CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER: NDA 20-726

STATISTICAL REVIEW(S)

Statistical Review and Evaluation (carcinogenicity)

DATE: 6/9/97 JUN 13 1997

NDA#: 20-726

APPLICANT: Ciba-Geigy Corporation

NAME OF DRUG: Femara (Letrozole)

DOCUMENTS REVIEWED: Volume 1.25 and Volume 1.14 Containing Data, Results, and Study Reports of the Rat and Mouse Studies, and One Volume Dated 07/22/96 Containing the Data Layout and Diskettes for these Studies.

I. Background

Dr. M. Brower (HFD-150) requested from the Division of Biometrics I a statistical review of the rat and mouse studies data as well as an evaluation of the sponsor's findings.

II. The Rat Study

II.a. Design

The product was studied for 104 weeks in Sprague Dawley rats. For each sex four groups of 60 animals each received the compound via gavage at level of 0.0, 0.1, 1.0 and 10.0 mg/kg/day. Animals dying during the study or at final sacrifice were necropsied and all tissues were examined for all animals.

II.b. Sponsor's Analyses of the Rat Study

Survival Analysis

There were 12 animals of each sex which were sacrificed early due to various technical reasons. The sponsor treated these as censored in the lifetable analyses. Kaplan-Meier estimates were plotted and a two-sided Mantel-Cox logrank test was used to test for trend. If the two-sided trend test was found statistically significant at $\alpha=.05$, a one-sided trend test was used excluding one dose at a time, starting with the highest, until a non-significant result was achieved or all comparisons were exhausted. The sponsor did not observe a significant trend in mortality for either sex.

Tumor Data Analysis

The sponsor analyzed all neoplastic lesions using a one-sided trend test adjusted for possible mortality differences. Fatal and incidental lesions were analyzed according to their context. The sponsor favored the use of logistic regression procedures as suggested by Dinse and Lagakos. The incidental and fatal components of the analyses were statistically combined. If the overall trend test was statistically significant the highest dose group was excluded and the analysis repeated. This approach was continued until a trend test was non-significant or all possible tests had been performed. In this manner, the sponsor observed a significant trend at p=.028 for benign hepatocellular adenomas of the liver in male rats. Malignant hepatocellular carcinomas of the liver also occurred in these animals but were not statistically significant and when combined with the hepatocellular adenomas the trend test was no longer significant. No other tumor findings indicated a positive trend among the male rats. Among the female rats benign gonadal stromal tumors of the ovary occurred only in the high dose and resulted in a significant test statistic with p=.01. No other neoplastic lesions exhibited a statistically significant trend test.

II.c. Reviewer's Analyses

Survival Analysis

The sponsor's survival analyses are acceptable. The summary mortality findings of the rats are given in Table 1 and the sponsor's survival curves are reproduced in Figures 1-2.

Tumor Data Analysis

The sponsor's statistical approach to analyzing tumor data is somewhat different than that used routinely in the Divisions of Biometrics. This reviewer therefore re-analyzed these data to ensure consistency across reviews. The tumor data were re-analyzed using pre-set time intervals and the methods described in the paper of Peto et al. (Guidelines for simple sensitive significance test for carcinogenic effects in long-term animal experiments, Long term and short term screening assays for carcinogens: A critical appraisal, International Agency for Research against Cancer Monographs, Annex to Supplement, WHO, Geneva, 311-426, 1980) and the method of the exact permutation trend test developed by the Division of Biometrics. The following criteria for the levels of significance ensure a false positive rate of about ten percent for the trend tests of the usual two-species two-sexes studies: Tumors with less than 1.00% occurrence in the control group are considered rare and a positive trend test is statistically significant when it reaches a p-value of < .025 (one-sided). Higher tumor occurrences in the control group are considered common for these animals and a positive trend is statistically significant when its p-value is less than .005 (one-sided). An approximate permutation trend test is used when fatal and incidental tumors of the same kind are combined and have overlapping time intervals. All tests are survival adjusted and treatment groups are weighted by the actual dose levels. For tissues where not all dose groups were fully necropsied only pairwise comparisons between the high and control groups were performed.

The exact permutation trend test for incidental hepatocelluar adenoma of the liver in male rats had a p-value of .110 which is well above the criterion of α=.025 for rare tumors. The p-value associated with ovarian gonadal stromal tumors was .010, reproducing the sponsor's findings and being significant at the α-level of .025 for rare tumors. This reviewer could not identify the reason why the findings in hepatocellular adenomas of the liver in males was much less significant than in the sponsor's analysis.

II.d. Validity of the Male Rat Study

As there are no statistically significant tumor trends among the male rats, this reviewer evaluated the validity of the study. For this, two questions need to be answered (Haseman, Statistical Issues in the Design, Analysis and Interpretation of Animal Carcinogenicity Studies, Environmental Health Perspectives, Vol 58, pp 385-392, 1984):

(i) Were enough animals exposed for a sufficient length of time to allow for late developing tumors?

(ii) Were the dose levels high enough to pose a reasonable tumor challenge in the animals?

The following are some rules of thumb as suggested by experts in the field: Haseman (Issues in Carcinogenicity Testing: Dose Selection, Fundamental and Applied Toxicology, Vol 5, pp 66-78, 1985) had found that on the average, approximately 50 % of the animals in the high dose group survived the two-year study. In a personal communication with Dr. Karl Lin of HFD-715, he suggested that 50 % survival of the usual 50 initial animals in the high dose group between weeks 80-90 would be considered as a sufficient number and adequate exposure. Chu, Cueto, and Ward (Factors in the Evaluation of 200 National Cancer Institute Carcinogen Bioassays, Journal of Toxicology and Environmental Health, Vol 8, pp 251-280, 1981) proposed that "To be considered adequate, an experiment that has not shown a chemical to be carcinogenic should have groups of animals with greater than 50 % survival at one year". From these sources, it appears that the proportions of survival at weeks 52, 80-90, and at two years are of interest in determining the adequacy of exposure and number of animals at risk.

In determining the adequacy of the chosen dose levels, it is generally accepted that the high dose should be close to the MTD. Chu, Cueto, and Ward (1981) suggest:

(i) "A dose is considered adequate if there is a detectable weight loss of up to 10 % in a dosed group relative to the controls."

(ii) "The administered dose is also considered an MTD if dosed animals exhibit clinical signs or severe histopathologic toxic effects attributed to the chemical."

(iii) "In addition, doses are considered adequate if the dosed animals show a slightly increased mortality compared to the controls."

In another paper, Bart, Chu, and Tarone (Statistical Issues in Interpretation of Chronic Bioassay Tests for Carcinogenicity, Journal of the National Cancer Institute 62, 957-974, 1979), stated that the mean body weight curves over the entire study period should be taken into consideration with the survival curves, when adequacy of dose levels is to be examined. In particular, "Usually, the comparison should be limited to the early weeks of a study when no or little mortality has yet occurred in any of the groups. Here a depression of the mean weight in the treated groups is a indication that the treatment has been tested on levels at or approaching the MTD."

Survival for the male rats was poorest among the control group. At terminal sacrifice it ranged from 22-45 % and at week 93 it was 38-63%:

Percent Survival of Male Rats

Period/Dose	0.0	0.1	1.0	10.0
0-92 weeks	38 %	50 %	63 %	53 %
0-104 weeks	22 %	30 %	45 %	28 %

It appears therefore that there were sufficient numbers of animals living long enough to manifest any late developing tumors.

The high dose animals showed lower body weight gains starting early in the study and supporting the notion that the high dose was close to the MTD (The sponsor's body weight curve is reproduced in Figure 3). On the other hand, the survival experience showed the poorest performance for the controls and therefore does not support the notion that the high dose may be close to the MTD. It is left to the expertise of the pharmacologist to evaluate whether clinical signs and severe histopathological effects have occurred among these animals to suggest that the high dose was close to the MTD. From a statistical point of view the findings for the male rats are inconsistent, inasmuch as the body weight data would support the notion of the high dose being close to the MTD whereas the survival data do not. The lack of significant tumor findings cannot clearly be interpreted as lack of carcinogenic activity of this compound in male rats.

III. The Mouse Study

III.a. Design

In this study 280 CD-1 mice (per sex) were treated via gavage for 104 weeks. The controls received the vehicle only and the actively treated groups of 70 animals each received the compound at 0.6, 6.0, and 60.0 mg/kg/day. The high dose was associated with high mortality and these animals were sacrificed at week 94. For the remaining animals terminal sacrifice was conducted during week 105.

III.b. Sponsor's Analyses of the Mouse Study

Survival Analysis

The sponsor used the same statistical methods for the mouse data as they had for the rat data. There were five males and six females which were treated as censored as their deaths were attributed to technical reasons. The mortality data of the females exhibited a dose-related trend at the p=.01. Trend tests involving only the lower dose groups did not reach statistical significance. The mortality data of the males exhibited an even stronger trend at p=.001 when all dose groups were included. The trend statistics involving only the lower dose groups did not reach statistical significance.

Tumor Data Analysis

The statistical methodolodgy applied to the mouse data is the same as was applied to the rat data. None of the neoplastic lesions among the male mice were considered to be treatment-related. Among the female mice ovarian granulosa-theca cell tumors showed a highly significant trend at p<.001.

III.c. Reviewer's Analyses

Survival Analysis

The sponsor's survival analyses seemed generally appropriate. The mortality experience is shown in Table 2 and Figures 4-5.

Tumor Data Analysis

The time intervals formed by this reviewer are slightly non-standard to accomodate the early sacrifice of the high dose animals in week 94. Specifically, the usual interval of 79-92 weeks was extended by one week.

The sponsor's statistical approach is somewhat different (logistic regression) than the one used in OEB. However, in the case at hand it has the advantage of using time as a covariate and therefore sidestepping the problem that early termination creates for the time-interval approach. On the other hand the false positive rate of the logistic regression has not yet been determined and it is not clear at which α-level significance should be declared. Additionally, the high mortality for both the males and females of the high dose may reflect that the high dose exceeded the MTD and that tumor trends should be investigated including only doses up to the mid-dose. Such an investigation would be different from the sponsor's trend test on the lower doses because theirs is conditional on a significant finding involving all dose groups.

When all four groups of animals are analyzed there were no statistically significant tumor trends with dose in the male mice but benign or malignant incidental ovarian granulosa-theca cell tumors were highly significant. These findings prompted an evaluation of the validity of the male mouse arm (see below).

If it is concluded that the high dose should not have been included in this study and that the mid dose is close to the MTD, then there again are no significant tumor trends among the male mice. However, among the female mice, besides the highly significant granulosa-theca tumors of the ovaries (fatal p=.0063, incidental p=.0000, combined p=.0000), combined hepatocellular adenomas and carcinomas of the liver are also statistically significant (p=.0021). The hepatocellular adenomas or carcinomas alone did not reach statistical significance.

III.d. Validity of the Male Mouse Study

Before concluding that the male mouse study showed no tumorigenic effect of femara the validity of the study needs to be determed following the statistical criteria outline above for the rat study.

The male mice survival experience is documented below.

Survival of Male Mice (percent)

Period/Dose	0.0	0.6	6.0	60.0
0-52 weeks	83 %	94 %	87 %	76 %
0-78 weeks	63 %	61 %	64 %	24 %
0-94 weeks	39 %	33 %	51 %	14 %

It is clear that the high dose groups did not have a sufficient number of animals surviving long enough to manifest any late developing tumors.

Establishing whether the high dose was close to the MTD is difficult. It has already been shown that the mortality was high and statistically significant rather than only numerically increased. In addition, the bodyweight data showed a statistically higher gain during the first 84 days for the dosed animals when compared to their controls (Figure 6). After that time, the high dose animals had lower body weights, showing a 18.4 percent gain deficit when compared to the controls at the end of their lives (week 94). The evaluation of non-neoplastic treatment related findings may help decide whether the high dose was close to the MTD.

