anwar Goheer of HFD-170, the reviewing pharmacologist of this NDA, requested Division of Biometrics to perform a statistical review of the protocol. This reviewer made comments on only the components of the protocol which are related to study design and methods of statistical analysis of the data.

B. Reviewer's Comments

The reviewer has the following comments on the sponsor's submitted protocol:

1. The sponsor has proposed the use of a study design with dual control groups. Because the difficulties encountered in the analysis of data from a study using this design, the sponsor should be advised to combine the two identical control groups into one.

   There are arguments for and against using two identical controls in a study. The arguments for using this type of design are to use the results from the two identical controls as a quality control mechanism for identifying unsuspected biases in the study; and to evaluate the biological significance of increases in tumor incidence in the treated groups (i.e., true increases versus noises). However, as to be described below, there are encountered difficulties in statistical analysis of data from a study with dual controls. Also there is a concern about the possible existence of extra-binomial within-study variability between the two identical groups which will result in an inflation of false positive rates. These difficulties and concerns form the basis for the arguments against the use of designs with two identical controls.

Depending on mortality and tumor rates, data from dual control groups may or may not be combined for statistical analysis. If comparisons of the dual controls show no major differences in mortality and tumor rate, then the data of the two controls are combined to form a single control group in subsequent analyses (Haseman, Hajian, Crump, Selwyn, and Peace, 1990). If the data show evidences of major differences in mortality or tumor incidence between the identical controls, then three tests - control 1 versus treated groups, control 2 versus treated groups, and control 1 plus control 2 versus treated groups - for each tumor/organ combination can be carried out.

There is a question of how to interpret the study results this case. There are two approaches to the question. The first one is that a trend or a difference in tumor rate is considered as significant only if it is significant at a pre-specified level of significance in all the three tests. The second one is that the trend or the difference is considered as significant as long as any one of the three tests shows a significant result at the level of significance. The first approach may be very conservative in the sense that the null hypothesis will be rejected much less likely than it should be. On the other hand, the second approach may result in a high false positive error.

Currently there is no information about appropriate levels of significance for the tests of the above two approaches in order to maintain the 10% overall false positive rate used by the Center: The exact consequences of the above two approaches of dealing with dual controls with differences in tumor incidence or mortality are unknown. Unless all the three tests yield consistent results, i.e., all significant or all not significant from the statistical perspective, the most prudent way of interpreting the test results under this circumstance may be either to regard the study as providing equivocal evidence of carcinogenicity or to consider the study as inadequate for meaningful evaluation (Haseman, Hajian, Crump, Selwyn, and Peace, 1990).

Another reason in favor of the use of identical dual control groups is that the old Eruopean guideline (Committee for Proprietary Medical Products (CPMP), 1983) recommends the use of this study design. However, the European community has recently decided to drop the recommendation based on the considerations that the use of one control group saving animals and money, and avoids confusing results, and that it also allows flexibility in approach to design (Spindler, van der Laan, Ceuppens, Harling, Ettlin and Lima, 2000).

2. The sponsor has proposed to microscopically examine [redacted] during the study and of all animals in the control and high dose groups. With complete histopathological data of only the control and high dose groups, statistical tests for dose-response trend can not be performed. The sponsor should be asked to [redacted] the animals in the low and medium dose groups as well.

3. The sponsor has not proposed detailed methods for analyzing the data from the study. In the protocol the sponsor has mentioned simply that relevant data will be analyzed statistically using the SAS (1996) package. Statistical analysis of data of a

carcinogenicity study is rather complicated. Some of the study data such as body weight, food consumption, and organ weight, can be analyzed by the packaged procedures in SAS. For analysis of tumor and survival data, the sponsor is advised to the use the following statistical procedures.

3.1. Adjustment of Tumor Rates for Intercurrent Mortality

Intercurrent mortality refers to all deaths unrelated to a tumor being analyzed for evidence of carcinogenicity. Like human beings, older rodents have a many fold higher probability of developing or dying of tumors than those of younger age. Therefore, in the analysis of tumor data, it is essential to identify and adjust for possible differences in intercurrent mortality (or longevity) among treatment groups to eliminate or reduce biases caused by these differences.

Before analyzing the tumor data, the intercurrent mortality data should be routinely tested first to see if the survival distributions of the treatment groups are different. It is recommended by Peto, et al. (1980) that, whether or not survival among treatment groups is significantly different, tumor rates should routinely be adjusted for survival when presenting experimental results. The Cox test (Cox, 1972; Thomas, Breslow, and Gart, 1977); the generalized Wilcoxon or Kruskal-Wallis test (Breslow, 1970; Gehan, 1965; Thomas, Breslow, and Gart, 1977); and the Tarone trend tests (Tarone, 1975) are routinely used to test for heterogeneity in survival distributions and significant dose-response relationships (trends) in survival.

3.2. Statistical Analysis of Tumor Data with Information about Cause of Death, Tumor Lethality, but Without Multiple Sacrifices

One way to choose appropriate survival-adjusted methods in the analysis of tumor data is to base on the role that a tumor plays in causing the animal's death. Tumors can be classified as "incidental," "fatal," and "mortality-independent (or observable)" according to the contexts of observation described in Peto, et al. (1980). Tumors that are not directly or indirectly responsible for the animal's death, but are merely observed at the autopsy of the animal after it has died of an unrelated cause, are said to have been observed in an incidental context. Tumors that kill the animal, either directly or indirectly, are said to have been observed in a fatal context. Tumors, such as skin tumors, whose detection occurs at times other than when the animal dies are said to have been observed in a mortality-independent (or observable) context. To apply a survival-adjusted method correctly based on this piece of information, it is essential that the context of observation of a tumor be determined as accurately and subjectively as possible.

Different statistical techniques have been proposed for analyzing data of tumors which contain the information of contexts of observation (cause of death) of tumors.

The prevalence method described in the paper by Peto, et al. (1980) should be used in testing for positive trends in prevalence rates of incidental tumors. This method focuses on the age-specific tumor prevalence rates to correct for intercurrent mortality differences among treatment groups in the test for positive trends or differences in incidental tumors.

It is recommended that the death-rate method described in Peto, et al. (1980) be routinely used to test for the positive trend or difference in incidence of tumors observed in a fatal context.

When a tumor is observed in a fatal context for some animals and is also observed in an incidental context for other animals in the experiment, data for the incidental and fatal tumors should be analyzed separately by the prevalence and the death-rate methods. Results from the different methods can then be combined to yield an overall result. The combined overall result can be obtained simply by adding together either the separate observed frequencies, the expected frequencies, and the variances, or the separate T statistics and their variances.

Tumors observed in a mortality-independent context, such as skin tumors and mammary gland tumors, which are visible and/or can be detected by palpation in living animals, are routinely analyzed using the onset-rate method. The onset-rate method for mortality-independent tumors and the death-rate method for fatal tumors are essentially the same in principle except that the endpoint in the onset-rate method is the occurrence of such a tumor (e.g., skin tumor reaching some prespecified size) rather than the time or cause of the animal's death.

The above methods are based on normal approximation in the calculation of p-values. It is well known that the approximation results may not be stable and reliable, and tend to underestimate the exact p-values when the total numbers of tumor occurrence across treatment groups are small. In this situation, the exact permutation trend test (40) should be used to test for the positive trend (9). The exact permutation trend test is a generalization of the Fisher's exact test to a sequence of $2x(r+1)$ tables.

There are issues in the determination of the three contexts of observation of tumors, especially the first two contexts of most of occult tumors. Some people argue that the determination whether a tumor causes an animal's death is a rather complicate and subjective process. Very often it is difficult for a pathologist to classify accurately and objectively a tumor type as straight causing or not causing the animal's death. In practice, there is a continuum between these two extremes. That is, many tumors contribute ultimately to an animal's death, but are not instantly (or even rapidly lethal. Such tumors technically are neither 'incidental' or 'fatal' and it is not clear how such tumors should be regarded. Also even the information of contexts of observation is reliable and available, it will be overly simplistic to assume that all tumors of a given type are 100% fatal or 100% incidental. It is likely that the tumor type is a mixture of incidental and fatal tumors, that is, it is fatal to some but non-fatal to the other animals.

