CENTER FOR DRUG EVALUATION AND RESEARCH

APPROVAL PACKAGE FOR:

APPLICATION NUMBER

20-449/S-018

Statistical Review(s)

STATISTICAL REVIEW AND EVALUATION

DEPARTMENT OF HEALTH AND HUMAN SERVICES
PUBLIC HEALTH SERVICE
FOOD AND DRUG ADMINISTRATION
CENTER FOR DRUG EVALUATION AND RESEARCH

NDA NO./SUPPLEMENT:	20-449 / SE 18
DATE RECEIVED BY THE CENTER:	February 1, 2002
DRUG NAME:	Taxotere™ (docetaxel)
INDICATION:	Patients with unresectable locally advanced or metastatic non-small cell lung cancer (NSCLC) who have not previously received chemotherapy for this condition.
DOCUMENTS REVIEWED:	NDA Vols. 1 – 4, 34 – 185 received on Feb. 1, 2002. Report received on Feb. 11, 2002. SAS codes received on June 17 & 28, 2002. Sponsor's presentation slides on March 13, 2002. Submission on Aug. 19, 2002 in response to Agency's questions.
BIOMETRICS DIVISION:	**DIVISION OF BIOMETRICS I (HFD-710)**
REVIEWING STATISTICIAN:	Peiling Yang, Ph.D.
TEAM LEADER:	Gang Chen, Ph.D.
DEPUTY DIVISION DIRECTOR:	Kooros Mahjoob, Ph.D.
MEDICAL DIVISION:	**ONCOLOGY DRUG PRODUCTS (HFD-150)**
PROJECT MANAGER:	Ms. Ann Staten
MEDICAL OFFICER:	Ramzi Dagher, M.D.

DISTRIBUTION: NDA 20449/SE18

HFD-150/Ms. Ann Staten
HFD-150/Dr. Donna Griebel
HFD-150/Dr. Ramzi Dagher
HFD-710/Dr. Peiling Yang
HFD-710/Dr. Gang Chen
HFD-710/Dr. Kooros Mahjoob
HFD-710/Dr. George Chi
HFD-700/Dr. Charles Anello

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

1 EXECUTIVE SUMMARY AND STATISTICAL FINDINGS

1.1 CONCLUSIONS AND RECOMMENDATIONS

Two test regimens (docetaxel in combination with either cisplatin or carboplatin) were compared to the vinorelbine+cisplatin combination. The primary endpoint was overall survival. The Kaplan-Meier median estimates of overall survival were 10.9 months for the docetaxel+cisplatin combination, 9.1 months for the docetaxel+carboplatin combination, and 10.0 months for the vinorelbine+cisplatin combination. There was no statistical evidence for survival superiority of docetaxel in combination with either cisplatin or carboplatin to the vinorelbine+cisplatin combination. However, it was estimated that the docetaxel+cisplatin combination preserved at least 62% of the effect of the vinorelbine+cisplatin combination. Assessment of the effect of the vinorelbine+cisplatin combination in non-inferiority analysis was based on the upper limit, 0.86, of the 95% confidence interval for the hazard ratio (vinorelbine+cisplatin / cisplatin) since there was only one single historical randomized trial appropriate for the assessment. There was no statistically significant finding in any secondary endpoint.

1.2 OVERVIEW OF STUDIES REVIEWED

The sponsor has submitted a supplemental New Drug Application to support the use of TAXOTERE® plus cisplatin (or plus carboplatin) for the treatment of patients with unresectable locally advanced or metastatic non-small cell lung cancer (NSCLC) who have not previously received chemotherapy for this condition. The proposed new indication is to be supported by the results of two independent phase III trials: TAX 326 and TAX . , in addition to a series of phase I- and II- studies.

Study TAX might serve as a supportive docetaxel single-agent trial in a future application for the first-line treatment of NSCLC which would include additional randomized controlled trials of the safety and efficacy of docetaxel as part of combination therapy. In response to the Agency's comments, the sponsor has submitted the pivotal study TAX326 for the proposed indication in this application.

Study TAX326 was a multicenter, open-label, three parallel group, randomized phase III study comparing the effects of the drug combinations (docetaxel + cisplatin, docetaxel + carboplatin) to the control regimen (vinorelbine + cisplatin) in chemotherapy-naive patients with NSCLC. A total of 140 centers participated the study: 52 of those were in the United States and Canada, 60 in Europe, 14 in South America, 14 in New Zealand and South Africa and 11 in Middle East. A total of 1220 patients (408, 407, 405, respectively) were included in the study. The primary endpoint was overall survival. Secondary endpoints included response rate, duration of response, time to disease progression, QoL, weight change, change of karnofsky performance status, change of pain score. Both non-inferiority and superiority analyses were performed by the sponsor. In the proposed label, the sponsor claimed superiority of the docetaxel+cisplatin

combination to the control (vinorelbine+cisplatin combination) in addition to claiming benefits of some secondary efficacy endpoints. This reviewer's did not agree to the sponsor's claims because of many statistical/technical issues involved, including change of the primary analysis, failure to control the familywise false positive rate due to multiple comparisons, multiple endpoints or multiple analyses, unequal lengths of treatment cycles between the treatment arms, etc. For more details, please refer to Reviewer's Comments in Section 2.2.9.

Based on this reviewer's analysis, there was no statistical evidence for survival superiority of the docetaxel in combination with either cisplatin or carboplatin to the vinorelbine+cisplatin combination. However, it was estimated that the docetaxel+cisplatin combination preserved at least 62% of the effect of the vinorelbine+cisplatin combination. Assessment of the effect of the vinorelbine+cisplatin combination in non-inferiority analysis was based on the upper limit, 0.86, of 95% confidence interval for the hazard ratio (vinorelbine+cisplatin / cisplatin) since there was only one single historical randomized trial appropriate for the assessment. There was no statistically significant finding in any secondary endpoint.

1.3 PRINCIPAL FINDINGS

Two test regimens (docetaxel+cisplatin and docetaxel+carboplatin) were compared to the control (vinorelbine+cisplatin). The primary endpoint was overall survival. The Kaplan-Meier median survival estimates were 10.9 months for the docetaxel+cisplatin combination, 9.1 months for the docetaxel+carboplatin combination, and 10.0 months for the active control.

It was to be noted that an interim analysis was performed at the nominal significance level of 0.005 (two-sided). In order to control the two-sided error rate at a level of 0.05 in a familywise manner on multiple analyses, the nominal significance level at the final analysis was reduced to 0.047. At the final analysis, since each test regimen was compared to the control, the Hochberg procedure, proposed by the sponsor and accepted by the Agency, was performed to adjust for multiple comparisons. The first step in the Hochberg procedure was to compare each p-value to 0.047. Since no statistical significance was found, Step 2 was carried out, where the smaller of the two p-values was compared to 0.0235 (= 0.047/2). Based on the stratified logrank test, the hazard ratio of the docetaxel+cisplatin combination to vinorelbine+cisplatin combination was estimated to be 0.892 with a nominal 97.65% confidence interval of (0.744, 1.070). This confidence interval was also considered the adjusted 95% confidence interval, adjusting for interim analysis and multiple comparisons.

Since only one historical trial was appropriate for the assessment of the active control effect in non-inferiority analysis, this reviewer considered the 95%-CI approach as the primary approach to estimating the effect of the vinorelbine+cisplatin combination. That is, the upper limit of the 95% confidence interval for the hazard ratio (vinorelbine + cisplatin / cisplatin) in the single historical trial was regarded the effect of the vinorelbine+cisplatin combination. This resulted in the hazard ratio (test regimen / active control) cutoff for non-inferiority being 1.078. Since the confidence interval for the hazard ratio entirely lied below the cutoff 1.078, there was statistical evidence for the alternative hypothesis that the docetaxel+cisplatin combination preserved at least 50% of the effect of the vinorelbine+cisplatin combination. Moreover, it was estimated that the docetaxel+cisplatin combination preserved at least 62% of the effect of the vinorelbine+cisplatin combination. All other analyses (supportive, exploratory, sensitivity) support the conclusion that

- There was no statistical evidence for superiority of docetaxel in combination with either cisplatin or carboplatin to the vinorelbine+cisplatin combination.
- There was statistical evidence that the docetaxel+cisplatin combination, but not the docetaxel+carboplatin combination, preserved at least 50% of the effect of the vinorelbine+cisplatin combination.

There was no statistical evidence for survival non-inferiority of the other test regimen, the docetaxel+carboplatin combination to the active control. There was also no statistical evidence for survival superiority of either test regimen to the active control.

2 STATISTICAL REVIEW AND EVALUATION OF EVIDENCE

2.1 BACKGROUND

The sponsor has submitted a supplemental New Drug Application to support the use of TAXOTERE® plus cisplatin (or plus carboplatin) for the treatment of patients with unresectable locally advanced or metastatic non-small cell lung cancer (NSCLC) who have not previously received chemotherapy for this condition. The proposed new indication is to be supported by the results of two independent phase III trials: TAX 326 and TAX in addition to a series of phase I- and II- studies.

Study TAX might serve as a supportive docetaxel single-agent trial in a future application for the first-line treatment of NSCLC which would include additional randomized controlled trials of the safety and efficacy of docetaxel as part of combination therapy. In response to the Agency's comments, the sponsor has submitted the pivotal study TAX326 for the proposed indication in this application. This study compared the efficacy and safety of docetaxel in combination with either cisplatin or carboplatin to the active control (vinorelbine+cisplatin) in the proposed indication.

2.2 STUDY TAX 326

Protocol Title: A Multicenter, Multinational, Randomized Phase III Study of Docetaxel (RP56976, TAXOTERE®) Plus Carboplatin Versus Vinorelbine Plus Cisplatin in Chemotherapy-Naive Patients with Unresectable Locally Advanced and/or Recurrent (Stage IIIB) or Metastatic (Stage IV) Non-Small Cell Lung Cancer.

2.2.1 Introduction

This was a multicenter, open-label, three parallel group, randomized phase III study comparing the effects of the drug combinations (docetaxel+cisplatin and docetaxel+carboplatin) to the control regimen (vinorelbine + cisplatin) in chemotherapy-naive patients with NSCLC. A total of 140 centers participated the study: 52 of those were in the United States and Canada, 60 in Europe, 14 in South America, 14 in New Zealand and South Africa and 11 in Middle East. The three regimens are summarized in the following:

[1] **Test Regimen A: Docetaxel and Cisplatin (D75+Cis)**
- Docetaxel, 75 mg/m^2 administered intravenously (IV) over 60 minutes on Day 1, immediately followed by
  Cisplatin, 75 mg/m^2 administered IV over 30-60 minutes.
- Dexamethasone premedication will be administered orally as described above.
- Patients should be adequately hydrated and receive an anti-emetic regimen.

- Each chemotherapy cycle will be repeated every 21 days.

[2] **Test Regimen B: Docetaxel and Carboplatin (D75+Cb6)**
- Docetaxel, 75 mg/m^2 administered IV over 60 minutes on Day 1, immediately followed by
  Carboplatin, AUC = 6 mg/ml•min administered IV over 30-60 minutes.
- Dexamethasone premedication will be administered orally as described above.
- Patients should receive an anti-emetic regimen.
- Each chemotherapy cycle will be repeated every 21 days.

[3] **Active Control: Vinorelbine and Cisplatin (V+Cis)**
- Vinorelbine, 25 mg/m^2 administered IV over 6-10 minutes into a free-flowing IV infusion of NS or D5W on Days 1, 8, 15 and 22, followed by
  Cisplatin, 100 mg/m^2 administered IV on day 1 only.
- It is recommended that following the weekly vinorelbine infusions, the vein be flushed with sufficient volume to prevent injection site reactions.
- Patients should be adequately hydrated and receive an anti-emetic regimen.
- Each chemotherapy cycle will be repeated every 28 days.

