CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

204275Orig1s000

STATISTICAL REVIEW(S)

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

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/BLA #:	NDA 204275
Drug Name:	Breo Ellipta (fluticasone furoate/vilanterol) Inhalation Powder
Indication(s):	Chronic Obstructive Pulmonary Disease (COPD)
Applicant:	GlaxoSmithKline (GSK)
Date(s):	Receipt date: July 12, 2012
	PDUFA date: May 10, 2013 (actual day May 12, 2013)
Review Priority:	Standard
Biometrics Division:	Division of Biometrics II
Statistical Reviewer:	Kiya Hamilton, Ph.D.
Concurring Reviewers:	Joan Buenconsejo, Ph.D., Team Leader
	Thomas Permutt, Ph.D. Division Director
Medical Division:	Division of Pulmonary, Allergy and Rheumatology Products
Clinical Team:	Sofia Chaudhry, M.D., Medical Reviewer
	Susan Limb, M.D., Team Leader
	Badrul A. Chowdhury, M.D. Ph.D., Medical Division Director.
Project Manager:	Angela Ramsey

Keywords: NDA, clinical studies, Missing Data, Multiplicity

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 EXECUTIVE SUMMARY

GlaxoSmithKline (GSK) proposes fluticasone furoate/vilanterol (FF/VI) inhalation powder, administered once daily for the long-term treatment of airflow obstruction in patients with chronic obstructive pulmonary disease (COPD) including chronic bronchitis and/or emphysema and to reduce exacerbations of COPD in patients with a history of exacerbations. GSK is requesting approval for dosage strength of fluticasone furoate 100 mg (FF) and vilanterol 25 mg (VI). Neither of the components is approved for treatment of COPD.

The clinical program for FF/VI includes multiple dose-ranging and dose-interval studies for the FF and VI monocomponents and for the FF/VI combination, four key efficacy and safety studies, as well as four additional active comparator studies. The focus of the statistics review is on the four efficacy and safety studies. All four studies were designed to demonstrate the efficacy of FF/VI and its components in terms of improvement in airflow obstruction and symptomatic endpoints, including reduction in the annual rate of moderate and severe COPD exacerbations (studies HZC102871 and HZC102970 only).

Lung function endpoints (weighted mean FEV_1 (0–4 h) and change from baseline in trough FEV_1) were the primary endpoints in studies HZC112206 and HZC112207 and the primary endpoint in studies HZC102970 and HZC102871 was annual rate of moderate and severe exacerbations. Of note, within each of the four primary studies, in order to account for multiplicity across treatment comparisons and key endpoints, a specific step-down testing procedure was applied, whereby inference for a test in the pre-defined hierarchy was dependent upon statistical significance having been achieved for the previous tests in the hierarchy.

Compared to placebo, both VI 25 and all dosage strengths of FF/VI showed efficacy with respect to the weighted mean FEV_1 (0–4 h) and change from baseline in trough FEV_1 (studies HZC112206 and HZC112207). These studies also demonstrated the contribution of VI to the FF/VI combination at all dosage strengths, based on the difference in weighted mean FEV_1 (0–4 h). However, neither study demonstrated the contribution of FF to the FF/VI combination at all dosage strengths based on trough FEV_1. Change from baseline in trough FEV_1 for VI 25 was 100 mL compared to 150 mL for FF/VI 100/25 and about 140 mL for FF/VI 200/25. Therefore, for the proposed dose of FF/VI 100/25, the difference when compared to VI 25 was about 50 mL (95% CI -6, 102). Since the confidence interval includes zero, this implies that the direction of the difference, if any, is not known with much confidence.

In both studies, the higher dose FF/VI combination did not have a larger effect on the primary endpoints (weighted mean FEV_1 or trough FEV_1) compared to the lower dose FF/VI combination.

Only one of the two exacerbation studies showed a statistically significant improvement for all FF/VI doses over VI 25 for annual rate of moderate and severe exacerbations. In study HCZ102970, the mean rate of moderate and severe exacerbation in the VI 25 group was about

one exacerbation per year. For the proposed dose of FF/VI 100/25, the rate of moderate and severe exacerbation was reduced by about a quarter of an event in one year.

The Pulmonary-Allergy Advisory Committee will convene on April 17, 2013 to discuss the efficacy and safety of Breo Ellipta (fluticasone furoate 100 mg and vilanterol 25 mg) administered once daily for the long-term treatment of airflow obstruction in patients with chronic obstructive pulmonary disease (COPD) including chronic bronchitis and/or emphysema and to reduce exacerbations of COPD in patients with a history of exacerbations.

2 INTRODUCTION

2.1 Overview

2.1.1 Class and Indication

GlaxoSmithKline (GSK) proposes fluticasone furoate/vilanterol inhalation powder (hereafter referred to as FF/VI), administered once daily for the long-term treatment of airflow obstruction in patients with chronic obstructive pulmonary disease (COPD) including chronic bronchitis and/or emphysema and to reduce exacerbations of COPD in patients with a history of exacerbations. It contains fluticasone furoate, an inhaled corticosteroid (ICS), hereafter referred to as FF, and vilanterol tridentate, a long acting beta$_2$-agonist (LABA), hereafter referred to as VI. GSK is requesting approval for dosage strength of fluticasone furoate 100 mg and vilanterol 25 mg. As neither of the components is approved for treatment of COPD, the clinical development program aimed to demonstrate the efficacy of FF and VI individually, their contribution to the combination, and the efficacy of the FF/VI combination.

2.1.2 History of Drug Development

GSK had several interactions with the Division of Pulmonary, Allergy, and Rheumatology Products regarding their FF/VI clinical development program for COPD (under IND 77,855). They also met with the Division to discuss their clinical development program for asthma, as well as their development program for each of the individual components (under IND 74,696 for the VI program and under IND 70,297 for the FF program). Pertinent parts of the statistical portion of the communications and interactions for the FF/VI COPD program are summarized herein.

The design and analysis of the phase 3 studies (Table 1) as well as the results from the Phase 2 dose-ranging and dose-interval studies were discussed at the End-of-Phase 2 meeting held on June 17, 2009. In this meeting the applicant discussed the primary endpoint, the annual rate of COPD moderate/severe exacerbations, for the two 52-week studies (HCZ102871 and HCZ102970, hereafter referred to as 2871 and 2970, respectively). The applicant stated that the rate would be calculated as the total number of moderate and/or severe exacerbations experienced by the patient during the treatment period and analyzed using a generalized linear model, assuming the Negative Binomial distribution, with the logarithm of time on treatment as an offset variable. While the Division informally agreed to the applicant's proposed primary analysis, we recommended that the applicant also analyze the exacerbation rates by Poisson regression as a sensitivity analysis. The applicant also discussed the primary endpoints, namely the trough FEV_1 for comparisons pertaining to the evaluation of the FF and VI components and weighted mean (based on the AUC) FEV_1 over 0–4 hours for comparisons pertaining to the evaluation of the VI component, for the two 6-month studies (HCZ112206 and HZC112207, hereafter referred to as 2206 and 2207, respectively). The applicant stated that for each of these endpoints, change from baseline would be analyzed using mixed models repeated measures (MMRM), with an unstructured variance-covariance matrix. Visit would be fitted as a categorical variable and a treatment by visit interaction term would be fitted to allow estimates of

treatment effect at each visit separately. While the Division informally agreed to the applicant's proposed approach, we also recommended that the applicant conduct sensitivity analyses using other missing data imputation methods and other covariance matrix structures. The applicant also proposed a hierarchy of statistical tests across the primary and pre-defined secondary endpoints in order to control for multiplicity. The Division at that time responded

> When there are multiple studies available and each study has multiple doses, the efficacy evidence will be evaluated collectively from the multiple studies and multiple doses. The error rate of approving an ineffective drug will be controlled if the dose- response relationship is reasonable and results across studies are consistent. The proposed hierarchical testing procedure protects against type I error in a rigid way and may lead to irrational conclusion when the dose- response was guessed incorrectly. In addition, this procedure does not add any value in the selection of the optimal doses, as the optimal doses should be selected based on the effect size, safety concerns, and risk/benefit ratio.

In the discussion that followed, the applicant agreed that the closed testing procedure protects Type I error in a rigid way and may lead to an irrational conclusion. However, the applicant still would like to use the procedure. The Division agreed the procedure was acceptable and recommended that the applicant not include the comparison between FF versus placebo in the testing procedure and to include the comparison between the FF/VI versus VI for trough FEV_1 in order to evaluate the contribution of FF. While the evidence of efficacy is evaluated collectively from the multiple studies, we agree with the applicant that a strong control of type 1 error should be in place for each individual studies.

A Type B pre-NDA meeting was held on July 13, 2011, to discuss the applicant's data to support the use of the FF/VI inhalation powder in the treatment of COPD and Asthma. The Division raised concerns regarding the lack of robust results to support the proposed bronchodilation indication and satisfy the Combination Rule for COPD population. Based on the preliminary review of the data from studies 2206 and 2207 at that time, only the lowest combination dose FF/VI 50/25 mcg showed a statistically significant benefit in terms of trough FEV_1 over VI 25 and there does not appear to be a replicated comparison of FF/VI 50/25 to placebo in the clinical program. Furthermore, trough FEV_1 data for FF/VI 100/25 and FF/VI 200/25 compared to VI were not supportive. The Division noted that the COPD exacerbation studies (2871 and 2970) may provide efficacy support for the addition of FF to VI, but positive exacerbation results may be problematic in the context of the negative lung function results. There was also a discussion of the proposed statistical methodology for examining subgroups as outlined in the summary Document Analysis Plans for the ISE (submitted on March 11, 2011 with serial No. 0291) and for the ISS (submitted on March 24, 2011 with serial No. 0296) for COPD in IND 77,855. The Division informally agreed that their approach was reasonable and noted that generally the results from individual studies to support any claims in the label are used.

> Pooled analyses are not usually very helpful in this regard with the exception of required analyses by age, sex and race. Additional analyses may be performed using pooled data; however, little weight will be given to the results from these analyses.

2.1.3 Specific Studies Reviewed

The clinical program for FF/VI includes multiple studies for the FF and VI monocomponents and for the FF/VI combination. The applicant submitted data from 12 dose-ranging and dose-interval studies for the FF and VI monocomponents and for the FF/VI combination, data from four key efficacy and safety studies, as well as data from four additional active comparator studies.

The focus of the statistics review is on the four key efficacy and safety studies (Table **1**). All four studies were phase 3, randomized, double-blind, parallel-group, multi-center studies in male and female patients at least 40 years of age at screening. The review will also include results from the active-comparator studies, except for study 3107 where the dose of the active comparator is not approved in the US for COPD (Table 2). Review of the dose-ranging and dose-interval studies can be found in the Clinical Review.

Table 1: Study Design for the Four Efficacy Studies

	Phase and Design	Length of the Study	Treatment Arms	Number of Patients per Arm	Study Population	Primary Efficacy Endpoints	% in US Sites
HZC112206	Phase 3, randomized, double-blind, parallel-group, multi-center	RI: 2 weeks TP: 24 weeks FU: 1 week	FF/VI 50/25 mcg FF/VI 100/25 mcg FF/VI 100 mcg VI 25 mcg Placebo	206 206 206 205 207	Moderate/severe COPD	Weighted mean Clinic Visit FEV_1 0–4 hours on Day 168 Change from baseline in Clinic Visit trough FEV_1 on Day 169	39%
HZC112207	Phase 3, randomized, double-blind, parallel-group, multi-center	RI: 2 weeks TP: 24 weeks FU: 1 week	FF/VI 100/25 mcg FF/VI 200/25 mcg FF 100 mcg FF 200 mcg VI 25 mcg Placebo	204 205 204 204 204 205	Moderate/severe COPD	Weighted mean Clinic Visit FEV_1 0–4 hours on Day 168 Change from baseline in Clinic Visit trough FEV_1 on Day 169	25%
HZC102871	Phase 3, randomized, double-blind, parallel-group, multi-center	RI: 4 weeks TP: 52 weeks FU: 1 week	FF/VI 50/25 mcg FF/VI 100/25 mcg FF/VI 200/25 mcg VI 25 mcg	408 403 402 406	Moderate/severe COPD	Annual rate of moderate and severe exacerbations	33%
HZC102970	Phase 3, randomized, double-blind, parallel-group, multi-center	RI: 4 weeks TP: 52 weeks FU: 1 week	FF/VI 50/25 mcg FF/VI 100/25 mcg FF/VI 200/25 mcg VI 25 mcg	412 403 409 409	Moderate/severe COPD	Annual rate of moderate and severe exacerbations	36%

- RI: Run-in period, TP: Treatment period, FU: Follow-up

Table 2 Study Design for the Active Comparator Studies

	Phase and Design	Length of the Study	Treatment Arms	Number of Patients per Arm	Study Population	Primary Efficacy Endpoints	% in US Sites
HZC112352	Phase 3b, randomized, double-blind, double-dummy, parallel-group, multi-center	RI: 2 weeks TP: 12 weeks FU: 1 week	FF/VI 100/25 mcg FP/salmeterol 250/50 mcg	259 252	COPD	Change from baseline trough in 24–hour weighted mean serial FEV_1 on Day 84	29%
HZC113109	Phase 3b, randomized, double-blind, double-dummy, parallel-group, multi-center	RI: 2 weeks TP: 12 weeks FU: 1 week	FF/VI 100/25 mcg FP/salmeterol 250/50 mcg	261 260	COPD	Change from baseline trough in 24–hour weighted mean serial FEV_1 on Day 84	28%
HZC113107	Phase 3, randomized, double-blind, double-dummy, parallel-group, multi-center	RI: 2 weeks TP: 12 weeks FU: 1 week	FF/VI 100/25 mcg FP/salmeterol 500/50 mcg	266 262	COPD	Change from baseline trough in 24–hour weighted mean serial FEV_1 on Day 84	0%
HZA113091	Phase 3, randomized, double-blind, double-dummy, parallel-group, multi-center	RI: 4 weeks TP: 24 weeks FU: 1 week	FF/VI 100/25 mcg FP/salmeterol 250/50 mcg	403 403	Persistent bronchial asthma	Weighted mean for 24–hour serial FEV1 at the end of the 24–week treatment period	30%

- RI: Run-in period, TP: Treatment period, FU: Follow-up

2.2 Data Sources

NDA 204-275 was submitted on July 12, 2012. The study reports including protocols, statistical analysis plan, and all referenced literature were submitted by the Applicant to the Agency.

3 STATISTICAL EVALUATION

3.1 Data and Analysis Quality

In general, the submitted efficacy data are acceptable in terms of quality and integrity. I was able to reproduce the primary and secondary efficacy endpoint analyses for each clinical study submitted. I was able to verify the randomization of the treatment assignments.

3.2 Evaluation of Efficacy

3.2.1 Study Design and Endpoints

The summary of the study designs and endpoints for the four key efficacy studies are given in Table **1**. All four studies were Phase 3, randomized, double-blind, parallel-group, multi-center studies in male and female patients at least 40 years of age at screening (Visit 1). The design and efficacy endpoints are explained in detail in the following paragraphs.

Studies 2206 and 2207 were designed similarly. Both studies consisted of 24 weeks of treatment and were designed to assess the efficacy and safety of FF/VI when administered once daily via the novel dry powder inhaler in patients with COPD. Study 2206 studied the dosage strengths of FF/VI 50/25 mcg and 100/25 mcg, FF 100 mcg, VI 25 mcg and placebo. Study 2207 studied the dosage strengths FF/VI 100/25 mcg, 200/25 mcg, FF 100 mcg, FF 200 mcg, VI 25 mcg and placebo. Studies 2871 and 2970 were designed similarly. These two studies were designed to evaluate the effects of once daily dosing in the morning with dosage strengths FF/VI (50/25, 100/25 and 200/25 mcg) versus one dosage strength of VI (25 mcg) in patients with COPD. For each of the four studies, following the run-in period, patients were randomized into treatment arms with stratification on smoking status (current smoker or previous smoker).