Alternative survival-adjusted statistical procedures such as the poly-k tests (Bailer and Portier, 1988) and the the ratio trend test (Bieler and Williams, 1993) which do not need such information have been developed and used for tumor data analysis because the complexity and subjectivity in the pathologist's determination of the cause of death of a tumor. The alternative procedures should be used to replace the procedures proposed in Peto, et al. (1980) in the analysis of tumor data in situations in which there is no such information available or such information although available is considered as not accurate enough.

C. References

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

Bieler GS, Williams RL: "Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity", Biometrics, 49, 793-801, 1993.

Breslow N: "A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship," Biometrics, 57, 579-594, 1970.

Committee for Proprietary Medical Products (1983), "Carcinogenic Potential, The Roles Governing Medical Products in the European Union", Vol. 38, Medical Products for Human Use, 63-71.

Cox DR: "Regression Models and Life Tables (with discussion)," Journal of Royal Statistical Society, Series B, 34, 187-220, 1972.

Haseman JK., Hajian G, Crump KS, Selwyn MR, Peace KE (1990), "Dual Control Groups in Rodent Carcinogenicity Studies", in Statistical Issues in Drug Research and Development, K. E. Peace, Editor, Marcel Dekker, New York.

Gehan EA: "A Generalized Wilcoxon Test for Comparing K Samples Subject to Unequal Patterns of Censorship," Biometrika, 52, 203-223, 1965.

Peto R, Pike MC, Day NE, Gray RG, Lee PN, Parish S, Peto J, Richards S, Wahrendorf J (1980), "Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-term Animal Experiments," In Long-term and Short-term Screening Assays for Carcinogens: An Critical Appraisal, World Health Organization.

Spindler P, van der Laan JW, Ceuppens P, Harling R, Ettlin R, Lima BS (2000), "Carcinogenicity Testing of Pharmaceuticals in the EU: A Workshop Report" submitted to Drug Information Journal.

Tarone RE: "Tests for Trend in Life Table Analysis," Biometrika, 62, 679-682, 1975.

Thomas DG, Breslow N, Gart JJ: "Trend and Homogeneity Analyses of Proportions and Life Table Data," Computer and Biomedical Research, 10, 373-381, 1977.

[/S/]

Karl K. Lin, Ph.D.
Expert Mathematical Statistician
(Applications in Pharmacology and Toxicology)

Concur: [/S/]

S. Edward Nevius, Ph.D.
Director, Division of Biometrics II

IND 58,653

cc: Original/NDA 20-733 File
HFD-170/ A Goheer, L Jean
HFD-715/Chron
HFD-715/E Nevius, T Permutt, K Lin
HFD-170 /S/

APPEARS THIS WAY ON ORIGINAL

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

Statistical Review and Evaluation

STABILITY STUDIES

NDA:	20-733/N-000-BC (9/3/02)
Name of drug:	Suboxone (buprenorphine/naloxone) sublingual tablets
Applicant:	Reckitt & Benckiser
Indication:	maintenance in —
Documents reviewed:	amended stability report (12 month data)
Project manager:	Sara Shepherd
Chemistry reviewer:	Ali Al-Hakim, Ph.D.
Dates:	letter 9/3/02
Statistical reviewer (team leader):	Thomas Permutt
Secondary reviewer:	Karl Lin, Ph.D.
Biometrics division director:	S. Edward Nevius, Ph.D.

Keywords: NDA review, stability

1 INTRODUCTION

This submission adds a 12-month timepoint to stability data at 25°C and 30°C for Suboxone (buprenorphine/naloxone) tablets, 2 mg and 8 mg of buprenorphine with 0.5 mg and 2 mg of naloxone, respectively, packed [redacted] in bottles. I previously reviewed data through [redacted] submitted 3/13/02. Expiration dating of [redacted] is requested for the [redacted] and [redacted] for the bottles.

[redacted] This review does not address those data, nor the question of whether data from the older formulation are applicable to the new one.

[redacted]

3 BOTTLES

Assays for buprenorphine and naloxone and for impurities were within specifications for all [redacted] batches at 12 months for both 2 mg and 8 mg tablets. Results of the FDA/STAT analysis are shown in the table below (p. 134). As indicated by the asterisks, in all cases except for the buprenorphine assay, batches were pooled based on nonsignificant differences in slope or in slope and intercept at level 0.25. The table does not distinguish cases where

common slopes and separate intercepts were used from those with common slopes and intercepts; a footnote indicating that the asterisks represent common slopes *and* intercepts appears to be incorrect. The decision to pool or not, however, appears to have been made correctly according to the usual criterion of testing at level 0.25.

Summary of the shelf life predictions made for Suboxone Z914 using the FDA STAT Package.

	Extrapolated Shelf Life (Months)					
	06001/246 Z914 A2	06001/249 Z914 B2	06001/252 Z914 C2	06001/247 Z914 A6	06001/250 Z914 B6	06001/253 Z914 C3
Specification Item						
Buprenorphine Assay						
Naloxone Assay						
Naloxone related impurities:						
Total Naloxone Impurities:						
Minimum Supportable Shelf-Life						

	25°C	30°C
Average Minimum Supportable Shelf-Life:		

* Indicates that the FDA STAT package has selected the Common Intercept and Common slope model to estimate the shelf-life. With this model, the data sets for each batch have similar values at the initial time-point and similar gradients, therefore the data is pooled. This means a single shelf life will be determined for all batches being tested and the data was plotted on a

The rest of the data was calculated using the Separate Intercept and Common slope model in which the shelf life is generated per batch, but the data from all batches is pooled for the calculation of the confidence intervals. This means a moreaccurate shelf-life is generated, but it may differ from batch to batch. Each batch was plotted on a separate graph.

In all cases the confidence limits remain within the specifications well beyond the requested. Level 2 dissolution testing was required only for batch at 12 months and met the standard. The extrapolation for , past real time is provisionally justified. It should be supported by real-time data when they become available.

APPEARS THIS WAY
ON ORIGINAL

BEST POSSIBLE COPY

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
9/25/02 03:16:11 PM
BIOMETRICS

Karl Lin
9/25/02 03:40:23 PM
BIOMETRICS
Concur with review

APPEARS THIS WAY
ON ORIGINAL

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

Statistical Review and Evaluation

STABILITY STUDIES

NDA:	20-733
Name of drug:	Suboxone (buprenorphine/naloxone) sublingual tablets
Applicant:	Reckitt & Benckiser
Indication:	[redacted]
Documents reviewed:	volumes 2, 5 of 5
Project manager:	Sara Shepherd
Chemistry reviewer:	Ali Al-Hakim, Ph.D.
Dates:	letter 3/13/02
Statistical reviewer (team leader):	Thomas Permutt
Secondary reviewer:	Karl Lin, Ph.D.
Biometrics division director:	S. Edward Nevius, Ph.D.

Keywords: NDA review, stability

1 INTRODUCTION

Suboxone is a combination product intended to mitigate the liability to abuse of buprenorphine. Naloxone is thought to have little effect when the product is administered sublingually, as directed. If the product were dissolved and injected, however, naloxone might precipitate rather than prevent opioid withdrawal symptoms. Compared to the single-entity buprenorphine product which is also under review (Subutex, NDA 20-732), therefore, Suboxone is thought to represent less risk of abuse by the patient or diversion and abuse by others.

This function of naloxone raises unusual issues with respect to stability. The potency of naloxone in the product when used as directed is expected and intended to be nil. Even with respect to abuse, the function is intended to be preventive, so that the amount of naloxone actually contained in any given dose of the product may not be at all critical. Still, specifications on naloxone content are a control on the quality of the product, particularly with respect to degradation products not individually controlled.

The submission reports data at [redacted] on two dosage strengths (2 and 8 mg of buprenorphine with 0.5 and 2 mg of naloxone, respectively, hereinafter identified by the buprenorphine content as 2 and 8 mg) in [redacted] package [redacted] and bottles). [redacted] batches of each were tested under three sets of storage conditions (25°C/60%RH, 30°C/60%RH, 40%/75%RH). Expiration dating of [redacted] and [redacted] in bottles is claimed.