IV. Summary and Conclusion

The rat study appears to be a well conducted and well analyzed study. The mortality experience for either sex showed no statistically significant trend with dose. For the female rats there was a statistically significant trend in ovarian gonadal stromal tumors. For the male rats no statistically significant trend in any tumor incidence rates was observed. In trying to assess the validity of the male arm this reviewer concluded that there were a sufficient number of animals living long enough to manifest any late developing tumors. It was more difficult to determine whether the high dose was close to the MTD. The weight gain data were supportive of this notion whereas there was no increased mortality associated with dose. A dose relationship with clinical signs and severe histopathologic toxic effects should help in deciding whether the high dose can be considered an MTD.

The mouse study suffered from high mortality in the high dose animals which resulted in a statistically significant trend test. Both the male and female mice of the high dose were terminated early at week 94. The age-adjusted analysis of tumor data showed a statistically significant trend of ovarian granulosa-theca cell tumors. Among the male mice no statistically significant trend in tumor incidence rates was observed. Evaluating the validity of the male arm it was noted that there were insufficient numbers of animals surviving from the high dose to manifest late developing tumors. Additionally the bodyweight data did not support the notion that the high dose was an MTD. The evaluation of clinical signs and severe histopathologic toxic effects may be decisive. As the high dose caused early and high mortality this reviewer re-analyzed the tumor data of the male mice for possible trends excluding the high dose. Again, no statistically significant trends were observed in tumor incidence rates.

Roswitha Kelly
Roswitha E. Kelly
Mathematical Statistician

Concur:

Clare Gnecco 6/9/97
Clare Gnecco, Ph. D.
Team Leader

6/13/97
George Chi, Ph.D.
Director, DB I

cc:Archival NDA 20-726, Femara (letrozol), Ciba Geigy.
HFD-150/Division File
HFD-150/Dr. Brower
HFD-150/Dr. Andrews
HFD-710/Chron.
HFD-710/Dr. Gnecco
HFD-710/Ms. Kelly
HFD-150/D. Spillman

This review consists of 9 pages of text, 2 tables and 6 figures.
RKELL/05/27/97/wp-femara2.rev

Table 1
INTERCURRENT MORTALITY RATES

FEMALE RATS

			mg/kg/day	
Weeks	0	0.1	1.0	10
0- 52	1/60 (2%)	5/60 (8%)	9/60 (15%)	5/60 (8%)
53- 78	10/59 (18%)	13/55 (30%)	9/51 (30%)	5/55 (17%)
79- 92	12/49 (38%)	15/42 (55%)	10/42 (47%)	12/50 (37%)
93-104	17/37 (67%)	6/27 (65%)	7/32 (58%)	19/38 (68%)
Term. Sac.	20/60 (33%)	21/60 (35%)	25/60 (42%)	19/60 (32%)

MALE RATS

			mg/kg/day	
Weeks	0	0.1	1.0	10
0- 52	4/60 (7%)	5/60 (8%)	6/60 (10%)	2/60 (3%)
53- 78	19/56 (38%)	13/55 (30%)	7/54 (22%)	10/58 (20%)
79- 92	14/37 (62%)	12/42 (50%)	9/47 (37%)	16/48 (47%)
93-104	10/23 (78%)	12/30 (70%)	11/38 (55%)	15/32 (72%)
Term. Sac.	13/60 (22%)	18/60 (30%)	27/60 (45%)	17/60 (28%)

Note: Except for Terminal Sacrifice, an entry of this table represents the number of animals dying or being sacrificed during the time interval divided by the number of animals entering the time interval. The entry in parenthesis is the cumulative mortality percent, i.e. the cumulative percent of animals dying up to the end of the time interval. The entry for Terminal Sacrifice represents the number of animals surviving till the end of the study divided by the initial number of animals. The entry in parentheses for this row represents the number of animals surviving to terminal sacrifice.

Table 2
INTERCURRENT MORTALITY RATES

FEMALE MICE

			mg/kg/day	
Weeks	0	0.6	6.0	60
0- 52	9/70 (13%)	10/70 (14%)	11/70 (16%)	21/70 (30%)
53- 78	26/61 (50%)	21/60 (44%)	20/59 (44%)	31/49 (74%)
79- 93	15/35 (71%)	15/39 (66%)	10/39 (59%)	9/18 (87%)
94-104	7/20 (81%)	8/24 (77%)	10/29 (73%)	9/70 (13%)
Term. Sac.	13/70 (19%)	16/70 (23%)	19/70 (27%)	---

MALE MICE

			mg/kg/day	
Weeks	0	0.6	6.0	60
0- 52	12/70 (17%)	4/70 (6%)	9/70 (13%)	17/70 (24%)
53- 78	14/58 (37%)	23/66 (39%)	16/61 (36%)	36/53 (76%)
79- 93	17/44 (61%)	20/43 (67%)	9/45 (49%)	7/17 (86%)
94-104	10/27 (76%)	11/23 (83%)	15/36 (70%)	10/70 (14%)
Term. Sac.	17/70 (24%)	12/70 (17%)	21/70 (30%)	---

Note: Except for Terminal Sacrifice, an entry of this table represents the number of animals dying or being sacrificed during the time interval divided by the number of animals entering the time interval. The entry in parenthesis is the cumulative mortality percent, i.e. the cumulative percent of animals dying up to the end of the time interval. The entry for Terminal Sacrifice represents the number of animals surviving till the end of the study divided by the initial number of animals. The entry in parentheses for this row represents the number of animals surviving to terminal sacrifice.