The primary endpoints for both studies 2206 and 2207 were weighted mean clinic visit FEV_1 0–4 hours post-dose on treatment Day 168 (Visit 11) and change from baseline in clinic visit trough (pre-bronchodilator and pre-dose) FEV_1, on treatment Day 169 (Visit 12). Trough FEV_1 on treatment Day 169 was defined as the mean of the FEV_1 values obtained 23 and 24 hours after dosing on treatment Day 168, measured at visit 12. If one of the two paired assessments was missing then trough FEV_1 was defined as the single 23 or 24 hour assessment. For inclusion in the calculation the 23- and 24-hour values must have been pre- the next day's dose.

Baseline FEV_1 was defined as the mean of the two assessments made 30 and 5 minutes pre-dose on Treatment Day 1. The -30 and 0 minutes pre-dose measurements must have had time of assessments less than or equal to the time of Day 1 dosing to be included in the baseline calculation; measurements after the time of Day 1 dosing were set to missing. If one of these two

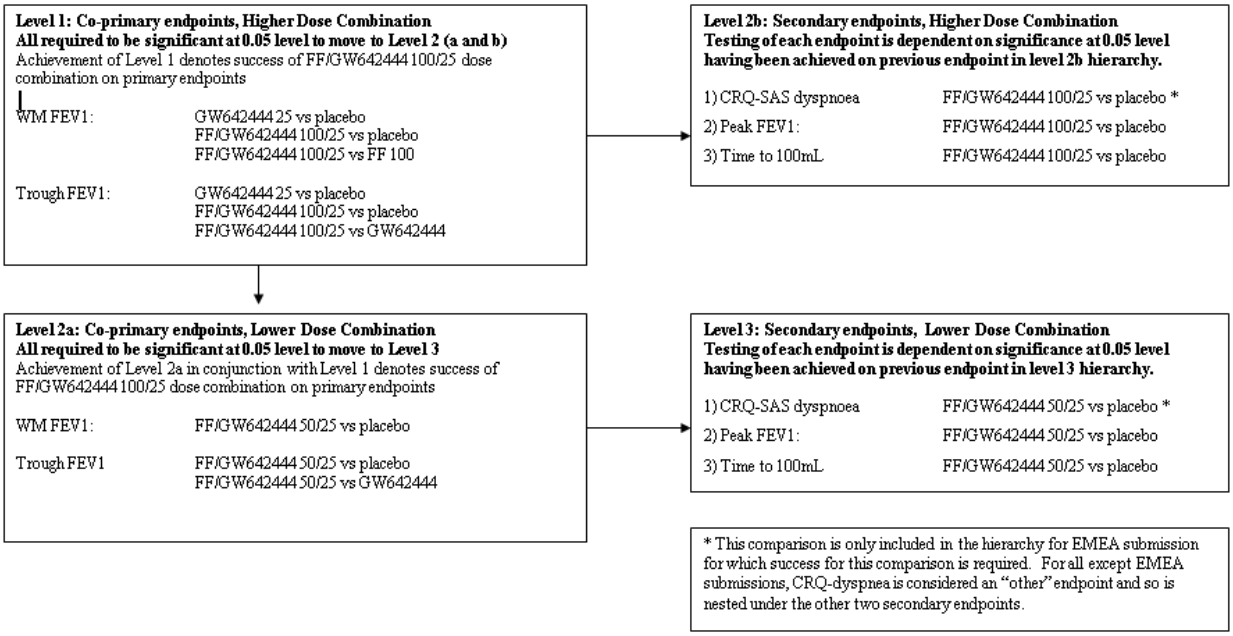
assessments was missing then baseline was defined as the single pre-dose FEV_1 value on Day 1. If both were missing then baseline was missing.

The weighted mean clinic FEV_1 was used to evaluate the contribution of VI and the trough FEV_1 was used to evaluate the contribution of FF in the intent-to-treat (ITT) population. The ITT population was defined as all patients who were randomized to and received at least one dose of randomized double-blind study medication in the treatment period. The secondary endpoints for studies 2206 and 2207 were peak FEV_1 on treatment Day 1 and time to onset (increase of 100 mL above baseline in FEV_1) on treatment day 1 in the ITT population.

The primary endpoint in both studies 2871 and 2970 was the annual rate of moderate and severe exacerbations. The secondary endpoints for both studies were time to first moderate and severe exacerbation, annual rate of exacerbations requiring systemic/oral corticosteroids, and change from baseline in trough FEV_1 at visit 11. COPD exacerbation was defined as an acute worsening symptom of COPD requiring the use of any treatment other than study medication or rescue albuterol/salbutamol. A moderate exacerbation was defined as worsening symptoms of COPD that required treatment with oral corticosteroids and/or antibiotics. A severe exacerbation was defined as worsening symptoms of COPD that required treatment with in-patient hospitalization. Albuterol/salbutamol was used as rescue medication.

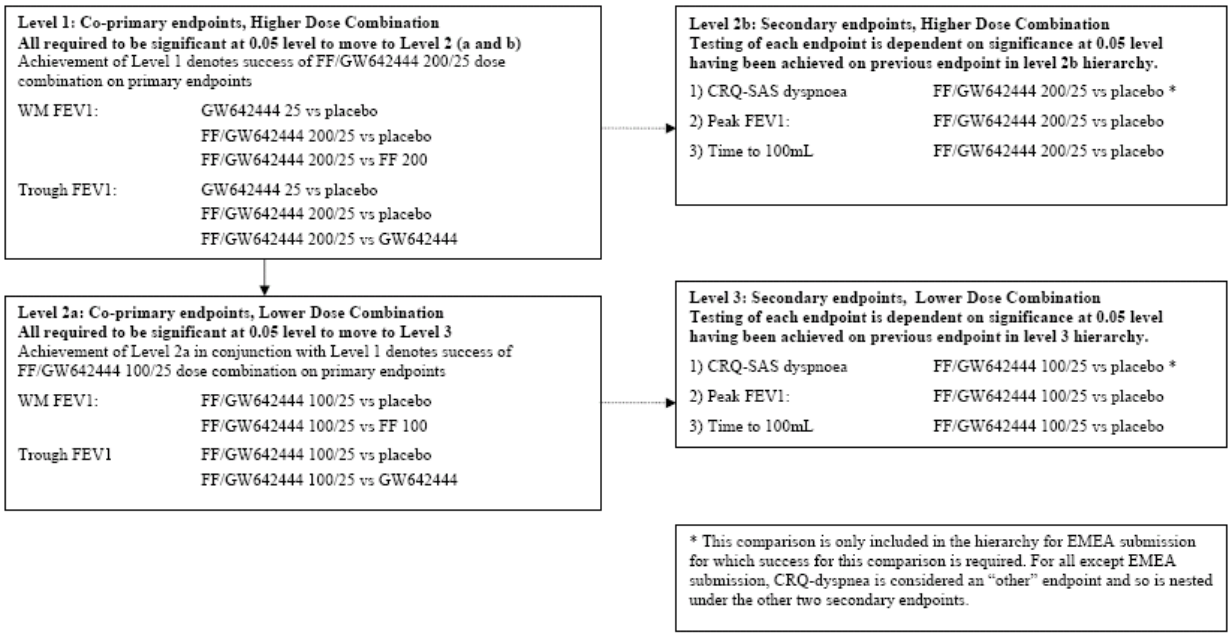
There was a strong control of the Type 1 error for the primary endpoints. Studies 2206 and 2207 used a step-down procedure to account for multiplicity across treatment comparisons and key endpoints (Figure 1 and Figure 2).

Figure 1: Statistical Testing Strategy Study 2206



Source: Clinical Study Report-Protocol Number HZC112206 Attachment 2, page 2043

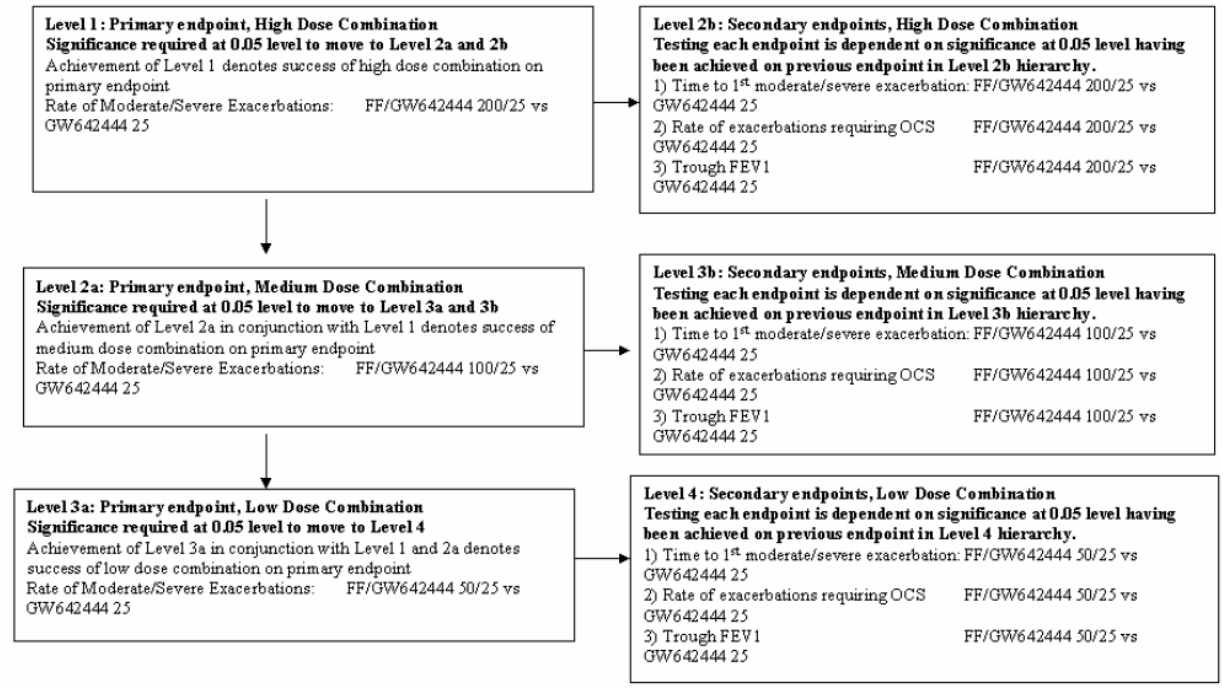
Figure 2: Statistical Testing Strategy Study 2207



Source: Clinical Study Report-Protocol Number HZC112207 Attachment 1, page 2029

A step-down testing approach (Figure 3) was used to account for multiplicity across treatment comparisons and key endpoints in both studies 2871 and 2970. Using this approach the inference for the primary efficacy endpoint for the FF/VI 100/25 combination dose versus VI 25 was dependent upon statistical significance at the 5% level having first been achieved for the primary efficacy endpoints for the FF/VI 200/25 versus VI 25. For a given FF/VI combination dose, the secondary endpoints were nested under the primary endpoint.

Figure 3: Statistical Testing Strategy Studies 2871 and 2970



Source: Protocol Amendment Protocol-Protocol Number HZC102871 Figure 1, page 64 and Clinical Protocol-Protocol Number HZC102970 Figure 1, page 63

The summary of the study designs and endpoints for the four active-comparator studies are given in Table 2. Studies HZC112352, HZC113109 and HZC113107, hereafter referred to as 2352, 3109 and 3107, respectively were designed similarly. All three studies consisted of 12 weeks of treatment and were designed to assess the efficacy and safety of FF/VI inhalation powder administered once daily in the morning versus FP/salmeterol inhalation powder administered twice daily on lung function in subjects with COPD. Studies 2352 and 3109 studied the dosage strengths of FF/VI 100/25 mcg and FP/salmeterol 250/50 mcg. Study 3107 studied the dosage strengthens FF/VI 100/25 mcg and FP/salmeterol 500/50 mcg. Because the dose of the active comparator FP/salmeterol 500/50 mcg is unapproved, the results from this study are not included in the review. Study HZA113091 hereafter referred to as 3091 was designed to evaluate the efficacy and safety of once daily in the evening treatment with FF/VI 100/25 mcg compared with twice daily FP/salmeterol 250/50 mcg (morning and evening) on lung function in subjects with persistent bronchial asthma over a 24-week treatment period. For each of the COPD studies (2352, 3107 and 3109), following the run-in period, patients were randomized into treatment

arms with stratification on the subject's reversibility (reversible or non-reversible) to albuterol (salbutamol).

The primary endpoint for studies 2352, 3107 and 3109 was change from baseline trough in 24–hour weighted mean serial FEV_1 on Day 84. The weighted mean was calculated from the pre-dose FEV_1 and post-dose FEV_1 measurements at 5, 15, 30 and 60 minutes and 2, 4, 6, 8, 12, 13, 14, 16, 20 and 24 hours on treatment Day 84. Baseline trough FEV_1 was the mean of the two assessments made 30 and 5 minutes pre-dose on treatment Day 1. The primary endpoint for study 3091 was weighted mean for 24 hour serial FEV_1, calculated from serial spirometry over 0–24 hours at the end of 168-day double-blind treatment period. The 24 hour serial FEV_1 included a pre-dose assessment within 5 minutes prior to dosing and post-dose assessments after 5, 15 and 30 minutes and 1, 2, 3, 4, 11, 12, 12.5, 13, 14, 16, 20, 23 and 24 hours.

3.2.2 Statistical Methodologies

For studies 2206 and 2207 the primary analyses for the primary endpoints, 0–4 hours post-dose weighted mean FEV_1 and trough FEV_1, were analyzed using mixed model repeated measures (MMRM) in the ITT population. The model covariates were baseline FEV_1, smoking status (stratum), Day (1, 14, 56, 84 and 168), center grouping, treatment, Day by baseline interaction and Day by treatment interaction. Additional analyses assessed whether the effect of the active treatment groups were modified by smoking status at screening, center grouping or baseline FEV_1. This was achieved by fitting separate repeated measures models identical to the primary analysis model but also including additional terms for the treatment by smoking status interaction, treatment by center grouping and treatment by baseline FEV_1 interaction, respectively. An assessment of whether the effect of the active treatment groups were modified by reversibility, percent predicted GOLD categories, and cardiovascular (CV) history/risk factors were also conducted by fitting separate repeated measures models, identical to the primary analysis model but also included additional terms for reversibility and the reversibility by treatment interaction, percent predicted and the percent predicted by treatment interaction, cardiovascular history/risk factors and the cardiovascular history/risk factors by treatment interaction respectively. If the interactions from any of these analyses were significant at the 10% level, further investigation and characterization of the interactions was undertaken. The applicant defined reversibility as an increase in FEV_1 of ≥12% and ≥200 mL following administration of albuterol/salbutamol. The applicant defined percent predicted GOLD categories as:

I: $FEV_1 \geq 80$ % predicted
II: $50\ \% \leq FEV_1 < 80$ % predicted
III: $30\ \% \leq FEV_1 < 50$ % predicted
IV: $FEV_1 < 30$ % predicted

The CV history/risk factors were defined as any patient with at least one of the following current or past medical conditions at screening:

- Coronary Artery Disease

- Myocardial Infarction
- Arrhythmia
- Congestive Heart Failure
- Hypertension
- Cerebrovascular Accident
- Diabetes Mellitus
- Hypercholesterolemia.

The secondary endpoint, peak FEV_1 on treatment Day1, for studies 2206 and 2207 was analyzed using an Analysis of Covariance (ANCOVA) model. The covariates included in this model were baseline FEV_1, smoking status, center grouping and treatment. The secondary endpoint, time to ≥100 mL increase from baseline in FEV_1, was analyzed using the log-rank test, stratified for smoking status for each of the treatment comparisons. Actual times of FEV_1 results were used. A Kaplan-Meier plot showing the survival curves for all treatment groups was produced. Median time to ≥100 mL increase from baseline in FEV_1 (taken from the Kaplan-Meier analysis) was also presented.

For studies 2871 and 2970 the primary endpoint, annual rate of moderate and severe exacerbations, was analyzed using a general linear model assuming the negative binomial distribution in the ITT population. The response variable was the number of recorded, on-treatment, moderate and severe exacerbations experienced per patient. The explanatory variables consisted of treatment group, smoking status at screening (stratification variable), baseline disease severity (as percent predicted FEV_1) and center grouping. The model also included the logarithm of time on treatment per patient (derived from exposure start and stop) as an offset variable. The same model was also used assuming a Poisson regression model on the ITT population. Subgroup analyses were conducted to explore the effect of treatment by covariate interactions. There were three models fitted for both the negative binomial and the Poisson regression models in the ITT population: (i) with the addition of an interaction term for treatment by smoking status; (ii) with the addition of an interaction term for treatment by center grouping; and (iii) with the addition of an interaction term for treatment by percent predicted FEV_1. Two additional models were fitted to investigate the effect of treatment by covariate interactions: (iv) with the addition of a covariate of CV history/risk factors and an interaction term for treatment by CV history/risk factors, and (v) with the addition of a covariate of reversibility (yes/no) and an interaction term for treatment by reversibility.