According to the original protocol, all patients were to be treated until evidence of progression disease, unacceptable adverse event(s) or completion of 6 treatment cycles. After completion of 6 cycles of treatment, patients might be continued on their randomized treatment at the discretion of the treating physician. After discontinuation of study treatment, all patients were to be followed at regularly scheduled (2 months) intervals until death.

The sponsor has made several changes in the development of the protocol and the Statistical Analysis Plan (SAP). Important correspondences between the Agency and the sponsor before the NDA submission are dated and summarized in the following:

[1] Original Protocol:
- Dated May 14, 1998; correspondence date: June 11, 1998, SN 678.
- Planned Start Date: 2Q98.
- Planned Recruitment End Date: 4Q99.
- Planned Follow-up: 4Q00.
- **Randomization**: would be stratified by disease stage.
- **Sample size**: 360 patients per group to have a power of about 90% for superiority test and a power of about 85% for non-inferiority test. Percent of retention and margin for non-inferiority test were not mentioned.
- No analysis methods were specified for endpoint except for the sample size estimation.
- **Interim Analysis**: After 50% of patients had completed chemotherapeutic treatment phase and having completed 9 months of follow-up, an interim analysis would be performed. Following the O'Brien-Fleming method, the logrank or Wilcoxon test must show a significant difference in overall survival at the nominal significance level of 0.005. This would adjust the Type I error rate at 0.047 for the final analysis (p. 8-53-28 of vol. 86).
- Quality of life assessment would be administered only in countries for which validated language translations were available.

[2] First Amendment:
- Dated January 29, 1999.
- No statistical conduct was involved.

[3] Second Amendment:

- Dated March 4, 1999.
- No statistical conduct was involved.

[4] Statistical Analysis Plan (SAP):

- Dated July 19, 2000; submission date: Aug. 2, 2000 (SN. 880).
- The total sample size would be 1080 based on the logrank test.
- Two comparisons would be made: each of the two test regimens versus the active control. No multiplicity adjustment would be made on the two comparisons.
- The margin for non-inferiority test would be 33%, but justification was not provided.
- The cutoff date for interim analysis would be July 10, 2000.
- The hypothesis for non-inferiority would be tested first. If the result were significant, the hypothesis for superiority would be tested.
- The primary analysis for **superiority** test in survival would be the logrank test stratified by stage of malignant disease and the geographic region of cancer treatment.
- The primary analysis for **non-inferiority** test in survival would be based on the proportional hazards model including treatment, the stratification factors: disease stage and region. In addition, the following covariates would also be included: (1) age, (2) performance status, (3) time from diagnosis of disease to randomization, (4) weight loss in the prior 6 months, (5) histological subtype, (6) level of albumin, (7) level of LDH, (8) sex, (9) baseline QoL score, (10) liver involvement and (11) bone involvement.
- **The non-parametric, covariance-adjusted logrank test stratified by the two pre-specified stratification factors would be a secondary analysis for survival data.** The covariates would jointly be: age, performance status, time from diagnosis of disease to randomization, history of weight loss in the prior 6 months, histology, level of albumin, level of LDH, sex, baseline QoL score and type of organ involved (liver and bone).
- Proportional hazards regression model in a stepwise regression fashion would also be performed to investigate the potential effects of covariates.
- **Quality of life assessments would be primarily based on Lung Cancer Symptom Scale (LCSS)** and secondarily on EurQoL (EQ5D) scores.
- **The Agency did not agree to the setting in non-inferiority test.**
- The Agency requested the sponsor clarify whether the trial would be considered negative upon failing one of the two comparisons.

[5] Interim Analysis Report:

- Dated August 28, 2000.
- The cut-off date for survival data was July 10, 2000.
- A total of 601 patients were included in the interim analysis.
- The primary analysis for non-inferiority test was based on the proportional hazards model as specified in the SAP. However, one covariate (level of albumin) pre-specified in the SAP was missing from the model in the interim analysis. The Sponsor concluded that both test regimens were non-inferior to the active control, each at the nominal significance level of 0.05.
- The primary analysis for superiority was stratified logrank test. The Sponsor concluded that no strong statistical evidence was seen that either test regimen was superior to the active control, each at the nominal significance level of 0.005.

[6] Meeting:

- Requested by the sponsor on Dec. 22, 2001.
- Held on February 22, 2001 (SN. 907).
- **The Agency did not agree to the sponsor's setting in non-inferiority test.**
- **The Agency responded to the sponsor that multiplicity adjustment should be made for the two comparisons and that the Hochberg method could be used for the adjustment**

as mentioned in the SAP: "*statistical significant will apply to both comparisons if both have p<0.05 or it will apply to one of them if it has p<0.028*" (p. 8-58-51 of vol. 91).

[7] Amendment to SAP:

- Submission date: March 19, 2001 (SN. 925).
- The sponsor proposed a closed testing procedure to identify the extent to which non-inferiority would be supportive at a level that was more stringent than the value of 0.75 as addressed in the initial primary hypothesis. The sponsor also indicated that no alpha adjustment would be made for the two comparisons, but if there were concern for the multiple comparisons, the sponsor would resolve that in the process with the Hochberg method. **The Agency did not agree to the Sponsor's proposal.**
- In view of the possibilities that the Hochberg method would be needed to manage the two comparisons for multiplicity adjustment, **the sponsor would like to change the primary analysis for superiority test from the stratified logrank test to the non-parametric covariance adjusted logrank test. The originally proposed test would become supportive. The Agency did not make any comments on this proposal.**

[8] Study Report:

- Dated January 27, 2002 on Report; submitted on February 1, 2002.
- Study Start Date: July 25, 1998.
- Study Complete Date: August 9, 2001.

2.2.2 Data Analyzed and Sources

Data used for review are from the electronic submission received on 2/01/02. The network path is "\\Cdsesub1\n20449\S_018\2001-02-01\crt\Datasets\tax326" in the EDR. The Sponsor's SAS codes are received on June 17 for the paper copy and June 28 for the electronic copy. The network path for the electronic copy is "\\Cdsesub1\n20449\S_018\2002-06-28\Programs".

2.2.3 Study Objectives

The primary objective of the study was to compare the effects of the two drug combinations, docetaxel plus cisplatin and docetaxel plus carboplatin, to the active control regimen of vinorelbine plus cisplatin on overall survival in chemotherapy-naive patients with unresectable locally advanced and/or recurrent (Stage IIIB) or metastatic (Stage IV) non-small cell lung cancer (NSCLC). The secondary objectives were to compare the time to progression, overall objective response rate, duration of response, quality of life and safety between the three treatment regimens.

2.2.4 Efficacy Endpoints

The primary efficacy endpoint was survival. Secondary efficacy endpoints included objective tumor response, duration of response, time to progression, quality of life, Karnofsky performance status, and weight changes.

2.2.5 Sample Size and Power Considerations

The protocol states that a total of 1080 patients (360 per arm) will be included in the study. This is based on the following assumptions:

- Of interest is each of the two test regimens compared to the active control. No multiplicity adjustment is made for the two comparisons.
- The postulated median survival time for each test regimen is 10.5 months and for the active control is 8 months.
- The patient accrual time is 18 months (changed to 10 months in SAP).
- The follow-up time after the last cycle is 12 months.
- By controlling the two-sided Type I error rate per-comparison at a level of 0.05, the proposed sample size will have about 90% power (changed to about 88% in SAP) to detect the desired effect in superiority test for each comparison. It will also have about 85% power in non-inferiority test. The SAP states that the power for non-inferiority is calculated assuming equal hazard rates (i.e., hazard ratio = 1).

The test statistic used in sample size estimation is not stated in the protocol, but is stated in SAP, based on the logrank statistic for superiority test (p. 8-58-39 of vol. 91).

Details regarding percent of retention and margin are not stated in the protocol but are referred to the future SAP (p. 8-53-80 of vol. 86 in the sponsor's Study Report). In SAP, the margin is chosen to be 1/3, which means that to establish non-inferiority it is required to show that the hazard ratio of a test regimen to the active control is less than 1.33. It is also equivalent to show that the hazard ratio of the active control to a test regimen is greater than 0.75 (the reciprocal of 1.33)." However, the choice of the margin was not agreed upon by the Agency.

2.2.6 Stratification and Randomization

The protocol states that randomization will be administered centrally and computer-generated randomization logs will be used for each stratum of disease stage, IIIB or IV, (p. 5-53-79 of vol. 86). However, in the SAP one more stratification factor is included: the geographic region of cancer treatment (p 8-58-46 of vol. 91). It is not stated in the SAP as to how the region will be divided.

2.2.7 Efficacy Analysis Methods

2.2.7.1 Analysis Populations Defined by Sponsor

[1] **ITT Population:** The primary analysis was to be performed on the intent-to-treat (ITT) population. However, the SAP states that patients who were misdiagnosed at baseline and identified later as not having NSCLC will be excluded from the ITT population. (p. 8-58-43 of Vol. 91 in SAP). This has resulted in a slightly different definition of the ITT population from the Agency's (all randomized patients).

[2] **Evaluable Population**: To be evaluable, a patient should satisfy the following conditions:

- Must be eligible;
- Must have started at least one infusion of study treatment;
- Must not have had any concomitant systemic anti-cancer therapy other than the study drugs during the course of the study. If such therapy is received, the follow-up until this time is evaluable and is censored afterward.

[3] **Population Evaluable for Response:** A patient included in the response-evaluable population should satisfy the following conditions:

- Must be eligible;
- Must have received a minimum of two cycles of treatment unless progression occurred before the second cycle, in which case the patient is considered as evaluable with an early progression;
- All baseline lesions must have assessed at least once after the second cycle, with the same method of measurement as baseline.

[4] **Population Evaluable for Safety:** To be evaluable for safety, a patient should have started at least one infusion, whether they are eligible or not.

2.2.7.2 Multiplicity Adjustment

The SAP states that a total of two comparisons (each of the test regimens compared to the active control) will be made. It further states that no multiplicity adjustment for the two comparisons will be made because the purpose of including the two test regimens (docetaxel combinations with different platinum compounds) in this study is to provide for the treatment choices between geographical regions of cancer treatment. The sponsor believes that cisplatin combination (test regimen A) is almost always necessary in regions with economic priorities, while a carboplatin combination (test regimen B) is highly preferred in other countries due to case of administration and less toxicity anticipated (p. 8-58-38 of Vol. 91). However, the SAP also states that

> *"if there is excessive controversy concerning the error rate per comparison having primary relevance and being 0.05, then the Hochberg method will be used to account for the two pairwise treatment comparisons. For this method, statistical significance will apply to both comparisons if both $p<0.05$ or it will apply to one of them if it has $p<0.028$."*

2.2.7.3 Survival

Survival was defined as the duration from the date of randomization to the date of death from any cause. Survival was censored at

- the last contact date for patients lost to follow-up without a known record of death,
- the cutoff date for patients alive for all follow-up in the study.

The sponsor's primary analysis for superiority test was different from that for non-inferiority test.

[1] **Superiority Test:**

- According to the sponsor's SAP, the primary analysis was to be the stratified logrank test by disease stage (IIIB vs. IV) and the geographic region (4 regions). The non-parametric, covariate-adjusted stratified logrank test would also be performed but would be considered as a supportive analysis. These covariates would be (1) age, (2) performance-status, (3) time from diagnosis of disease to randomization, (4) history of weight loss in the prior 6 months, (5) histology, (6) level of albumin, (7) level of LDH, (8) sex, (9) baseline QoL score and (10) type of organ involved (liver and bone). Proportional hazards regression model (also known as Cox model) in a stepwise regression fashion would also be performed to investigate the potential effects of covariates. The SAP does not state the possible categories for each covariate.
- The Agency did not agree to the sponsor's intention of not controlling the overall Type I error rate for the two comparisons. Therefore, in the Amendment to SAP the sponsor proposed to change the primary analysis from the stratified logrank test to the non-parametric covariate-adjusted stratified logrank test and the originally proposed test would become supportive. The Agency did not make any comments on this proposal.