SURVIVAL CURVES

19JUL95
surv##

CGS 20267:104 WK ORAL CARC. STUDY IN RATS (MIN 924172)
SEX=Female

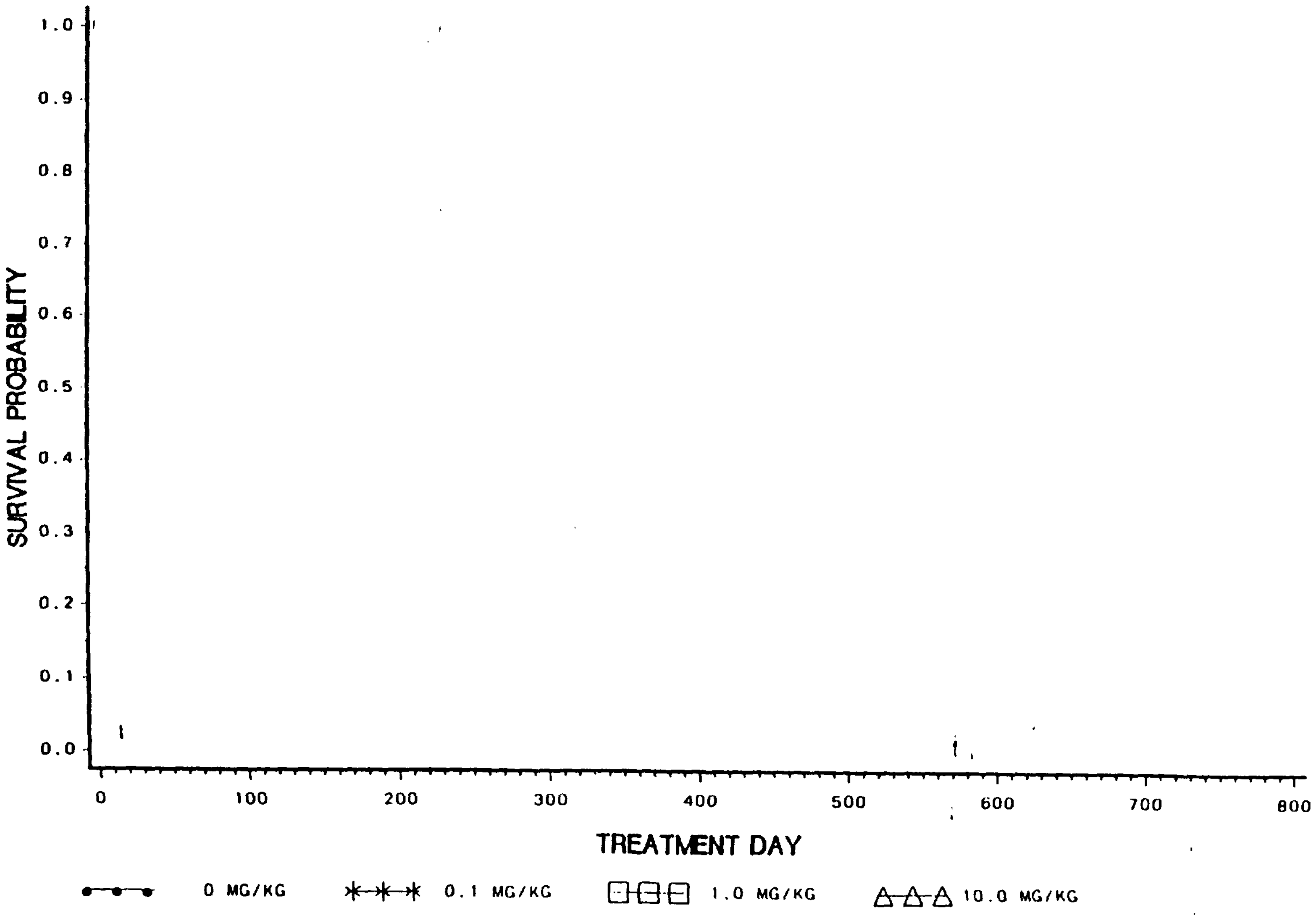


Figure 1

SURVIVAL CURVES

CGS 20267:104 WK ORAL CARC. STUDY IN RATS (MIN 924172)
SEX=Male

19JUL95
surv##

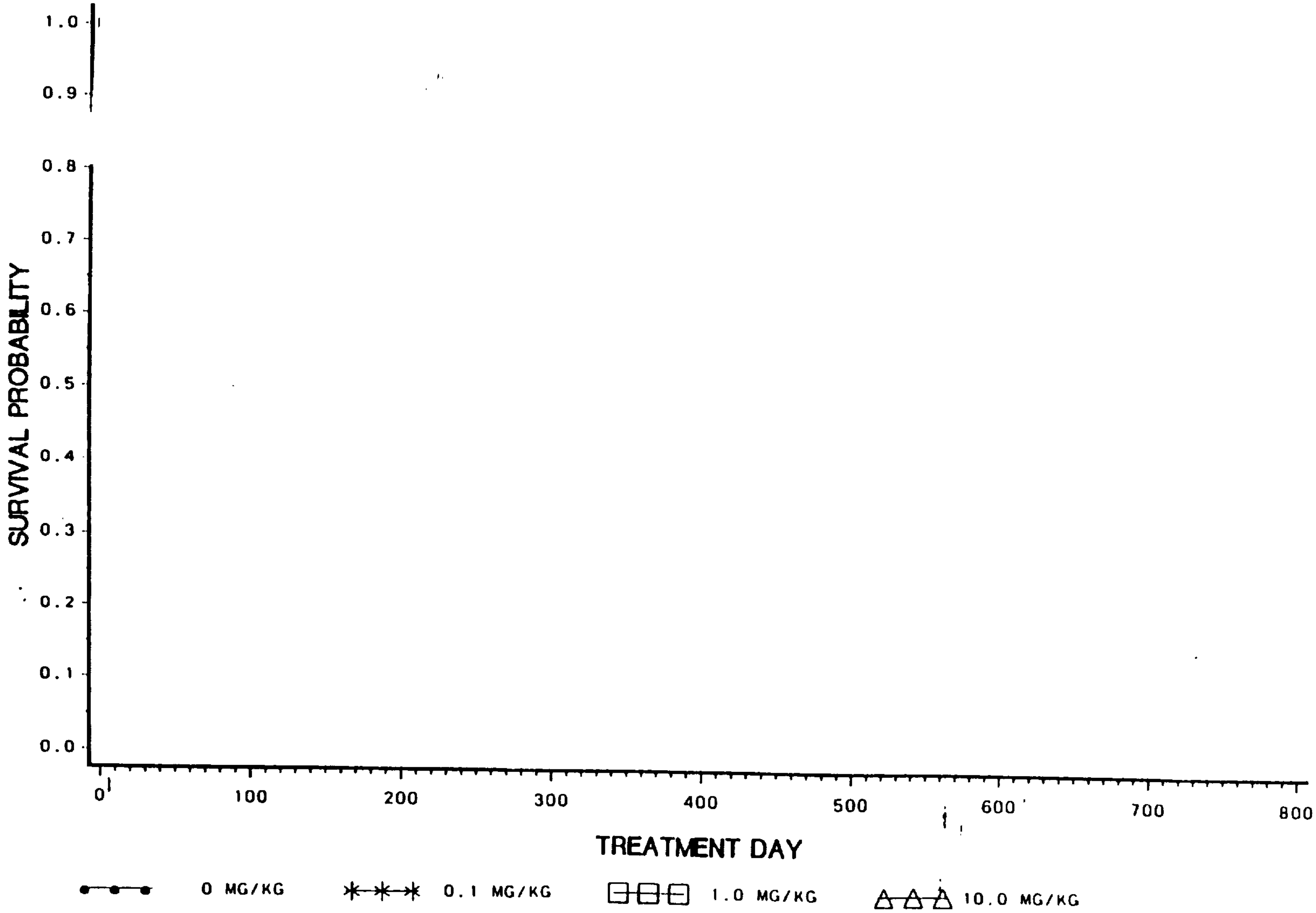
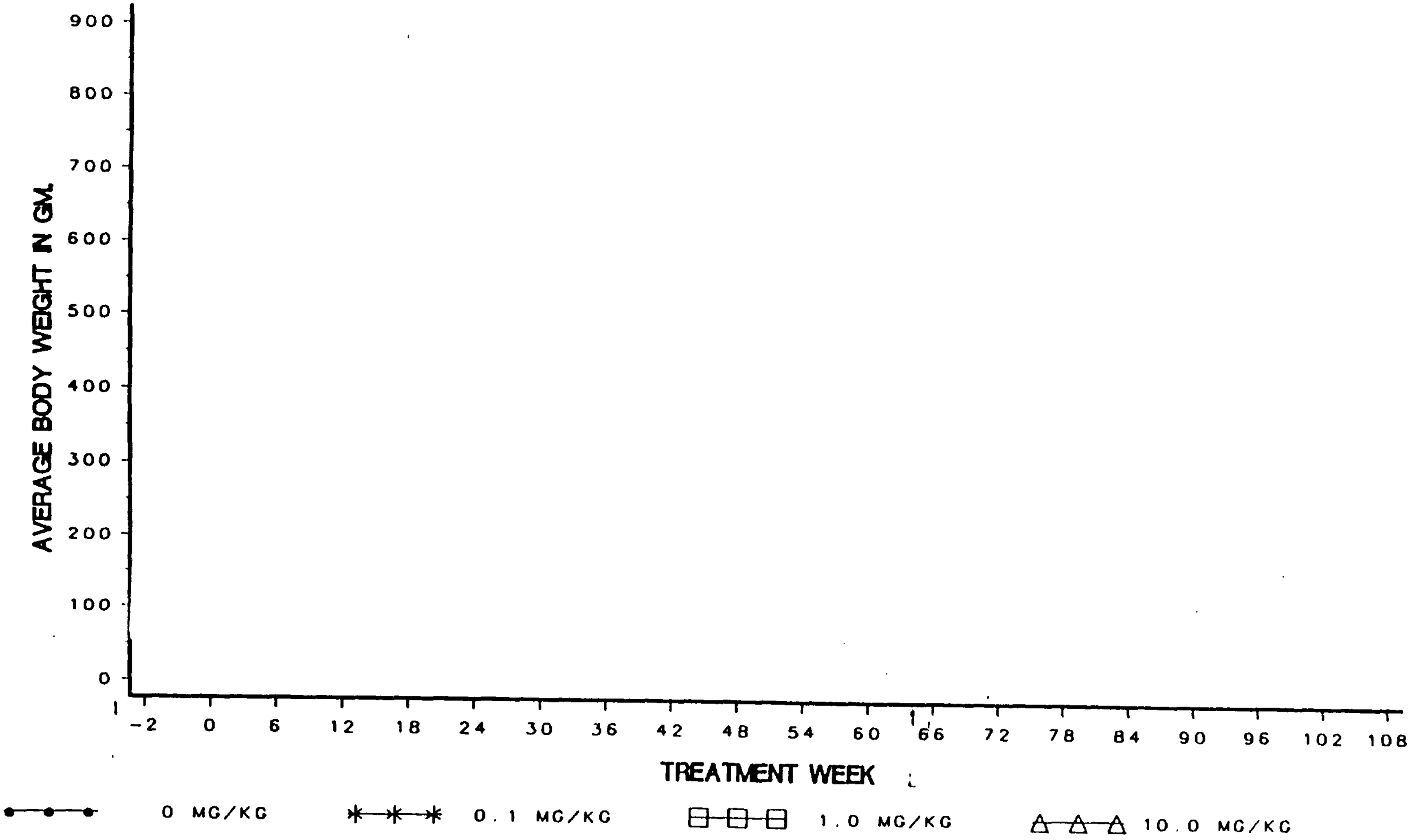