The secondary endpoint, time to first moderate or severe exacerbation, in studies 2871 and 2970 was analyzed using a Cox proportional hazard model, with the exact method for handling ties in times of first exacerbation in the ITT population. The covariates included in the model were treatment group, smoking status at screening, baseline disease severity (percent predicted FEV_1) and center grouping. Annual rate of exacerbations requiring systemic/oral corticosteroids was analyzed using a generalized linear model assuming a negative binomial distribution. The response variable was the annual rate of exacerbations requiring systemic/oral corticosteroids for each patient. The explanatory variables were treatment group, smoking status at screening, baseline disease severity and center grouping. The model also included the logarithm of time on treatment per patient (derived from exposure start and stop) as an offset variable. The secondary endpoint, trough FEV_1 at visit 11 (week 52), was analyzed using mixed–models repeated–

measures with a repeated effect of visit within each patient and an unstructured covariance matrix. The response variable was change from baseline in trough FEV_1 at visits 3 to 11 with explanatory variables: treatment group, smoking status at screening (stratum variable), visit by baseline and visit by treatment interaction. Similar to the primary efficacy endpoint, additional models were fitted which explored the effect of treatment by covariate interactions: (i) with the addition of an interaction term for treatment by smoking status; (ii) with the addition of an interaction term for treatment by center grouping; and (iii) with the addition of an interaction term for treatment by baseline FEV_1.

In studies 2206 and 2207, the applicant pre-specified four additional analyses to explore missing data for the primary endpoints in the ITT population. One of the sensitivity analyses conducted by the applicant was the last observation carried forward (LOCF) for both primary endpoints. If the data was missing for the endpoint then the last non-missing post-baseline value was imputed. The LOCF analysis was performed using an ANCOVA model with covariates baseline FEV_1, smoking status, center grouping, and treatment. The Division generally does not accept LOCF as an imputation strategy because this implies patients who discontinue treatment will have the same outcome over time. This may lead to a biased standard error estimates since we are ignoring inherent uncertainty in the imputed values. In addition, this approach may not be conservative in terms of the patient's imputed outcome. For example, if a patient discontinued due to adverse events but had a good FEV_1, we will then be imputing a good score when in fact this patient was not successfully treated.

The applicant also applied two multiple imputation approaches, which they referred to as missing at random (MAR) and copy differences from control (CDC), to show how different assumptions influence the results obtained in the primary analysis. The multiple imputation methods allowed post-discontinuation missing observations to be imputed by fitting a Bayesian multivariate normal model for the data (including the same covariates as for the primary MMRM analysis) within each treatment using a Markov Chain Monte Carlo approach and quasi-independent samples drawn from the posterior distributions for the parameters of the multivariate normal distribution for each arm. Joint distribution of the pre- and post-withdrawal data was constructed based on the applicant's pre-specified assumptions concerning the post-withdrawal data (i.e., MAR and CDC). Conditional distribution of post-withdrawal given pre-withdrawal data and also covariates values for the individual subjects was then constructed using the joint distribution. This approach allowed the creation of completed datasets.

The MAR approach is based on the means and variance-covariances structures using patients in the same treatment group as the withdrawn patient. The main difference is that this approach uses separate covariance parameter estimation for each arm and also separate regression parameters using baseline covariates within each arm. Since the MAR approach assumes missing at random mechanism, this is concerning given that we are assuming that the behavior of the post-withdrawal data can be predicted from the observed variables. Like LOCF, this approach may not be conservative given that patients who discontinued from treatment may have the worse post-withdrawal outcome (e.g., they may be the more severe population) than patients who continued treatment.

An alternative method is the CDC approach. This is based on the assumption that patients who withdrew from the treated group would have followed the same trend over time (difference in mean value between time points) as those in the placebo group. According to the article provided by the applicant[1], a patient's mean profile in the treated group following withdrawal tracks that of the mean profile in the placebo group, but starting from the benefit already obtained. Post-withdrawal data in the placebo group are imputed under the MAR approach. Therefore, the placebo patients who withdrew are handled the same way as those who continued treatment. While this approach provided a specific assumption about the treated patients who withdrew from the study, it is unclear whether the assumption is suitable given that placebo patients who completed the trial may be more likely to be doing better than those placebo patients who discontinued. Furthermore, this approach may not account for patients who may have worse post-withdrawal outcomes (e.g. they may be the more severe population) that potentially decline over time compared to those who continued treatment.

To shed light on the nature and pattern of missing data, data for the 0–4 hours weighted mean FEV_1 and the trough FEV_1 endpoints were examined through cohorts of patients where the cohorts are defined based on the scheduled visits that were completed by each patient. The cohorts helped to show if there were any differences between the treatment groups in the mean values at each visit within and across cohorts. Such comparisons may be of use in speculating whether or not the MAR assumption is reasonable and whether the pre-specified primary and sensitivity analyses are adequate to address the missing data problem.

In studies 2871 and 2970, the exacerbation data was summarized in terms of recorded (i.e., not imputed) on-treatment exacerbations only and imputed year rates and counts of moderate and severe exacerbations. Supplementary analyses used imputed yearly rates and counts of moderate and severe exacerbations using a linear equation that accounted for the number of recorded on-treatment exacerbations and which quarter the exacerbation fell into (Table 3). The calculation of imputed exacerbation rates was based on treatment period intervals in order to avoid obtaining high imputed rates if a subject withdrew very early from the study after experiencing an exacerbation. Since treatment courses for moderate/severe exacerbations were to be ≤4 weeks when possible, imputed numbers of exacerbations for subjects who withdrew from the study were based on 4-week intervals of the treatment period.

Table 3: Exacerbation Quarters

Period	Period Start	Period End
Quarter 1	day 1	day 91
Quarter 2	day 92	day 182
Quarter 3	day 183	day 273
Quarter 4	day 274	day 364
N/A	day 365	N/A

Like the primary analysis, this approach assumes that there is no relationship between the response and the missing outcome i.e., the method assumes that the event rate after withdrawal

[1] Carpenter, Roger and Kenward. Analysis of Longitudinal Trials with Protocol Deviation: A Framework for Relevant, Accessible Assumptions, and Inference via Multiple Imputation

from trial is the same as the event rate on study treatment. This is often not the case, particularly when the reason for missing data is treatment-related.

For studies 2352 and 3109 the primary analysis for the primary endpoint, change from baseline trough in 24–hour weighted mean serial FEV_1 on Day 84, was analyzed using an ANCOVA model with covariates baseline FEV_1, reversibility stratum, smoking status (at screening), country and treatment. For study 3091 the primary analysis for the primary endpoint, weighted mean serial FEV_1 over 0–24 hour post-dose at the end of the 24-week treatment (Day 168), was analyzed using an ANCOVA model with covariates baseline FEV_1, region, sex, age, and treatment group. All analyses were conducted on the ITT population.

3.2.3 Patient Disposition, Demographic and Baseline Characteristics

The summary of the patient disposition in studies 2206 and 2207 is given in Table 4 and Table 5 respectively and studies 2871 and 2970 are shown in Table 5. Study 2206 had about 30% of the patients withdraw from the study. Study 2207 had about 25% of the patients withdraw from the study. Note that the applicant assumed that approximately 27% of patients would withdraw before the end of the treatment period in studies 2206 and 2207. The primary reasons for discontinuation were adverse advent (AE) with 7% to 9% in the FF/VI groups and 7% to 12% in the VI group and lack of efficacy with 3% to 6% in the FF/VI groups, 6% to 10% in the placebo group, 5% to 7% in the VI groups and 2% to 9% in the FF group. For both studies, lack of efficacy was higher in the placebo groups compared to the other treatment groups. Protocol violations accounted for 1% to 3% overall for the discontinuations.

About 25% of the patients withdrew in study 2871 and about 27% of the patients withdrew in study 2970 (Table 6). The primary reasons for discontinuation was AE (7% overall in both studies) and withdrawal of consent (6% overall in both studies). Lack of efficacy accounted for 4% to 5% of the discontinuation. Lack of efficacy due to exacerbations accounted for 3% in both studies.

Table 4: Study 2206 Summary of Patient Disposition

	Number (%) of Patients				
	FF 100	VI 25	FF/VI 50/25	FF/VI 100/25	Placebo
Randomized	206	205	206	206	207
Completed	145 (70)	142 (69)	147 (71)	151 (73)	138 (67)
ITT	206	205	206	206	207
PP	204	191	195	197	196
Discontinued	61 (30)	63 (31)	59 (29)	55 (27)	69 (33)
Adverse Event	23 (11)	24 (12)	17 (8)	14 (7)	15 (7)
Lack of Efficacy	18 (9)	15 (7)	12 (6)	12 (6)	20 (10)
Exacerbation	16 (8)	13 (6)	9 (4)	12 (6)	17 (8)
Protocol Deviation	4 (2)	2 (<1)	1 (<1)	4 (2)	3 (1)
Patient Reached Protocol-defined Stopping Criteria	5 (2)	8 (4)	13 (6)	9 (4)	11 (5)
Study closed/terminated	0	0	0	0	0
Lost to Follow-up	0	2 (<1)	1 (<1)	3 (1)	4 (2)
Investigator discretion	2 (<1)	5 (2)	5 (2)	4 (2)	5 (2)
Patient Withdrew Consent	9 (4)	7 (3)	10 (5)	9 (4)	11 (5)

Source: Clinical Study Report-Protocol Number HZC112206 Table 6, page 72

Table 5: Study 2207 Summary of Patient Disposition

	Number (%) of Patients					
	FF 100	FF 200	VI 25	FF/VI 100/25	FF/VI 200/25	Placebo
Randomized	204	204	204	204	205	205
Completed	155 (76)	160 (79)	161 (79)	144 (71)	158 (77)	146 (71)
ITT	204	203	203	204	205	205
PP	193	190	191	193	194	198
Discontinued	49 (24)	43 (21)	42 (21)	60 (29)	47 (23)	59 (29)
Adverse Event	12 (6)	15 (7)	15 (7)	17 (8)	19 (9)	18 (9)
Lack of Efficacy	5 (2)	6 (3)	11 (5)	8 (4)	7 (3)	12 (6)
Exacerbation	2 (<1)	5 (2)	11 (5)	7 (3)	7 (3)	12 (6)
Protocol Deviation	7 (3)	2 (<1)	3 (1)	8 (4)	4 (2)	7(3)
Patient Reached Protocol-defined Stopping Criteria	12 (6)	7 (3)	7 (3)	15 (7)	12 (6)	7 (3)
Study closed/terminated	1 (<1)	0	0	0	1 (<1)	0
Lost to Follow-up	2 (<1)	0	0	2 (<1)	1 (<1)	3 (1)
Investigator discretion	1 (<1)	6 (3)	3 (1)	1 (<1)	1 (<1)	4 (2)
Patient Withdrew Consent	9 (4)	7 (3)	3 (1)	9 (4)	2 (<1)	8 (4)

Source: Clinical Study Report-Protocol Number HZC112207 Table 6, page 71

Table 6: Summary Patient Disposition Study 2871 and Study 2970

	Number (%) of Patients			
	VI 25	FF/VI 50/25	FF/VI 100/25	FF/VI 200/25
Study 2871				
Randomized	409	408	403	402
Completed	294 (72)	315 (77)	312 (77)	301 (75)
ITT	409	408	403	402
PP	390	393	381	381
Discontinued	115 (28)	93 (23)	91 (23)	101 (25)
Adverse Event	22 (5)	25 (6)	29 (7)	31 (8)
Withdrew Consent	34 (8)	18 (4)	17 (4)	22 (5)
Lack of Efficacy	24 (6)	16 (4)	11 (3)	18 (4)
Exacerbation	15 (4)	10 (2)	4 (<1)	13 (3)
Protocol Deviation	8 (2)	7 (2)	8 (2)	7 (2)
Patient Reached Protocol-defined Stopping Criteria	10 (2)	14 (3)	13 (3)	10 (2)
Study closed/terminated	2 (<1)	0	1 (<1)	0
Lost to Follow-up	11 (3)	7 (2)	6 (1)	5 (1)
Investigator discretion	4 (<1)	6 (1)	6 (1)	8 (2)
Study 2970				
Randomized	409	412	403	409
Completed	284 (69)	303 (74)	291 (72)	306 (75)
ITT	409	412	403	409
PP	382	391	379	386
Discontinued	125 (31)	109 (26)	112 (28)	103 (25)
Adverse Event	25 (6)	32 (8)	35 (9)	30 (7)
Withdrew Consent	30 (7)	22 (5)	25 (6)	25 (6)
Lack of Efficacy	35 (9)	14 (3)	16 (4)	14 (3)
Exacerbation	20 (5)	8 (2)	9 (2)	7 (2)
Protocol Deviation	7 (2)	11 (3)	9 (2)	8 (2)
Patient Reached Protocol-defined Stopping Criteria	11 (3)	13 (3)	12 (3)	9 (2)
Study closed/terminated	1 (<1)	1 (<1)	0	0
Lost to Follow-up	6 (1)	8 (2)	6 (1)	10 (2)
Investigator discretion	10 (2)	8 (2)	9 (2)	7 (2)

Source: Clinical Study Report-Protocol Number HZC102871 Table 4, page 55 and HZC10290 Table 4, page 54

The demographics and baseline characteristics in studies 2206 and 2207 are summarized in Table 20 and Table 20, respectively for the ITT population (see appendix). The patients' mean age was about 62 to 63 years in the two studies. Most of the patients were White (72% ~ 94%) and male (67% ~ 72%) in these two studies. The mean body mass index (BMI) of the patients was 26.1 kg/m^2 to 26.5 kg/m^2 which indicated that the patients were slightly overweight in both studies.

The demographics and baseline characteristics in studies 2871 and 2970 are summarized in Table 22 and Table 23, respectively for the ITT population (see appendix). The patients' mean age was about 63.6 to 63.7 years in these two studies. Most of the patients were White (82% ~ 88%) and male (59% ~ 55%) in these two studies. The BMI of the patients was 26.69 kg/m^2 to 27.05 kg/m^2 which indicated that the patients were slightly overweight in both studies.

Less than 11% of patients withdrew from the three active-comparator studies (7% in study 2352, 9% in 3109, and 11% in 3091). The reasons for discontinuation varies from withdraw of consent, protocol deviation, lack of efficacy, and adverse events, but generally they were well-balanced across treatment groups. For studies 2352 and 3109 the patients' mean age was about 61 to 62 years. Majority of the patients were White (94% ~ 97%) and male (64% ~ 68%) in these three studies. The BMI of the patients was 27.3 kg/m^2 to 27.5 kg/m^2 which indicated that the patients were slightly overweight in these studies. In the asthma study, study 3091, the patients are younger with a mean age of 43 years. Most of the patients were White (59%) and female (61%). The median height was 163 cm and the median weight was 70.5 kg.

3.2.4 Results and Conclusions

3.2.4.1 Lung Function Studies (Studies 2206 and 2207)

In both studies, the VI 25 treatment group showed a statistically significant improvement in the weighted mean FEV_1 compared to the placebo group, with a 103 mL improvement in study 2206 (Table 7) and a 185 mL improvement in study 2207 (Table 8).

In study 2206, the FF/VI 100/25 treatment group showed a statistically significant improvement over the placebo group (with a 173 mL improvement), as well as over the FF 100 treatment group (with a 120 mL improvement). This statistically significant improvement supports the demonstration of the benefit of FF/VI 100/25 over FF 100 on lung function in study 2206. In study 2207, the FF/VI 200/25 treatment group showed a statistically significant improvement over the placebo group with a 209 mL improvement, as well as over the FF 200 treatment group with a 168 mL improvement. This statistically significant improvement supports the demonstration of the benefit of FF/VI 200/25 over FF 200 to lung function, similar to study 2206 but in a different dosage. In both studies, the higher dose FF/VI combination did not have a larger effect on the weighted mean FEV_1 compared to the lower dose FF/VI combination.