[2] **Non-inferiority Test:**
- According to the SAP, the primary analysis was to be based on the Cox model, which was considered by the sponsor as supportive in the superiority test.
- The Agency did not agree to the setting in the non-inferiority test. Therefore, in the Amendment to SAP, the sponsor proposed a closed testing procedure to identify the extent to which non-inferiority would be supportive at a value that was more stringent than a threshold of 0.75 for the hazard ratio as addressed in the initial primary hypothesis. The Agency did not agree to the sponsor's proposal.

[3] **Testing Order:** Non-inferiority test would be performed first for each of the two comparisons. If the result were significant, superiority test would be performed.

2.2.7.4 Time to Disease Progression (TTP)

Time to disease progression (TTP) was defined as the duration from the date of randomization to the date of documentation of disease progression or to the date of death without a documented disease progression. Time to progression was censored at

- the last contact date for patients lost to follow-up,
- the cutoff date for patients reaching the time point of study without a known record of progression, and
- the date of this subsequent therapy for patients who received a subsequent chemotherapy, radiotherapy or surgery and did not have a documented progression.

According to the sponsor's SAP, the stratified logrank test by the two stratification factors was to be performed.

2.2.7.5 Objective Response Rate

The overall response rate was defined as the sum of complete and partial response rate. According to the sponsor's SAP, best overall response rate was to be compared using the Fisher's exact test on the ITT as well as the evaluable-for-response (defined by the sponsor) population.

2.2.7.6. Duration of Response

According to the sponsor's SAP (p.8-58-56), duration of response was defined as the time from the randomization date to the first documentation of progression or death. Duration of response was censored at the date of the last assessment before that therapy for patients receiving further anticancer therapy. Duration of response was to be described in responders (CR or PR) only by the Kaplan-Meier method.

2.2.7.7 Quality of life (QoL)

According to the sponsor's SAP (p. 8-58-53), the primary instrument for the QoL assessment was the Lung Cancer Symptom Scale (LCSS) and the secondary instrument was EuroQol Scale, a global assessment of patient's health status.

[1] LCSS contains two parts:
- **Patient scale,** consisting of 9 items with each item having 100 mm visual analog scale.
- **Observer scale,** consisting of 6 items with each having a five-point scale.

According to the sponsor's SAP, the primary LCSS score would be the global QoL item (item 9) rated by the patient, and the total patient score as well as the observer score would be secondary.

[2] EuroQoL contains two parts:
- A global health assessed on a scale of 0 to 100.
- Five specific questions regarding mobility, self care, usual activities, pain/discomfort and anxiety/depression on a 3-point scale.

According to the sponsor's SAP, the global score would be primary.

[3] The sponsor's **analysis population** would be all randomized patients with an evaluable baseline assessment and at least one evaluable assessment during the dosing period would be included in the analysis of each scale.

[4] The sponsor's **analysis methods**:
- A fixed interval of 21/28 days (an ideal treatment cycle) would be used for all QoL analyses. Such an interval was termed a period by the sponsor.
- An analysis of covariance (ANCOVA) was to be performed on the subset of ITT patients who have paired QoL assessments at baseline and the last available assessment during the treatment phase. This model would include the covariates specified in the superiority test.
- A longitudinal analysis using a mixed model with change from baseline as the response would also be performed to account for all the available assessments during the study, including follow-up visits. In this analysis, the assumed covariance structure would be the compound symmetry.

[5] The sponsor's SAP states two levels of missing data:

- **Missing item(s) in the given evaluation.** If more than 1/3 of the items were missing, the QoL assessment would not be evaluable; otherwise, the missing values would be replaced by the overall mean of the given item.
- **The totally missing evaluation at the targeted time interval.** The assumption of randomly missing information would be evaluated through a longitudinal analysis with a mixed model approach. If the assumption of non-informative missingness holds, an analysis of covariance model would be used for the endpoint analysis of the global item of LCSS as well as the subscale total scores rated by patient and observer/nurse, with the set of covariates stated in the superiority test.

2.2.7.8 Performance Status, Body Weight and Pain Control

According to the sponsor's SAP, mean changes from baseline in the Karnofsky Performance Status ratings would be compared between the treatment groups. Of interest would be the changes at the early three cycles. In addition, changes at the last assessment would also be compared using a t-test.

According to the sponsor's SAP, mean weight changes from baseline during the study treatment period would be compared. In addition, all weight changes would be categorized into three groups: weight loss < 5%, weight loss ≥ 5% but < 10%, and weight loss ≥ 10% relative to baseline. Fisher's exact test would be used in the categorical data.

According to the sponsor's SAP, the pain score would be assessed through the LCSS by the patient and by an observer (a nurse or caregiver at the clinical site). Changes from baseline in the pain score would be analyzed using a multivariate mixed model (longitudinal analysis).

2.2.8 Interim Analysis

According to the sponsor's SAP: The purpose of the interim analysis was to evaluate if there was an overwhelming convincing difference in the primary outcome measure of survival between the treatment groups, and if so, the number of patients exposed to the inferior treatment could be minimized by switching treatment. An interim analysis was to be performed when 50% of patients had completed the chemotherapeutic treatment phase and a 9-month follow-up. Survival would be the only efficacy endpoint to be analyzed. Two comparisons would be made: each of the test regimens versus to the active control. For each of the two comparisons the nominal significance level used at the interim analysis was 0.005 and at the final analysis was 0.047 based on the logrank test (p. 8-53-80 of vol. 86 and p. 8-58-60 of vol. 91 in the sponsor's Study Report).

2.2.9 Sponsor's Results and Reviewer's Comments

A total of 1220 patients were randomized to the three regimens: 408 to test regimen A (D75+Cis), 407 to test regimen B (D75+Cb6) and 405 to the active control (V+Cis). The sponsor excluded 2 of the 1220 patients from the ITT population because the two patients did not have NCSLC. Table 1 is a summary of patient populations defined by the sponsor.

Table 1: Sponsor-Defined Populations

Population	Test Regimen A (D75+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)	Total
Randomized Patients	408	407	405	1220
Sponsor-defined ITT Population	408	406	404	1218
Evaluable for Efficacy	402	398	381	1181
Evaluable for Response	366	363	336	1065
Evaluable for Safety	406	401	396	1203

2.2.9.1 Results of Interim Analysis

A total of 601 patients (210, 199 and 201, respectively) were included in the interim analysis with a total of 420 (139, 143 and 138, respectively) deaths observed by the cutoff date July 10, 2000.

In the interim analysis both non-inferiority and superiority tests were performed. For non-inferiority test, the primary analysis was based on the proportional hazards model (Cox model) as specified in the SAP. However, one covariate (level of albumin) pre-specified in the SAP was not included in the model of the interim analysis. The sponsor concluded that both test regimens were non-inferior to the active control, each at the nominal significance level of 0.05 (i.e., the error rate was controlled at a level of 0.05 per comparison at interim analysis). For superiority test, the primary analysis was stratified logrank test. The sponsor concluded that no strong statistical evidence was seen that either test regimen was superior to the active control, each at the nominal significance level of 0.005.

Reviewer's Comments:

[1] Since the purpose of the interim analysis was to see if there was an overwhelming convincing difference in the primary outcome measure of survival between the treatment groups, non-inferiority test should not be performed and no conclusion regarding non-inferiority should be drawn. If performed, same analysis (refer to Bullet [2] in Section 2.2.9.3) and same nominal significance level as in the superiority test should be used for a consistent comparison.

2.2.9.2 Baseline Characteristics

The sponsor's results of baseline and disease characteristics are in Table 2 and Table 3, respectively. Gender, karnofsky performance status, race and region of study treatment were balanced among the three treatment groups. The percentage of missing values for "percent weight loss in the last 6 months before randomization" approached 17% overall. Of those reported, the test regimen B (D75+Cb6) group had more patients and a higher percentage (37.9%) of patients with > 5% weight loss in the preceding 6 months, as compared to the other treatment groups. Disease characteristics at baseline were similar across the 3 treatment groups. Most patients were randomized to study treatment within 3 months of initial lung cancer diagnosis.

Table 2: Sponsor's Baseline Characteristics (on Sponsor-Defined ITT Population)

[Source: Table 18 on p. 8-49-162 in Sponsor's Final Study Report]

Baseline Characteristics	Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
Age (years)			
Mean	60.1	58.9	59.6
Median	61	59	61
Min, Max	30, 81	23, 87	35, 80
Gender			
Female	114 (27.9%)	114 (28.1%)	102 (25.2%)
Male	295 (72.1%)	292 (71.9%)	302 (74.8%)
Race			
Black	13 (3.2%)	13 (3.2%)	6 (1.5%)
Caucasian	360 (88.2%)	353 (86.9%)	360 (89.1%)
Hispanic	23 (5.6%)	31 (7.6%)	26 (6.4%)
Oriental	5 (1.2%)	2 (0.5%)	7 (1.7%)
Other	7 (1.7%)	7 (1.7%)	5 (1.2%)
Karnofsky Performance Status			
70			
80	15 (3.7%)	16 (3.9%)	16 (4.0%)
90	157 (38.5%)	154 (37.9%)	153 (37.9%)
100	171 (41.9%)	170 (41.9%)	167 (41.3%)
	65 (15.9%)	66 (16.3%)	68 (16.8%)
Percent Weight Loss in the Last 6 Months			
0 – 5	214 (52.5%)	198 (48.8%)	195 (48.3%)
5 – 10	63 (15.4%)	78 (19.2%)	56 (13.9%)
≥ 10	62 (15.2%)	76 (18.7%)	78 (19.3%)
Missing	69 (16.9%)	54 (13.3%)	75 (18.6%)
Geographic Regions			
Unites States / Canada	115 (28.2%)	115 (28.3%)	113 (28.0%)
Europe / Lebanon / Israel	197 (48.3%)	197 (48.5%)	197 (48.8%)
South America / Mexico	63 (15.4%)	62 (15.3%)	62 (15.3%)
South Africa / Australia / New Zealand	33 (8.1%)	32 (7.9%)	32 (7.9%)

Table 3: Sponsor's Disease Characteristics (on Sponsor-Defined ITT Population)

[Source: Table 19 on p. 8-49-162 in Sponsor's Final Study Report]

Baseline Characteristics	Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
Staging			
Stage IIIB	135 (33.1%)	132 (32.5%)	133 (32.9%)
Stage IV	273 (66.9%)	274 (67.5%)	271 (67.1%)
Histologic Subtypes			
Adenocarcinoma	181 (44.4%)	169 (41.6%)	164 (40.6%)
Squamous Cell Carcinoma	132 (32.4%)	136 (33.5%)	140 (34.7%)
Large Cell Carcinoma	41 (10.0%)	48 (11.8%)	49 (12.1%)
Bronchioalveolar Carcinoma	15 (3.7%)	11 (2.7%)	11 (2.7%)
Other	39 (9.5%)	42 (10.3%)	40 (9.9%)
Time from Diagnosis to Randomization (months)			
Mean	4.6	4.2	3.8
Median	0.9	0.9	0.9
Minimum	0.1	0.0	0.0
Maximum	164.3	117.8	187.1
Number of Organs Involved at Baseline			
One	70 (17.2%)	64 (15.8%)	73 (18.1%)
Two	199 (48.8%)	199 (49.0%)	180 (44.6%)
Three or More	139 (34.1%)	143 (35.2%)	151 (37.4%)
Selected Metastatic Sites			
Adrenal Gland	44 (10.8%)	51 (12.6%)	59 (14.6%)
Bone	66 (16.2%)	86 (21.2%)	82 (20.3%)
Brain	9 (2.2%)	5 (1.2%)	6 (1.5%)
Liver	54 (13.2%)	54 (13.3%)	44 (10.9%)
Lung	389 (95.3%)	370 (91.1%)	378 (93.6%)
Lymph Nodes	269 (65.9%)	267 (65.8%)	256 (63.4%)
Pleura	76 (18.6%)	80 (19.7%)	77 (19.1%)
Skin	2 (0.5%)	6 (1.5%)	11 (2.7%)

2.2.9.3 Analyses of Survival (Primary Efficacy Endpoint)

The cutoff date for survival was April 3, 2001 (p. 8-49-132 of vol. 82). The SAP (submitted on August 2, 2000) states the following two important issues:

- Two comparisons (each of test regimens vs. active control) will be made, but no multiplicity adjustment will be made; i.e., the Type I error rate will be controlled at 0.05 level for each comparison.
- Both non-inferiority and superiority tests will be performed. The primary analysis for superiority test will be the stratified logrank test, and for non-inferiority test will be based on proportional hazards model (Cox model) with pre-specified covariates.