Figure 2

12OCT95
CLXXW

BODY WEIGHT CURVES

CGS 20267: 104 WK ORAL CARC. STUDY IN RATS (MIN924172)
CONTROLS VS TREATED
SEX-MALE



PRE-DOSE TIME(WEEKS < 0) IS NOT PLOTTED IN SAME SCALE AS POST-DOSE TIME

Figure 3

Survival Function Estimates

CGS 20267: 104-Week Oral Carcinogenicity Study in Mice (MIN 934020)
Sex=Female

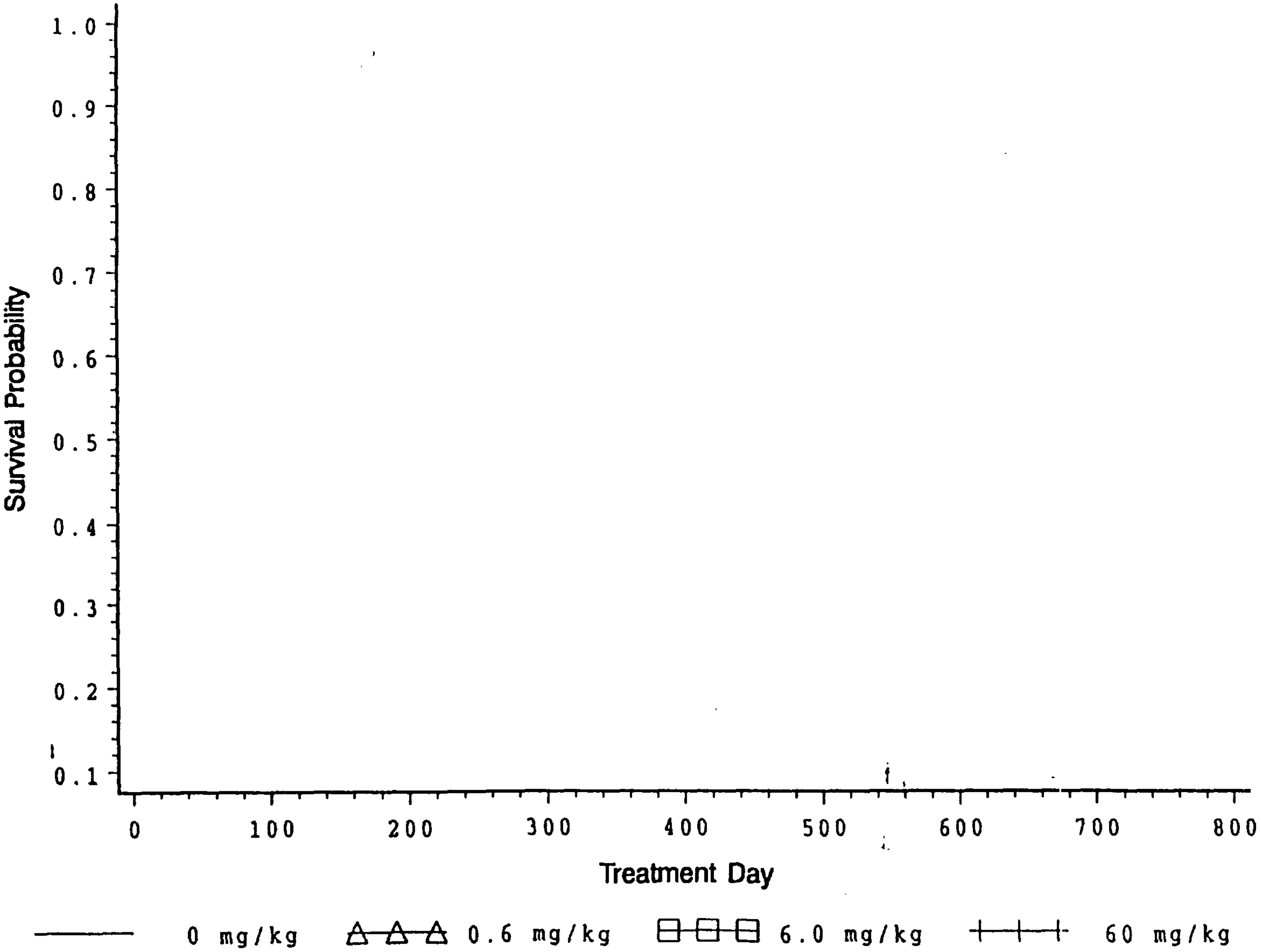


Figure 4

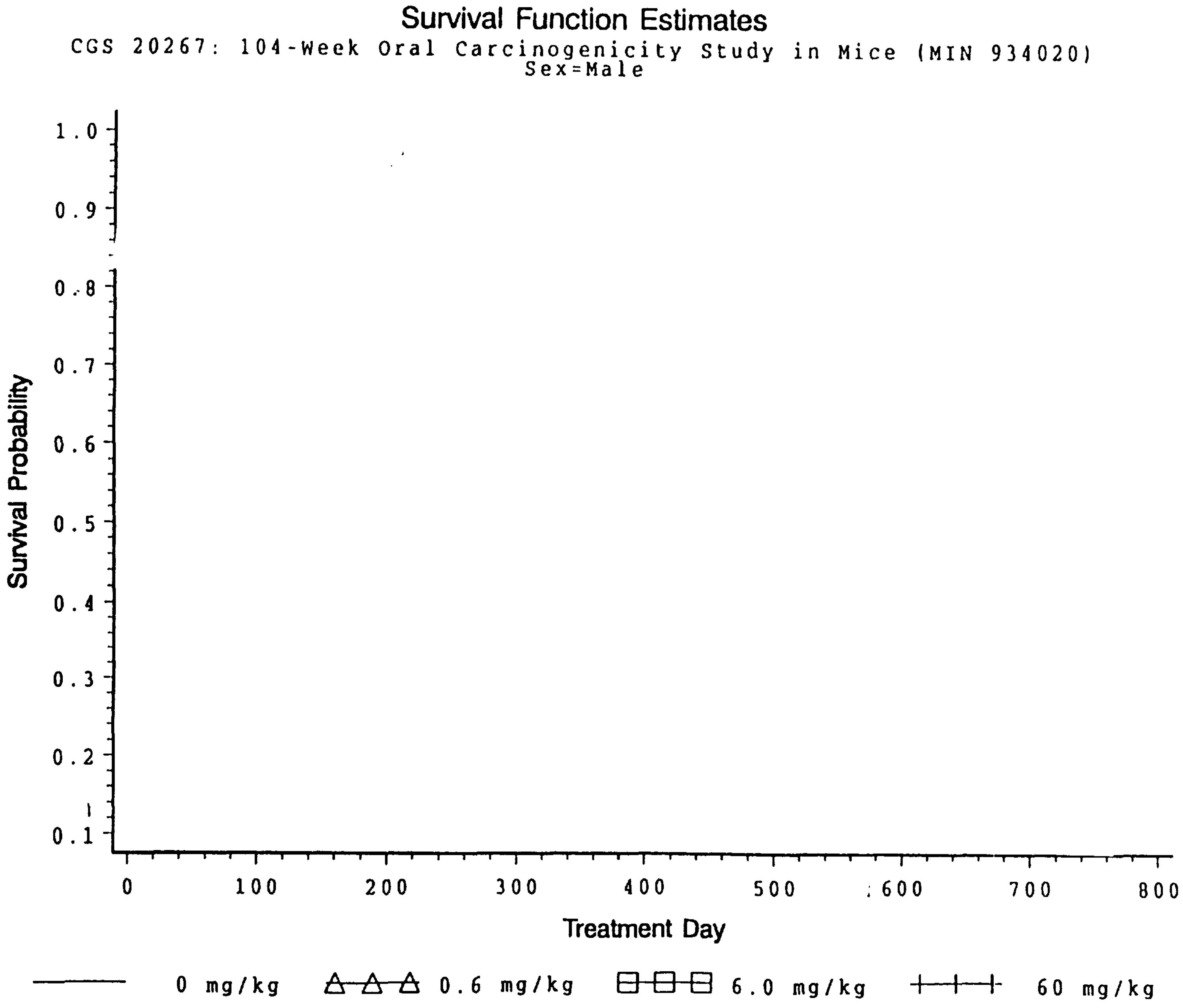
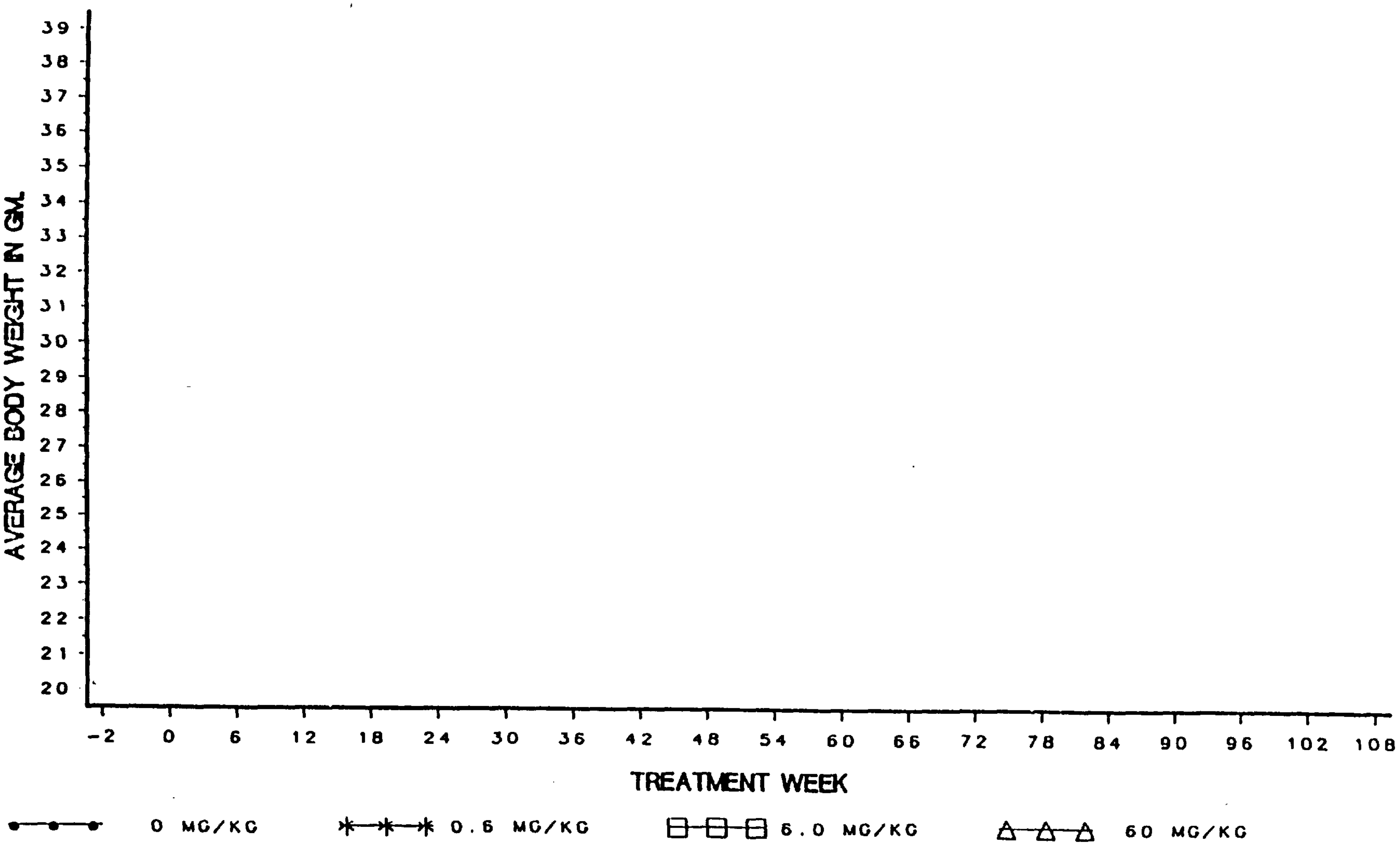

Survival Function Estimates
CGS 20267: 104-Week Oral Carcinogenicity Study in Mice (MIN 934020)
Sex=Male
Survival Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
200
300
400
500
600
700
800
Treatment Day
0 mg/kg
0.6 mg/kg
6.0 mg/kg
60 mg/kg

Figure 5

5MAY95
CLXXW

BODY WEIGHT CURVES

CGS 20267: 104-WK ORAL CARCINOGENICITY IN MICE (MIN934020)
CONTROLS VS TREATED
SEX-MALE



PRE-DOSE TIME(WEEKS < 0) IS NOT PLOTTED IN SAME SCALE AS POST-DOSE TIME

JUN 25 1997

Statistical Reviews and Evaluation

NDA#: 20-726

Applicant: Ciba-Geigy Corporation

Name of Drug: Femara (Letrozole, CGS 20267)

Indication: Second-line endocrine therapy in postmenopausal patients with advanced breast cancer

Documents Reviewed: Vols. 1.10, and 1.71 - 1.117 dated July 25, 1996
and data submitted on February 4, 1997

Medical Officer: Genevieve Schechter, M.D.

Major Statistical Issues:

(i) discrepancy in results by a logistic regression with and without covariates
(ii) discrepancy in results by nonparametric survival and Cox model
(iii) a correlation issue and a missing mechanism in a longitudinal analysis
(iv) unadjusted multiple comparisons for all of the primary efficacy endpoints

I. Background

Six phase Ib/IIa trials (AR/BC1, AR/PS1, Protocol 01, AR/ST1, AR/ES1, and NJ03) were conducted in postmenopausal advanced breast cancer patients to demonstrate initial efficacy and tolerability of the estrogen suppresive doses. Then pivotal phase IIb/III trials (AR/BC2, AR/BC3, and Protocol 02) were initiated "to confirm clinical efficacy and tolerability of th e potentially most effective and best tolerated maximally estrogen suppresive doses with relation to standard available second-line hormonal therapy, since placebo-controlled trials would be unethical in a cancer population."

In this review statistical analyses will be focussed mainly on AR/BC2 study and the results from AR/BC3 will be summarized briefly at the end of this review.

One controlled clinical trial, AR/BC2, is included in this NDA submission because 'on January 18, 1996, the FDA agreed to Ciba's proposal that an NDA filing based on the positive results of the single large phase IIb/III trial AR/BC2 was acceptable and approvable" Therefore, the AR/BC2 trial will be evaluated in this review.

II. AR/BC2 Trial

The AR/BC2 trial is a double-blind, randomized, parallel-group, multicenter (91 centers in ten countries), comparative study in subjects (postmenopausal patients with locally advanced or loco-regionally recurrent or metastasizing breast cancer who progressed on tamoxifen treatment, which is 'currently the first-line therapy of choice in advanced breast cancer') utilizing daily oral doses of 0.5 mg and 2.5mg of letrozole versus megestrol acetate 160 mg once daily. The treatment will be continued for patients who respond (complete response, partial response) or have disease stabilization (no change) until 'disease progression or until any reason necessitated discontinuation.'

Two data sets were defined in this trial; one is derived from a trial defined as 'core trial', and the other defined as 'extension trial'. The first one is 'based on data collected up to 9 months after the end of enrollment (core trial)', and the second one is based on 'data collected on a further 6 months (extension trial) for patients still on trial treatment and for those still alive after trial discontinuation'. The reason for the two data sets in this trial is stated as follows; 'since the median time to progression is about 6 months in postmenopausal patients being treated with second-line hormone therapy for advanced breast cancer, each patient should in theory have the possibility of at least 9 months of treatment...[and the period in the core trial is considered to be] adequate to collect sufficient data regarding tumor response'.

In this review these reviewers presents his analysis of the data sets submitted by the sponsor. The results of his analyses of the data sets derived by the medical officer can be found in her clinical review.

III. The Results From the Sponsor

"The statistical analyses were based on the intent-to-treat approach, where the intent-to-treat population was defined as all patients enrolled unless documented as never having taken at least one dose of trial treatment." The data for the statistical analysis were based on the data "collected from the start of the trial on 25-Mar-93 until the cut-off date for analysis, namely "last patient/last visit" date, 26-Jun-95".

(I) The Primary Variables:

The primary variables in this trial are tumor response (peer reviewed confirmed), duration of response, time to progression, time to treatment failure, and time to death. In this review we focus on the three variables, tumor response, time to progression, and time to death.

(i) Tumor Response

The sponsor distinguishes two types of tumor response; one is called "peer reviewed confirmed best overall tumor response", referred to as overall response, and the other is called "peer reviewed confirmed overall complete or partial tumor response", referred to as objective response.

The tumor response was evaluated "at baseline, 3 months after the start of the trial treatment and every 3 months thereafter or when the patient discontinued treatment at or after 3 months".

Reviewers' TABLE 3.1 shows three types of response rate in each arm. The numbers in ( ) indicate percentages of response rates in each type of response rate of each treatment group. As noticed t he rate of overall response is the smallest among the three types of rates and the reponse rate of each type of response in 2.5 mg is the highest among the three treatment arms. The statistical analyses will be based on overall response confirmed by peer reviewers.

Reviewrs' TABLE 3.1: Types of Response Rate in Each Treatment Arm

	Treatment Arm		
	0.5 mg	2.5 mg	MA
Total Sample Size	188	174	189(190)
Overall Response*	22 (11.7%)	41 (23.6%)	31 (16.3%)
Objective Response**	31 (16.5%)	47 (27.0%)	39 (20.5%)
Investigator's Assessment***	26 (13.8%)	43 (24.7%)	34 (17.9%)

Note:(i) overall response rate was derived by peer reviewed confirmed best overall objective tumor response.
(ii) objective response rate was derived by peer reviewed confirmed best overall objective tumor reponse (whether confirmed or not)
(iii) investigator's assessment was derived by confirmed tumor response.
Note: the figures are adapted from Sponsor's Table 8.1-1.1, Table 8.1-1.3, and Table 8.1-1.4.