In both studies, the results for trough FEV_1 also showed a statistically significant improvement for the VI 25 treatment group compared to the placebo group, with a 67 mL improvement in study 2206 and a 100 mL improvement in study 2207.

In study 2206, the FF/VI 100/25 treatment group showed a statistically significant improvement in trough FEV_1 over the placebo group but failed to show statistically significant improvement over the VI 25 group. The same was observed in study 2207 where FF/VI 200/25 treatment group also failed to show statistical significant improvement over VI 25. In both studies, a numerical improvement was observed comparing FF/VI to VI 25 (48 mL in study 2206 and 32 mL in study 2207). In both studies, the higher dose FF/VI combination did not have a larger effect on the trough FEV_1 compared to the lower dose FF/VI combination.

Because multiple endpoints and multiple arms were being evaluated in both studies, hierarchical order for testing the null hypotheses was pre-specified by the applicant (Figures 1 and 2) with the high dose combination tested first (level 1) before the low dose combination (level 2a) or the secondary endpoints (level 2b and level 3). In both studies, achievement of level 1 in the hierarchical step-down approach at the 5% significance level was not met since the FF/VI treatment group did not achieve statistical significance over the VI 25 treatment group for the primary endpoint trough FEV_1 at day 169. In the strictest sense of alpha spending, all the alpha has been spent at level 1. Therefore, the p-values reported by the applicant from their analyses of the lower dosages are nominal p-values (Tables 7 and 8).

Table 7: Study 2206 Primary Efficacy Results (ITT Population)

	FF 100 N=206	VI 25 N=205	FF/VI 50/25 N=206	FF/VI 100/25 N=206	Placebo N=207
0–4 hrs Weighted Mean FEV_1 (L) at Day 168					
n[1]	206	205	205	206	207
LS Mean	1.29	1.34	1.43	1.41	1.24
LS Mean Δ	0.08	0.13	0.22	0.20	0.03
Drug vs Placebo					
Difference	0.053	0.103	0.192	0.173	
95% CI	0.003,0.104	0.052, 0.153	0.141,0.243	0.123, 0.224	
p-value	0.040*	<0.001	<0.001*	<0.001	
Drug vs FF 100					
Difference				0.120	
95% CI				0.07, 0.17	
p-value				<0.001	
Drug vs VI 25					
Difference			0.090	0.071	
95% CI			0.039,0.140	0.021,0.121	
p-value			<0.001*	0.006*	
Trough FEV_1 (L) at Day 169					
n[1]	202	202	204	206	205
LS Mean	1.28	1.32	1.38	1.36	1.25
LS Mean Δ	0.07	0.10	0.17	0.15	0.04
Drug vs Placebo					
Difference	0.033	0.067	0.129	0.115	
95% CI	-0.022,0.088	0.012,0.121	0.074,0.184	0.06,0.17	
p-value	0.241*	0.017	<0.001*	<0.001	
Drug vs FF 100					
Difference				0.082	
95% CI				0.028,0.136	
p-value				0.003*	
Drug vs VI 25					
Difference			0.062	0.048	
95% CI			0.008,0.117	-0.006,0.102	
p-value			0.025*	**0.082**	

Source: Clinical Study Report-Protocol Number HZC112206 Table 19, page 91 and Table 21, page 96.
1 Number of patients with analyzable data for 1 or more time points
* Nominal p-values
Black font = Level 1 of the testing hierarchy, Red font = Level 2a of the testing hierarchy, Blue font = additional analyses

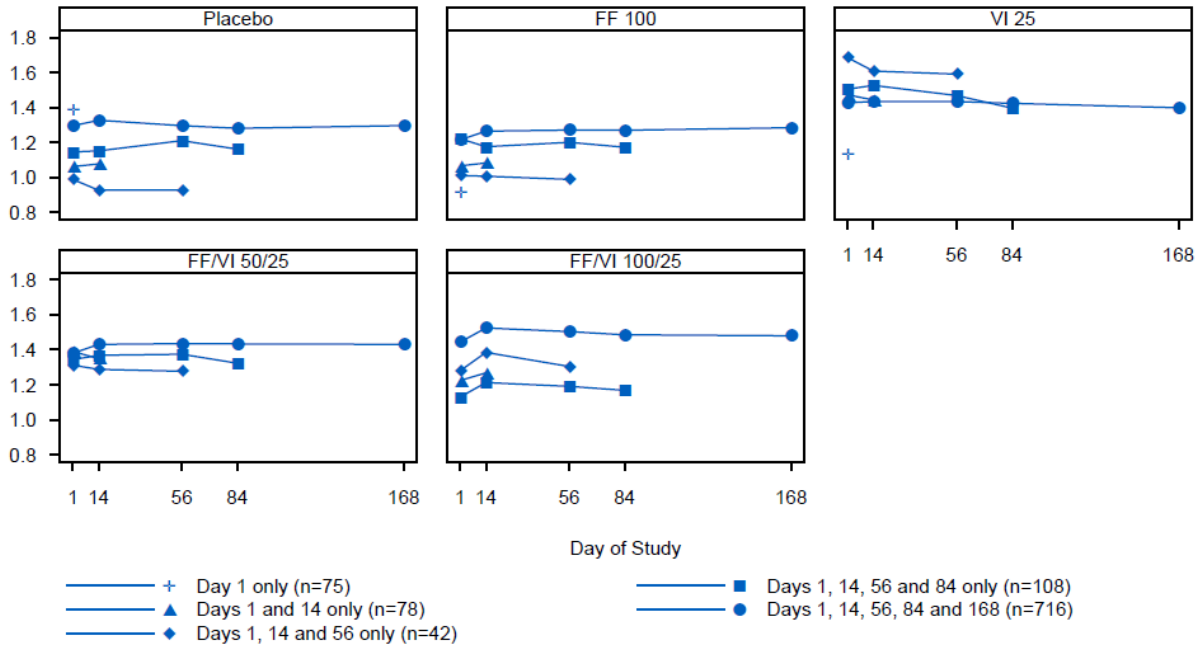
Table 8: Study 2207 Primary Efficacy Results (ITT Population)

	FF 100 N=204	FF 200 N=203	VI 25 N=203	FF/VI 100/25 N=204	FF/VI 200/25 N=205	Placebo N=205
0–4 hrs Weighted Mean FEV_1 (L) at Day 168						
n[1]	203	203	202	203	205	205
LS Mean	1.38	1.37	1.52	1.55	1.54	1.33
LS Mean Δ	0.03	0.03	0.17	0.20	0.20	-0.01
Drug vs Placebo						
Difference	0.046	0.041	0.185	0.214	0.209	
95% CI	-0.006,0.098	-0.011,0.093	0.133, 0.237	0.161,0.266	0.157, 0.261	
p-value	0.085*	0.123*	<0.001	<0.001	<0.001	
Drug vs FF 100						
Difference				0.168		
95% CI				0.116, 0.220		
p-value				<0.001		
Drug vs FF 200						
Difference					0.168	
95% CI					0.117, 0.219	
p-value					<0.001	
Drug vs VI 25						
Difference				0.029	0.024	
95% CI				-0.023,0.081	-0.027,0.075	
p-value				0.274*	0.357*	
Trough FEV_1 (L) at Day 169						
n[1]	202	202	202	200	204	202
LS Mean	1.39	1.36	1.45	1.49	1.48	1.35
LS Mean Δ	0.05	0.01	0.10	0.15	0.14	0.004
Drug vs Placebo						
Difference	0.044	0.008	0.100	0.144	0.131	
95% CI	-0.008,0.097	-0.044,0.060	0.048,0.151	0.091,0.197	0.08,0.18	
p-value	0.095*	0.756*	<0.001	<0.001*	<0.001	
Drug vs FF 100						
Difference				0.100		
95% CI				0.047,0.152		
p-value				<0.001*		
Drug vs FF 200						
Difference					0.123	
95% CI					0.072,0.174	
p-value					<0.001*	
Drug vs VI 25						
Difference				0.045	0.032	
95% CI				-0.008,0.097	-0.019,0.083	
p-value				**0.093***	**0.224**	

Source: Clinical Study Report-Protocol Number HZC112207 Table 19, page 89 and Table 21, page 95.
1. Number of patients with analyzable data for 1 or more time points
* Nominal p-values
Black font = Level 1 of the testing hierarchy, Red font = Level 2a of the testing hierarchy, Blue font = additional analyses

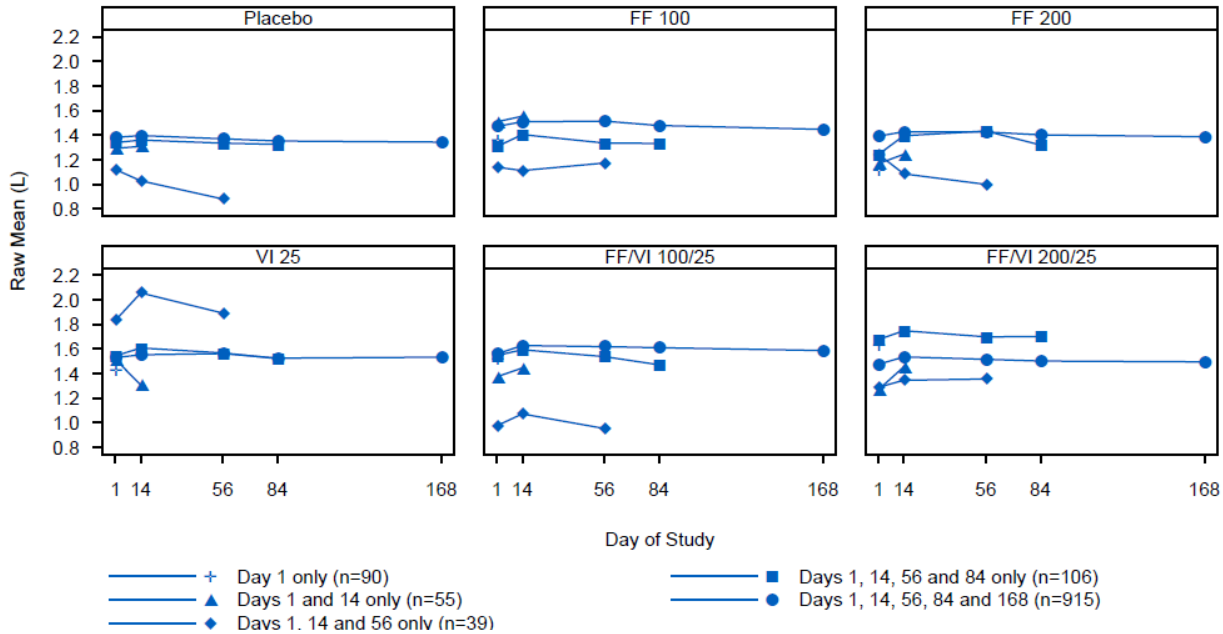
A large percentage of patients withdrew from studies 2206 (30%) and 2207 (25%). The primary reasons for the discontinuations were adverse events and lack of efficacy. The observed FEV_1 scores (0–4 hours weighted mean, Figure 4 and Figure 5, or trough, Figure 6 and Figure 7) for patients in the active arm appeared to be better than those in the placebo arm. Although cohorts who discontinued early appeared to have worse observed scores than those who discontinued later or those who completed the study, this is not as concerning because this happened in almost all treatment arms. The pre-specified primary analysis method and the sensitivity analyses have limitations since these approaches do not account for patients who may get worst post-withdrawal. Nonetheless, it is reassuring that the results of the LOCF, MAR and the CDC multiple imputations analyses (applying various missing data assumptions) conducted by the applicant were all consistent in magnitude and direction to the primary analysis (MMRM) and that the dropout rates and the reasons for discontinuations were well-balanced across the active treatment arms.

Figure 4: Study 2206- Raw Mean 0–4 hours Weighted Mean FEV_1 (L) at Each Visit by Cohort



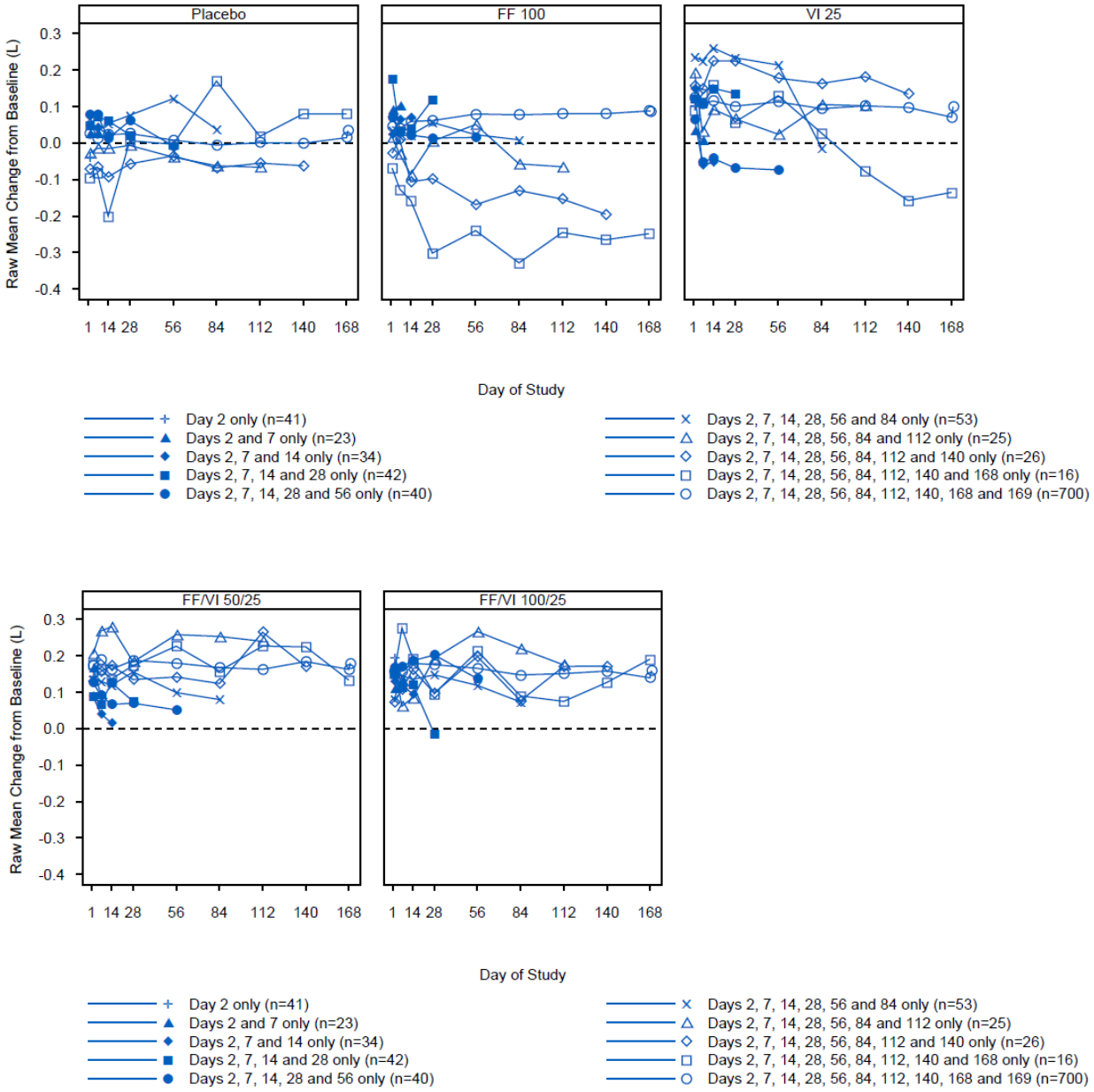
Source: Clinical Study Report-Protocol Number HZC112206 Figure 6.09, page 640

Figure 5: Study 2207- Raw Mean 0–4 hours Weighted Mean FEV_1 (L) at Each Visit by Cohort