The Agency replied that (1) multiplicity adjustment for the two comparisons was needed unless the trial would be claimed as negative upon failing one of the two comparisons, and (2) the setting (percent of retention, margin, etc.) for non-inferiority test was not acceptable.

Following the Agency's reply the sponsor requested a meeting for discussion and/or clarification. The meeting was held on February 22, 2001. At the meeting the Agency commented that (1) controlling the Type I error rate per comparison was not acceptable and the Hochberg method mentioned in SAP for controlling the familywise error rate was acceptable, and (2) the setting for non-inferiority was not acceptable.

Following this meeting, the sponsor submitted the Amendment to SAP on March 19, 2001, where the sponsor proposed a methodology regarding the setting in non-inferiority test. In addition, the sponsor also requested change of the primary analysis from the stratified logrank test to the non-parametric covariate-adjusted stratified logrank test in testing superiority of survival because of the possibility of using the Hochberg method to control the familywise error rate as opposed to the per-comparison error rate. The Agency replied that the proposal on the non-inferiority setting was not acceptable. However, the Agency did not comment on the proposal to change of the primary analysis. Subsequently, the primary analysis was changed to the non-parametric one by the sponsor.

A total of 1220 patients were randomized. The sponsor's analyses were performed on the 1218 patients, who were determined after randomization to have NSCLC. Overall, 78% of patients died before the study was complete. Percent of deaths among the three treatment groups was nearly the same, as seen in Table 4.

Table 4: Sponsor's Descriptive Results of Survival

	Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
Number of Deaths	307 (75.2%)	319 (78.6%)	323 (80.0%)
Number of Censored	101 (24.8%)	87 (21.4%)	81 (20.0%)
-- Alive	93 (22.8%)	75 (18.5%)	70 (17.3%)
-- Loss to Follow-up	8 (2.0%)	12 (3.0%)	11 (2.7%)

Because one interim analysis was performed, at which the nominal significance level was 0.005, the nominal significance level for the final analysis was reduced to 0.047 accordingly. The sponsor's primary analysis was the non-parametric covariate-adjusted stratified logrank test, where covariates and stratification factors included in the model are listed in Table 5. The sponsor's results for each of the two comparisons based on this adjusted analysis are summarized in Table 6. Based on the sponsor's primary analysis, the p-value for comparing test regimen A (D75+Cis) to the active control (V+Cis) was 0.044 and the p-value for comparing test regimen B (D75+Cb6) to the active control (V+Cis) was 0.66. Since the p-value 0.044 was less than 0.047, the sponsor concluded that test regimen A (D75+Cis) was associated with a longer time to survival as compared to the active control (p. 8-49-174 in the sponsor's Study Report). Since the hazard ratio cannot be obtained in a non-parametric logrank test setting, the sponsor employed a stratified proportional hazards model adjusted for the same set of covariates to estimate the hazard ratios. The estimated hazard ratio of the active control to test regimen A was 1.183 with a 95% confidence interval of (1.008, 1.388). The estimated hazard ratio of active control to test regimen B) was 1.048 with a 95% confidence interval of (0.894, 1.229). Although, superiority of test regimen B to the active control could not be established (p-value = 0.66), the sponsor concluded that the non-inferiority was achieved not only with a threshold of 0.75, but with a threshold up to 0.89 for the hazard ratio (sponsor's Study Report on p. 8-49-187).

In the sponsor's NDA presentation held on March 13, 2002 and the Study Report (starting on p. 8-49-289 in Section 10.4), the sponsor provided results of literature search regarding the preservation of the control effect in the non-inferiority test. Only one[1] of the four papers provided by the sponsor was appropriate for estimating the effect of the active control relative to the historical control. Based on the information provided in this paper, in which cisplatin (Cis) was considered as the historical control, the sponsor provided the information that the hazard ratio of V+Cis to Cis was estimated to be 0.74 with a 95% confidence interval of (0.65, 0.86) and concluded the following:

(1) Preservation of the control effect based on the HR was more than 100% for D+Cis and 66% for D+Cb.
(2) D+Cis preserved more than 75% (with 99% probability) of the effect of size of V+Cis even under very stringent criteria; i.e., adjusting for multiplicity and using the upper 95% CI for the control effect.
(3) D+Cis was non-inferior to V+Cis.

Table 5: Covariates in Sponsor's Primary Analysis of Survival in Superiority Test

Covariates		
1)	Age	≤ 60 vs. > 60 years
2)	Performance status	KPS = 100 vs. < 100
3)	Time from diagnosis of disease to randomization	> vs. ≤ 60 days
4)	Weight loss in the prior 6 months	< vs. ≥ 5%
5)	Histological subtype	(a) Adenocarcinoma vs. others (b) Squamous cell carcinoma vs. others (c) LCU vs. others
6)	Level of albumin	≤1 vs. >1 (only in SAS program)
7)	Level of LDH	LDH ≤ vs. > ULN
8)	Sex	Female vs. Male
9)	Baseline QoL score	(a) LCSS QOL Today > vs. ≤ 60 (b) EQ5D Global Health Status > vs. ≤ 60
10)	Liver involvement	No vs. yes
11)	Bone involvement	No vs. yes
12)	**Prior radiotherapy[a]**	**Yes vs. no**
13)	**Prior surgery[a]**	**Yes vs. no**
Stratification Factors		
1)	Disease stage	IIIB vs. IV
2)	Geographical region of cancer treatment	US/CANADA vs. Europe/Morocco/Israel vs. South Africa/Australia/New Zealand vs. South America/Mexico

[a] These two covariates were not pre-specified in the SAP, but were included in the model in a post-hoc manner.

[1] Wozniak et al., "Randomized Trials Comparing Cisplatin With Cisplatin Plus Vinorelbine in the Treatment of Advanced Non-Small-Cell-Lung Cancer: A Southwest Oncology Group Study"; *Journal of Clinical Oncology*, Vol. 16, No. 7, 1998.

Table 6: Sponsor's Primary Analysis Results of Survival (on Sponsor-Defined ITT Population)

Sponsor-Defined ITT Population	Comparison 1		Comparison 2	
	Test Regimen A (D75+Cis)	Active Control (V+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)
Adjusted [a] Median (m.)	11.3	10.1	9.4	9.9
1-Year Survival Rate (95% CI)	46% (42%, 51%)	41% (36%, 46%)	38% (33%, 43%)	40% (35%, 45%)
2-Year Survival Rate (95% CI)	21% (16%, 25%)	14% (10%, 18%)	18% (13%, 22%)	14% (10%, 18%)
P-Value [a]	0.044		0.66	
Hazard Ratio [b] (95% CI)	1.183 (1.008, 1.388)		1.048 (0.894, 1.229)	

[a] Based on non-parametric covariate-adjusted stratified logrank test

[b] Hazard ratio of the active control to the test treatment. Stratified proportional hazards model (known as Cox model) adjusted for covariates. A hazard ratio of greater than 1 indicates the test regimen is associated with a longer survival time as compared to the active control.

Reviewer's Comments And Analyses:

[1] **Issues of Change of Primary Analysis for Superiority Test.** Since the Agency requested that multiplicity adjustment for the two comparisons (each test regimen to the active control) be made, the sponsor proposed the Hochberg procedure[2] and changed the primary analysis from the stratified logrank test to the non-parametric covariate-adjusted stratified logrank test. Although the Agency did not comment on the change of the primary analysis, **the change is not acceptable because**

(a) The stratified logrank test had already been performed at the interim analysis. The rationale for change of the primary analysis at the final analysis due to multiplicity adjustment is not acceptable.
(b) Most covariates included in the adjusted analysis are not acceptable to the Agency.
(c) The nominal significance levels for the interim and the final analyses were calculated based on the stratified logrank test, not the test adjusted for covariates. To control the false positive rate, nominal significance levels might need to be adjusted with change of analysis because the correlation in test statistics between the interim and the final analysis might be changed.

[2] **Inconsistent Analyses for Superiority and Non-inferiority Tests.** In the sponsor's statistical analysis plan, the stratified proportional hazards model adjusted for covariates was proposed for the non-inferiority test. Typically, identical analysis should be used for both superiority and non-inferiority tests for consistent results. Since the methods were not clearly addressed in the sponsor's statistical analysis plan, the Agency was unaware of the inconsistency and, hence, did not comment on this. The stratified proportional hazards

[2] The Hochberg procedure for two comparisons is illustrated as following: Step 1: each p-value is compared to a significance level α. If both p-values are less than α, one may infer significant findings in both comparisons; else, proceed to Step 2. Step 2: if the smaller of the two p-values is less than $\alpha/2$, one may infer significant finding in the comparison corresponding to the smaller p-value; else, stop.

model adjusted for covariates is different from the stratified non-parametric covariate-adjusted logrank test. Although both are covariate-adjusted analyses, the assumptions required for the analyses are different. The former requires constant hazard ratio in each of possible combinations of levels of covariates; the latter requires balanced distributions of covariates between treatment arms. Although the Agency did not comment on the proposed analysis for the non-inferiority test, **the sponsor's proposed analysis is not acceptable** because

(a) Same analysis should be used for both superiority and non-inferiority tests to avoid contradictory results. An example of a paradox is as following:

- In non-inferiority analysis, not only non-inferiority but also superiority is concluded. This would lead to the inference that "the test regimen preserved more than 100% of the active control effect". In contrast, in superiority analysis, superiority (and even non-inferiority) can not be concluded, contradictory to the quoted inference.

(b) Same reason as in (b) in the previous bullet.

[3] **Issues of Sponsor's Superiority Analysis.** In the sponsor's study report, the survival curves as well as survival rates at 1 and 2 years were based on the stratified non-parametric covariate-adjusted logrank analysis, instead of Kaplan-Meier estimates, and the estimated hazard ratios 1.183 and 1.048 were based on the stratified Cox-proportional hazards model. In addition, two covariates (prior radiotherapy: yes or no, prior surgery: yes or no) included in the adjusted analyses were not pre-specified in the statistical analysis plan and classifications of possible outcomes for some covariates were not pre-specified either. Moreover, based on the p-values reported by the sponsor, which were 0.044 and 0.657, there was no evidence for superiority of either test treatment using the Hochberg procedure for multiple comparisons.

[4] **Issues of Sponsor's Non-inferiority Analysis.**

(a) Based on the current available methodology in non-inferiority analysis, it is required to estimate the active control effect relative to a historical control. In this NDA submission, only one provided reference (by Wozniak et al., 1998) was appropriate for estimating the active control effect relative to the historical control, cisplatin. This has made the inference less reliable.