Logistic regresion analyses were applied to compare the overall response rate between two treatment groups. Reviewers' TABLE 3.2 shows the results derived from the logistic regression analyses with and without covariates. In the adjusted analyses eleven covariates, which were specified in Statisitcal Analysis Plan dated April 11, 1995, were adjusted through a logistic regression model

Reviewers' TABLE 3.2: The Results derived from a logistic regression with and without covariates

(Sponsor's Table 8.1-1.2)

	Treatment Comparison		
	0.5mg vs 2.5mg	0.5mg vs MA	2.5mg vs MA
Adjusted Odds Ratio	0.37	0.55	1.81
95%CI	(0.20, 0.68)	(0.29, 1.04)	(1.01, 3.24)
p-value	0.0011	0.0608	0.0454
Unadjusted Odds Ratio	0.43	0.68	1.57
95%CI	(0.24, 0.76)	(0.38, 1.22)	(0.93, 2.64)
p-value	0.0028	0.1883	0.0873

Note: 0.5 mg = 0.5 mg letrozole
2.5 mg = 2.5 mg letrozole
MA = 160 mg megestrol acetate

Odds are defined as the ratio of the probability of response over the probability of nonresponse in a treatment group, and an odds ratio is defined as the ratio of the odds of two treatments. In the 0.5 mg vs 2.5 mg comparison, the 2.5 mg treatment arm is considered as the base and in the other comparisons, the megestrol acetate treatment group is treated as the base.

<u>0.5mg vs 2.5 mg of letrozole</u>

There was a statisitcally significant effect in favor of 2.5 mg letrozole over 0.5 mg letrozole in both unadjusted and adjusted analyses. The odds ratio of overall response with 0.5 mg letrozole over 2.5 mg letrozole was 0.43 (95% CI: 0.24, 0.76, P=0.0028 in unadjusted analysis).

<u>0.5 mg of letrozole vs megestrol acetate</u>

There was no statistically significant effect in favor of megestrol acetate over 0.5 mg letrozole in unadjusted analysis (P=0.1883), but there was a marginally statistically significant effect in favor of megestrol acetate over 0.5 mg letrozol with an odds ratio of 0.55 (95%CI:0.29, 1.04, P=0.0608) in the adjusted logistic regression analysis.

<u>2.5 mg letrozole vs megestrol acetate</u>

There was no statistically significant effect in favor of 2.5 mg over megestrol acetate in unadjusted analysis (P=0.0873). On the other hand, there was a statistically significant effect in favor of 2.5 mg letrozole over megestrol acetate with an odds ratio of 1.81 (95%CI:1.01, 3.24, P=0.0454) in the adjusted logistic regression analysis.

It is noted that adjusted analyses always gave smaller p-values than unadjusted analyses, specially in the comparison of 0.5 mg letrozole to megestrol and 2.5 mg letrozole to megestrol (from P=0.1883 in unadjusted analysis to P=0.0608 in adjusted analysis, and from P=0.0873 in unadjusted analysis to P=0.0454 in adjusted analysis, respectively). These reviewers' statistical comments will be discussed in Section IV, reviewers' comments regarding covariate adjustment.

Note that no adjustments to the significance level were made for multiple comparisons.

(ii) Time to Progression

Time to progression was calculated by subtracting the earlist date of documented progression or death from "either a malignant cause or from an unknown cause" from the first day of treatment defined as "date of randomization/dispensation of medicaton". Time to progression was censored if the subject remained on a treatment arm at the date of "last patient/last visit", 06/26/95, without any evidence of disease progression, of if the subject 'was withdrawn from the trial for any reason other than progressive disease".

Reviewers' Table 3.3 shows the total sample size and the number of censored subjects in each treatment arm for the time to progression (TTP) analysis.

Reviewers' TABLE 3.3: Total Sample Size and the Number of Censored Subjects in Each Treatment Arm (Adapted from Sponsor's Table 8.1 - 1.7) for TTP

	Treatment Arm		
	0.5 mg	2.5 mg	MA
Total Sample Size	188	174	189(190)
# of Censored Patient	69	66	56

Reviewers' Table 3.4 shows the results from both unadjusted and adjusted analyses for eleven covariates by Cox Regression analyses. Note that these eleven covariates were not specified in the protocol dated October 27, 1992. Instead eight prognostic factors and stratified logrank tests were specified at this time. In the statistical analysis plan dated April 11, 1995, these eleven covariates and Cox Regression analyses were stated.

Reviewers' TABLE 3.4: Unadjusted and Adjusted Relative Risks with Corresponding 95% CI and P-Values (Adapted from Sponsor's Table 8.1 - 1.6) for TTP

	Treatment Comparison		
	0.5 mg vs 2.5 mg	0.5 mg vs MA	2.5 mg vs MA
Adjusted Relative Risk	1.37	1.10	0.84
95% CI	(1.05, 1.80)	(0.85, 1.42)	(0.65, 1.09)
p-value	0.0219	0.4743	0.1919
Unadjusted Relative Risk	1.26	0.98	0.77
95%CI	(0.97, 1.64)	(0.77, 1.26)	(0.60, 1.00)
p-value	0.0813	0.8973	0.0481

In the comparison of 0.5 mg vs 2.5 mg letrozole statistical significance and marginally statistical significance in relative risk, favoring the 2.5 mg letrozol over the 0.5 mg letrozole was found in Cox Regression analysis with both 11 covariates adjusted and unadjusted analyses, (P=0.0219 and P=0.0813, respectively).

In the comparison of 0.5 mg letrozole vs megestrol acetate treatment arms no statistical significance was found in either adjusted or unadjusted analyses (P=0.4743 and P=0.8973, respectively).

In the comparison of 2.5 mg letrozole vs megestrol acetate treatment arms statistical significance in relative risk, favoring the 2.5 mg letrozole arm, was found in the unadjusted analysis (P=0.0481) and no statistical significance in relative risk was found in the adjusted analysis (P=0.1919). Note that p-values changed from 0.0481 (unadjusted) to 0.1919 (adjusted), favoring the 2.5 mg letrozole arm over the megestrol acetate arm by unadjusted analysis and no difference between the two treatment arms by adjusted analysis. Also note that for this treatment comparison the tumor response adjusted analysis with 11 covariates by logistic regression analysis caused the p-value to change to p=0.0454 from p=0.0873 by unadjusted analysis. Therefore, the direction of the p-value was changed, i.e., covariate adjustment is in favor for tumor response rate but the adjustment is against for time to progression.

The evaluation of these discreapncies will be discussed in reviewer's comments section.

(II) The Secondary Variables

Secondary variables - performance status, severity of pain, and quality of life - were repeatedly measured over time. The sponsor summarized the number and percentage of patients falling into each category on each variable at each visit. No major difference in performance status, the severity of pain, and quality of life were apparent between treatment arms, over visit. If a formal analysis was performed at each visit, we face a multiple testing problem.

In this type of repeated measurements setting, we have to face a correlation issue among observations within each subject and a missing data problem (after 6 months more than 50% of patients dropped out of the study). We do not expect a huge treatment effect in these variables if one existed so that, even without consideration of the two issues in analyses, we may not detect a small, even moderate treatment effect. Details will be discussed in Reviewers' Comments section.

IV. Reviewers' Comments:

In this section four major statistical issues found in this submission will be discussed: (i) discrepant results in logistic regression analyses with and without covariate adjustments for the tumor response variable, (ii) discrepant results in Cox regression analyses with and without covariates adjustment for the time to progression variable (iii) repeated measurement design issues (a correlation problem within a subject and a missing data problem) in quality of life variables, and (iv) unadjusted multiple comparisons for the primary efficacy variables.

(1) Tumor Response

These reviewers note that in the final protocol dated October 27, 1992, six prognostic factors (performance status, age class, disease-free interval, previous chemotherapy, previous response to hormone therapy, previous or concomitant bisphosphonates) were specified for examination of prognostic influence and that in the statistical analysis plan dated April 11, 1995, twelve covariates were specified. However, in the final submission eleven covariates were used in the logistic regression model. As noticed in Appendices 4.1.1-4.1.3, by changing six prognostic factors to twelve the associated p-values change from p=0.0041 to p=0.0015 in the model comparing 0.5 mg letrozole to 2.5 mg letrozole, from p=0.1590 to p=0.0636 in the comparison of 0.5 mg letrozole to megestrol acetate, and from p=0.0757 to p=0.0472 in the comparison of 2.5 mg letrozole to megestrol acetate. Addition of five extra covariates in a logistic regression analysis improved p-values from nonsignificance to marginal significance in the contrast of 0.5 mg letrozole to megestrol acetate, and from marginal significance to significance in the contrast of 2.5 mg letrozole to megestrol acetate.

Reviewers' Appendices 4.1.1-4.1.3 show the results from various adjusted logistic regression analyses and unadjusted analysis on each treatment comparison. As noted in these appendices these reviewers selected five different sets of covariates out of **2049** possible combination from the eleven covariates applied in this NDA submission. In Appendices 4.1.1-4.1.3 the results in adjusted _11 come from a logistic regression analysis with the eleven covariates selected by the sponsor of this submission. The results in adjusted_6 comes from a logistic regression analysis with a set of six covariates specified in the protocol dated Oct. 27. 1992. The results in adjusted_S come from a logistic regression with covariates (selected by a forward stepwise procedure without treatment effect) and also with treatment effect. The results in adjusted_FDA come from a logistic regression with a set of covariates suggested by a survey of medical reviewers at the Division of Oncology at FDA, a Medline search by Dr. Grant Williams, and a textbook, *Clinical Oncology* by Abelloff *et al.* Also, this reviewer consulted with Drs. Martin and Beitz as experts within FDA in evaluation of breast cancer drugs for their perspective regarding important prognostic factors. The results in adjusted_FDA* come from a logistic regression with the same covariates employed in adjusted_FDA, but using a different category of receptor status, (ER+/PR+ and ER+/PR) vs Unknown category instead (ER/PR+)vs (ER/PR) vs (Unknown). Note that in these analyses no interaction terms are added in the model, which indicate the assumption was made that the odds ratio between treatments is homogeneous in each prognostic factor.

This reviewer identified the following statisitcal issues and problematic areas in the sponsor's analyses of tumor response;

(i) Misspecification of a logistic regression model

The estimated treatment effect in a logistic regression model is derived from a maximum likelihood approach. In this setting if covariates are correctly specified in a model, an estimated treatment effect will converge in probabilty to a true effect, and a asymptotic normality will be hold with an estimated variance obtained by an inverse of Fisher's information matrix. On the other hand, if covariates are omitted or misspecified in a logistic regression model, an estimated treatment effect may not converge in probabilty to a true effect. In addition an estimated variance of the treatment effect in a asymptotic normal distribution may not be correct since the variance is calculated by an inverse of a Fisher's information, which is incorrect by omitting or misspecifying covariates. In this situation a score test or a Wald test may not be appropriate for a hypothesis testing.

Gail *et al* (1988) investigated the effect of misspecified covariates in a logistic regression model and Cox regresion model in the context of a score test. The authors suggested a robust variance calculated from residuals by fitting a model without treatment effect. This robust variance is equivalent to a robust variance suggested by Lin and Wei (1988), called "sandwich" estimator in a Cox regression model. These reviewers recommend application of this robust variance to obtain a robust result.

Reviewers' Appendix 4.1.2 shows how the estimated treatment coefficient and the associated odds ratio depend upon the selected set of prognostic factors. The estimated standard errors of the estimated treatment coefficients were fairly consistent across the different sets of covariates, which were derived from an inverse of Fisher's information matrix, defined as the second derivative of a score function with respect to the parameter of interest. These two factors affected p-values which changed from 0.0522 to 0.1911. It is recommended that "sandwich" estimators as well as naive estimators (an inverse of Fisher's information) should be reported to evaluate how robust the reported results are.

(ii) Stability Issue

No interaction terms were added in the logistic regression model. This strategy is based on assuming that the odds ratio of treatment effect is homogeneous. If the odds ratio varies across strata within a prognostic factor, for example, the odds ratio of two treatments is dependent on hormone receptor status (e.g., an odds ratio in the ER/PR+ category is different from that in the Unknown status category), a score derived from a maximum likelihood with the homogeneity assumption may not be correct in the sense that the estimated odds ratio of treatment effect may not converge in probability to the true odds ratio, and the associated variance estimated by the inverse of Fisher's information may not be appropriate to use for an hypothesis test.