Assays are reported for buprenorphine, for [redacted] identified buprenorphine-related impurities, and for the total of buprenorphine-related impurities; also for naloxone, for [redacted] identified naloxone-related impurities, and for the total of naloxone-related impurities. Statistical analysis was not performed for some of these impurities for the tablets in bottles because they were not detected or were below the limit of quantitation. Two replicate determinations are reported at each timepoint. The program FDA/STAT was used by the applicant to estimate expiration dates based on these data. The [redacted] batches were pooled if neither the slopes nor the intercepts were significantly different at level 0.25; this is indicated by asterisks in the tables. If the intercepts were significantly different but not the slopes, a model with common slope was used; this is not indicated in the tables. I checked the calculations in several cases and found them to be correct. The tables in this review are copied from the submission.

[redacted]

1 Page(s) Withheld

4 BOTTLES 2 MG

Results for the bottles are given in the table below (v. 5, p. 411). For the 2 mg tablets, stability again appears to be limited by the naloxone assay, but at 25° the statistical results indicate dating to [redacted] However, this is a long extrapolation from [redacted] of real time. Dating of [redacted] would appropriately reflect the uncertainty of long extrapolation. The 30° data also support dating at least to [redacted]

Summary of the shelf life predictions made for Suboxone Z914 using the FDA STAT Package.

	Extrapolated Shelf Life (Months)					
	06001/248 Z914 A2	06001/249 Z914 B2	06001/252 Z914 C2	06001/247 Z914 A8	06001/250 Z914 B8	06001/253 Z914 C8
Specification Item						
Buprenorphine Assay						
Naloxone Assay						
Naloxone related impurities						
[ ]						
Total Naloxone impurities:						
Minimum Supportable Shelf-Life						

5 BOTTLES 8 MG

For the 8 mg tablets in bottles, only [redacted] are indicated by the results at 25°, or [redacted] at 30°, and there is no explanation for the claim of [redacted] Extrapolation to [redacted] months appears to be justified.

6 CONCLUSIONS AND RECOMMENDATIONS

[ ]

For tablets in bottles, statistical analysis indicates dating to [redacted] as claimed, for the 2 mg tablets and [redacted] for the 8 mg tablets at 25°. However, it might be preferable to limit extrapolation to [redacted] beyond real time, or [redacted] for both strengths, and dating to [redacted] is supported by the 30° data as well as the 25° data. In any case, extrapolated dating should be considered provisional, to be confirmed by real-time data when available.

APPEARS THIS WAY
ON ORIGINAL

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
7/1/02 01:26:16 PM
BIOMETRICS

Karl Lin
7/1/02 03:10:39 PM
BIOMETRICS
Concur with review

APPEARS THIS WAY
ON ORIGINAL

Statistical Review and Evaluation

NDA 20-732, 20-733
Names of drugs: Subutex (buprenorphine sublingual tablets); Suboxone (buprenorphine/naloxone sublingual tablets)
Applicant: Reckitt & Colman
Indication:
Documents reviewed: NDA 20-733 volumes 1, 93, 111-113
Project manager: Tony Chite
Medical officer: Chang Lee, M.D.
Dates: NDA 20-733 received 7 June 1999, user fee goal 7 December 1999 (priority, 6 months)

Reviewer: Thomas Permutt

INTRODUCTION

Subutex and Suboxone are intended as alternatives to methadone or levo-alpha-acetylmethadol (LAAM), the only opioid drugs permitted by regulation under the Narcotic Addict Treatment Act to be used for maintenance therapy in opiate addicts. Buprenorphine is a mixed opioid agonist/antagonist. Naloxone, when given intravenously, is an antagonist, but it has very low oral bioavailability. It is believed that the addition of naloxone will deter abuse of buprenorphine by dissolving and injecting the tablets, as this would precipitate withdrawal symptoms in opiate addicts. It is hoped that the naloxone would have little effect on the safety and efficacy of the product used sublingually as directed because of low absorption of naloxone, although the kinetics with sublingual administration may be different than with oral administration. The combination, then, is not expected to be more effective than the single-ingredient product; rather, it is expected to be safer with respect to diversion and abuse but hoped to be about equally effective.

NDA 20-732 for Subutex was found to be approvable 30 June 1998. A resubmission was received 29 July 1999. NDA 20-733 for Suboxone, an original submission, was received 7 June 1999. As the two applications rely on the same clinical studies, they will be reviewed together here.

The three principal studies were all reviewed under NDA 20-732. My reviews are attached. The studies were:

- a methadone-controlled trial of buprenorphine sublingual *solution* (090 or CR88/130),
- a dose-controlled trial of buprenorphine sublingual *solution* (999a or CR92/099), and
- an incomplete (no naloxone arm) factorial trial of combination, mono-ingredient and placebo tablets (1008A).

An additional study (9912 or CR 99/102) will also be discussed briefly.

I noted two deficiencies in the review of the factorial trial. First, there was no analysis by race and sex. Second, the trial was halted on the advice of a monitoring board, but the documentation of this decision was insufficient to allay concerns about multiplicity arising from interim analysis. Both these deficiencies are addressed in the present submissions.

FACTORIAL TRIAL: DEMOGRAPHICS

As discussed in my earlier review, the primary endpoint from the standpoint of approvability is the percentage for each patient of urine samples free of opiates (other than buprenorphine, naloxone and their metabolites). This was what I called the big-denominator percentage, counting all missing samples as failures.

The table below shows the means and standard errors, along with sample sizes, of these percentages by treatment and sex and by treatment and race. Analysis by age was not done because there were no patients over 65 in the trial, nor are there many in the target population for this drug.

	Suboxone	Subutex	placebo
male	17 ± 3 (68)	21 ± 4 (70)	6 ± 2 (71)
female	20 ± 4 (41)	19 ± 5 (35)	6 ± 3 (38)
white	17 ± 3 (65)	20 ± 4 (62)	7 ± 2 (70)
black	19 ± 4 (32)	22 ± 6 (35)	4 ± 2 (25)
other	17 ± 6 (12)	15 ± 9 (8)	4 ± 3 (14)

after tables 13.3.1.2, 13.3.1.3, NDA 20-733 volume 93

Both active treatments were more effective than placebo in males and in females, and in whites, blacks and others.

FACTORIAL TRIAL: INTERIM ANALYSIS

Study 1008A was terminated early on the recommendation of a data monitoring board. This decision was not well documented in the report submitted previously to NDA 20-732. The present submission addresses this issue in detail.

There was no prospective plan for interim analysis. The data monitoring board first met when about half the patients had been entered. At that time the institutional review board suggested that the study be terminated because of ethical concerns arising out of already

nominally significant differences between treatments (coded A, B or C). The data monitoring board rejected the suggestion, partly on the grounds that there were many additional patients already under study but not reported in the interim analysis. Even if recruitment were terminated, these additional patients would have to be included in analysis and might, the board felt, change the results substantially.

It was agreed instead to reconvene around the time when all patients then entered would have finished the efficacy phase of the trial. At the time of the second meeting 301 of 384 proposed patients had been randomized. A statistical report to the data monitoring board pointed out that the results exceeded what it called Haybittle-Peto boundaries of three standard errors. It also calculated that if the results for the remaining patients were the same for all treatment groups, treatments B and C would still be significantly different from A. This was somewhat confusingly described as a "worst-case" analysis. It was the worst that could be expected if the treatment effects were in fact in the right direction, but that is what the study was intended to show; and even if so, random variation could produce worse results than this. In any event, the data monitoring board asked for the treatments to be unmasked. Finding that the two active groups were alike and very different from placebo, they recommended termination of the study.

In my opinion, the documentation in this submission of the interim analysis procedure, taken together with the strength of the results, is adequate to conclude that the difference between the treatments is not an artifact of multiple interim testing but may be interpreted at face value.

CONCLUSIONS AND RECOMMENDATIONS

In my previous review of NDA 20-732 (Subutex) as amended, I recommended against final approval until two deficiencies in documentation were corrected, one concerning demographic subgroups and the other concerning interim analysis. Both these deficiencies have been satisfactorily corrected. I believe that there is substantial evidence that sublingual buprenorphine is effective in promoting partial abstinence in opiate addicts.