(b) Since in this reference (by Wozniak et al., 1998) the logrank test stratified by center and disease stage was used, the sponsor should use the same analysis, not the stratified proportional hazards model adjusted for covariates, for a meaningful comparison.

(c) When estimating the percent of the active control effect preserved by the test treatments, the sponsor did not consider the variability of the estimated active control effect.

(d) Since the nominal significance level for the final analysis was 0.047 due to one interim analysis, inferences drawn should be based on 95.3%, instead of 95%, confidence intervals.

[5] **Reviewer's Analyses.**

(a) **Determination of Non-inferiority Margin.** The non-inferiority margin depends on the active control effect. Since it is unknown, several methods have been proposed to

estimate the effect. Of those used by the Agency included the 95%-CI approach[3] and the Rothmann et al. approach[4]. The 95%-CI approach estimates the active control effect by the upper limit of the 95% confidence interval for the hazard ratio (the active control / historical control). In the Rothmann et al. approach, the active control effect does not result from a fixed confidence level. Depending on historical trials, it may result from a relatively low confidence level compared to 95%. In general, the 95%-CI approach tends to be more conservative than the Rothmann et al. approach. In this NDA, only one historical trial was available for estimating the active control effect, making the Rothmann et al. approach considerably liberal. Therefore, this reviewer considered the 95%-CI approach to be the primary approach to estimating the active control effect in non-inferiority analysis and the other approach exploratory. Assume that non-inferiority requires the test regimens to preserve at least 50% of the active control effect. Based on the 95%-CI approach, the cutoff for the hazard ratio (test regimen / V+Cis) was 1.078, corresponding to a non-inferiority margin of 7.8%. That is, non-inferiority is established if there is strong statistical evidence for the hazard ratio (test regimen / V+cis) to be less than 1.078. For more detailed derivation, please refer to Section 3.1.1.

(b) **Reviewer's Primary Analysis.** This reviewer considered the stratified logrank test as the primary analysis for both superiority and non-inferiority tests because of issues discussed in Bullets [1], [2] and [4](b). In addition, this reviewer included all randomized patients in the ITT population, resulting in 2 more patients than the sponsor-defined ITT population. The results of this reviewer's primary analysis are summarized in Table 7. The non-inferiority test based on the Hochberg procedure to control the familywise false positive rate (one-sided) at a level of 0.0235 for multiple comparisons proceeds as following:

Step 1: Construct 95.3% confidence intervals (corresponding to a nominal two-sided significance level of 0.047, equivalent to one-sided significance level of 0.0235) for the hazard ratio of each test regimen to the active control. If both confidence intervals entirely lie below 1.078 (refer to the previous bullet), then non-inferiority evidence is shown in both test regimens. Otherwise, proceed to Step 2 to determine if non-inferiority evidence is shown in one of the two test regimens.

Step 2: If there is non-inferiority evidence, it should be in the comparison resulting in a smaller p-value. Base on the data, it is in the comparison of test regimen A to the active control. Thus, construct a 97.65% confidence interval (corresponding to a nominal two-sided significance level of 0.0235, equivalent to one-sided significance level of 0.01275) for the hazard ratio of test regimen A to the active control. If the confidence interval entirely lies below 1.078, then non-inferiority evidence is shown for test regimen A.

[3] CBER memo. "Excerpts from a CBER Memorandum Discussing Aspects of Active Comparator Trials of Thrombolytics in AMI".

[4] Rothmann, Li, Chen, Chi, and Tsou (2001). "Non-inferiority methods for mortality trials". *ASA Proceedings of the Biopharmaceutical Section.*

As seen in Table 7, not both 95.3% confidence intervals for the hazard ratios entirely lied below 1.078, so one should proceed to Step 2. The 97.65% confidence interval for the hazard ratio of D75+Cis to V+Cis was (0.737, 1.059), entirely below 1.078. This suggested statistical evidence for non-inferiority of test regimen A (D75+Cis). Based on the 97.65% confidence interval, the effect of test treatment A to the historical control (Cis) was estimated to be 0.910 (=1.059*0.86). Since the effect of the active control was estimated to be 0.86 after incorporating the variability of the point estimate, it was estimated that test treatment A (D75+Cis) preserved 62% (= *ln* 0.910 / *ln* 0.86) of the active control effect. However, it should be cautioned that the results were based on only one historical trial and, in which trial, patient characteristics may not be identical to those in this registration trial. Therefore, the non-inferiority results might differ had more historical trials been available. Regarding the superiority test, there was no statistically significant evidence for superiority in either test regimen (p-values: 0.122, 0.657),

Table 7: Reviewer's Primary Analysis of Stratified Logrank Test (on All Randomized Patients)

	Comparison 1	Comparison 2
	Test Regimen A (D75+Cis) vs. Active Control (V+Cis)	Test Regimen B (D75+Cb6) vs. Active Control (V+Cis)
P-value [a]	0.122	0.657
Estimated Hazard Ratio [b]	0.884	1.036
95.3% CI [c]	(0.754, 1.036)	(0.885, 1.212)
97.65% CI [d]	(0.737, 1.059)	Not needed.

[a] From the superiority test "H_0: hazard ratio = 1 vs. H_1: hazard ratio ≠ 1".
[b] Hazard ratio of test treatment to the active control. A hazard ratio of less than 1 indicates that the test treatment is associated with a longer time to survival.
[c] Corresponding to a nominal significance level of 0.047.
[d] Corresponding to a nominal significance level of 0.0235.

(c) **Reviewer's Supportive Analysis on Non-inferiority.** The results were consistent with those based on the 95%-CI approach; i.e., there was no statistical evidence for superiority in either test regimen, but there was statistical evidence for non-inferiority in test regimen A (D75+Cis). It was estimated that test regimen A (D75+Cis) preserved 78% of the active control effect. For a detailed derivation, please refer to Section 3.2. It is noted that the percent of retention was larger using the Rothmann et al. approach because the Rothmann et al. approach resulted in using a relatively low confidence level (less than 30%) for the active control effect. Because of only one historical trial, the variability of the active control effect may be underestimated.

(d) **Reviewer's Sensitivity Analyses on Different Populations.** To explore the robustness of the results on the ITT population, this reviewer preformed the stratified logrank test on evaluable patients as defined by the sponsor, on the population consisting of those with disease stage remaining unchanged after randomization, and on the population for which some patients' disease status was adjudicated by the Medical Officer, Dr. Ramzi Dagher. The results in general indicated a consistent conclusion except when performed on the population consisting of those with disease stage remaining unchanged after randomization, the percent of the active control effect that test

treatment D75+Cis preserved was estimated to be only 49% based on the 95%-CI approach. For more detailed analyses, please refer to Section 3.3.

(e) **Reviewer's Supportive Analyses Based on Other Test Procedures.** The medical officer, Dr. Ramzi Dagher, considered only three of the covariates as potentially relevant, which were performance status, gender and weight loss at baseline. The conclusion based on the stratified proportional hazards model adjusted for these three covariates was consistent with that based on this reviewer's primary analysis, stratified logrank test; i.e., statistical evidence in non-inferiority was only seen in test regimen A (D75+Cis) and there was no statistical evidence for superiority in either test regimen. This reviewer also performed the unstratified logrank test to explore the influence of the stratification factors, but found consistent results as in the stratified logrank test. It was estimated that that test treatment D75+Cis preserved at least 59% of the active control effect based on either of the two test procedures. For more detailed analyses, please refer to Section 3.4.

(f) **Reviewer's Kaplan-Meier Curves.** This reviewer obtained the Kaplan-Meier estimates of median survival as seen in Table 8. The Kaplan-Meier survival curves are described in Figure 1, and Figure 2.

Table 8: Reviewer's Kaplan-Meier Estimates of Median Survival

	Test Regimen A (D75+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)
All randomized patients	408	407	405
# censored	87 (21.4%)	101 (24.8%)	81 (20.0%)
Median (in months)	10.9	9.1	10.0

Figure 1: Reviewer's Kaplan-Meier Survival Curves (Test Regimen A vs. Active Control)

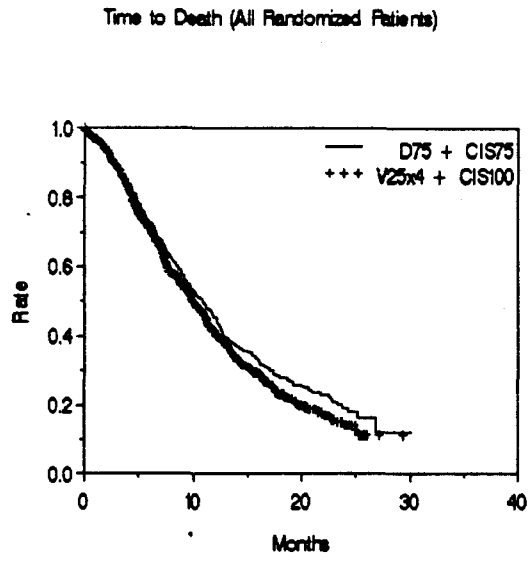
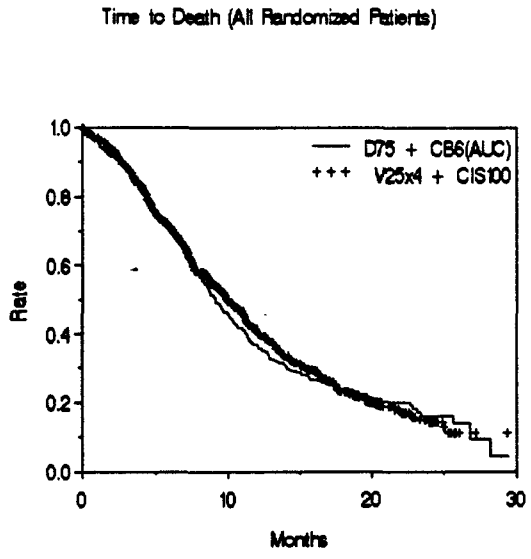


Figure 2: Reviewer's Kaplan-Meier Survival Curves (Test Regimen B vs. Active Control)



(g) **Reviewer's Exploratory Subgroup Analyses by Selected Prognostic Factors.** Several subgroup analyses were explored by this reviewer and are summarized in Section 3.6. There was no evidence against non-inferiority of test regimen A (D75+Cis) to the active control (V+Cis) in these exploratory analyses.

(h) **Reviewer's Exploratory Analyses of Comparison between Two Test Regimens.** Please see Section 3.5.

2.2.9.4 Analyses of Response Rates and Duration of Response (Secondary Efficacy Endpoints)

Response rate was analyzed on the sponsor-defined ITT population as well as the sponsor-defined response-evaluable population.

The sponsor's results are summarized in Table 9 and Table 10. As seen in Table 9, among the three treatment arms, test regimen A (D75+Cis) yielded the numerically highest, and test regimen B (D75+Car) the numerically lowest, overall response rate, whether on the sponsor-defined ITT or the sponsor-defined response-evaluable population.

In comparison of test regimen A (D75+Cis) to the active control (V+Cis), the p-value based on Fisher's exact test was 0.029 on the sponsor-defined ITT population and 0.074 on the sponsor-defined response-evaluable patients. In comparison of test regimen B (D75+Cb6) to the active control (V+Cis), the p-value was relatively large on either population.

The sponsor defined the duration of response as the time from the date of randomization and obtained the adjusted median duration of survival, based on non-parametric covariate-adjusted stratified logrank test, as in Table 11.