In this NDA submission eleven prognostic factors were adjusted for in a logistic regression model. These prognostic factors are **body mass index** (a binary variable, <30 vs ≥30), **age class** (ordinal variable, ≤55, 56-69, ≥70 years), **hormone receptor status** (3 categories, ER/PR+, ER or PR, Unknown), **dominant site of disease** (visceral, bone, and soft tissue), **number of anatomical sites involved** (ordinal variable,1 to 3), **disease free interval** (binary variable,<2 years vs ≥2 years), **previous anti-estrogen therapy** (categorical variable, 4 categories), **response to therapeutic anti-estrogen therapy** (categorical variable, 4 categories), **previous chemotherapy** (categorical variable, 3 categories), **previous or concomitant bisphosphonates** (binary variable), and **performance status** (ordinal variable, 3 categories). Therefore, we have 124416 cells with these eleven covariates and twice this number for the two treatment comparison.

The sponsor states "Although the total number of patients enrolled in the trial was adequate for the main objective of the trial, if interactions were to be examined, the numbers of patients in resultant sub-groups would be limited." Even without interaction terms these reviewers feel that eleven covariates are too many to adjust for. Such a large number may cause a stability problem in parameter estimation - point estimates of the odds ratio, standard error, and associated p-value. For example, in the comparison of 0.5 mg letrozole and megestrol acetate the estimated odds ratio varies from 0.544 to 0.675 with associated p-values of 0.0522 to 0.1911, respectively, depending on which covariates were selected for adjustment. Notice that (Reviewers' Appendix

4.1.3) when a forward selection method is employed without treatment effect in the model, two covariates, number of anatomical sites involved and performance status, give a smaller adjusted p-value (P=0.0522) compared to the sponsor's 11 covariate model (P=0.0636).

(iii) Results from Parsimonious Models

As previously pointed out we may face the possibility of incorrect parameter estimates resulting in inappropriate adjusted p-values for treatment effect. These reviewers attempted to reduce the number of important prognostic factors in order to obtain a more parsimonious model so that results obtained would be robust in the sense that they do not depend on major assumptions such as homogeneity of the odds ratio and factors chosen must have clinical significance. These reviewers did a survey to select the most important three covariates in order from the sponsor's eleven covariates by polling medical reviewers in the Division of Oncology, FDA. Also, these reviewers consulted with Drs. Beitz and Martin as internal experts in the evaluation of breast cancer drugs. Dr. Grant Williams performed a Medline search for this reviewer. This reviewer also referred to a textbook, *Clinical Oncology* by Abeloff *et al.* The following are the three selected covariates: hormone receptor status, dominant site of disease, and response to therapeutic anti-estrogen therapy in order of importance.

These reviewers applied stratified analyses for these three variables. Hormone receptor status was considered as the most important prognostic factor. Therefore response status among two treatments were stratified by three categories of the hormone receptor status - ER/PR+, denoted as RS=1, ER or PR+, denoted as RS=2, and Unknown, denoted as RS=3. Homogeneity of the odds ratio was examined across the three categories. If the homogeneity assumption was satisfied, then 2 x 2 tables were combined to produce one odds ratio by the Mantel-Haenszel relative risk estimator. If the homogeneity assumption did not appear to hold (existence of interaction), then 2 x 2 tables were not combined. In this case it was concluded that effect modification was observed differentially among the categories within the examined prognostic factor. To examine the homogeneity assumption, Breslow-Day and Zelen tests and "eyeballing" were applied. The reason for this approach is, according to the textbook, Epidemiology in Medicine, by Hennekens and Buring (1987), "the determination of whether effect modification is present in the data. In most circumstances, this decision should be based on simply "eyeballing" the data to judge the observed patterns of variation. This should be performed in the context of evidence from other investigations to achieve a biologic understanding of the nature of the association under study. If a more formal statistical evaluation of the uniformity of the stratum-specific estimates is desired,..........Again , however, statistical testing to determine the presence or absence of effect modification should only be used as a guide, since statistical significance is so heavily influenced by sample size."

Reviewers' Appendices 4.1.4 - 4.1.6 present the results from the stratified analyses on the two prognostic factors - hormone receptor status and dominant site of disease. The reason why the third prognostic factor, response to therapeutic anti-estrogen therapy, was not applied here is that the sample size of cells for tumor responders in the 2x2 table was very small (empty cells) so that the results may not be stable.

Reviewers' Appendix 4.1.4 shows the results for the comparison of 0.5 mg letrozole and 2.5 mg letrozole. Homogeneity testing indicated that effect modification may exist. The estimated odds ratio in ER/PR+ and ER or PR+ were similar but for the Unknown category it was three times more than that for ER/PR+ or ER or PR+. Therefore, effect modification existed so that the 2 x 2 tables in ER/PR+ and ER or PR+ were combined together, but the 2 x 2 table for the Unknown category needed to be considered on its own. Then the effect of dominant site of disease was examined for the combined category, ER/PR+ and ER or PR+. Homogeneity tests indicated no sign of effect modification among the three strata for the combined category. Combining the three 2 x 2 tables, the estimated odds ratio was 0.1706 with 95%CI: (0.05, 0.4723) and P=0.0002 (exact test). This means that, after controlling for dominant site of disease within the combined hormone receptor status group, the odds of responding in 0.5 mg letrozole were only 0.17 times those on the 2.5 mg letrozole group.

Reviewers' Appendix 4.1.5 shows the results in the comparison of 0.5 mg letrozole and megestrol acetate. The same phenomena were observed. Effect modification was observed in hormone receptor status so that the 2 x 2 tables in ER/PR+ and ER or PR+ groups were combined, but not the 2x2 table in the Unknown category. Homogeneity tests indicated no sign of effect modification among strata of dominant site of disease for the combined hormone receptor status group. When one combines the three 2x2 tables, the estimated odds ratio is 0.2684 with 95%CI: (0.08, 0.7622) and P=0.0101 (exact test). This means that the odds of responding on 0.5 mg letrozole are only 0.27 times those on megestrol acetate after controlling for dominant site of disease in the combined hormone receptor status. Note that eyeballing indicated effect modification among the three strata. In the bone category there was no statistically significant difference in odds ratio. In the combined soft and visceral tissue category the estimated odds ratio was 0.1493 with 95%CI:(0.025, 0.598) and P=0.0036 (exact test).

Reviewers' Appendix 4.1.6 shows the results in the comparison of 2.5 mg letrozole and megestrol acetate. Effect modification was observed in hormone receptor status. In this case it was not appropriate to combine the two strata of ER/PR+ and ER or PR+. The stratum of ER/PR+ was stratified further by the three categories of dominant site of disease. No effect modification was found in ER/PR+ stratum by dominant site of disease so that the three 2x2 tables were combined. The estimated odds ratio was 3.105 with 95%CI:(1.093, 9.627) and P=0.0313. This indicated that the odds of responding on 2.5 mg were 3.105 times higher than on megestrol acetate. Note that if the two receptor strata were combined with the three stratification levels of dominant site of disease, no statistical significance in odds ratio was found.

(2) Time to Progression

Adjusted Analysis:

The theory for a Cox regression model, the partial likelihood approach, was developed on the assumption that we have a correctly specified model in terms of proportional hazards and correctly selected prognostic factors in the model (Cox, 1972, 1975). Under this **true** model, Andersen and Gill (1982) showed by applying the counting process context that estimated coefficients converge to true values in distribution to a multivariate normal with mean 0 and a covariance matrix consistently estimated by a Fisher's information matrix derived from the partial likeliood.

(i) Misspecification of Cox Model (Covariate Adjustment)

Unfortunately, we do not know the true Cox model. Therefore an applied Cox model can be considered as a "working" parametric model with some misspecifications. Several approaches have been suggested for handling misspecified models (e.g., Gail et al, 1988; Huber 1967; Kent 1982; White 1982). Lin and Wei (1989) investigated the misspecified Cox proportional hazard model and proposed a "sandwich" estimator for the covariance matrix of estimated coefficients in the misspecified Cox model for testing purposes. This "sandwich" estimator is derived from M estimation theory and these authors modified the middle part of the "sandwich" estimator for the Cox model. Since a misspecified model is estimated, the estimated coefficients of the working model will not converge to the true parameter values; instead, they converge to some value, hopefully near the true value. They proved that the estimated coefficients converge to a value, ß*, where ß is the true value, in distribution to a multivariate normal with mean 0 and a "sandwich" covariance matrix.

Reviewers' Appendices 4.2.1-4.2.3 show the results from several sets of prognostic factors in Cox regression models for each treatment comparison. It is to be noted that estimated standard errors are fairly stable across sets of covariates in each treatment comparison. On the other hand, the estimated treatment effect depends upon the selected set of prognostic factors. For example, Reviwers' Appendix 4.2.3 shows the results for the comparison of 2.5 mg letrozole to megestrol acetate. The estimated treatment effect in this comparison changed from -0.175 to -0.258. On the other hand, the estimated standard errors derived from a type of inverse Fisher's information were stable across models, staying around 0.131. Therefore taking into account these two phenomena, p-values derived from a Wald test depended upon the estimated treatment effect (P-values changed from 0.0488 to 0.1927). In this sense it is recommended that a 'sandwich" estimator should be reported along with a regular estimate (naive estimate).

(ii) Stability Issue

As mentioned in the tumor response section, if the eleven covariates were employed in Cox regression model, we have 248832 cells in a comparison of two treatment effect. Even though we have a large number of patients in each arm, the number of covariates may affect stability of the parameter estimates. All Cox regression analyses in Reviewers' Appendices 4.2.1-4.2.3 rely on the key proportional hazards assumption for Cox model. If this assumption is violated, we do not have valid inference derived from a Cox model. A violation of the proportional hazards assumption in a Cox regression model will cause a severe loss of efficiency (Lagakos and Schoenfeld, 1984). Such a violation could have occured in the following parsimonious models.

(iii) Results from Parsimonious Models

To obtain a more robust result with fewer, yet clinically meaningful prognostic factors, two covariates - hormone receptor status and dominant site of disease- were adjusted for by a stratified logrank test. The stratified logrank test was mentioned in the protocol dated October 27, 1992, but dropped in the statistical analysis plan dated April 11, 1995. Therefore, the sponsor did not submit results of any stratified logrank analysis in this application.

Reviewers' Table 4.2.1 shows the results from the stratified logrank tests and associated Cox regression models with the two covariates - hormone receptor status and dominant site of disease - along with the results from unadjusted and adjusted with 11 covariates analyses. Also, the estimated relative risk from the Cox models is also reported.

Reviewers' TABLE 4.2.1 Results from Stratified Logrank Tests along with Corresponding Cox Models for Time to Progression

	0.5 mg vs 2.5 mg	0.5 mg vs MA	2.5 mg vs MA
Estimated RR	1.29	1.01	0.79
Stratified Logrank	P = 0.0684	P = 0.6453	P = 0.0604
Cox	P = 0.0599	P = 0.9145	P = 0.0777
Unadjusted	P = 0.0813	P = 0.8973	P = 0.0488
Adjusted_11	P = 0.0219	P = 0.4743	P = 0.1919

Note: Adjusted_11 stands for results from Cox model with 11 prognostic factors.

It is noted that the results among the three analyses are similar within each treatment comparison, but the results from the 11 covariates adjustment were different, particularly in the comparison of 2.5 mg letrozole to megestrol acetate compared to the other three analyses. Reviewers' Appendix 4.2.4 shows the total sample size, events occured and the number of subjects censored within each treatment comparison stratified by hormone receptor staus and dominant site of disease. Notice that 68% of the subjects were censored in the group of soft tissue site of disease with ER/PR+ hormone receptor status in the comparion of 2.5 mg letrozole to megestrol acetate. In other categories we have adequate sample size to do a logrank test and a weighted stratified logrank test.. If one more prognostic factor is adjusted through a stratified logrank test, the sample size within each cell will become too small to do a logrank test and weighted average for a stratified logrank test. This indicates that any result obtained may not be stable.