Suboxone (NDA 20-733) is a fixed-ratio combination drug product. Sublingual buprenorphine has been shown to be effective in the intended use, and naloxone is added to deter abuse. Both active ingredients therefore make a contribution to the safe and effective use of the combination, as is required by the policy on combination drugs. Direct, quantitative evidence of the efficacy of the combination product comes only from a single study, however.

[/S/] 10/6/99

Thomas Permutt, Ph.D.
Mathematical Statistician (Team Leader)

[/S/] 9/6/99

Concur: Michael Welch, Ph.D.
Deputy Director, Division of Biometrics II

APPEARS THIS WAY
ON ORIGINAL

ATTACHMENT

Statistical Review and Evaluation

NDA 20-732

Date of review: 9 October 1997
By: Thomas Permutt

Name of drug: Subutex (buprenorphine) sublingual tablets
Applicant: Reckitt & Colman
Indication: —
Documents reviewed: volumes 1.52-1.59, received HFD-170 1 April 1997
electronic data
medical officer's review
Project manager: Bonnie McNeal
Medical reviewer: Monte Scheinbaum, Ph.D., M.D.

INTRODUCTION

Buprenorphine is an opioid analgesic. This NDA deals with a sublingual tablet proposed for —. Two clinical trials carried out in 1988 and 1992 are characterized by the applicant as pivotal efficacy studies. These two trials studied a sublingual solution of buprenorphine rather than the tablet that is proposed for marketing. This review discusses the two studies from the standpoint of efficacy. It draws no conclusions concerning the relevance of studies of the solution to approvability of the tablet; this problem is taken up in the medical and clinical pharmacology reviews.

Safety is discussed in the medical officer's review. There were a number of deaths and serious adverse events in the clinical trials of buprenorphine. They were considered by the medical officer to be consistent in kind and in frequency with expectations for the population under treatment.

An amendment with some data from clinical trials of the tablet was submitted 5 September 1997. The amendment will be the subject of a separate review.

STUDY 999A (CR92/099)

Study 999a, also called CR92/099, was a double-blind, parallel-group, 16-week trial of four doses of sublingual buprenorphine solution (1 mg, 4 mg, 8 mg, or 16 mg daily) in heroin addicts at 12 centers in the United States. Approximately 60 patients per center were randomized to the four treatment groups in approximately equal numbers, making about 180

patients per treatment. The primary purpose of the study was to demonstrate effectiveness of the 8 mg dose compared to the 1 mg dose, which was believed (but seems not to be presently believed by the applicant) to be essentially inactive. In either case, showing a clear difference in effects at different doses would be sufficient to establish activity of the drug.

THE PROTOCOL

The protocol discussed statistical analysis at length. There are indications of discomfort with the specified approach, however, even within the protocol. Different statistical methods were proposed "for medical review" and "for statistical review." FDA personnel were mentioned by name, and there was an extended discussion of a method that was finally dismissed as inappropriate.

Four primary measures of efficacy were proposed. Retention in the study, regardless of use of heroin, was considered to be a positive outcome in itself. Urine samples with less than 300 ng/ml morphine (a metabolite of heroin but not of buprenorphine) were to be classified as "clean," and clean urines were a second positive outcome. Global ratings of condition from week to week by the patient and by the clinician were also considered primary outcomes. No method for jointly interpreting the four outcomes was proposed.

A proportional hazards model was proposed to analyze retention in treatment, with effects for age, sex and center as well as treatment. Global ratings would "be analyzed by two factor analysis of variance of difference scores" (from baseline), but it is not clear what other factor besides treatment was meant, nor how results from individual weeks were to be jointly interpreted. Possibly a repeated-measures model was contemplated, so that the between-subjects effects would be based on average scores for each subject over time, perhaps for completers only since imputation of missing data was not discussed.

The proposed analyses of the urine data were the most complicated and seemingly the most contentious. Disagreement appears to have centered on the handling of dropouts. The protocol attributed to FDA a suggestion that missing samples be scored as not clean, but rejected that approach in favor of a multiple-failure survival model that appears to treat dropouts as uninformatively censored.

The protocol made very clear that the primary efficacy comparisons were to be between the 1 mg and 8 mg doses, with secondary information on dose response gained from the other doses. Thus, no concerns about multiplicity should arise with respect to dose comparisons.

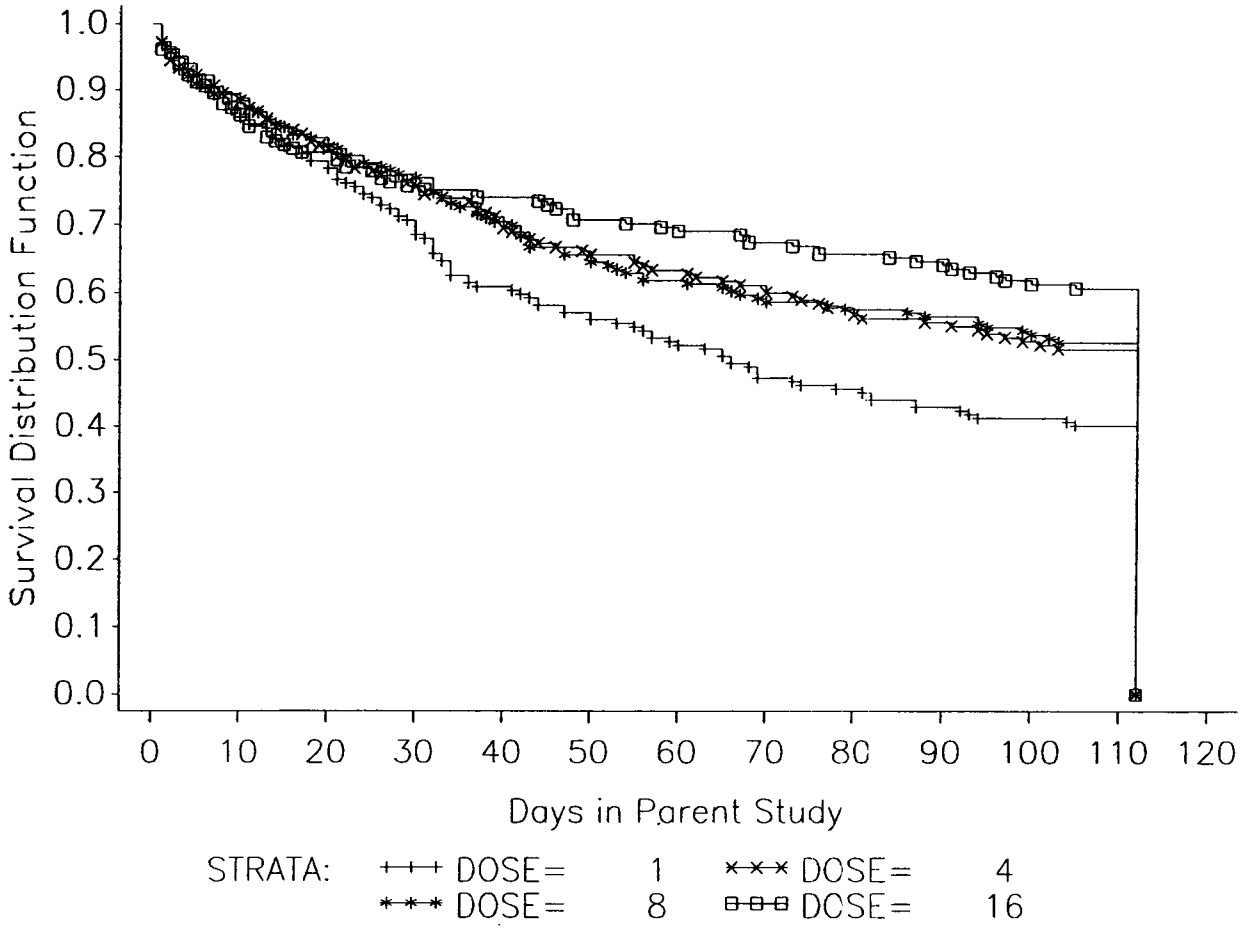
APPLICANT'S ANALYSIS—INTRODUCTION

None of the analyses discussed in the protocol are reported in the NDA. No reasons for the deviation are given. Instead, the application begins again with new primary variables and new methods of analysis. These new methods appear reasonably well justified, but by no means uniquely so. Thus, the text of the application does nothing to allay concerns that new methods

might have been chosen because other methods did not show what was wanted.