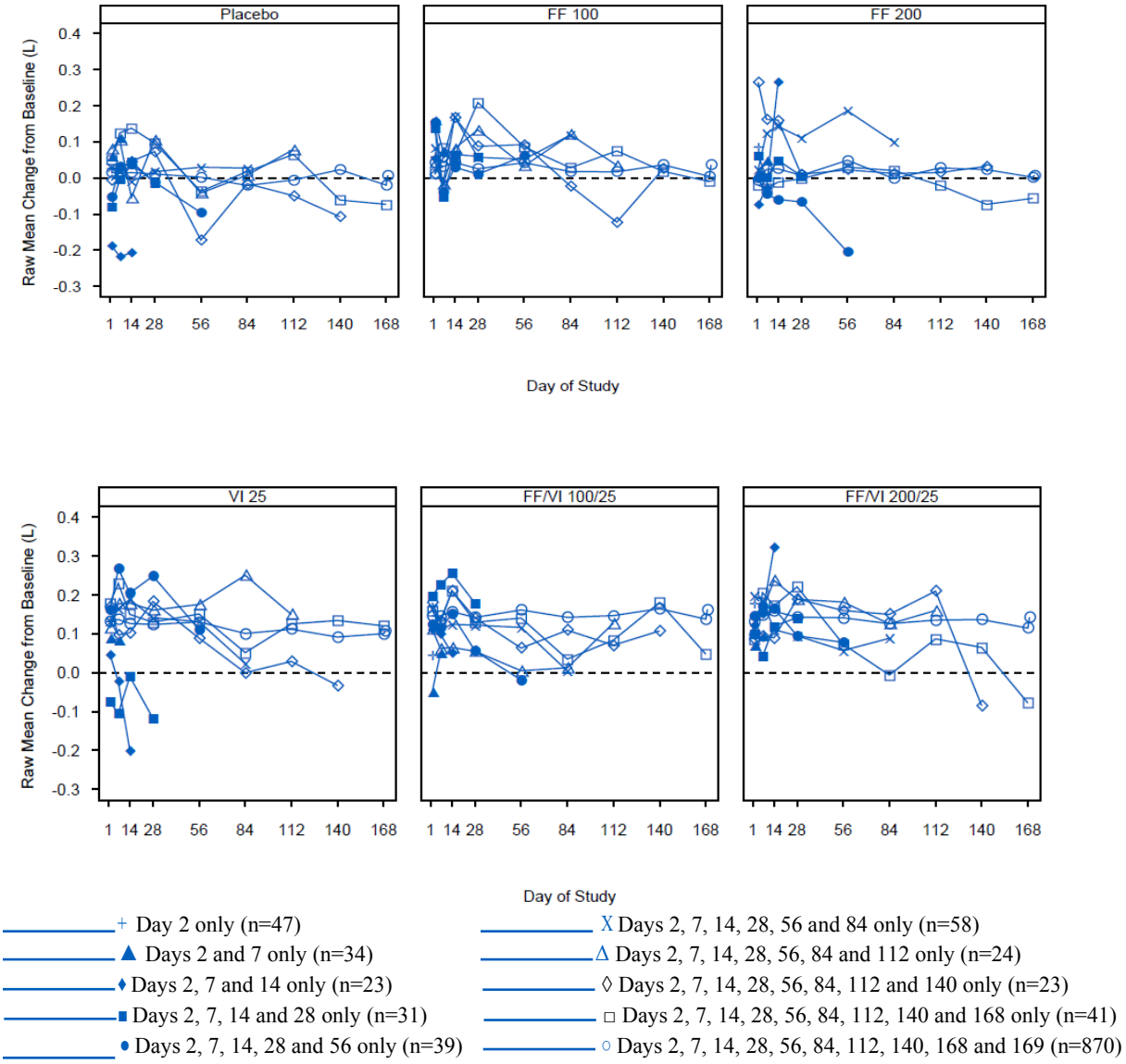
Source: Clinical Study Report-Protocol Number HZC112207 Figure 6.09, page 566

Figure 6: Study 2206-Raw Mean Change from Baseline in Trough FEV_1 (L) at Each Visit by Cohort



Source: Clinical Study Report-Protocol Number HZC112206 Figure 6.19, page 651

Figure 7: Study 2207-Raw Mean Change from Baseline in Trough FEV_1 (L) at Each Visit by Cohort



Source: Clinical Study Report-Protocol Number HZC112207 Figure 6.19, page 577

To complete the review, the results for the secondary endpoints, peak FEV_1 (Table 9 and Table 10) and time to 100 mL increase from baseline in FEV_1 (**Table 11**and **Table 12**) are shown for studies 2206 and 2207, respectively. These results are described for descriptive purposes only and the p-values reported are nominal p-values. The results from both studies were consistent in that FF/VI combination with at least a 140 mL improvement from placebo in peak FEV_1. The median time to onset at Day 1, which was defined a 100 mL increase from baseline in FEV_1, was 16 to 17 minutes post-dosing for all the FF/VI combination groups as well as VI 25 in both studies 2206 and 2207.

Table 9: Study 2206 Peak FEV_1 at Day 1-ITT Population

	FF 100	VI 25	FF/VI 50/25	FF/VI 100/25	Placebo
Randomized[1]	206	205	205	206	207
LS Mean	1.33	1.46	1.47	1.46	1.32
LS Mean Δ	0.12	0.25	0.25	0.25	0.11
Drug vs Placebo					
Difference	0.012	0.142	0.148	0.139	
95% CI	-0.015,0.039	0.114,0.169	0.120,0.175	0.112,0.166	
p-value*	0.393	<0.001	<0.001	<0.001	
Drug vs FF 100					
Difference				0.127	
95% CI				0.100,0.154	
p-value*				<0.001	
Drug vs VI 25					
Difference			0.006	-0.003	
95% CI			-0.022,0.033	-0.030,0.025	
p-value*			0.672	0.844	

Source: Clinical Study Report-Protocol Number HZC112206 Table 25, page 104.
* p-values are nominal

Table 10: Study 2207 Peak FEV_1 at Day 1-ITT Population

	FF 100 N=204	FF 200 N=203	VI 25 N=203	FF/VI 100/25 N=204	FF/VI 200/25 N=205	Placebo N=205
N[1]	203	202	201	203	205	204
LS Mean	1.49	1.47	1.61	1.61	1.60	1.46
LS Mean Δ	0.14	0.13	0.27	0.27	0.26	0.12
Drug vs Placebo						
Difference	0.024	0.007	0.147	0.152	0.141	
95% CI	-0.006,0.055	-0.023,0.037	0.117,0.177	0.122,0.182	0.111,0.171	
p-value*	0.111	0.635	<0.001	<0.001	<0.001	
Drug vs FF 100						
Difference				0.128		
95% CI				0.100,0.158		
p-value*				<0.001		
Drug vs FF 200						
Difference					0.134	
95% CI					0.104,0.164	
p-value*					<0.001	
Drug vs VI 25						
Difference				0.005	-0.006	
95% CI				-0.025,0.036	-0.036,0.024	
p-value*				0.725	0.699	

Source: Clinical Study Report-Protocol Number HZC112207 Table 25, page 103.
* all p-values are nominal

Table 11: Study 2206 Log-Rank Analysis of Time to 100 mL or More Increase from Baseline in 0-4 h Post-Dose FEV1 at Day 1 (ITT Population)

	FF 100 N=206	VI 25 N=205	FF/VI 50/25 N=205	FF/VI 100/25 N=206	Placebo N=207
Number of Events, n(%)	97 (43)	175 (85)	174 (85)	175 (85)	90 (43)
Number Censored, n(%)	109 (53)	30 (15)	31 (15)	31 (15)	117 (57)
Median time (min)	NA	16	17	17	NA
Drug vs Placebo p-value*	0.697	<0.001	<0.001	<0.001	
Drug vs FF 100 p-value*				<0.001	
Drug vs VI 25 p-value*			0.762	0.848	

Source: Clinical Study Report-Protocol Number HZC112206 Table 27, page 106.
* p-values are nominal

Table 12: Study 2207 Log-Rank Analysis of Time to 100 mL or More Increase from Baseline in 0-4 h Post-Dose FEV1 at Day 1 (ITT Population)

	FF 100 N=204	FF 200 N=203	VI 25 N=203	FF/VI 100/25 N=204	FF/VI 200/25 N=205	Placebo N=205
Number of Events, n(%)	118 (58)	106 (52)	180 (90)	172 (85)	177 (86)	101 (50)
Number Censored, n(%)	85 (42)	96 (48)	21 (10)	31 (15)	28 (14)	103 (50)
Median Time (min)	231	242	17	16	17	NA
Drug vs Placebo p-value*	0.086	0.538	<0.001	<0.001	<0.001	
Drug vs FF 100 p-value*				<0.001		
Drug vs FF 200 p-value*					<0.001	
Drug vs VI 25 p-value*				0.777	0.427	

Source: Clinical Study Report-Protocol Number HZC112207 Table 27, page 105.
* all p-values are nominal

3.2.4.2 Exacerbation Studies (Studies 2871 and 2970)

Neither study 2871 nor study 2970 included a placebo group since it was not appropriate to include a placebo control arm for the duration of one year in patients with a history of exacerbations. Treatment with FF/VI at all strengths provided a statistically significant improvement over the VI 25 group in study 2970, but FF/VI 200/25 failed to show a statistically significant improvement over the VI 25 group in study 2871 (Table 13). In study 2871, there was a numeric improvement with FF/VI at all strengths with 13%, 34%, and 15% reduction in the annual rate of moderate and severe exacerbations for FF/VI 50/25, FF/VI 100/25 and FF/VI 200/25, respectively. For the FF/VI 100/25 group in both studies, the rate of moderate and severe exacerbation was reduced by about a quarter to a third of an event in one year. The results from the Poisson analysis were consistent in magnitude and direction with the negative binomial results in the ITT population.

Achievement of level 1 in the hierarchical step-down approach at the 5% significance level was not met in study 2871 since the FF/VI 200/25 treatment group did not achieve statistical significance over the VI 25 treatment group for the primary endpoint, annual rate of moderate and severe exacerbations (Figure 3). Therefore, the p-values reported by the applicant from their analyses of the lower dosages in study 2871 are nominal p-values (Table 13).

Table 13: Study 2871 and Study 2970 analysis of Moderate and Severe Exacerbations Negative Binomial Model-ITT Population

	VI 25	FF/VI 50/25	FF/VI 100/25	FF/VI 200/25
Study 2871				
N	409	408	403	402
n	407	404	401	398
LS Mean Annual Rate	1.05	0.92	0.70	0.90
Column vs. VI 25				
Ratio		0.87	0.66	0.85
95% CI		0.72, 1.06	0.54, 0.81	0.70, 1.04
p-value		0.181*	<0.001*	0.109
Percent Reduction		13	34	15
95% CI		-6, 28	19, 46	-4, 30
Study 2970				
N	409	412	403	409
n	402	411	401	407
LS Mean Annual Rate	1.14	0.92	0.90	0.79
Column vs. VI 25				
Ratio		0.81	0.79	0.69
95% CI		0.66, 0.99	0.64, 0.97	0.56, 0.85
p-value		0.040	0.024	<0.001
Percent Reduction		19	21	31
95% CI		1, 34	3, 36	15, 44

Source: Clinical Study Report-Protocol Number HZC102871 Table 13, page 67 and Protocol Number HZC102970 Table 13, page 66.
* nominal p-values

Like the lung function studies, a large proportion of patients withdrew from studies 2871 (25%) and 2970 (27%). The dropout rate was slightly higher in the VI 25 group but the reasons for discontinuation were generally well-balanced. The applicant attempted to address the missing data problem by imputing the annual rates and counts of moderate and severe exacerbations using a linear equation that accounted for the number of recorded on-treatment exacerbations and which quarter the exacerbation occurred. Like the primary analysis, this approach assumes that there is no relationship between the response and the missing outcome i.e., the method assumes that the event rate after withdrawal from trial is the same as the event rate on study treatment. This is often not the case particularly when the reason for missing data is treatment-related. In fact, it is difficult to predict the number of exacerbations one may have post-withdrawal except to collect the actual exacerbation data after patient withdraws from the study. Therefore, the applicant's reported rates are crude estimates based on the assumption that the same event rates occur between pre- and post-withdrawal.

Examining the exacerbation data in other ways can be informative. One such analysis is the time to first moderate or severe exacerbation. Compared to the primary endpoint (i.e., annual rate of moderate and severe exacerbation), the number of missing data can be smaller since many patients may have had their first exacerbation prior to withdrawal. In study 2871, of the 25% of patients who withdrew from the study or treatment, about 54% had missing exacerbation data. Therefore, only 14% of the ITT population had missing exacerbation data. In study 2970, of the 27% of patients who withdrew from study or treatment, about 59% had missing exacerbation data. Therefore, only 16% of the ITT population had missing exacerbation data. Assigning patients with missing data as having an exacerbation at the time of withdrawal, the results were consistent with the Applicant's findings (Table 14).

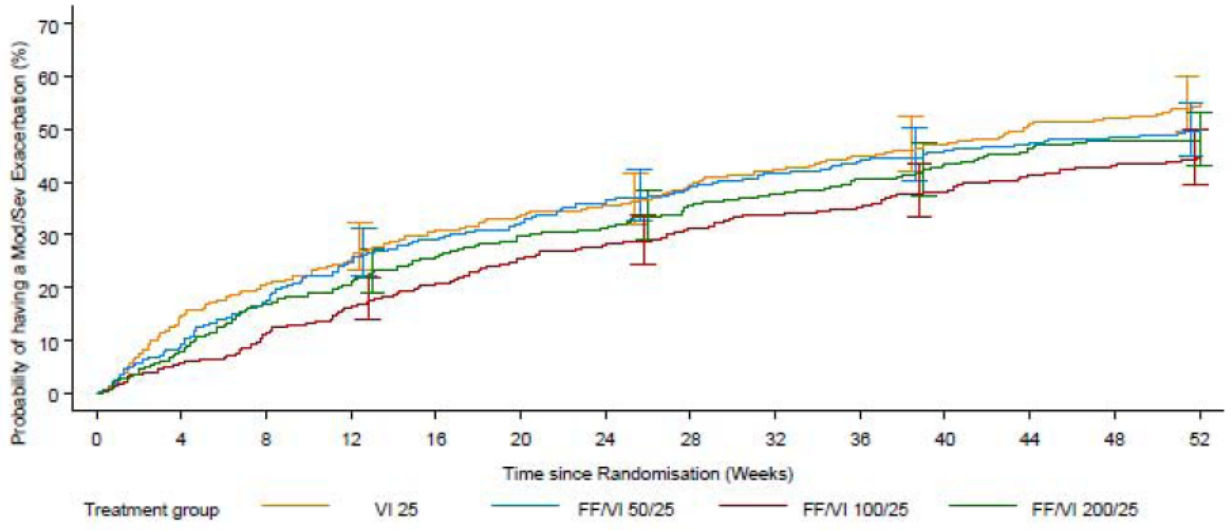
Table 14: Study 2871 and Study 2970 Analysis of Time to First Moderate or Severe On-treatment Exacerbations ITT Population

	Study 2871				Study 2970			
	VI 25	FF/VI 50/25	FF/VI 100/25	FF/VI 200/25	VI 25	FF/VI 50/25	FF/VI 100/25	FF/VI 200/25
Applicant's Results								
N	409	408	403	402	409	408	403	409
n	407	404	401	398	402	411	401	407
Column vs. VI 25								
Hazard Ratio		0.92	0.72	0.85		0.87	0.80	0.66
95% CI		0.76, 1.13	0.59, 0.89	0.69, 1.04		0.71, 1.06	0.66, 0.99	0.54, 0.82
Reviewer's Results								
Column vs. VI 25								
Hazard Ratio		0.88	0.78	0.84		0.89	0.83	0.71
95% CI		0.73, 1.04	0.65, 0.93	0.7, 1.00		0.75, 1.05	0.69, 0.98	0.59, 0.84

Source: Clinical Study Report-Protocol Number HZC102871 Table 16, page 72 and Protocol Number HZC102970 Table 16, page 70.