Reviewers' Table 4.2.1 indicates that there exists a marginally statistical significance in RR in the comparison of 0.5 mg vs 2.5 mg of letrozole and in the comparison of 2.5 mg letrozole vs megestrol, favoring the 2.5 mg letrozole treatment arm. There did not exist statisitcal significance in RR in the comparison of 0.5 mg letrozole vs megestrol.

These reviewers further investigated the homogeneous hazard ratio assumption across strata in each treatment comparison which fitted Cox models assumed in Reviewer's Table 4.2.1. Reviewer's Appendix 4.2.5 indicates that there existed fairly constant relative risks across dominant site of disease categories among ER/PR+ and ER or PR+ hormone receptor status and there seemed to exist an interaction across dominant site of disease categories in Unknown hormone recptor status in the comparison of 0.5 mg letrozole vs 2.5 mg letrozole. Note that RR=0.347 in bone compared to RR=1.347 and 1.561 in soft and visceral, respectively, in the Unknown hormone receptor status group in the comparison of 0.5 mg letrozole vs 2.5 mg letrozole. The same trends were observed in the comparison of 0.5 mg letrozole vs megestrol acetate, i.e., fairly constant relative risks across dominant site of disease categories among ER/PR+ or ER or PR+ hormone receptor status groups were observed in the same direction, but, on the other hand, the opposite direction in relative risks was observed in the Unknown hormone receptor status group across dominant site of disease categories.

In the comparison of 2.5 mg letrozole vs megestrol acetate, similar relative risks were observed across dominant site of disease categories in ER/PR+ hormone receptor status and in ER or PR+ hormone receptor status, but the magnitude of the relative risks between ER/PR+ and ER or PR+ were different, indicating that the two categories of ER/PR+ and ER or PR+ of hormone receptor status may not appropriately be combined together. There seemed to exist an interaction of relative risks across dominant site of disease categories in Unknown hormone receptor status. Therefore it is reasonable to report three results from stratified logrank tests with dominant site of disease strata combined within each hormone receptor status separately.

Reviewers' Table 4.2.2 shows the results from stratified logrank tests along with corresponding Cox models and estimated relative risks. As noted, the results from stratified logrank tests and the corresponding Cox models are very similar because homogeneous hazard ratio assumptions seemed to be satisfied within each category reasonably combined together across dominant site of disease strata.

Reviewers' TABLE 4.2.2 Results from Apparently Reasonable Models

	0.5 mg vs 2.5 mg	0.5 mg vs MA	2.5 mg vs MA
HRS	Stratified by Dominant Sites of Disease		

RS = 1	P = 0.0084
	P_Cox = 0.0079
	Estimated RR = 1.60
		P = 0.0834
		P_Cox = 0.1017
		Estimated RR = 1.32
			P = 0.0345
			P_Cox = 0.04886
			Estimated RR = 0.63

RS = 2			P = 0.6700
			P_Cox = 0.8573
			Estimated RR =1.05

RS = 3	P = 0.7012
	P_Cox = 0.7947
	Estimated RR = 0.95
		P = 0.1573
		P_Cox = 0.1371
		Estimated RR = 0.74
			P = 0.1773
			P_Cox = 0.1774
			Estimated RR = 0.76

Note: - HRS stands for hormone receptor status so that RS=1 stands for ER/PR+, RS=2 stands for ER or PR+, and RS=3 stands for Unknown receptor status.
- P_Cox stands for a P-value from the Cox models corresponding to the stratified logrank tests

In the comparison of 0.5 mg letrozole vs megestrol acetate, there existed statistical significance in relative risk, estimated as 1.60, favoring 2.5 mg letrozole over 0.5 mg letrozole treatment in a combined category of ER/PR+ and ER or PR+ hormone receptor status adjusting for dominant site of disease categories (P=0.0084). On the other hand, no statistical significance in relative risk estimated as 0.95 was found in the Unknown receptor category adjusting for dominant site of disease categories in the comparison of 0.5 mg vs 2.5 mg letrozole treatment arms.

In the comparison of 0.5 mg letrozole vs megestrol acetate there existed marginally statistical significance in relative risk, estimated as 1.32, favoring the megestrol acetate treatment arm in the combined category of ER/PR+ and ER or PR+ hormone receptor status adjusting for dominant site of disease categories (P=0.0834). On the other hand, no statistical significance in relative risk, estimated as 0.74, was found in Unknown receptor category adjusting for dominant site of disease categories in the comaprison of 0.5 mg letrozole vs megestrol acetate treatment arms.

In the comparison of 2.5 mg letrozole vs megestrol acetate treatments there existed statistical significance in relative risk, estimated as 0.63, favoring 2.5 mg letrozole over megestrol acetate treatment arm in ER/PR+ hormone receptor status adjusting for dominant site of disease strata

(P=0.0345). On the other hand, no statistical significance was found in relative risks, estimated as 1.05 and 0.76, in categories of ER or PR+ and Unknown hormone receptor status adjusting for dominant site of disease categories, respectively (P=0.6700 and P=0.1773).

(II) Secondary Variables

Quality of Life (QOL)

For this assessment we focus on three variables of the quality of life data, global quality of life, pain score in QLQ-C30 by the European Organization for Research and Treatment of Cancer (EORTC), and Performance Status by WHO.

Quality of life variables were measured at baseline and one month, two months, three months, six months, nine months and so on post baseline. The sponsor analyzed these data by change from the baseline values at each time point. It is to be noted that these variables were obtained in repeated measurements setting.

In general, we will face two challenges in an analysis of repeated measurements: (i) a correlation problem within each subject and (ii) a missing data problem. For the first challenge the linear mixed effects model (Laird and Ware, 1982) and the GEE approach (Liang and Zeger, 1986) were developed to deal with a correlation problem among observations per subject. In a classical univariate repeated ANOVA, a particular correlation structure, known as a compound symmetry structure, must be assumed for a valid F test of interaction of treatment and time, or a multivariate analysis would be applied. But in a multivariate approach, a distributional assumption must be valid with a correct mean and a correct variance structure. In addition, we may encounter a singular covariance matrix which adds an additional level of complexity.

In order to cope with these problems, the linear mixed effects model introduces a random factor via a **Z** matrix, a subset of the design matrix, in a framework of the maximum likehood approach, and the GEE approach introduces a concept of a "working" correlation in a framework of M estimation theory, deriving a robust variance, known as a "sandwich estimator", originally termed so by Lin and Wei (1989).

The second key issue is the missing data problem. Reviewers' Table 4.3.1 shows the missing data pattern over time in the Global quality of life variable. By Visit 5 more than 50% of the patients had dropped out from the study on each treatment arm.

Reviewers' TABLE 4.3.1: Missing Data Pattern Over the Study Period in the Global Quality of Life Variable / Sample Size Changes

Visit	0.5 mg (N=180)	2.5 mg (N=170)	MA (N=185)
Visit 0 (Baseline)	180	170	185
Visit 1 (Month 1)	164	166	174
Visit 2 (Month 2)	150	156	153
Visit 3 (Month 3)	126	138	139
Visit 4 (Month 6)	82	90	97
Visit 5 (Month 9)	60	67	62
Visit 6+	39	47	32

The sponsor analyzed the data by the change from the baseline for the patients who were on the study at each time point. This type of analysis is called "observed cases" (OC) analysis, based on the assumption that missing data would be caused by a "missing completely at random" (MCAR) mechanism. This missing mechanism is a very strong assumption so that in reality it would be very difficult to justify its validity, particularly in an oncology trial.

These reviewers applied a growth curve analysis to cope with the correlation issue in a longitudinal analysis. This reviewer employed three types of linear model: (i) GEE with three different "working" correlation assumptions; independent, compound symmetry, and AR-1, (ii) a linear mixed effects model, known as a 'Laird and Ware" model (Laird and Ware, 1982), with three different random coefficients; intercept (corresponding to compound symmetry), slope , and intercept and slope, and (iii) a two stage model to obtain robust results. The details of these approaches are described in appendix 4.3.1.

These reviewers employed the concept of a "Pattern-Mixture Model", advocated by Little (1993 and 1995) to judge whether the observed missing mechanism is ignorable or nonignorable. These reviewers did not attempt to produce one estimate or derive one p-value when the observed missing mechanism was judged to be nonignorable because (i) the derived results by modeling the possible missing mechanism in a likelihood function will be very sensitive to the proposed missing mechanism, and (ii) there is no data to verify the assumed missing mechanism. The employed approach here is outlined in figure 4.3.1.

These reviewers requested from the sponsor exact dates as to when the repeated meaurements were taken. These data were requested on Aug. 22, 1996, and received on Oct. 30, 1996. The reason for this request is that when we have a measurement error (I.e., assessments not performed at the precise time specified) in an independent variable in a linear model, it is well known that the estimated coeffficient will be biased toward the null in a classical measurement error setting, but a measurement error problem will not affect parameter estimates in a Berkson model (Fuller, 1986). In general, measurement error models have an identifiability problem because there are too many parameters to estimate in such a model. Therefore sometimes a ratio of the variance of the measurement error and the variance of the error of a linear model is known or a small measurement error assumption will be imposed. In our setting we know the actual times when QOL parameters were measured so that it is natural to estimate a variance of the measurement error at each visit. Even after we adjust for the measurement error, the associated standard error of the estimated coefficients would be larger than one without the measurement error problem, which indicates that we may lose statistical power.

In our setting, we do not expect an extremely large change from baseline value over time, rather we expect a modest or small change from baseline. If we allow a measurement error at each occasion: (1) the estimated coeffficients will be biased toward the null and (2) the adjustment of the measurement error problem will cause a larger confidence interval of the parameter estimates. Thus, we may not detect the modest or small change due to the measurement error. Thus, these reviewers recommend a smaller window around each visit to minimize the measurement error problem to avoid the bias to the null by treating a clinical trial design as a balanced design or we can use actual time by considering the trial as an unbalanced design.

These reviewers used reported actual time for each subject considering the trial design as an unbalance and imcomplete design.

In the following analyses, completers are defined as subjects who stayed on study at least 6 months, and patients who dropped out from the study before 6 months were defined as dropouts. Reviewers' Table 4.3.2 shows frequency of response status for each time category. Note the category definitions, maxi=6, indicates patients who stayed on the study up to 9 months, maxi=5 indicates patients who stayed on the study up to 6 months and maxi ≤ 4 indicates patients who dropped out from the study before 6 months.

Reviewers' TABLE 4.3.2: Frequency of Response Status on Each Time Category Based on Global Quality of Life parameter in the QLQ-C30

0.5 mg letrozole			
CBRP	Maxi = 6	Maxi = 5	Maxi ≤ 4
CR	3	0	0
PR	18	1	0
SD	24	2	2
PD	15	19	60

Unknown	3	2	22
2.5 mg letrozole			
CBRP	Maxi = 6	Maxi = 5	Maxi ≤ 4
CR	10	0	0
PR	29	2	0
SD	15	4	1
PD	6	17	69
Unknown	8	1	9
Megestrol			
CBRP	Maxi = 6	Maxi =5	Maxi ≤ 4
CR	7	0	0
PR	20	3	0
SD	23	3	1
PD	13	26	60
Unknown	1	2	17

Note: maxi=6 indicates subjects who stayed on the study at least 9 months
maxi=5 indicates subjects who stayed on the study up to 6 months
maxi≤4 indicates subjects who dropped out from the study before 6 months

As noticed from the above table, patients with CR or PR stayed on at least up to 6 months.