However, the application includes data submitted electronically in a small computer. After review of these data, I believe that the methods of analysis are not critical. Other analyses support the applicant's main conclusions so strongly as to remove concerns about post hoc choice of methods. The analyses specified in the protocol may have been considered unsound or ambiguous, and the applicant may have felt that it was more dignified to let the new analysis speak for itself rather than to rejoin the argument about the methods in the protocol. I believe it is likely that such considerations as these, rather than more favorable results, guided the choice of methods in the study report.

The global evaluations have been relegated to secondary status in the application. The four new primary outcomes are retention in treatment and three different functions of the urine data. The first of these three is the number of clean samples, divided by the total number of samples that should have been provided over the whole course of treatment. The second is the same numerator with a different denominator: the number of samples that should have been provided before a patient dropped out. In either case, missed samples while the patient remained in treatment are considered the same as dirty samples. In the first case, the score for a patient who dropped out would be the same as if all later samples were dirty; in the second case, as if they were clean and dirty in the same proportions as before the patient dropped out. These two analyses thus reflect in principle, if not in detail, the two kinds of analysis discussed in the protocol. The third analysis of the urine data is to look for a string of 13 consecutive nonmissing, clean samples for each patient, counting the numbers of patients who had and who did not have such a string. As samples were collected three times a week, such a string would probably represent about a month of continuous abstinence from heroin.



RETENTION IN TREATMENT

Figure 1 shows the survival curves for retention in treatment for the four dose groups. Fifty-three percent of the 8 mg group completed the study, as compared to 40 percent in the 1 mg group. The applicant reports a p-value for this pair of treatments of less than 0.01 by a proportional-hazards model, incorrectly described as "log-rank analysis." The model included effects of center, age, sex, and sex-by-age interaction, and possibly center-by-treatment interaction: the text says center-by-treatment interaction was included, but the appendix includes details only of a model without it. The applicant also reports a p-value of less than 0.02 comparing the fractions of completers, by logistic regression with the same covariates; in this case the appendix includes dozens of models with different sets of covariates.

A log-rank test (with no covariates) done by me gave a p-value of 0.02. The 2×2 chi-square test (completers vs. noncompleters × 1 mg vs. 8 mg) also has a p-value of 0.02. It is therefore fairly clear that the difference between the 1 mg and 8 mg groups is not an artifact of the choice of statistical methods, nor is it likely to be due to chance. There is convincing evidence of an

effect, albeit a rather modest one: only about 1 in 8 patients (53 − 40 = 13 percent) finished treatment on 8 mg but would not have finished on 1 mg.

CLEAN URINES

In this section I will discuss the three different analyses of the urine morphine data. These are: the number of clean urines as a fraction of the number of samples which should have been provided *before dropout*, which I will call the small-denominator percentage; the same number as a fraction of the number that should have been provided over the whole study (the big-denominator percentage); and the number of patients with a string of 13 clean samples. I will also discuss an analysis of the data for completers only.

SMALL-DENOMINATOR PERCENTAGE

The number of clean urines was counted for each patient, and divided by the number of samples there would have been if patients had given three samples per week until they dropped out. The means of these quotients (expressed as percentage) for the four dose groups, in increasing order, were 15, 25, 27 and 32; the standard error of the mean was approximately 2 in each case. The sponsor again analyzed these data by a complicated model with many covariates, in this case applying first an "empirical logistic" transformation to "normalize" the data. The transformation is neither well justified nor adequately explained (how are zeroes treated?). In view of the sample sizes, however, there need be little concern about the sampling distribution of the most straightforward statistic, the two-sample t-statistic. With a difference of 12 and standard errors of 2 the t-test gives a very small p-value for the comparison of the 1 mg and 8 mg groups. Thus, patients in the 8 mg group were more likely to give a clean sample while they remained in treatment than patients in the 1 mg group. Again the smallness of the effect is noteworthy: most of the patients in all the groups appear to have been using heroin most of the time.

BIG-DENOMINATOR PERCENTAGE

The mean numbers of clean urines as a percentage of all possible samples (i.e., treating all missing samples the same as dirty samples) were 12, 20, 22 and 28 in the four dose groups in increasing order. The standard error of each mean was 2. The sponsor's analysis, the same as for the small denominator, is again not well justified or explained. Again, however, the 1 mg and 8 mg groups are very significantly different by the t-test. Here again, the magnitude of the effect is not large, but its statistical significance is not in doubt.

CONSECUTIVE CLEAN URINES

The four dose groups had 7/184, 21/180, 17/186 and 34/181 patients with a string of 13 consecutive clean urines. The applicant reports a p-value of 0.04 for the comparison of 8 mg

to 1 mg by logistic regression, this time without covariates. I get the same p-value with a chi-square test, but 0.06 with Fisher's exact test.

If this variable were not so closely related to the others, I would have some concerns about any claim based on post hoc analysis of borderline significance. In fact, however, this is saying the same thing as the other analyses of urine data, in a possibly more intuitive and clinically relevant way. That is, patients on 8 mg were more likely to abstain from heroin than on 1 mg. Once again, many more patients in all groups failed than succeeded by this criterion, but 8 mg significantly increased the number of successes compared to 1 mg.

URINE DATA FOR COMPLETERS

The applicant emphasizes the four primary analyses as comparisons between randomized groups, with conservative handling of missing values. I agree, except insofar as the small-denominator percentage requires an implausible assumption (dropouts are the same after quitting treatment as they were before) to be directly interpretable. Even so, the agreement among all four of these measures indicates clearly that 8 mg was more effective than 1 mg in getting patients to remain in treatment and give clean samples. However, except for the small-denominator percentage, these four seemingly different measures all place a heavy emphasis on retention in treatment. A patient who dropped out would have a lower big-denominator percentage than one who stayed in, even if their heroin use were the same, and also would have more chance of giving 13 consecutive clean samples. In this respect the small-denominator data are reassuring, in that they indicate that the patients on 8 mg not only stayed in longer but also used less heroin while they stayed in.

Further reassurance comes from analysis of the percentage of clean urines for those who never dropped out. In this case the big and small denominators are the same; nonterminal missing samples were still treated the same as dirty samples. The four dose groups had mean clean percentages of 25, 35, 36 and 44, with standard errors of 3 in each group. The applicant declined to do any significance test because of the possibility of selection bias in analysis of completers only. I think this is sound policy, but I do note a substantial trend in the right direction. That is, the 8 mg patients were not only less likely to drop out, but they were also less likely to use heroin while in treatment than the 1 mg patients.

GLOBAL EVALUATIONS

The global evaluations by the patient and by the clinician were primary outcomes according to the protocol, but are now considered secondary by the applicant. On their own these outcomes would indeed not be strong indicators of effectiveness, so that they may reasonably be considered secondary. Nevertheless, because of their inclusion in the protocol, it is important to consider them briefly.

Ratings by both the patient and clinician were better for the 8 mm group than for the 1 mm

group, whether for completers only or including dropouts with last observation carried forward. The differences were modest, ranging from 7 to 10 points on a 100-point scale, but the standard error of each difference was only about 4. Thus, the global rating data are consistent with the more objective measures of outcome.

SUBGROUPS

Women were about as likely as men to remain in treatment, but were much less likely to give clean samples, regardless of dose. However, 8 mg was more effective than 1 mg for women as for men.

	1 mg	8 mg
completed:		
men	47/116 (41%)	65/120 (54%)
women	27/68 (40%)	33/66 (50%)
small-denom. percent clean (mean ± s. e.):		
men	18 ± 3	32 ± 3
women	10 ± 2	17 ± 3
big-denom. percent clean (mean ± s. e.):		
men	14 ± 2	26 ± 3
women	8 ± 2	14 ± 3

The dose effect was somewhat more pronounced for blacks and Hispanics than for non-Hispanic whites.