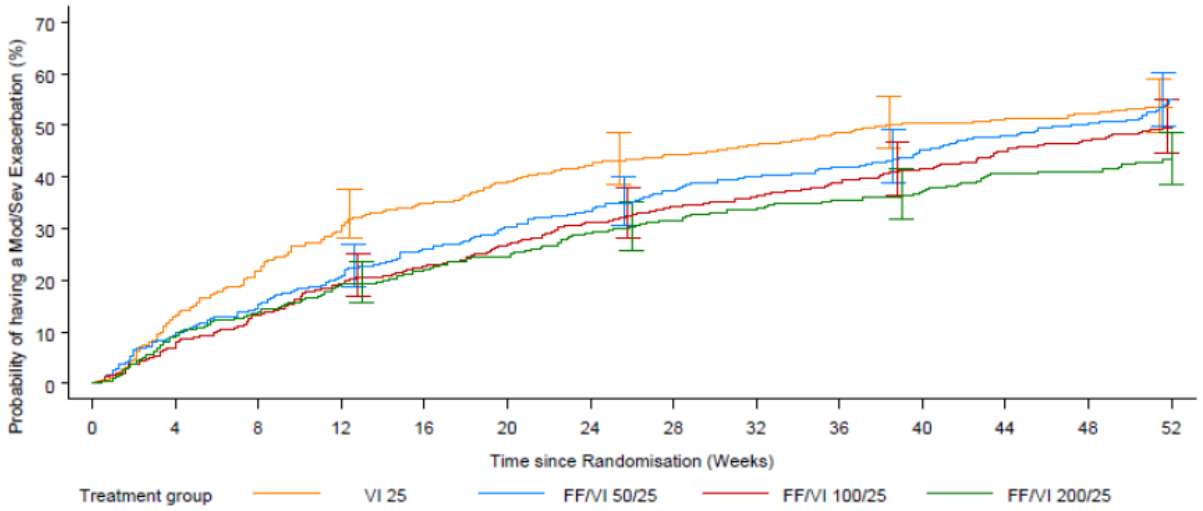
The time to first moderate or severe exacerbation showed a numerical treatment benefit for FF/VI 100/25 over VI 25 alone in both trials (Figure 8 and Figure 9). The findings are the same (figures not shown) for imputed data.

Figure 8: Kaplan-Meier Plot of Time to First Moderate or Severe Exacerbation – Study 2871



Source: Clinical Study Report-Protocol Number HZC102871 Figure 4, page 73

Figure 9: Kaplan-Meier Plot of Time to First Moderate or Severe Exacerbation – Study 2970



Source: Clinical Study Report-Protocol Number HZC102970 Figure 4, page 71

At the Pre-NDA meeting held last July 13, 2011, the Agency raised concerns regarding the lack of robust results to support the proposed bronchodilation indication from the two lung function studies (studies 2206 and 2207). The applicant proposed that the contribution of FF be

demonstrated in these exacerbation studies by the difference in exacerbation rates. Since these studies also measured trough FEV_1, they could further define the contribution of FF to changes in lung function. As noted in Section 2.1.2, the Division noted that the COPD exacerbation studies (2871 and 2970) may provide efficacy support for the addition of FF to VI, but positive exacerbation results may be problematic in the context of the negative lung function results observed in studies 2206 and 2207.

Because FF/VI 200/25 failed to show a statistically significant improvement over the VI 25 group in study 2871 for the primary endpoint, the pre-specified multiplicity plan does not allow the test of hypotheses at the lower dosages or secondary endpoints. Nonetheless, in study 2871, all three FF/VI dosage strengths showed numerical improvement compared to VI 25 for trough FEV_1 (Table 15); both FF/VI 200/25 and 100/25 had about 60 mL improvement over VI 25 and FF/VI 50/25 had a 41 mL improvement over VI 25.

On the other hand, in the positive exacerbation study 2970, there was no statistically significant improvement over VI 25 for dosages FF/VI 200/25 or FF/VI 100/25 for trough FEV_1. All three dosage strengths showed numerical improvement of about 20 to 30 mL over VI 25.

Table 15: Studies 2871 and 2970 Trough FEV_1 (L) at Week 52/Visit 11-ITT Population

	VI 25	FF/VI 50/25	FF/VI 100/25	FF/VI 200/25
Study 2871				
N	409	408	403	402
N	392	395	388	387
LS Mean (SE)	1.18 (0.0114)	1.22 (0.0112)	1.24 (0.0112)	1.24 (0.0114)
Column vs. VI 25				
Difference		0.041	0.058	0.064
95% CI		0.009, 0.072	0.027, 0.090	0.033, 0.096
p-value		0.011*	<0.001*	<0.001*
Study 2970				
N	409	412	403	409
N	387	387	381	391
LS Mean (SE)	1.22 (0.0116)	1.25 (0.0113)	1.24 (0.0115)	1.24 (0.0113)
Column vs. VI 25				
Difference		0.034	0.024	0.026
95% CI		0.003, 0.066	-0.008, 0.056	-0.006, 0.057
p-value		0.034	0.143	0.115

Source: Clinical Study Report-Protocol Number HZC102871 Table 18, page 75 and Protocol Number HZC102970 Table 18, page 73.
* nominal p-values

In summary, only one of the two exacerbation studies showed a significant improvement for all FF/VI doses over VI 25 for annual rate of moderate and severe exacerbations. In both studies, the mean rate of moderate and severe exacerbation in the VI 25 group was about 1 exacerbation per year. For the proposed dose of FF/VI 100/25, the rate of moderate and severe exacerbation was reduced by about a quarter to a third of an event in one year.

3.2.4.3 Active Comparator Studies (Studies 2532, 3109 and 3091)

In study 2532, 7% of patients discontinued from the study; however, there were an additional 8% of patients without Day 84 primary endpoint data. Similarly, in study 3109, only 9% of patients discontinued from the study, but an additional 6% (4% in FF/VI group and 8% in FP/Salmeterol group) of patients had missing Day 84 primary endpoint data. Therefore, the results presented (Table 16) by the applicant included only about 85% of the ITT population (i.e., observed case analysis). Using only observed cases in the analysis will likely introduce bias. In many cases, the use of observed cases only may not preserve the baseline comparability between treatment groups achieved by randomization. In addition, excluding patients who dropped out that are related to outcome may introduce bias and influence the results. To examine the effect of missing data, a zero change from baseline was assigned to the missing data (i.e., baseline imputation). This assumed that patients who dropped out from treatment or study did not improve and reverted back to their original baseline score. The results were consistent with the Applicant's results (Table 17). In study 3109, there was a significant improvement in weighted mean FEV_1 in the FF/VI 100/25 OD treatment group compared to FP/Salmeterol 250/50 mcg BID. Although the difference did not reach statistical significance in study 2532, there was a numeric improvement of about 25 mL in favor of FF/VI 100/25 treatment group.

Table 16: Applicant's Analysis of Weighted-Mean FEV_1 (L) up to 24 Hours on Day 84 (Completer's)

	Study 2352		Study 3109	
	FF/VI 100/25 OD PM N=259	FP/salmeterol 250/50 mcg BID N=252	FF/VI 100/25 OD PM N=260	FP/salmeterol 250/50 mcg BID N=259
N	219	217	228	213
LS Mean	1.475	1.447	1.513 (0.015)	1.433 (0.016)
LS Mean Change	0.142 (0.018)	0.114 (0.018)	0.174 (0.015)	0.094 (0.016)
FF/VI 100/25 mcg vs. FP/salmeterol 250/50 mcg	0.029		0.08	
95% CI	(-0.022, 0.080)		(0.037, 0.124)	
p-value	0.267		<0.001	

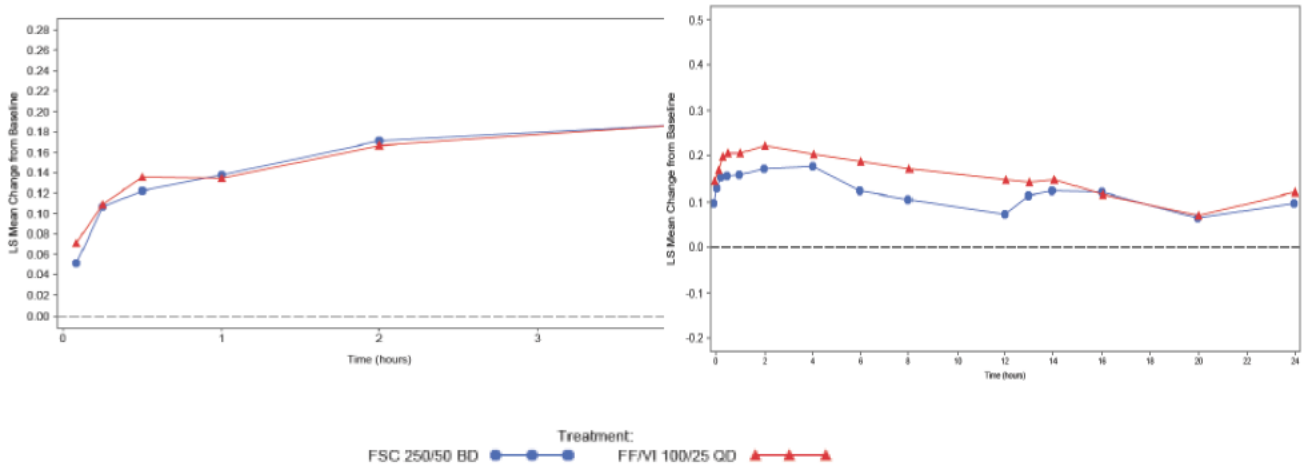
Source: Clinical Study Report HCZ112352, Table 13 page 51; Clinical Study Report HCZ113109, Table 13 page 53

Table 17: Reviewer's Analysis of Weighted-Mean FEV_1 (L) up to 24 Hours on Day 84 (ITT Population)

	Study 2352		Study 3109	
	FF/VI 100/25 OD PM N=259	FP/salmeterol 250/50 mcg BID N=252	FF/VI 100/25 OD PM N=260	FP/salmeterol 250/50 mcg BID N=259
n	259	251	260	259
LS Mean	1.48	1.45	1.52	1.44
LS Mean Change	0.13	0.11	0.20	0.12
FF/VI 100/25 mcg vs. FP/salmeterol 250/50 mcg	0.025		0.08	
95% CI	(-0.020, 0.069)		(0.04, 0.12)	
p-value	0.278		<0.001	

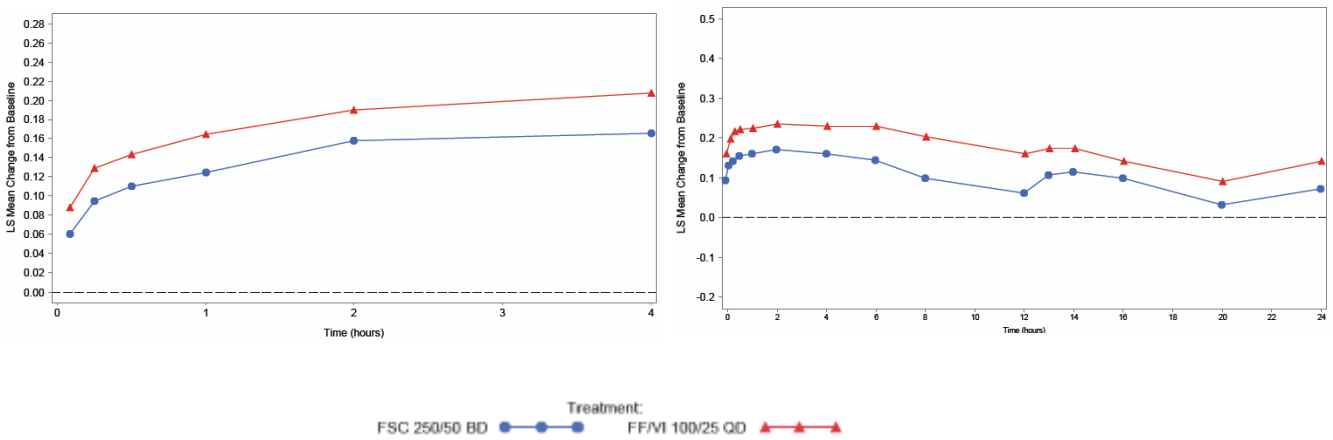
Serial FEV_1 at Day 1 and at Day 84 were also examined by the applicant. Twenty-four FEV_1 measurements were recorded at Day 84 and 4 hour measurements were recorded at Day 1. The applicant's results from applying repeated measures model at Day 1 and Day 84 are presented in Figure 10 and Figure 11. The model includes the same covariates as the primary endpoint, and missing data were not implicitly imputed in the analysis. The results were consistent with the primary analysis, in that, there is a clear separation of the curves favoring FF/VI in Study 3109 (Figure 11) as early as Day 1 and Day 84. In Study 2352, there was a small separation during the first 12 hours on Day 84 favoring FF/VI (a once a day dosing) compared to FP/Salmeterol (a twice a day dosing), but none was observed on Day 1 (Figure 10). The findings are the same (figures not shown) for the observed data.

Figure 10: LS Mean Change from baseline in FEV_1 (L) on Day 1 and Day 84 (ITT Population) – Study 2352



Source: Clinical Study Report HCZ112352, Figure 2 page 52 and Figure 4 page 55
Note: Scale in the y-axis is slightly different between the two figures.

Figure 11: LS Mean Change from baseline in FEV_1 (L) on Day 1 and Day 84 (ITT Population) – Study 3109



Source: Clinical Study Report HCZ112352, Figure 2 page 53 and Figure 4 page 57
Note: Scale in the y-axis is slightly different.

In the asthma study, study 3091, there were 11% of patients who discontinued treatment or from study. Unlike the COPD studies where 6% to 8% additional patients have missing Day 84 data, in this study only 2% additional patients have missing Day 168 data. Assigning a zero change from baseline to the missing data, the results were still consistent with the applicant's findings (Table 18). There was no significant difference observed in weighted mean FEV_1 between the FF/VI 100/25 group and FP/Salmeterol 250/50 group. There was a numeric improvement of about 22 to 37 mL in favor of FP/salmeterol treatment group in this patient population.

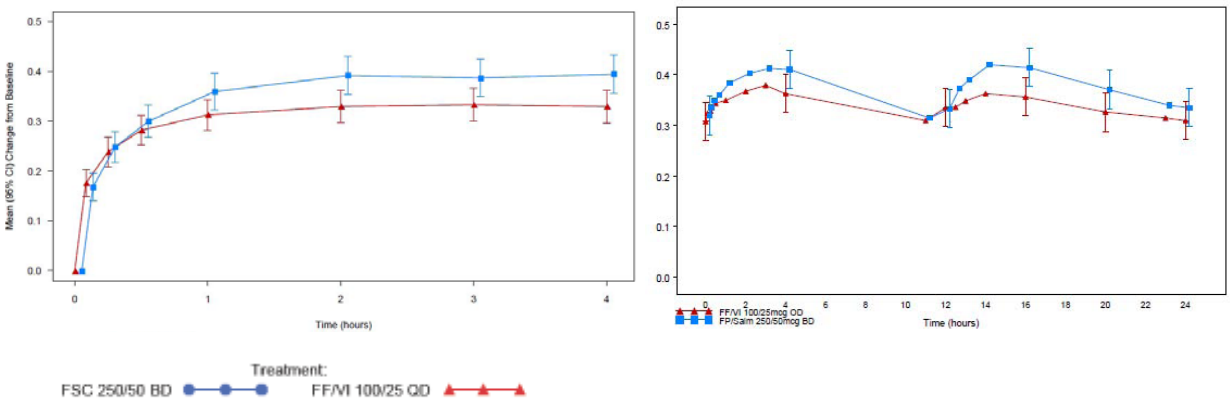
Table 18: Analysis of Weighted-Mean FEV_1 (L) up to 24 Hours on Day 84 (ITT Population) – Study 3091

	Applicant's		Reviewer's	
	FF/VI 100/25 OD PM N=403	FP/salmeterol 250/50 mcg BID N=403	FF/VI 100/25 OD PM N=260	FP/salmeterol 250/50 mcg BID N=259
N	352	347	401	401
LS Mean	2.364	2.400	2.34	2.36
LS Mean Change	0.341 (0.018)	0.377 (0.019)	0.31	0.33
FF/VI 100/25 mcg vs. FP/salmeterol 250/50 mcg	-0.037		-0.022	
95% CI	(-0.088, 0.015)		(-0.070, 0.027)	
p-value	0.162		0.380	

Source: Clinical Study Report HCA113091 Table 12 page 49

There is a separation of curves between FF/VI (a once a day dosing) and FP/Salmeterol (a twice a day dosing) favoring the FP/salmeterol group. The profiles appear to be similar at Days 1 and 168. The findings were the same (figures not shown) for the observed data.