Table 9: Sponsor's Descriptive Results of Response Rate

Population [a]	Response Rate	Test Regimen A (D75+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)
ITT	**Overall (CR+PR) [95% CI[b]]**	**129/408 (31.6%) [27.1%, 36.4%]**	**97/406 (23.9%) [19.8%, 28.3%]**	**99/404 (24.5%) [20.4%, 29.0%]**
	Complete (CR)	8 (2.0%)	5 (1.2%)	8 (2.0%)
	Partial (PR)	121 (29.7%)	92 (22.7%)	91 (22.5%)
Response-Evaluable	**Overall (CR+PR) [95% CI[b]]**	**127/366 (34.7%) [29.8%, 39.8%]**	**96/363 (26.4%) [22.0%, 31.3%]**	**95/336 (28.3%) [23.5%, 33.4%]**
	Complete (CR)	8 (2.0%)	5 (1.4%)	7 (2.1%)
	Partial (PR)	119 (32.5%)	91 (25.1%)	88 (26.2%)

[a] Both populations are sponsor-defined.
[b] Nominal 95% confidence interval.

Table 10: Sponsor's P-values for Analysis of Overall Response Rate

Population[b]	Comparison 1	Comparison 2
	Test Regimen A (D75+Cis) Vs. Active Control (V+Cis)	Test Regimen B (D75+Cb6) Vs. Active Control (V+Cis)
ITT	0.029	0.870
Response-Evaluable	0.074	0.611

Note: P-values are based on Fisher's exact test for H_0: equal response rates vs. H_1: unequal response rates.
[b] Both populations are sponsor-defined.

Table 11: Sponsor's Adjusted Median Duration of Response from Date of Randomization

Population	Comparison 1		Comparison 2	
	Test Regimen A (D75+Cis)	Active Control (V+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)
Sponsor-defined evaluable population	32 weeks	34 weeks	31 weeks	35 weeks

Reviewer's Comments:

[1] **Response Rate.** The p-values for testing equal response rates between the test regimen A (D75+Cis) and the active control (V+Cis) were 0.029 and 0.074 on the sponsor-defined ITT and the response-evaluable populations, respectively. Although a value of 0.029 seemed small, there was no statistically significant evidence that test regimen A yielded a higher overall response rate compared to the active control when the Hochberg procedure was employed for multiple comparisons with the control. In summary, there was no statistically significant evidence that either test regimen was associated with a higher overall response rate as compared with the active control.

[2] **Duration of Response.** The duration of response defined by the sponsor is not acceptable. It should be defined as the duration from the date of response, not the date of randomization. In addition, the median duration of response should be obtained by the Kaplan-Meier estimates (also pre-specified in the sponsor's statistical analysis plan), not the non-parametric covariate-adjusted stratified logrank test (post-hoc). The latter may lead to different estimates of the median duration of response for the active control arm as seen in Table 11. This reviewer's descriptive summary of duration of response is in Table 12. As seen in this table, the median duration of response was around 150 days for test regimen A (D75+Cis), 141 days for test regimen B (D75+Cb6) and 173 days for the active control (V+Cis). It is to be noted that duration of response is considered on responders only. It should not be compared between treatment groups because the respective responder subgroups are treatment-outcome dependent.

Table 12: Reviewer's Results of Duration of Response and Response Rate

Population		Test Regimen A (D75+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)
All randomized patients	Proportion of responders [Adjusted [b] 95% CI]	129/408 (= 31.6%) [26.5%, 36.8%]	97/407 (= 23.8%) [19.2%, 28.6%]	99/405 (= 24.4%) [19.8%, 29.2%]
	Proportion censored for duration	37/129 (= 28.7%)	21/97 (= 21.7%)	22/99 (= 22.5%)
	Median [c] duration [Adjusted 95% CI]	21.3 [18.1, 24.3]	20.1 [16.6, 23.7]	24.7 [21.6, 26.1]
Response-evaluable [a]	Proportion of responders [Adjusted 95% CI]	127/366 (= 34.7%) [29.2%, 40.3%]	96/363 (= 26.4%) [21.4%, 31.6%]	95/336 (= 28.3%) [22.9%, 33.8%]
	Proportion censored for duration	36/127 (= 28.3%)	21/96 (= 21.9%)	21/336 (= 22.3%)
	Median duration [Adjusted 95% CI]	21.6 [18.1, 25.4]	20.1 [16.6, 23.4]	24.7 [21.0, 27.1]

Note: This reviewer defined duration of response from the date of response.
[a] Sponsor-defined response-evaluable population.
[b] Adjusted for multiple comparisons on this endpoint based on the Hochberg procedure.
[c] Kaplan-Meier estimates in weeks.

2.2.9.5 Analysis of Time to Disease Progression (TTP)

During the study, most of the patients had a determination of disease progression. The sponsor reported that most assessments were performed at the end of every other treatment cycle within the first 26 weeks. Since each chemotherapy cycle was repeated every 3 weeks in these test regimen groups and every 4 weeks in the active control group, the sponsor commented that the assessments of time to progression might have been biased due to the differences in the target tumor assessment intervals. The sponsor further commented that it was more likely that TTP was inflated in the treatment group with longer cycles (i.e., the active control group).

Table 13: Sponsor's Descriptive Summary of Time to Disease Progression on Sponsor-Defined Population

Population	Comparison 1		Comparison 2	
	Test Regimen A (D75+Cis) [N = 408]	Active Control (V+Cis) [N=404]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N = 404]
Censored patients	92 (23%)	91 (23%)	60 (14.8%)	91 (23%)
Median TTP (weeks) [95% CI[a]]	22 [21, 25]	23 [21, 27]	20 [19.0, 22.5]	22 [19.0, 25.0]

[a] Nominal 95% confidence interval based on the sponsor's stratified non-parametric covariate-adjusted analysis.

Reviewer's Findings and Comments:

[1] The medians obtained by the sponsor were not based on the Kaplan-Meier method; they were adjusted medians using the sponsor-proposed non-parametric covariate-adjusted stratified logrank test. The sponsor's approach resulted in different median estimates in the same active control group (V+Cis) between the two comparisons (see bolded numbers in Table 13). This reviewer's descriptive results of median TTP on all randomized patients based on the Kaplan-Meier method is summarized in Table 14. It is noted in this table that there is substantially lower percentage of censoring in patients randomized to test regimen B (D75+Cb6).

[2] Because of unequal durations of a cycle between the test regimen groups and the control group, a bias may be introduced in analysis of TTP. Therefore, comparisons between test regimens with the active control are not appropriate.

Table 14: Reviewer's Descriptive Summary of Time to Disease Progression on All Randomized Patients

	Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 407]	Active Control (V+Cis) [N = 405]
Censored patients	92 (22.5%)	60 (14.7%)	92 (22.7%)
Median TTP (weeks) [Adjusted [a] 95% CI]	21.4 [19.3, 24.6]	19.4 [18.1, 21.3]	22.1 [18.1, 25.6]

[a] Adjusted for multiple comparisons on this endpoint based on the Hochberg procedure.

2.2.9.6 Analyses of Quality of Life (QoL)

QoL instruments were completed prior to the first treatment infusion (within 14 days or less before initiation of chemotherapy), during chemotherapy administration (prior to each new cycle), at the end of study-treatment and during the follow-up period (every two months). The instruments were administered only in countries where a translated version of the QoL with validation was available (sponsor's Study Report on p. 8-49-266 in vol. 82).

Two validated instruments were used: LCSS and EuroQoL. The sponsor considered the global QoL item "Quality of Life Today" as the primary LCSS endpoint and the "Global Health State" item as the primary EuroQoL endpoint and performed two analyses, longitudinal analysis and analysis of covariance, for each endpoint. The sponsor used a fixed interval of 21 days (i.e., the sponsor-defined period) for all QoL analyses because the length of treatment cycle in these test regimen groups was different from that in the active control group. The sponsor's analysis results are summarized in Table 15. The Sponsor concluded that an improvement in "Global Health State" was seen in both test regimen groups as compared to the active control group in both analyses and an improvement in "Quality of Life Today" was seen in test regimen B as compared to the active control in both analyses.

Table 15: Sponsor's Results of QoL Analyses

Instrument / Endpoint	Analysis	Test Regimen A (D75+Cis) vs. Active Control (V+Cis) [N = 313]	Test Regimen B (D75+Cb6) vs. Active Control (V+Cis) [N = 307]
LCSS / Quality of Life Today	Longitudinal	0.064	0.016
	Covariance	0.216	0.012
EuroQoL / Global Health State	Longitudinal	0.016	< 0.001
	Covariance	0.014	< 0.001

Reviewer's Comments:

[1] The results based on the sponsor's QoL analyses might be biased due to a high rate of missing data and other issues related to multiple endpoints and multiple analyses as summarized in the following:

(a) **Issue of Missing Data:** Many patients in countries where translations of QoL instruments were not available did not participate in QoL evaluation. In addition, a proportion of patients who participated in the evaluation had missing values at baseline assessment, much more at post baseline assessments. This reviewer summarizes the distribution of the number of patients with values available at baseline for each primary score on each instrument as in Table 16 (for LCSS) and Table 17 (for EqurQOL), respectively. It was observed from these two tables that near 28% of patients did not participate in QoL evaluations and around 8% of patients, although participated, had missing values at baseline on the primary score. This indicated that the sponsor's QoL results were mainly driven by at most 64% of the ITT population provided that no missing values at post baseline assessments. When missing at post baseline assessments was considered, the percent of missing values was even greater. It was not clearly pre-specified in the sponsor's SAP how to handle missing values and what was considered an evaluable patient for QoL analysis (more detailed "evaluable" definition in the sponsor's Study Report than in the SAP). With such a high rate of missing data, results based on the sponsor's QoL analyses might be biased and can not be extended to the randomized ITT population.

Table 16: Reviewer's Summary of Baseline Values on LCSS Evaluations

	Test Regimen A (D75+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)
ITT Population	408	407	405
Patients participating in LCSS evaluations	295/408 (= 72.3%)	290/407 (= 71.3%)	291/405 (= 71.9%)
Patients who had baseline values of the Global QoL Today item (i.e., bp9 ≠ .)	261/408 (= 64.0%)	253/407 (= 62.2%)	259/405 (= 64.0%)

Note: Based on the sponsor's ulcss.xpt data set.

Table 17: Reviewer's Summary of Baseline Values on EuroQoL Evaluations

	Test Regimen A (D75+Cis)	Test Regimen B (D75+Cb6)	Active Control (V+Cis)
ITT Population	408	407	405
Patients participated in EuroQoL evaluations	295/408 (= 72.3%)	291/407 (= 71.5%)	288/405 (= 71.1%)
Patients who had baseline values of the global health item (i.e., be6 ≠ .)	268/408 (= 65.7%)	268/407 (= 65.8%)	265/405 (= 65.4%)

Note: Based on the sponsor's ueq5d.xpt data set.

(b) **Issue of Multiple Endpoints:** According to the sponsor's SAP, LCSS was the primary instrument to assess quality of life and the primary LCSS score was the global QoL item rated by the patient; i.e., variable '*p9*' in the data set "ulcss.xpt". EuroQoL scale was the secondary instrument and the primary EuroQoL score was the global score; i.e., var '*e6*' in the data set "ueq5d.xpt". Since the primary instrument was LCSS, analysis based on the secondary instrument EqurQOL could only be exploratory if the sponsor failed to show statistical significance based on the primary instrument LCSS. Based on the sponsor analysis results from the LCSS instrument as summarized in Table 15, no statistical significance was found in comparison of the test regimen A (D75+Cis) to the active control (V+Cis). Hence, analysis based on the secondary instrument should not be considered. If the sponsor intended to make a claim based on results from either instrument, then a procedure regarding how to allocate alpha to the two instruments should be pre-specified in the SAP in order to control the false positive rate with regard to the QoL endpoint. Moreover, the false positive rate should also be controlled in a familywise manner on all the secondary endpoints of which the sponsor intended to be make a claim.