	1 mg	8 mg
completed:		
white non-Hispanic	35/87 (40%)	46/91 (51%)
black non-Hispanic	22/44 (50%)	22/39 (56%)

Hispanic	16/51 (31%)	28/53 (53%)
small-denom. percent clean (mean ± s. e.):		
white non-Hispanic	15 ± 3	24 ± 3
black non-Hispanic	17 ± 4	34 ± 6
Hispanic	12 ± 4	28 ± 4
big-denom. percent clean (mean ± s. e.):		
white non-Hispanic	12 ± 2	20 ± 3
black non-Hispanic	13 ± 4	26 ± 5
Hispanic	9 ± 3	22 ± 4

I did not do any analysis by age because the population was fairly homogeneous with respect to age, 88 percent of the patients being between 25 and 50 years old.

To check consistency of effects across centers, I fit analysis of variance models with effects of dose (1 vs. 8 mg), center and center-by-dose interaction for both percentages and for retention to the end of the trial (as 1 or 0). In each case the center main effect was highly significant, suggesting differences in populations at the different centers. The interactions were not significant, and the sums of squares for interaction were smaller than for dose, indicating absence of qualitative interactions.

CONCLUSIONS

Sublingual buprenorphine solution was effective in keeping heroin addicts in treatment and in reducing their use of heroin. The effects were modest but clearly statistically significant. A daily dose of 8 mg was more effective than 1 mg, the primary comparison. Also, 16 mg was somewhat more effective than 8 mg, but not much difference was seen between 4 mg and 8 mg.

STUDY 090 (CR88/130)

Study 090 was a comparison of sublingual buprenorphine solution (8 mg q.d.) to two doses of oral methadone (20 mg or 60 mg q.d.) in maintenance and then detoxification of heroin

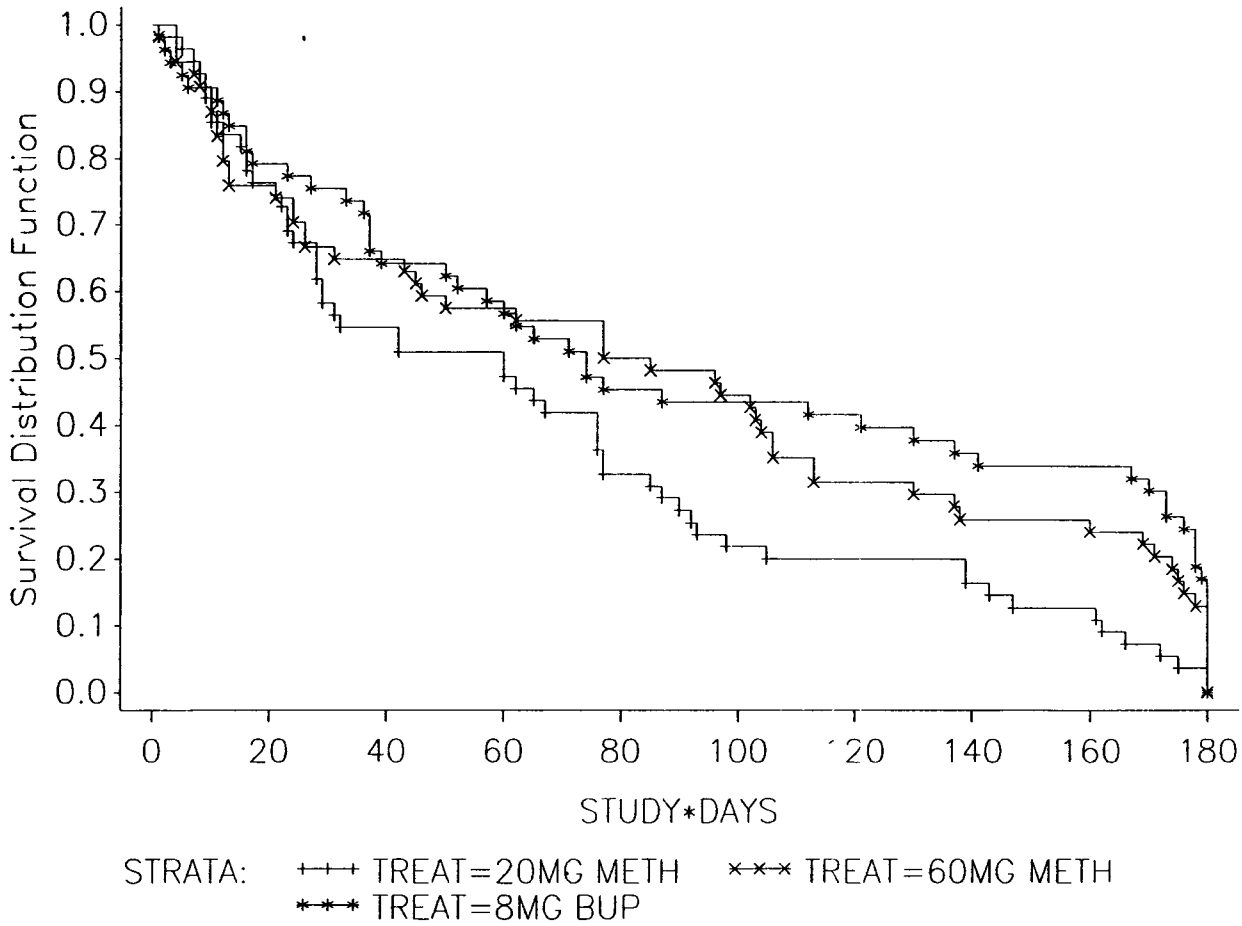
addicts. The submission focuses on the 17-week (including one week of induction) maintenance phase. The study was conducted at the National Institute on Drug Abuse's Addiction Research Center in Baltimore. One hundred sixty-two patients were randomized in approximately equal numbers to the three treatments. The study is described as a double-blind, double-dummy trial, but may not be quite so. Patients were monitored by an unblinded clinician for adverse events. If these were judged to be intolerable, the dose was halved; the patient continued in treatment but was considered a dropout for analysis of efficacy. Thus, an unblinded observer could affect the efficacy data. In fact, however, only four patients were dropped in this way, one on buprenorphine and three on the higher dose of methadone, so that the results could not have been substantially affected.

The protocol was vague as to endpoints and statistical methods. The report explains that the trial was not foreseen as an essential part of an NDA. The report focuses on the same outcome measures with the same analysis as study 999a. As stated above, I believe that analysis is sound. Furthermore, analyzing both trials by the same methods strengthens their corroboration of each other, and also tends to alleviate concerns that the methods could have been tailored to the results. Finally, the results are again convincing enough that the choice of methods appears not to be critical.

In this trial there does arise a question of multiple comparisons among the three treatment groups. I think it is reasonable to consider the comparison of buprenorphine to the lower dose of methadone as the primary analysis for the purpose of demonstrating efficacy. That is, if buprenorphine were not better than methadone 20 mg, it would be hard to see this as a successful trial regardless of the other comparisons. From the standpoint of hypothesis testing, therefore, I think that this comparison can be interpreted without adjustment. The comparison with the higher dose of methadone is also of qualitative interest. Neither the submission nor this review, however, formally addresses the question of superiority or equivalence of buprenorphine to methadone 60 mg, which should therefore not be considered to have been established in this study.

RETENTION IN TREATMENT

Figure 2 shows the survival curves for retention in treatment for the three treatment groups.



Twenty-two of 53 patients on buprenorphine completed the maintenance phase, as compared to 11 of 55 on methadone 20 mg. This difference was statistically significant ($p = 0.02$, chi-square test). The two survival curves were also significantly different for the maintenance phase by the log-rank test ($p=0.04$). These and all significance tests in the review of this study were done by me. I emphasize these tests in the review because they are straightforward and methodologically sound. As in study 999a, the applicant's analysis involved adjustments that may have been determined after the fact, are not convincingly justified, and are not necessary in my view. I therefore do not discuss the applicant's calculations in detail, but I note that they lead to the same conclusions.