Figure 12: LS Mean Change from baseline in FEV_1 (L) on Day 1 and Day 168 (ITT Population) – Study 3091



Source: Clinical Study Report HZA113091, Figure 3 page 52 and Figure 4 page 53
Note: Scale in the y-axis is the same.

In summary, studies 2532 and 3109 provided an additional benchmark comparison for FF/VI. The results of these studies demonstrated a similar or slightly increased mean change from baseline for FF/VI 100/25 compared to FP/Salmeterol 250/50. In the asthma study (study 3091), FP/Salmeterol 250/50 numerically outperformed FF/VI.

3.3 Evaluation of Safety

Safety evaluations for this submission will be evaluated by the Medical Reviewer, Sofia Chaudhry, M.D. Please refer to her review for more details regarding the safety findings.

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

The applicant evaluated the consistency of the treatment effect on the primary efficacy endpoints for studies 2206, 2207, 2871 and 2970 across subgroups by adding treatment-by-subgroup interaction into the primary analysis models. The statistical significance of the interaction term indicated whether the treatment effect was different among the subgroups. If any interaction p-value was less than 0.1 then further investigations were carried out.

The prespecified subgroup analyses that were considered included the following.

1. age (≤64 years and ≥65 years)
2. race (African American/African Heritage, American Indian or Alaskan Native, Asian, Native Hawaiian or other Pacific Islander, White and Mixed Race)
3. gender
4. region (US, European Union, other)
5. reversibility
6. percent predicted GOLD categories
7. smoking status
8. baseline FEV_1
9. center grouping
10. cardiovascular (CV) history/risk factors

In study 2206, there was a nominal significant quantitative interaction between treatment and reversibility for the primary endpoints, weighted mean FEV_1 0–4 hours at Day 168 (p=0.003) and change from baseline in clinic visit trough FEV_1 on treatment Day 169 (p=0.018), as well as, for weighted mean FEV_1 0–4 hours at Day 168 in study 2207 (p=0.004) (Table 24, Table 25 and Table 26, respectively). For both endpoints in study 2206, as expected, the magnitude of the treatment effect was greater in the reversible patients than in the non-reversible patients. For study 2207, the magnitude of the treatment effect was smaller in the non-reversible patients relative to the reversible patients in the FF 100, FF 200, VI 25 and FF/VI 100/25 groups. In the FF/VI 200/25 group the magnitude of the effect was larger in the non-reversible group. Both effects were in the same direction for both endpoints. On the other hand, a nominal significant quantitative interaction between treatment and smoking status at screening (p=0.065), as well as, treatment and baseline FEV_1 for change from baseline in clinic visit trough FEV_1 on treatment Day 169 (p=0.096) was observed in study 2207 (Table 26). The treatment effects in former smokers were smaller than those of the current smokers in the VI versus placebo and FF/VI 100/25 versus placebo. There was a larger treatment effect seen in former smokers compared to current smokers for the FF/VI 200/25 versus the placebo group. In general, for the treatment by baseline FEV_1 interaction, larger effects were seen with the VI 25 and FF/VI 200/25 groups compared with the placebo group in those with baseline FEV_1 values above the median of 1.3L than in those with baseline FEV_1 values below the median. In both studies, no evidence of interaction was found with treatment and age, gender, race, region, center grouping, GOLD category, baseline disease severity (pre-dose Day 1 percent predicted FEV_1) or CV history.

For study 2871, there was a nominal significant quantitative interaction between treatment and reversibility for the negative binomial model (p=0.093) (Table 28). There was a greater reduction in the annual rate of moderate and severe exacerbation for FF/VI 50/25 and FF/VI 200/25 compared to VI in the reversible subjects than in the non-reversible subjects, however the effect was opposite in the FF/VI 100/25 versus VI group. This interaction was not observed in study 2970. Instead, there was a significant interaction between treatment and smoking status for the negative binomial model in study 2970 (p=0.065) (Table 30). There was a greater reduction in the annual rate of moderate and severe exacerbation for FF/VI 50/25 and FF/VI 200/25 compared to VI in former smokers than in the current smokers, however the effect was opposite in the FF/VI 100/25 versus VI group. For study 2871, there was a nominal significant quantitative interaction between treatment and smoking status at screening (p=0.060) (Table 29). There was a greater reduction in the LS mean treatment differences for VI in all three FF/VI doses in trough FEV_1 for former smokers compared to current smokers. In study 2970 there was a nominal significant quantitative interaction between treatment reversibility for trough FEV_1 (p=0.062) (Table 31). There was a greater LS mean treatment difference for VI in all three FF/VI doses in trough FEV_1 for reversible subjects compared to non-reversible subjects. No evidence of interaction was found with treatment and age, gender, race, region, baseline disease severity (pre-dose Day 1 percent predicted FEV_1), center grouping, Gold category, or CV history in either study. Similar results were seen for the Poisson analysis.

In summary, there was some evidence of a quantitative interaction between treatment and reversibility, and between treatment and smoking in lung function and in exacerbation. The magnitude of effect appears to be greater in reversible patients and in current smokers in some of

the combination dose groups, but appears to be smaller in other combination dose groups. In the absence of a consistent effect, it is difficult to draw any definitive conclusion.

5 SUMMARY AND CONCLUSIONS

In study 2206, VI 25 showed a significant improvement compared to placebo for weighted mean 0–4 hours FEV_1 (Table 19). VI also showed a significant improvement compared to placebo for trough FEV_1. However, FF/VI 100/25 did not show a significant improvement over VI 25 for trough FEV_1, failing to show the contribution of FF in the FF/VI combination. This is in agreement with the applicant's conclusion. Change from baseline in trough FEV_1 for VI 25 was 100 mL compared to 150 mL for FF/VI 100/25; therefore, the difference, if any, was about 50 mL (95% CI: -6 mL, 100 mL).

Study 2207 showed similar results, but at the higher dosage of FF/VI, 200/25. VI also showed a significant improvement from placebo for trough FEV_1. However, FF/VI 200/25 did not show a significant improvement over VI 25 for trough FEV_1, failing to show the contribution of FF in the FF/VI combination. This is also in agreement with the applicant's conclusion. Change from baseline in trough FEV_1 for VI 25 was also 100 mL compared to 150 mL for FF/VI 100/25 and about 140 mL for FF/VI 200/25; therefore, the difference, if any, was about 45 mL (95% CI: -8 mL, 97 mL) and 32 mL (-19 mL, 83 mL), respectively.

Only one of the two exacerbation studies showed a significant improvement for all FF/VI doses over VI 25 for annual rate of moderate and severe exacerbations. In study 2970 there was a significant improvement for all FF/VI doses over VI 25 for annual rate of moderate and severe exacerbations. Study 2871 did not show a significant improvement for FF/VI 200/25 compared to VI 25 for annual rate of moderate and severe exacerbations, thus failing to show the contribution of FF in the FF/VI combination. However, there was a numeric improvement with FF/VI at all strengths with 13%, 34%, and 15% reduction in the annual rate of moderate and severe exacerbations for FF/VI 50/25, FF/VI 100/25 and FF/VI 200/25 respectively in study 2871. For the FF/VI 100/25 group in both studies, the rate of moderate and severe exacerbation was reduced by about a quarter to a third of an event in one year. Exploratory analyses of the change in trough FEV_1 showed a significant improvement at all FF/VI dosage strengths compared to VI 25 in study 2871 but not in study 2970. When compared to VI 25, the numeric improvements at all FF/VI dosage strengths were below 35 mL in study 2970 and about 50–60 mL in study 2871 that is consistent with the findings in studies 2206 and 2207.

Active comparator studies 2532 and 3109 provided an additional benchmark comparison for FF/VI. The results of these studies demonstrated a similar or slightly increased mean change from baseline for FF/VI 100/25 compared to FP/Salmeterol 250/50. In study 3091 (asthma study), FP/Salmeterol 250/50 numerically outperformed FF/VI.

In summary, there was evidence of efficacy for the VI 25 and all dosage strengths of FF/VI in the weighted mean FEV_1 (0–4 h) and change from baseline in trough FEV_1 when compared to placebo (studies 2206 and 2207). These studies also successfully demonstrated the contribution of VI 25 in the FF/VI at all dosage strengths, based on the difference in weighted mean FEV_1

(0–4 h). However, neither study demonstrated the contribution of FF in the FF/VI combination at all dosage strengths based on trough FEV_1. Change from baseline in trough FEV_1 for VI 25 was 100 mL compared to 150 mL for FF/VI 100/25 and about 140 mL for FF/VI 200/25; therefore for the proposed dose of FF/VI 100/25, the difference was about 50 mL (95% CI: -6, 102). Since the confidence interval includes zero, this implies that the direction of the difference, if any, was not known with much confidence. In both studies, the higher dose FF/VI combination did not have a larger effect on the primary endpoints (weighted mean FEV_1 or trough FEV_1) compared to the lower dose FF/VI combination.

Only one of the two exacerbation studies showed a significant improvement for all FF/VI doses over VI 25 for annual rate of moderate and severe exacerbations. In this study, the mean rate of moderate and severe exacerbation in the VI 25 group was about 1 exacerbation per year. For the proposed dose of FF/VI 100/25, the rate of moderate and severe exacerbation was reduced by about a quarter of an event in one year.

Table 19: Summary of Efficacy Findings

	Study 2206		Study 2207		Study 2871		Study 2970	
	WMFEV **Diff** **P-Value**	**Trough** **Diff** **P-Value**	**WMFEV** **Diff** **P-Value**	**Trough** **Diff** **P-Value**	**%Reduction Exacerbation** **P-value**	**Trough at Week 52** **Diff** **P-Value**	**%Reduction Exacerbation** **P-Value**	**Trough at Week 52** **Diff** **P-Value**
VI 25 vs PBO	103 mL <0.001	67 mL 0.017	185 mL <0.001	100 mL <0.001				
FF/VI 200/25 vs PBO			209 mL <0.001	131 mL <0.001				
FF/VI 200/25 vs FF 200			168 mL <0.001					
FF/VI 200/25 vs VI				32 mL 0.224	15% 0.109	64 mL <0.001*	31% <0.001	26 mL 0.115
FF/VI 100/25 vs PBO	173 mL <0.001	115 mL <0.001	214 mL <0.001	144 mL <0.001*				
FF/VI 100/25 vs FF 100	120 mL <0.001		168 mL <0.001					
FF/VI 100/25 vs VI		48 mL 0.082		45 mL 0.093*	34% <0.001*	58 mL 0.001*	21% 0.024	24 mL 0.143
FF/VI 50/25 vs PBO	192 mL <0.001	129 mL <0.001*						
FF/VI 50/25 vs VI		62 mL 0.025*			13% 0.181*	41 mL 0.007*	19% 0.040	34 mL 0.034*

Key: * = nominal p-value; red font = p-value greater than 0.05

5.1 Labeling Recommendations

The focus of the labeling review is on Sections 6 and 14. The applicant included information from the two 6-month lung function studies and two 12-month exacerbation studies.

Edits to the label are pending based on the outcome of the Pulmonary-Allergy Advisory Committee meeting to be convened on April 17, 2013. Based on the preliminary review of the label, we have the following general comments for consideration:

Section 6:

- Include information about pneumonia and possibly bone fractures since these were considered important safety findings by our clinical colleagues.

Section 14:

- Add the dose-ranging studies
- 14.1 Lung Function
  - Remove some of the figures
  - Remove results (b) (4) except for peak and onset
- 14.2 Exacerbations
  - Change (b) (4) to risk difference
  - Remove (b) (4)
  - Remove (b) (4)

6 ADVERSE REACTIONS

(b) (4)

6.1 Clinical Trials Experience

6-Month Trials: (b) (4)

6 Page(s) of Draft Labeling have been Withheld in Full as B4 (CCI/TS) immediately following this page

APPENDICES

Table 20: Study 2206-Summary of Demographics Characteristics-ITT Population

	FF 100 N=206	VI 25 N=205	FF/VI 50/25 N=206	FF/VI 100/25 N=206	Placebo N=207
Age (years)					
Mean (SD)	62.7 (9.47)	63.4 (9.58)	62.8 (9.13)	62.3 (8.49)	62.1 (8.80)
Sex n (%)					
Female	74 (36)	65 (32)	71 (34)	69 (33)	66 (32)
Male	132 (64)	140 (68)	135 (66)	137 (67)	141 (68)
Race and Racial Combinations, n (%)					
African American/African Heritage	3 (1)	7 (3)	6 (3)	9 (4)	7 (3)
American Indian or Alaska Native	0	0	1 (<1)	1 (<1)	1 (<1)
Asian	64 (31)	57 (28)	43 (21)	46 (22)	44 (21)
Central/South Asian Heritage	0	1 (<1)	0	0	0
White	139 (67)	141 (69)	156 (76)	150 (73)	155 (75)
Ethnicity, n (%)					
Hispanic or Latino	9 (4)	6 (3)	12 (6)	9 (4)	10 (5)
Not Hispanic or Latino	197 (96)	199 (97)	194 (94)	197 (96)	197 (95)
Height (cm)					
Mean (SD)	166.1 (8.46)	167.7 (9.09)	167.7 (9.24)	167.9 (9.66)	168.8 (8.16)
Weight (kg)					
Mean (SD)	71.4 (17.32)	72.2 (18.51)	73.7 (18.68)	76.5 (22.51)	74.5 (18.45)
BMI (kg/m^2)					
Mean (SD)	25.7 (5.44)	25.6 (5.98)	26.1 (5.73)	26.9 (6.80)	26.0 (5.61)

Source: Clinical Study Report-Protocol Number HZC112206 Table 8, page 76

Table 21: Study 2207-Summary of Demographic Characteristics-ITT Population

	FF 100	FF 200	VI 25	FF/VI 100/25	FF/VI 200/25	Placebo
Age (years)						
Mean (SD)	61.8 (8.28)	61.8 (9.02)	61.2 (8.62)	61.9 (8.79)	61.1 (8.67)	61.9 (8.14)
Sex n (%)						
Female	54 (26)	52 (26)	52 (26)	60 (29)	68 (33)	53 (26)
Male	150 (74)	151 (74)	151 (74)	144 (71)	137 (67)	152 (74)
Race and Racial Combinations, n (%)						
African American/African Heritage	2 (<1)	5 (2)	3 (1)	4 (2)	2 (<1)	0
American Indian or Alaska Native	0	1 (<1)	0	2 (<1)	0	0
Asian	5 (2)	14 (7)	4 (2)	8 (4)	11(5)	8 (4)
Japanese/East Asian Heritage/~ South East Asian Heritage	5(2)	14 (7)	4 (2)	8 (4)	11 (5)	8(4)
White	197 (97)	183 (90)	196 (97)	190 (93)	192 (94)	197 (96)
Ethnicity, n (%)						
Hispanic or Latino	1 (<1)	0	0	1 (<1)	0	0
Not Hispanic or Latino	203 (>99)	203 (100)	203 (100)	203 (>99)	205 (100)	205 (100)
Height (cm)						
Mean (SD)	171.7 (9.01)	169.7 (8.34)	171.2 (8.43)	171.1 (9.09)	170.3 (9.24)	170.9 (8.66)
Weight (kg)						
Mean (SD)	80.3 (19.38)	77.3 (20.24)	77.0 (17.18)	77.3 (18.81)	75.4 (16.08)	78.8 (17.08)
BMI (kg/m^2)						
Mean (SD)	27.1 (5.71)	26.7 (6.35)	26.2 (5.21)	26.2 (5.12)	25.9 (4.86)	26.9 (5.36)