Pain Score:

Reviewers' Appendices 4.3.2 - 4.3.4 present a summary of the results of pain score analysis in the QLQ-C30 by the EROTC on each treatment arm. For the 0.5 mg letrozole arm the pain score did not change over a time for completers, but the score declined over a time for dropouts. This indicates that a possible mising mechanism is nonignorable, and also that pain was decreased only in dropouts, not in completers. Note that the pain score in both completers with CR or PR and with SD, PD, or Unknown did not change over time (Reviewer's Appendix 4.3.2).

For the 2.5 mg letrozole arm the pain score did not change over time for completers, but the score declined over time for dropouts. This indicates that a possible missing mechanism is nonignorable, and also that pain decreased only in dropouts, not in completers. Note that the pain score in both completers with CR or PR and with SD, PD, or Unknown did not change over time (Reviewer's Appendix 4.3.3).

For the MA arm the pain score did not change over time for completers, but the score declined over time for dropouts. This indicates that a possible missing mechanism is nonignorable, and also that pain decreased only in dropouts, not in completers. Note that as mentioned in comments in Reviewer's Appendix 4.3.4 the pain score declined in completers with maxi=5 and that the pain score in completers with CR or PR has a quadratic time trend but with SD, PD, or Unknown there is a linear decline time trend.

Overall, the pain score did not change over time in completers in each treatment arm, but a linear decline was detected for dropouts in each arm. The linear decline patterns are similar.

Global Quality of Life Score:

For the 0.5 mg letrozole arm the quality of life score did not change over time for completers, but the score declined over time for dropouts. This indicates that a possible missing mechanism is nonignorable, and also that the quality of life became worse over time in dropouts, not in completers. Note that the quality of life in both completers with CR or PR and with SD, PD, or Unknown did not change over time (Reviewer's Appendix 4.3.5).

For the 2.5 mg letrozole arm the quality of life score did not change over time for completers, but the score declined over time for dropouts. This indicates that a possible missing mechanism is nonignorable, and also that the quality of life became worse over time in dropouts, not in completers. Note that as pointed out in comments in Appendix 4.3.6 both linear and quadratic terms are found to be statistically significant in completers with maxi=6 and in completers with CR or PR.

For the MA arm the quality of life score did not change over time for completers, but the score declined over time for dropouts. This indicates that a possible missing mechanism is nonignorable, and also that the quality of life became worse over time in dropouts, not in completers. Note that the quality of life in both completers with CR or PR and with SD, PD, or Unknown did not change over time. Note that as mentioned in comments in Reviewer's Appendix 4.3.7 a linear decline was found to be statistically significant in completers with maxi=5.

Overall, the quality of life score did not change over time in completers in each treatment arm, but a linear decline was detected for dropouts in each arm. The linear decline patterns are similar.

Performance Status by WHO:

In longitudinal analyses the score of performance status by WHO was treated as a continuous variable, although the score ranged from 0 to 4 on an ordinal categorical scale. Therefore interpretation of the results must be cautious. The GEE approach is a marginal approach so that when we have a missing data problem, the "missing completely at random" (MCAR) assumption will be required. But in our clinical trial setting it is unrealistic to assume this type of missing mechanism. Note that in a linear setting, the MCAR assumption will not be required.

For the 0.5 mg letrozole arm the performance status score did not change over time for completers, but the score increased over time for dropouts. This indicates that a possible missing mechanism is nonignorable, and also that the performance status became worse over time in dropouts, not in completers. Note that the performance status score in both completers with CR or PR and with SD, PD, or Unknown did not change over time (Reviewer's Appendix 4.3.8).

For the 2.5 mg letrozole arm the performance status score did not change over a time for completers, but the score increased over time for dropouts. This indicates that a possible missing mechanism is nonignorable, and also that the performance status became worse over time in dropouts, not in completers. Note that the performance status score in both completers with CR or PR and with SD, PD, or Unknown did not change over time (Reviewer's Appendix 4.3.9).

For the MA arm the performance status score increased in both conpleters and dropouts. The difference between the slopes (0.00148 in completers vs 0.00829 in dropouts) is statistically significant. This indicates that a possible missing mechanism is nonignorable, and also that the performance status became worse more rapidly over time in dropouts compared to completers. Note that the performance status score in completers with CR or PR did not change over time, but in completers with SD, PD, or Unknown it was found to become worse (Reviewer's Appendix 4.3.10).

Overall, the performance status score did not change over time in completers in both the 0.5 mg and the 2.5 mg treatment arms, but a linear increase indicating worsening was detected in completers on the MA arm. Similar increasing linear trends are found in the three treatment arms among dropouts.

V. Summary and Conclusions of AR/BC2 Trial:

Four statistical issues were considered in this review: (1) discrepant results from a logistic regression model with and without covariate adjustment in the tumor response variable, (2) discrepant results from a Cox regression model with and without covariate adjustment in the time to progression variable, (3) a correlation issue and a missing data mechanism issue in the repeated measurements setting in quality of life variables, and (4) a multiple comparison problem.

(1)-(2) covariate adjustment issue in both logistic regression and Cox regression models

Reviewers' Appendices 4.1.1 - 4.1.3 for the logistic regression setting and reviewers' Appendices 4.2.1 - 4.2.3 for the Cox regression setting indicate that p-values depended on whether covariates were adjusted for in a logistic regression model or not, and if covariate adjustments were performed, which covariates were adjusted for in a logistic regression model. For example, in the comparison of 0.5 mg letrozol and megestrol in a logistic regression model, p-values changed from 0.0522 (adjusted by covariates selected by a forward stepwise procedure without treatment effect in the model) to 0.1911 in unadjusted analysis. In the Cox regression model , p-values changed from 0.1927 (adjusted by the 11 covariates) to 0.0488 in unadjusted analysis in the comparison of 0.5 mg letrozol and megestrol acetate. This suggests that p-values derived from covariate adjustments were not robust and that the results were not in the same direction in the sense that covariate adjustments in both models do not necessarily provide smaller p-values than ones in unadjusted analyses. Two issues were discussed at length in this review - misspecified logistic and Cox regression models by covariate adjustments (robust variance estimator was suggested) and a stability problem in parameter estimation (parsimonious models were investigated).

Two covariates - hormone receptor status and dominant site of disease - were adjusted for through Mantel-Haenzel and exact procedures for the tumor response variable. Stratified logrank tests were applied to assess the time to progression variable. Theses two covariates were selected by a survey and consultation within the Division of Oncology, CDER, FDA and via a Medline search.

In the comparison of 0.5 mg letrozole vs 2.5 mg letrozole treatment arms statistically significant results were found in two variables - tumor response rate and time to progression - in hormone receptor status (ER/PR+ or ER or PR+) categories with adjustment for dominant site of disease (P=0.0002 by an exact procedure for the tumor response variable and P=0.0084 in the time to progression variable). These results were in favor of the 2.5 mg letrozole treatment over the 0.5 mg letrozol treatment. No statistically significant results were found in either variable for the Unknown hormone receptor status category.

In the comparison of 0.5 mg letrozole treatment and megestrol acetate treatment arms a statistically significant result was found in the tumor response variable and marginally statistically significant result was found in the time to progression variable in hormone receptor status (ER/PR+ or ER or PR+) categories with adjustment for the dominant site of disease prognostic factor (P=0.0101 by an exact procedure for the tumor response and P=0.0834 in the time to progression variable).

These results were in favor of megestrol acetate treatment over 0.5 mg letrozole treatment. No statistically significant results were found in either variable for the Unknown hormone receptor status category.

In the comparison of 2.5 mg letrozole vs megestrol acetate treatments a statistically (or marginally) significant results were found in two variables - tumor response and time to progression - in hormone receptor status, ER/PR+ category, with adjustment for the dominant site of disease prognostic factor (P=0.0313 by an exact procedure in the tumor response variable and P=0.0345 in the time to progression variable). The results were in favor of the 2.5 mg letrozole treatment arm over megestrol acetate treatment. No statistically significant results were found in either variable for ER or PR+ and Unknown hormone receptor status categories.

(3) Correlation and missing data mechanism issues in secondary variables

The secondary variables such as performance status and quality of life assessment were repeatedly measured over time. More than 50% of patients dropped out of the study before 9 months. If these data were analyzed at each visit by comparing a mean of each variable to baseline, the following assumptions were necessary: (1) the correlation among observations per subject is independent and (2) the missing data mechanism at work in the study was the so-called "missing completely at ramdom" (MCAR) type. These two assumptions are very unlikely to hold in clinical trial settings.

These reviewers applied a growth curve approach to the secondary variables, specifically to performance status, pain score and global quality of life score in quality of life questions by QLQ-C30, EORTC. In this approach a linear mixed effects model (Laird and Ware, 1982) and a generalised estimating equation approach (Liang and Zeger, 1986) were employed for the correlation issue and the concept of a "pattern-mixture model" (Little, 1993 and 1995) was applied for the evaluation of the missing data mechanism.

The most plausible observed missing data mechanism was judged to be a "nonignorable" missing mechanism in the three treatment arms so that separate analyses were performed for completers and for dropouts.

In the pain score and global quality of life variables no statistical significance in a time trend over the study period was found in completers across the three treatment arms and in dropouts across the three treatment arms. Note that the time trend for completers was different from the one for dropouts.

On the other hand, in the performance status variable different time trends were found for completers across the three treatment arms. In the two letrozole treatment arms - 0.5 mg and 2.5 mg - performance status was not changed over the study period in completers, but in the megestrol acetate treatment arm performance status score was increased over the study period in completers, indicating that performance status became worse in the megestrol acetate arm. No apparent difference in the time trend in dropouts across the three treatment arms was observed. Performance status became worse over the study period in dropouts across the three treatment arms. Note that this variable was treated as a continuous variable in the analyses, even though this is an ordinal variable so that the interpretation of the results should be cautious.

(4) Multiple Comparison Problems

Three treatment comparisons were performed by the sponsor. In the protocol dated October 27, 1992, there was no statement regarding which treatment comparison was of primary interest. Therefore, strictly speaking, a multiple comparison adjustment should be applied. A conservative adjustment procedure such as a Bonferroni adjustment uses p=0.017 (0.05/3) as the criteria at the 0.05 significance level since the false positive error rate increases when more than a single comparison is undertaken. However, in the statistical analysis plan dated April 11, 1995, treatment contrasts were stated as follows: "The three treatment contrasts, in decreaing order of importance are: 2.5 mg letrozole vs 160 mg megestrol acetate, 0.5 mg letrozole vs 160 mg megestrol acetate, 0.5 mg letrozole vs 2.5 mg letrozole". This reviewer notes that the database was frozen on June 26, 1995. Since this comparisons' ranking in order of importance was stated very close to the study's database closure, the argument for claiming prospective identification of the primary comparison is weak. In addition, it is also noted that this is the single 'protocol' study in the application presented with complete data analysis. Thus, this reviewer feels that a conservative statistical adjustment is appropriate.

VI. Brief Summary of AR/BC3 Trial

The AR/BC3 trial was "an open, randomized, multicenter, comparative between patients, out-patient, Phase IIb/III trial in postmenopausal women with advanced breast cancer, who had previously progressed with an anti-estrogen (e.g. tamoxifen) given as adjuvant therapy and/or first-line treatment for advanced disease."

The study population consisted of "postmenopausal women with locally advanced or loco-regionally recurrent or metastasizing breast cancer who previously progressed on or following anti-estrogens given as adjuvant therapy and/or treatment for advanced disease." Patients were randomized to one of three treatments; once daily doses of 0.5 mg letrozole or 2.5 mg letrozole or twice daily 500mg aminoglutethimide plus daily 30 mg HC or 37.5 mg CA. The responders (complete or partial response) or patients with stable disease (no change) stayed in the study until disease progression or until any other reason necessitated discontinuation.

(I) The Primary Variables

The primary variables in this trial are tumor response (peer reviewed, confirmed), duration of response, time to progression, time to treatment failure, and time to death. In this review we