ATTACHMENT

CLEAN URINES

Clean urines in this case were defined as less than 300 ng/ml of opiates or metabolites of opiates, presumably excluding methadone and buprenorphine and their metabolites. The small-denominator percentages were (mean ± standard error) 41 ± 5 for buprenorphine, 24 ± 4 for methadone 20 mg, and 30 ± 5 for methadone 60 mg. The first two groups were significantly different ($p = 0.008$, t-test). The big-denominator percentages were 34 ± 5 for buprenorphine, 18 ± 4 for methadone 20 mg, and 23 ± 4 for methadone 60 mg. Again, the first two groups were significantly different ($p = 0.01$, t-test). Ten of 53 patients on buprenorphine, 4 of 55 on methadone 20 mg, and 7 of 54 on methadone 60 mg gave 13 consecutive clean samples. The first two groups were significantly different ($p = 0.02$, chi-square or Fisher's exact test).

SUBGROUPS

Men and blacks were less likely to remain in treatment on methadone 20 mg, so that the efficacy of buprenorphine relative to methadone 20 mg was somewhat greater than for women and whites. However, buprenorphine appeared to be effective both for men and women and both for whites and blacks. Again, I did not do separate analyses by age because the study population consisted mainly of young adults.

	buprenorphine 8 mg	methadone 20 mg
completed:		
men	15/38 (39%)	5/38 (13%)
women	7/15 (47%)	6/17 (35%)
small-denom. percent clean (mean ± s. e.):		
men	47 ± 6	33 ± 5
women	49 ± 7	25 ± 6
big-denom. percent clean (mean ± s. e.):		
men	35 ± 6	15 ± 4
women	34 ± 8	16 ± 5

		buprenorphine 8 mg	methadone 20 mg
completed:			
	white	9/33 (27%)	9/31 (29%)
	black	12/18 (67%)	2/23 (8%)
small-denom. percent clean (mean ± s. e.):			
	white	45 ± 7	30 ± 6
	black	53 ± 8	32 ± 6
big-denom. percent clean (mean ± s. e.):			
	white	28 ± 6	17 ± 4
	black	43 ± 8	14 ± 3

CONCLUSIONS

Sublingual buprenorphine solution (8 mg q.d.) was more effective than methadone (20 mg q.d.) in keeping heroin addicts in treatment and in reducing their use of opiates while in treatment. The effectiveness of buprenorphine was in the same range as methadone 60 mg q.d., but neither superiority nor equivalence has been demonstrated.

SUMMARY CONCLUSIONS

Buprenorphine sublingual solution (8 mg q.d.) was shown in two active-controlled studies to be more effective than comparator drugs (buprenorphine 1 mg or methadone 20 mg) in keeping addicts in maintenance treatment and off heroin. It appeared to be effective both in men and in women, and in whites, blacks and Hispanics. Buprenorphine was also compared to methadone 60 mg, but neither difference nor equivalence was convincingly demonstrated. A dose-response relationship was shown, with buprenorphine 16 mg somewhat more effective than 8 mg. No recommendation is made concerning the relevance of studies of the solution to approvability of the tablet formulation.

ATTACHMENT

Thomas Permutt, Ph.D.
Mathematical Statistician
Team Leader,
Division of Anesthetic, Critical Care
and Addiction Drug Products

concur:
Nancy Smith, Ph.D.
Director, Division of Biometrics III

ATTACHMENT

Statistical Review and Evaluation

NDA 20-732 (amendment)

Date of review: 13 January 1998
By: Thomas Permutt

Name of drug: Subutex (buprenorphine) sublingual tablets
Applicant: Reckitt & Colman
Indication: —
Document reviewed: volumes 3.1, 3.6, 5 September 1997
Project manager: Bonnie McNeal
Medical reviewer: Monte Scheinbaum, Ph.D., M.D.

The principal clinical trials in NDA 20-732 were discussed in my review of 9 October 1997. I concluded that they offered substantial evidence of efficacy of the sublingual solution that was tested, but that they might not be sufficient to approve the proposed marketing of a different formulation, a tablet. This amendment to the NDA includes a brief report of a placebo-controlled trial involving both buprenorphine sublingual tablets and combination buprenorphine-naloxone tablets. The study (1008A) was primarily designed as a test of the combination product. However, it also appears to contain the strongest direct evidence of the efficacy of the single-drug tablet.

The report is a copy of the last periodic report to the study's data and safety monitoring board. The board recommended the study be terminated because of the differences in efficacy among the three treatments, and this recommendation was accepted.

The principal result was a difference in the average percentage of urine samples that were negative for opiates: 5 ± 2 percent (mean ± standard error) for Group A, 20 ± 3 percent for Group B, and 20 ± 3 percent for Group C. Group A was identified in the amendment (but not in the report to the DSMB) as placebo, but Groups B and C remain unidentified.

Whichever group turns out to be the single-drug tablet, the difference from placebo is surely statistically significant. The brief report, however, does not constitute the full report of a clinical investigation required by statute. Data on individual patients are not included; there is no discussion of the protocol and planned analysis; there is no analysis by race or sex. If this study had been identified as a pivotal trial in the original application, the inadequacy of the report would have been reason to refuse to file the application. It is also insufficient, therefore, to substantially support a conclusion of efficacy of the sublingual tablet.

I do not mean to imply any conclusion about the study itself, but only about the insufficiency

of this early report. The study should have a full review if the application is amended further by a full report.

Thomas Permutt, Ph.D.
Mathematical Statistician
Team Leader,
Division of Anesthetic, Critical Care
and Addiction Drug Products

concur:
Nancy Smith, Ph.D.
Director, Division of Biometrics III

APPEARS THIS WAY
ON ORIGINAL

ATTACHMENT

Statistical Review and Evaluation—Correction

NDA 20-732

Date of review: 27 January 1998
By: Thomas Permutt

Name of drug: Subutex (buprenorphine) sublingual tablets
Applicant: Reckitt & Colman
Indication: —
Document reviewed: volume 1.53, received HFD-170 1 April 1997
electronic data
medical officers' reviews
Project manager: Bonnie McNeal
Medical reviewers: Monte Scheinbaum, Ph.D., M.D.; Celia Winchell, M.D. (team leader)

In her secondary review as medical team leader, Dr. Winchell noted some numerical discrepancies between Dr. Scheinbaum's review and mine of 9 October 1997 with respect to numbers of "clean" (opiate-negative) urines in study 090 (CR88/130). My numbers were incorrectly drawn from tables in the NDA referring to *cocaine*-negative urines. The comparisons between treatments are qualitatively similar, and my conclusions are unaffected. The incorrect paragraph follows:

> Clean urines in this case were defined as less than 300 ng/ml of opiates or metabolites of opiates, presumably excluding methadone and buprenorphine and their metabolites. The small-denominator percentages were (mean ± standard error) 41 ± 5 for buprenorphine, 24 ± 4 for methadone 20 mg, and 30 ± 5 for methadone 60 mg. The first two groups were significantly different (p = 0.008, t-test). The big-denominator percentages were 34 ± 5 for buprenorphine, 18 ± 4 for methadone 20 mg, and 23 ± 4 for methadone 60 mg. Again, the first two groups were significantly different (p = 0.01, t-test). Ten of 53 patients on buprenorphine, 4 of 55 on methadone 20 mg, and 7 of 54 on methadone 60 mg gave 13 consecutive clean samples. The first two groups were significantly different (p = 0.02, chi-square or Fisher's exact test).

It should read:

Clean urines in this case were defined as less than 300 ng/ml of opiates or metabolites of opiates, presumably excluding methadone and buprenorphine and their metabolites. The small-denominator percentages were (mean ± standard error) 48 ± 5 for buprenorphine, 31 ± 4 for methadone 20 mg, and 42 ± 5 for methadone 60 mg. The first two groups were significantly different (p = 0.003, t-test). The big-denominator percentages were 34 ± 5 for buprenorphine, 15 ± 3 for methadone 20 mg, and 27 ± 4 for methadone 60 mg. Again, the

first two groups were significantly different ($p = 0.001$, t-test). Fourteen of 53 patients on buprenorphine, 1 of 55 on methadone 20 mg, and 7 of 54 on methadone 60 mg gave 13 consecutive clean samples. The first two groups were significantly different ($p = 0.0002$, chi-Fisher's exact test).