(c) **Issue of Multiple Analyses:** Although the sponsor pre-specified the primary score to be analyzed for each instrument, several statistical procedures (analyses) for the primary score were performed, for example, "longitudinal" and "covariance" as in Table 15. The former incorporated all the available assessments during the study in the model and the latter only compared the last available assessment with the baseline values. All the covariates (as used in the sponsor's primary analysis of the survival endpoint) were included in both longitudinal and covariance analyses. Other analyses were also proposed in the SAP. It was not clear which was the primary analysis. In addition, some of the covariates included in the analyses were not appropriate. If the sponsor intended to make a claim based on results from any of the performed analyses, then a procedure regarding how to allocate alpha to these analyses should be pre-specified in the SAP in order to control the false positive rate of the QoL endpoint.

(d) **Issue of Different QoL Assessment Times:** QoL was assessed in each cycle. However, cycle lengths were different between test regimen arms (3 weeks) and the control arm (4 weeks). Inconsistent assessment time points may introduce bias in QoL analysis as pointed out in analysis of the TTP endpoint.

2.2.9.7 Analysis of Weight Change from Baseline

Weight changes from baseline during the 6 treatment periods were compared using a t-test. In addition, all weight changes were categorized into 3 levels: weight loss ≥ 5% but < 10%, weight loss ≥ 10%, and the Cochran-Mantel-Haenszel (CMH) test was used. The sponsor concluded that results of both analyses were in favor of both test regimens as compared to the active control (all p-values < 0.001).

Reviewer's Comments:

The results based on the sponsor's analyses were biased and might be misleading because of issues related to multiple analyses and unequal cycle lengths as summarized in the following:

[1] **Issue of Multiple Endpoints:** It is noted from the proposed label that the sponsor intended to make a claim on several secondary endpoints. The false positive rate should be controlled in a familywise manner on all the secondary endpoints of which the sponsor intended to make a claim.

[2] **Issues of Multiple Analyses:** Some analyses were mentioned, but only one (Fisher's Exact test) was clearly pre-specified in the sponsor's SAP. The p-value (<0.001) the sponsor obtained was not based on this pre-specified analysis but rather a post-hoc analysis. If the sponsor intended to make a claim based on one of the performed analyses, these analyses should be clearly pre-specified and alpha adjustment due to multiple analyses should be made to control the false positive rate.

[3] **Issue of Unequal Cycle Lengths:** Weight was assessed in each cycle and the weight loss was measured by the sponsor as the change from baseline during the study treatment. It is noted that timing for the last available weight assessment varied from patient to patient. In addition, since the cycle length for the control group was one week longer than the treatment group, patients in the control group were more likely to lose more weight than those in the treatment groups in any given cycle, especially in a later cycle.

2.2.9.8 Analysis of Change of Karnofsky Performance Status (KPS)

Changes in performance status from baseline were compared at the early 6 periods using a longitudinal logistic regression model. For this analysis, the change was categorized into two levels: decrease of 20% on KPS or more versus otherwise. The Sponsor concluded that the results were in favor of the test regimens as compared with the active control (p-values were 0.028, <0.001, respectively).

Reviewer's Comments:

The results based on the sponsor's analyses are not reliable and might be misleading because of issues related to multiple analyses, post-hoc analyses, and unequal cycle lengths as summarized in the following:

[1] **Issue of Multiple Endpoints:** Please refer to Reviewer's Comment [1] in Section 2.2.9.7.

[2] **Issue of Multiple Analyses:** The sponsor states in SAP that of interest will be the changes at the early three cycles; in addition, changes at the last assessment will also be compared using a t-test. Since the primary interest is at the early three cycles, if no statistical significance is found at the early three cycles, no other analysis should be considered. If the sponsor intends to make a claim based on any of performed analyses, these analyses should be pre-specified in the SAP and the false positive rate should be controlled for multiple analyses.

[3] **Issue of Post-Hoc Analysis:** The inference made by the sponsor was based on a post-hoc analysis, where (i) comparison at the early 6 periods and (ii) the longitudinal logistic analysis and the cutoff for change of KPS (20% decrease or not) were used in a post-hoc manner.

[4] **Issue of Unequal Cycle Lengths:** Please refer to Reviewer's Comment [3] in Section 2.2.9.7.

2.2.9.9 Analysis of Pain Score

The sponsor concluded an improvement in pain through the whole study in test regimen A as compared to the active control using a multivariate Mixed model (longitudinal analysis).

Reviewer's Comments:

The results based on the sponsor's pain score analyses were not reliable and might be misleading because of issues related to multiple endpoints, multiple analyses and poor quality of data. Please refer to this reviewer's comments in Section 2.2.9.6.

2.2.10 Reviewer's Conclusions and Recommendation

Based on the stratified logrank test, the hazard ratio of the docetaxel+cisplatin combination to vinorelbine+cisplatin combination was estimated to be 0.892 with a nominal 97.65% confidence interval of (0.744, 1.070). This confidence interval was also considered the adjusted 95% confidence interval, adjusting for interim analysis and multiple comparisons. Since only one historical trial was appropriate for estimating the active control effect in non-inferiority analysis this reviewer considered the 95%-CI approach as the primary approach to estimating the active control effect. This resulted in the hazard ratio cutoff for non-inferiority being 1.078. It was estimated that D75+Cis preserved at least 62% of the active control effect. Except from one sensitivity analysis where the percent of the active control effect preserved by the D75+Cis combination was 49%, results from all other analyses (such as analyses performed on different patient populations than the ITT population, different approaches in estimating the active control effect, different test procedures than stratified logrank test) lead to the same conclusion that

- There was no statistical evidence for superiority of either test regimen.
- There was statistical evidence that the D75+Cis combination, but not the D75+Cb6 combination, preserved at least 50% of the effect of the V+Cis combination.

There was no statistical significance in any secondary endpoint after adjusting for multiplicity (multiple comparisons, multiple endpoints, multiple analyses, unequal cycle length, a high rate of missing data, etc.).

3 APPENDIX

3.1 DETERMINATION OF NON-INFERIORITY MARGIN

In the sponsor's presentation, the hazard ratio of the active control (V+Cis) relative to the historical control cisplatin (Cis) was estimated as 0.74 with the corresponding 95% confidence interval of (0.65, 0.86) based on the Wozniak et al. paper.

3.1.1 The 95%-CI Approach

The 95%-CI approach considers the upper limit of the 95% confidence interval of the hazard ratio (V+Cis / Cis) to be the estimated active control effect. Since the 95% confidence interval was (0.65, 0.86), the estimated active control effect was 0.86. In order for the test regimens to preserve at least 50% of the active control effect, the hazard ratio of test regimens to cisplatin needed to be no larger than 0.927 (= $e^{0.5*ln(0.86)}$. Note that the log-scale was used for a logical calculation). Therefore, the hazard ratio of test regimens to the V+Cis combination needed to be no lager than 1.078 (= 0.927/0.86). That is, the cutoff for the hazard ratio (test regimen / V+Cis) in the non-inferiority test was 1.078, corresponding to a non-inferiority margin of 7.8%.

3.1.2 The Rothmann et al. Approach

The active control effect and Δ (i.e., the cutoff for the log hazard ratio used in non-inferiority hypothesis test) were calculated through log-scale. The log of the hazard ratio was estimated to be $ln(0.74) = -0.301$, and the 95% confidence interval for the log hazard ratio was ($ln(0.65)$, $ln(0.86)) = (-0.431, -0.151)$. The standard deviation was estimated to be $(ln(0.86)-ln(0.65))/(2*z_{0.975}) = 0.071$, where $z_{0.975}$ denotes the 97.5%th quantile of a standard normal distribution. It follows from the Rothmann et al. approach that

$$\Delta = -0.5(-0.301) - z_{1-\alpha/2}(4/D + 1/4*0.071^2)^{0.5} + z_{1-\alpha/2}(4/D)^{0.5},$$

where D denotes the number of events in the current trial and α corresponds to a two-sided nominal significance level for the final analysis. In comparison 1 (test regimen A vs. active control), D = 321+324 = 645. In comparison 2 (test regimen B vs. active control), D = 306+324 = 630. Using the formula above, it leads that

(1) At $\alpha = 0.047$ level, Δ was rounded off to be 0.135 in both comparisons. Hence, the cutoff for the hazard ratio in non-inferiority test was $e^{\Delta} = e^{0.135} = 1.145$ (corresponding to a non-inferiority margin of 14.5%). The active control effect was estimated to be 0.763 (the smaller the better), corresponding to the upper limit of a 25.4% confidence interval for the hazard ratio (V+Cis/ Cis).

(2) At $\alpha = 0.0235$ level which was needed for the Hochberg approach, Δ was rounded off to be 0.133 in both comparisons. Hence, the cutoff for the hazard ratio in non-inferiority test was e^{Δ}, nearly $e^{0.133} = 0.142$ in both comparisons (corresponding to a non-inferiority margin of 14.2%). The active control effect was estimated to be 0.766 (the smaller the better), corresponding to the upper limit of a 28.6% confidence interval for the hazard ratio of (V+Cis / Cis).

3.2 REVIEWER'S SUPPORTIVE ANALYSIS OF NON-INFERIORITY

The supportive analysis was based on the Rothmann et al. approach to estimating the active control effect as seen in Section 3.1.2. Given that the test regimens needed to preserve 50% of the active control effect for demonstration of non-inferiority, the margins (or cutoff for the hazard ratio) were derived as in Section 3.1.2. The non-inferiority test using the Hochberg procedure to control the familywise false positive rate (one-sided) at a level of 0.0235 for multiple comparisons proceeds as following.

Step 1: Construct 95.3% confidence intervals (corresponding to a nominal two-sided significance level of 0.047, equivalent to a one-sided significance level of 0.0235) for the hazard ratio of each test treatment to the active control. If both confidence intervals entirely lie below 1.145, then non-inferiority evidence is shown for each test treatment. Otherwise, there is no non-inferiority evidence in at least one comparison and one should proceed to Step 2 to determine whether lack of non-inferiority evidence is shown in only one comparison or in both comparisons.

Step 2: If there is non-inferiority evidence, it should be in the comparison resulting in a smaller p-value. This occurred in comparing D75+Cis to V+Cis, so construct a 97.65% confidence interval (corresponding to a nominal two-sided significance level of 0.0235, equivalent to a one-sided significance level of 0.01275) for the hazard ratio of D75+Cis to V+Cis. If the confidence interval entirely lies below 1.142, then non-inferiority evidence is shown for D75+Cis.

By using the results in Table 7 it can be seen that not both 95.3% confidence intervals entirely lied below 1.145, so one should proceed to Step 2. The 97.65% confidence interval in comparing D75+Cis to V+Cis was (0.737, 1.059), entirely below 1.142. This suggested non-inferiority evidence for the D75+Cis combination. Based on the 97.65% confidence interval, the effect of D75+cis to Cis was estimated to be 0.811 (=1.059*0.766). Since the effect of the active control was estimated to be 0.766 after incorporating the variability of the point estimate, it was estimated that test treatment A (D75+Cis) preserved at least 78% (= *ln* 0.811 / *ln* 0.766) of the active control effect. However, it should be cautioned that the results were based on only one historical trial and, in which trial, patient characteristics might not be identical to those in this registration trial. Therefore, the non-inferiority results might differ had more historical trials been available.

3.3 REVIEWER'S SENSITIVITY ANALYSES ON DIFFERENT POPULATIONS

To explore the robustness of the results on the ITT population, this reviewer preformed the stratified logrank test on evaluable patients and results are summarized in Table 18. The conclusion based on this patient population (1181 patients) was the same as that based on all randomized patients (1220 patients) except that the percentage of the active control effect that test treatment A (D75+Cis) preserved dropped from 62% (on all randomized patients) to 53% (= *ln*(1.073*0.86) / *ln* 0.86, on evaluable population). The same test was also performed on another patient population consisting of those with disease stage remaining unchanged after randomization. A total of 1169 were included in this population and the results, as summarized in Table 19, were consistent with those based on the ITT and evaluable populations except that the

percent of the active control effect that test treatment D75+Cis preserved was 49% (= *ln*(1.080*0.86) / *ln* 0.86). For those patients whose disease stages were changed from randomization, the Medical Officer, Dr. Ramzi Dagher, adjudicated their stages and a sensitivity analysis was performed on the updated data. Based on the result as summarized in Table 20, the percentage of the active control effect that test treatment D75+Cis preserved was 55% (= *ln*(1.070*0.86) / *ln* 0.86).

Table 18: Reviewer's Sensitivity Analysis of Stratified Logrank Test (on Evaluable Population)

	Comparison 1	Comparison 2
	Test Regimen A (D75+Cis) Vs. Active Control (V+Cis)	Test Regimen B (D75+Cb6) Vs. Active Control (V+Cis)
P-value [a]	0.159	0.651
Estimated Hazard Ratio [b]	0.891	1.037
95.3% CI [c]	(0.758, 1.048)	(0.883, 1.218)
97.65% CI [d]	(0.741, 1.073)	Not needed.

[a] From the superiority test "H_0: hazard ratio = 1 vs. H_1: hazard ratio ≠ 1".
[b] Hazard ratio of test treatment to the active control. A hazard ratio of less than 1 indicates that the test treatment is associated with a longer time to survival.
[c] Corresponding to a nominal significance level of 0.047.
[d] Corresponding to a nominal significance level of 0.0235.

Table 19: Reviewer's Sensitivity Analysis of Stratified Logrank Test (on Patients Whose Disease Stage Remained Unchanged since Randomization)

	Comparison 1	Comparison 2
	Test Regimen A (D75+Cis) Vs. Active Control (V+Cis)	Test Regimen B (D75+Cb6) Vs. Active Control (V+Cis)
P-value [a]	0.187	0.682
Estimated Hazard Ratio [b]	0.898	1.034
95.3% CI [c]	(0.763, 1.056)	(0.881, 1.213)
97.65% CI [d]	(0.746, 1.080)	Not needed.

[a] From the superiority test "H_0: hazard ratio = 1 vs. H_1: hazard ratio ≠ 1".
[b] Hazard ratio of test treatment to the active control. A hazard ratio of less than 1 indicates that the test treatment is associated with a longer time to survival.
[c] Corresponding to a nominal significance level of 0.047.
[d] Corresponding to a nominal significance level of 0.0235.

Table 20: Reviewer's Sensitivity Analysis of Stratified Logrank Test (Based on MO-Adjudicated Disease Stage)

	Comparison 1	Comparison 2
	Test Regimen A (D75+Cis) Vs. Active Control (V+Cis)	Test Regimen B (D75+Cb6) Vs. Active Control (V+Cis)
P-value [a]	0.154	0.844
Estimated Hazard Ratio [b]	0.892	1.016
95.3% CI [c]	(0.761, 1.046)	(0.868, 1.189)

97.65% CI [d]	(0.744, 1.070)	Not needed.

[a] From the superiority test "H_0: hazard ratio = 1 vs. H_1: hazard ratio ≠ 1".
[b] Hazard ratio of test treatment to the active control. A hazard ratio of less than 1 indicates that the test treatment is associated with a longer time to survival.
[c] Corresponding to a nominal significance level of 0.047.
[d] Corresponding to a nominal significance level of 0.0235.

3.4 REVIEWER'S SUPPORTIVE ANALYSES FROM OTHER TEST PROCEDURES.

The medical officer, Dr. Ramzi Dagher, considered only three of the covariates as potentially relevant, which were performance status, gender and weight loss at baseline. The results are summarized in Table 21. The conclusion based on the stratified proportional hazards model adjusted for these three covariates was consistent with that based on this reviewer's primary analysis, stratified logrank test; i.e., statistical evidence for non-inferiority was only seen in test regimen A (D75+Cis) and there was no statistical evidence for superiority in either test regimen. In addition, it was estimated that test regimen A (D75+Cis) preserved 69% (= *ln* (1.048*0.86) / *ln* 0.86) of the active control effect.

This reviewer also performed the unstratified logrank test to explore the influence of the stratification factors, but found consistent results (see Table 22) as in the stratified logrank test. It was estimated that test regimen A (D75+Cis) preserved 59% (= *ln* (1.063*0.86) / *ln* 0.86) of the active control effect.

Table 21: Reviewer's Supportive Analysis of Stratified Cox Model Adjusted for KPS, Gender and Weight Loss at Baseline (on All Randomized Patients)

	Comparison 1	Comparison 2
	Test Regimen A (D75+Cis) vs. Active Control (V+Cis)	Test Regimen B (D75+Cb6) vs. Active Control (V+Cis)
P-value [a]	0.092	0.837
Estimated Hazard Ratio [b]	0.874	1.107
95.3% CI [c]	(0.745, 1.025)	(0.868, 1.191)
97.65% CI [d]	(0.729, 1.048)	Not needed.

[a] From the superiority test "H_0: hazard ratio = 1 vs. H_1: hazard ratio ≠ 1".
[b] Hazard ratio of test treatment to the active control. A hazard ratio of less than 1 indicates that the test treatment is associated with a longer time to survival.
[c] Corresponding to a nominal significance level of 0.047.
[d] Corresponding to a nominal significance level of 0.0235.

Table 22: Reviewer's Supportive Analysis of Unstratified Logrank Test (on All Randomized Patients)

	Comparison 1	Comparison 2
	Test Regimen A (D75+Cis) vs. Active Control (V+Cis)	Test Regimen B (D75+Cb6) vs. Active Control (V+Cis)
P-value [a]	0.134	0.731
Estimated Hazard Ratio [b]	0.887	1.027
95.3% CI [c]	(0.757, 1.039)	(0.878, 1.202)
97.65% CI [d]	(0.741, 1.063)	Not needed.

[a] From the superiority test "H_0: hazard ratio = 1 vs. H_1: hazard ratio ≠ 1".
[b] Hazard ratio of test treatment to the active control. A hazard ratio of less than 1 indicates that the test treatment is associated with a longer time to survival.
[c] Corresponding to a nominal significance level of 0.047.
[d] Corresponding to a nominal significance level of 0.0235.

3.5 Reviewer's Exploratory Analysis of Comparison Between Two Test Regimens

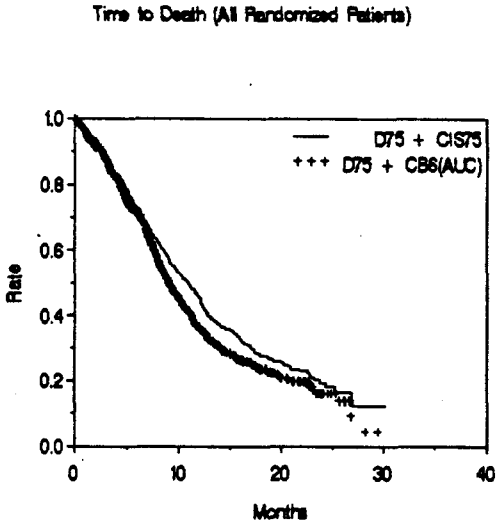
In order to explore whether test regimen A (D75+Cis) was associated with a longer survival time as compared to test regimen B (D75+Cb6), this reviewer performed stratified and unstratified logrank tests. As seen in Table 23, the corresponding p-values were 0.043 and 0.065, respectively. The Kaplan-Meier curves are in Figure 3. These results might suggest some statistical evidence that test regimen A was associated with a longer survival time as compared to test regimen B.

Table 23: Reviewer's Exploratory Analysis of Comparison between Two Test Regimens (on All Randomized Patients)

Analysis		Test Regimen A (D75+Cis) vs. Test Regimen B (D75+Cb6)
Stratified logrank	P-value [a]	0.043
	Estimated Hazard Ratio [b] (95% CI [c])	0.850 (0.725, 0.995)
Unstratified logrank	P-value	0.065
	Estimated Hazard Ratio (95% CI)	0.863 (0.737, 1.009)

[a] From the superiority test "H_0: hazard ratio = 1 vs. H_1: hazard ratio ≠ 1".
[b] Hazard ratio of test treatment A to test treatment B. A hazard ratio of less than 1 indicates that test treatment A is associated with a longer time to survival.
[c] Corresponding to a nominal significance level of 0.05.

Figure 3: Reviewer's Kaplan-Meier Survival Curves (Test Regimen A vs. Test Regimen B)



3.6 EFFICACY FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

Subgroup analyses of overall survival by certain prognostic factors (disease stage, region, age, sex and race) were explored by this reviewer and are summarized in the following tables. There was no evidence against non-inferiority of test regimen A (D75+Cis) to the active control (V+Cis) in these exploratory analyses.

Table 24: Reviewer's Descriptive Summary of Survival by Disease Stage

Subgroup: Stage		Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
IIIB	# of Patients	135	133	133
	# of censored	44	32	29
	Median (in days)	380	316	314
	(95% CI)	(285, 439)	(275, 368)	(309, 377)
IV	# of Patients	273	274	272
	# of censored	57	55	52
	Median (in days)	314	259	288
	(95% CI)	(276, 353)	(233, 286)	(233, 331)

Table 25: Reviewer's Descriptive Summary of Survival by Region

Subgroup: Region		Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
USA/Canada	# of Patients # of censored Median (in days) (95% CI)	115 27 341 (291, 386)	115 21 253 (209, 304)	114 21 300 (227, 366)
Europe/ Morocco/ Israel	# of Patients # of censored Median (in days) (95% CI)	197 49 312 (268, 371)	197 42 275 (240, 313)	197 37 297 (260, 345)
South Africa/ Australian/ New Zealand	# of Patients # of censored Median (in days) (95% CI)	33 9 330 (165, 476)	32 7 346 (241, 494)	32 4 286 (190, 344)
South America/ Mexico	# of Patients # of censored Median (in days) (95% CI)	63 16 369 (234, 477)	46 17 339 (242, 375)	43 19 364 (267, 471)

Table 26: Reviewer's Descriptive Summary of Survival by Age

Subgroup: Age		Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
≤ 60	# of Patients # of censored Median (in days) (95% CI)	204 44 314 (268, 54)	226 47 286 (253, 323)	198 30 286 (235, 318)
> 60	# of Patients # of censored Median (in days) (95% CI)	204 57 359 (284, 399)	181 40 275 (233, 309)	207 51 340 (280, 383)

Table 27: Reviewer's Descriptive Summary of Survival by Sex

Subgroup: Sex		Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
Female	# of Patients # of censored Median (in days) (95% CI)	114 30 351 (285, 388)	115 25 304 (242, 351)	102 26 345 (275, 450)
Male	# of Patients # of censored Median (in days) (95% CI)	294 71 324 (278, 370)	292 62 275 (248, 293)	303 55 297 (259, 334)

Table 28: Reviewer's Descriptive Summary of Survival by Race

Subgroup: Race		Test Regimen A (D75+Cis) [N = 408]	Test Regimen B (D75+Cb6) [N = 406]	Active Control (V+Cis) [N=404]
Caucasian	# of Patients	360	353	361
	# of censored	92	77	70
	Median (in days)	336	277	300
	(95% CI)	(297, 370)	(255, 309)	(267, 344)
Others	# of Patients	48	54	44
	# of censored	9	10	11
	Median (in days)	271	244	311
	(95% CI)	(194, 386)	(192, 339)	(225, 348)

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

/s/

Peiling Yang
11/19/02 12:06:28 PM
BIOMETRICS

Gang Chen
11/20/02 01:19:54 PM
BIOMETRICS

Kooros Mahjoob
11/20/02 02:50:09 PM
BIOMETRICS