Thomas Permutt, Ph.D.
Mathematical Statistician (Team Leader)

concur:
Michael Welch, Ph.D.
Acting Director, Division of Biometrics III

ATTACHMENT

Statistical Review and Evaluation

NDA 20-732 (amendment 2)

Date of review: 25 March 1998
By: Thomas Permutt

Name of drug: Subutex (buprenorphine) sublingual tablets
Applicant: Reckitt & Colman
Indication: —
Document reviewed: amendment (3 volumes) dated 20 March 1988, no serial number
Project manager: Bonnie McNeal
Medical reviewer: Monte Scheinbaum, Ph.D., M.D.

INTRODUCTION

This submission is a report of study 1008A, which was briefly reported earlier. That preliminary report was the subject of my review of 13 January 1998. The study was a three-arm, double-blind, four-week, parallel-group trial of buprenorphine/naloxone combination sublingual tablets, buprenorphine tablets and placebo tablets. It was designed primarily to support a marketing application for the combination product, which has not yet been submitted. The report is now submitted to the NDA which is under review for the buprenorphine (mono) tablet.

Three hundred twenty-six opiate abusers were randomized in approximately equal numbers to treatment with buprenorphine 16 mg (two 8 mg sublingual tablets) q.d., buprenorphine 16 mg and naloxone 4 mg q.d., or placebo. Sixty-five percent were male; 60 percent were white, 29 percent black and 8 percent Hispanic. The study was carried out at eight sites in the United States. The centers were not very imbalanced in size: the largest had 16 patients on buprenorphine (mono) and the smallest, nine.

The protocol called for 384 patients, but the study was stopped early on the recommendation of a Data Monitoring Board. The protocol also specified the comparison of the combination product to placebo as primary. Two primary measures of efficacy were contemplated: the percentage of urine samples free of opioids (other than buprenorphine), and craving for opiates reported by the patients.

CONCERNS RELATED TO INTERIM ANALYSIS

The submission describes the decision to stop the study in these terms:

> Enrollment into Study 1008A was closed at the recommendation of a Data Monitoring Board and the CSPCC's Human Rights Committee. On 17 July 1997, the committee recommended that recruitment into the study be stopped because the efficacy study had achieved its goal of determining

> that the buprenorphine/naloxone combination product was superior to placebo. These differences were highly significant ($p<0.001$) for both of the primary outcome measures. Furthermore, it was determined that the probability of finding no differences between the two treatments under the worst case scenario if recruitment were allowed to go until its regularly scheduled time was $p<0.005$.

No further explanation is given. The protocol described a recommendation that was to be made to the board as to formal interim analysis, protecting the overall probability of error, but it is not reported whether the board followed this recommendation. It is not reported how many other interim analyses were performed. No explanation is given of the "worst case scenario" and the computation under it. The report to the board, which was submitted in the previous amendment, gave data on 203 patients, compared to the 326 who appear to have been enrolled by the time recruitment was actually stopped.

The protocol clearly indicates that the sponsor was aware of the problems connected with interim analysis and intended to use an acceptable method to deal with them. Furthermore, the results are strong enough to suggest that they are unlikely to be an artifact of interim analysis. In this review, I will base my conclusions and recommendations on the assumption that the problems of interim analysis have been dealt with correctly. I would hesitate to recommend final approval, however, without documentation of this.

CLEAN URINES

A straightforward analysis of urine samples negative for opiates ("clean") was proposed as primary in the protocol. Urine samples were collected three times a week. The number of clean samples for each patient was divided by the total number of samples that should have been provided, and the quotient was expressed as a percentage. This is tantamount to imputing a value "not clean" to any missing sample. This method is appropriately conservative, and is also consistent with the analysis of the two trials of buprenorphine sublingual solution discussed in my review of 9 October 1997. The means for the treatment groups were compared by pairwise two-sample z-tests.

The mean percent clean was 16 ± 2 (mean ± standard error) for the combination, 19 ± 3 for buprenorphine alone, and 5 ± 2 for placebo. The difference between buprenorphine and placebo was statistically significant, with several zeroes. The centers were variable: the best mean for buprenorphine was 35 percent and the worst 7 percent. The treatment appeared to be effective at all centers, however, with the ratio of the buprenorphine mean to the placebo mean being nearly 2 or better at each center. Analyses by race, sex and age were not submitted. The target population may not include substantial numbers of elderly people, so that analysis by age may be unnecessary, but analyses by race and sex are essential.

In view of the strength of the results, any theoretical concern about multiple, pairwise comparisons can be discounted. Buprenorphine was unquestionably better than placebo with respect to this preplanned, sound criterion of efficacy.

CRAVING

Patients were asked daily to rate their craving for opiates on a 100 mm visual analog scale, from "no craving" to "maximum craving ever experienced." The method for analyzing these data was described as follows:

> Assuming the missing data occurred completely at random, using either the BMDP-5V or the SAS PROC MIXED procedures would be appropriate. This program provides for a flexible choice of mean (covariate) structure, specified in terms of between-subject and within-subject covariates. This procedure uses the expectation-maximization (EM) algorithm where each EM step increases the likelihood of the unknown parameters given the observable data. In practice, each E step often corresponds to a form of imputation of the missing data, thus, providing a link between maximum likelihood and imputation methods.

Beyond this vague description of what analysis might be carried out and what the advantages would be, the submission does not indicate precisely what analysis *was* performed. Results are reported by week of the study, and "baseline" also appears in tables, so that it seems most likely that a model with dummy variables for week and with baseline as a covariate was used. This amounts approximately to averaging the data for a week for each patient and then comparing the means between groups, adjusting for baseline, for each treatment group for each week. This would be a reasonable approach, except that the assumption that data are missing at random is implausible: patients who do not show up for their daily dose of opiate maintenance are likely to have a different experience, particularly as concerns craving, than those who do. Imputation of baseline (relatively high) scores for missing data might be safer.

The "adjusted" average craving scores diverged from about 60 in each treatment group at baseline to 29 ± 3 (standard error), 38 ± 3, and 56 ± 3 for the combination, monotherapy, and placebo groups, respectively. The differences between the two buprenorphine arms and the placebo arm, taken at face value, were again highly significant statistically. While the report is ambiguous as to precisely what was done, the results are extreme enough that it seems unlikely that *any* analysis consistent with the description could produce qualitatively different results.

In any event, regardless of the protocol, I do not think this analysis is primary as far as approvability of the drug product is concerned. It seems to me that if the active drug were not different from placebo with respect to observable abstinence from opiate abuse (which it is), craving would not matter very much. Conversely, if the drug were successful in promoting abstinence but subjects failed to report a difference in craving, this would still be seen as evidence of effectiveness. I believe craving was called "primary" to set it apart from other, secondary endpoints from the standpoint of scientific and promotional claims; but, as regards approvability, it should be regarded as secondary.

CONCLUSIONS AND RECOMMENDATIONS

I earlier reviewed two studies (Study 999a or CR92/099, and Study 090 or CR88/130) showing efficacy of sublingual buprenorphine in a different formulation. The new Study 1008A appears from this submission to give substantial evidence of efficacy of the tablet formulation which is

proposed for marketing. Safety is discussed in the medical officer's review.

Considering both the earlier studies of the solution and the present study of the buprenorphine (mono) tablet, I believe that sufficient information exists to find that the tablet is effective in reducing the use of narcotics by addicts. I do not recommend such a finding at this time, however, in the absence of further documentation. First, the procedure followed by the Data Monitoring Board in recommending termination of enrollment should be explained. Second, analysis of the primary outcome (clean urines) by race and sex should be supplied.

Thomas Permutt, Ph.D.
Mathematical Statistician
Team Leader,
Division of Anesthetic, Critical Care
and Addiction Drug Products

concur:
Michael Welch, Ph.D.
Acting Director, Division of Biometrics III