Source: Clinical Study Report-Protocol Number HZC112207 Table 8, page 75

Table 22: Study 2871- Summary of Demographic Characteristics-ITT Population

n(%)		VI 25 N=409	FF/VI 50/25 N=408	FF/VI 100/25 N=403	FF/VI 200/25 N=402	Total N=1622
Age (years)	n	409	408	403	402	1622
	Mean	63.6	63.6	63.6	63.8	63.6
	SD	9.43	9.06	9.06	9.30	9.21
	Min-Max	40-87	40-88	41-88	41-90	40-90
Sex	n	409	408	403	402	1622
	Female	170 (42)	163 (40)	172 (43)	153 (38)	658 (41)
	Male	239 (58)	245 (60)	231 (57)	249 (62)	964 (59)
Race	n	408	408	403	401	1620
	White	331 (81)	334 (82)	332 (82)	324 (81)	1321 (82)
	African American/African Heritage	9 (2)	8 (2)	6 (1)	9 (2)	32 (2)
	Asian	39 (10)	37 (9)	37 (9)	41 (10)	154 (10)
	Other	29 (7)	29 (7)	28 (7)	27 (7)	113 (7)
Ethnicity	n	409	408	403	402	1622
	Hispanic or Latino	78 (19)	73 (18)	72 (18)	76 (19)	299 (18)
	Not Hispanic or Latino	331 (81)	335 (82)	331 (82)	326 (81)	1323 (82)
Body Mass Index (kg/m^2)	n	407	408	402	402	1619
	Mean	26.17	26.94	27.14	26.52	26.69
	SD	5.596	5.771	6.144	6.191	5.936
	Min-Max	14.7-44.9	14.6-47.1	15.5-58.2	12.4-54.4	12.4-58.2

Source: Clinical Study Report-Protocol Number HZC102871 Table 6, page 58

Table 23: Study 2970- Summary of Demographics Characteristics-ITT Population

n(%)		VI 25 N=409	FF/VI 50/25 N=412	FF/VI 100/25 N=403	FF/VI 200/25 N=409	Total N=1633
Age (years)	n	409	412	403	409	1633
	Mean	63.6	63.7	64.0	63.5	63.7
	SD	9.29	9.56	9.28	8.84	9.24
	Min-Max	40-85	40-85	40-88	40-86	40-88
Sex	n	409	412	403	409	1633
	Female	174 (43)	181 (44)	181 (45)	191 (47)	727 (45)
	Male	235 (57)	231 (56)	222 (55)	218 (53)	906 (55)
Race	n	409	412	403	409	1633
	White	360 (88)	359 (87)	353 (88)	359 (88)	1431 (88)
	African American/African Heritage	9 (2)	14 (3)	7 (2)	9 (2)	39 (2)
	Asian	4 (<1)	3 (<1)	5 (1)	3 (<1)	15 (<1)
	Other	36 (9)	36 (9)	38 (9)	38 (9)	148 (9)
Ethnicity	n	409	412	403	409	1633
	Hispanic or Latino	70 (17)	68 (17)	74 (18)	73 (18)	285 (17)
	Not Hispanic or Latino	339 (83)	344 (83)	329 (82)	336 (82)	1348 (83)
Body Mass Index (kg/m^2)	n	409	412	403	408	1632
	Mean	27.31	27.10	26.97	26.82	27.05
	SD	6.184	5.737	5.638	5.979	5.886
	Min-Max	14.5-63.2	15.1-51.6	14.9-50.4	13.7-56.5	13.7-63.2

Source: Clinical Study Report-Protocol Number HZC102970 Table 6, page 57

Table 24 Subgroup Analysis for 0-4 Hours Weighted Mean FEV_1 (L) at Day 168 by Reversibility for Study 2206 (ITT Population)

	FF 100	VI 25	FF/VI 50/25	FF/VI 100/25	Placebo
	N=206	N=205	N=206	N=206	N=207
Not Reversible					
LS Mean (SE)	1.284 (0.0224)	1.328 (0.0218)	1.380 (0.0227)	1.395 (0.0219)	1.236 (0.0228)
Drug vs Placebo					
Difference	0.048	0.092	0.145	0.160	
95% CI	-0.014, 0.111	0.030, 0.154	0.081, 0.208	0.098, 0.222	
p-value	0.132	0.004	<0.001	<0.001	
Drug vs VI 25					
Difference			0.052	0.067	
95% CI			-0.009, 0.114	0.007, 0.128	
p-value			0.097	0.029	
Drug vs FF 100					
Difference				0.111	
95% CI				0.050, 0.173	
p-value				<0.001	
Reversible					
LS Mean (SE)	1.306 (0.0304)	1.373 (0.0325)	1.510 (0.0299)	1.453 (0.0311)	1.244 (0.0312)
Drug vs Placebo					
Difference	0.062	0.129	0.266	0.209	
95% CI	-0.023, 0.148	0.040, 0.217	0.182, 0.351	0.122, 0.295	
p-value	0.153	0.004	<0.001	<0.001	
Drug vs VI 25					
Difference			0.138	0.080	
95% CI			0.051, 0.224	-0.008, 0.168	
p-value			0.002	0.076	
Drug vs FF 100					
Difference				0.146	
95% CI				0.061, 0.232	
p-value				<0.001	

Source: Clinical Study Report-Protocol Number HZC112206 Table 6.74, page 1377-1386

Table 25 Subgroup Analysis for Trough FEV_1 (L) at Day 169 by Reversibility for Study 2206

	FF 100	VI 25	FF/VI 50/25	FF/VI 100/25	Placebo
	N=206	N=205	N=206	N=206	N=207
Not Reversible					
LS Mean (SE)	1.278 (0.0244)	1.313 (0.0236)	1.346 (0.0247)	1.355 (0.0238)	1.246 (0.0246)
Drug vs Placebo					
Difference	0.032	0.067	0.100	0.109	
95% CI	-0.036, 0.100	0, 0.134	0.031, 0.168	0.042, 0.176	
p-value	0.359	0.050	0.004	0.001	
Drug vs VI 25					
Difference			0.033	0.042	
95% CI			-0.034, 0.100	-0.024, 0.108	
p-value			0.340	0.208	
Drug vs FF 100					
Difference				0.077	
95% CI				0.010, 0.144	
p-value				0.024	
Reversible					
LS Mean (SE)	1.289 (0.0330)	1.328 (0.0352)	1.428 (0.0324)	1.386 (0.0338)	1.257 (0.0343)
Drug vs Placebo					
Difference	0.031	0.070	0.171	0.129	
95% CI	-0.062, 0.125	-0.026, 0.167	0.078, 0.263	0.034, 0.223	
p-value	0.511	0.153	<0.001	0.008	
Drug vs VI 25					
Difference			0.100	0.058	
95% CI			0.006, 0.194	-0.037, 0.154	
p-value			0.036	0.231	
Drug vs FF 100					
Difference				0.098	
95% CI				0.005, 0.190	
p-value				0.039	

Source: Clinical Study Report-Protocol Number HZC112206 Table 6.75, page 1387-1406

Table 26 Subgroup Analysis for 0–4 Hours Weighted Mean FEV1 (L) at Day 168 by Reversibility for Study 2207 (ITT Population)

	FF 100	FF 200	VI 25	FF/VI 100/25	FF/VI 200/25	Placebo
	N=204	N=203	N=203	N=204	N=205	N=205
Not Reversible						
LS Mean	1.368	1.351	1.479	1.503	1.512	1.326
(SE)	(0.0224)	(0.0215)	(0.0222)	(0.0227)	(0.0222)	(0.0225)
Drug vs Placebo						
Difference	0.042	0.025	0.153	0.176	0.186	
95% CI	-0.020,0.104	-0.036,0.086	0.091,0.215	0.114,0.239	0.124,0.248	
p-value	0.187	0.424	<0.001	<0.001	<0.001	
Drug vs VI 25						
Difference				0.023	0.033	
95% CI				-0.039,0.086	-0.029,0.095	
p-value				0.460	0.293	
Drug vs FF 100						
Difference				0.135		
95% CI				0.072,0.197		
p-value				<0.001		
Drug vs FF 200						
Difference					0.161	
95% CI					0.101,0.222	
p-value					<0.001	
Reversible						
LS Mean	1.403	1.423	1.599	1.642	1.609	1.338
(SE)	(0.0344)	(0.0364)	(0.0330)	(0.0351)	(0.0344)	(0.0346)
Drug vs Placebo						
Difference	0.065	0.085	0.260	0.304	0.271	
95% CI	-0.031,0.161	-0.014,0.183	0.166,0.354	0.207,0.400	0.175,0.366	
p-value	0.184	0.092	<0.001	<0.001	<0.001	
Drug vs VI 25						
Difference				0.043	0.010	
95% CI				-0.051,0.138	-0.083,0.104	
p-value				0.369	0.829	
Drug vs FF 100						
Difference				0.239		
95% CI				0.142,0.335		
p-value				<0.001		
Drug vs FF 200						
Difference					0.186	
95% CI					0.087,0.284	
p-value					<0.001	

Source: Clinical Study Report-Protocol Number HZC112207 Table 6.68, page 1243-1262

Table 27 Subgroup Analysis for Trough FEV_1 (L) at Day 169 by Smoking Status for study 2207 (ITT Population)

	FF 100	FF 200	VI 25	FF/VI 100/25	FF/VI 200/25	Placebo
	N=204	N=203	N=203	N=204	N=205	N=205
Current Smoker						
LS Mean	1.398	1.345	1.457	1.504	1.443	1.347
(SE)	(0.0247)	(0.0248)	(0.0260)	(0.0263)	(0.0259)	(0.0265)
Drug vs Placebo						
Difference	0.051	-0.002	0.110	0.157	0.096	
95% CI	-0.020,0.122	-0.073,0.069	0.037,0.183	0.084,0.230	0.023,0.169	
p-value	0.157	0.958	0.003	<0.001	0.010	
Drug vs VI 25						
Difference				0.047	-0.014	
95% CI				-0.026,0.119	-0.086,0.058	
p-value				0.205	0.705	
Drug vs FF 100						
Difference				0.106		
95% CI				0.035,0.176		
p-value				0.003		
Drug vs FF 200						
Difference					0.098	
95% CI					0.028,0.168	
p-value					0.006	
Former Smoker						
LS Mean	1.382	1.368	1.433	1.477	1.518	1.348
(SE)	(0.0287)	(0.0280)	(0.0265)	(0.0279)	(0.0264)	(0.0271)
Drug vs Placebo						
Difference	0.033	0.020	0.085	0.129	0.169	
95% CI	-0.044,0.111	-0.056,0.096	0.010,0.159	0.053,0.205	0.095,0.243	
p-value	0.397	0.605	0.026	<0.001	<0.001	
Drug vs VI 25						
Difference				0.044	0.085	
95% CI				-0.031,0.120	0.011,0.158	
p-value				0.250	0.024	
Drug vs FF 100						
Difference				0.095		
95% CI				0.017,0.174		
p-value				0.017		
Drug vs FF 200						
Difference					0.149	
95% CI					0.074,0.225	
p-value					<0.001	

Source: Clinical Study Report-Protocol Number HZC112207 Table 6.69, page 1263-1302

Table 28 Subgroup Analysis for Annual Rate of Moderate and Severe Exacerbations by Reversibility for Study 2871(ITT Population)

	VI 25 N=409	FF/VI 50/25 N=408	FF/VI 100/25 N=403	FF/VI 200/25 N=402
Not Reversible				
LS Mean Annual Rate	0.94	0.88	0.61	0.91
Drug vs VI 25 Ratio		0.93	0.64	0.97
95% CI		0.74,1.19	0.50,0.83	0.76,1.23
p-value		0.576	<0.001	0.794
Percent Reduction		7	36	3
95% CI		-19, 26	17, 50	-23, 24
Reversible				
LS Mean Annual Rate	1.32	1.04	0.91	0.80
Drug vs VI 25 Ratio		0.79	0.69	0.61
95% CI		0.56,1.11	0.49,0.98	0.42,0.88
p-value		0.177	0.037	0.008
Percent Reduction		21	31	39
95% CI		-11, 44	2, 51	12,58

Source: Clinical Study Report-Protocol Number HZC102871 Table 6.48, page 714-715

Table 29 Subgroup Analysis for Trough FEV_1 (L) at Week 52 by Smoking Status for Study 2871 (ITT Population)

	VI 25 N=409	FF/VI 50/25 N=408	FF/VI 100/25 N=403	FF/VI 200/25 N=402
Former Smoker				
LS Mean (SE)	1.167 (0.0150)	1.224 (0.0146)	1.251 (0.0148)	1.267 (0.0147)
Drug vs VI 25 Difference		0.057	0.084	0.100
95% CI		0.016,0.098	0.042,0.125	0.059,0.142
p-value		0.007	<0.001	<0.001
Current Smoker				
LS Mean (SE)	1.197 (0.0176)	1.215 (0.0174)	1.220 (0.0171)	1.208 (0.0181)
Drug vs VI 25 Difference		0.018	0.023	0.012
95% CI		-0.030,0.067	-0.025,0.071	-0.038,0.061
p-value		0.454	0.342	0.639

Source: Clinical Study Report-Protocol Number HZC102871 Table 6.49, page 716-733

Table 30 Subgroup Analysis for Annual Rate of Moderate and Severe Exacerbations by Smoking Status for Study 2970 (ITT Population)

	VI 25 N=409	FF/VI 50/25 N=412	FF/VI 100/25 N=403	FF/VI 200/25 N=409
Former Smoker				
LS Mean Annual Rate	1.19	0.90	0.98	0.66
Drug vs VI 25 Ratio		0.76	0.82	0.55
95% CI		0.57,1.01	0.62,1.09	0.41,0.74
p-value		0.056	0.175	<0.001
Percent Reduction		24	18	45
95% CI		-1, 43	-9, 38	26, 59
Current Smoker				
LS Mean Annual Rate	1.09	0.94	0.81	0.94
Drug vs VI 25 Ratio		0.86	0.74	0.86
95% CI		0.64,1.16	0.55,1.01	0.64,1.16
p-value		0.330	0.055	0.324
Percent Reduction		14	26	14
95% CI		-16, 36	-1, 45	-16, 36

Source: Clinical Study Report-Protocol Number HZC102970 Table 6.47, page 714-715

Table 31 Subgroup Analysis of Trough FEV_1 (L) by Reversibility for Study 2970 (ITT Population)

	VI 25 N=409	FF/VI 50/25 N=412	FF/VI 100/25 N=403	FF/VI 200/25 N=409
Not Reversible				
LS Mean (SE)	1.213 (0.0135)	1.231 (0.0131)	1.229 (0.0134)	1.229 (0.0130)
Drug vs VI 25 Difference		0.017	0.015	0.016
95% CI		-0.019,0.054	-0.022,0.053	-0.021,0.053
p-value		0.354	0.417	0.396
Reversible				
LS Mean (SE)	1.204 (0.0204)	1.279 (0.0195)	1.258 (0.0195)	1.250 (0.0199)
Drug vs VI 25 Difference		0.075	0.055	0.046
95% CI		0.020,0.130	-0.001,0.110	-0.010,0.102
p-value		0.008	0.052	0.107

Source: Clinical Study Report-Protocol Number HZC102970 Table 6.48, page 716-733

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

/s/

KIYA HAMILTON
03/18/2013

JOAN K BUENCONSEJO
03/18/2013
I concur

THOMAS J PERMUTT
03/19/2013
concur

STATISTICAL REVIEW AND EVALUATION

Biometrics Division: VI

NDA NO.:	204-275
SERIAL NO.:	S000
DATE RECEIVED BY THE CENTER:	July 12, 2012
DRUG NAME:	Fluticasone furoate/vilanterol inhalation powder
DOSAGE FORM:	Aerosol
INDICATION:	
SPONSOR:	GlaxoSmithKline
DOCUMENTS REVIEWED:	July 12, 2012, February 7, 2013, March 12, 2013
NAME OF STATISTICAL REVIEWER:	Meiyu Shen, Ph.D.
NAME OF PROJECT MANAGER:	Youbang Liu

_____

Meiyu Shen, PhD, Mathematical Statistician

Concur: _____

Yi Tsong, Ph.D.
Deputy Director, DBVI

TABLE OF CONTENTS

1 STATISTICAL REVIEW AND EVALUATION OF EVIDENCE

1.1 Introduction and Background

The sponsor proposes use of Parametric tolerance interval testing (PTIT) for Dose Content Uniformity and Dose Content Uniformity through Life. Office of New Drug Quality and Assessment sent the request to CMC statistical team in Division of Biometric VI for evaluating the adequacy of the proposed test for control of this product on September 21, 2012. On January 7, 2013, the agency sent the information request regarding the PTIT method and alternative sample size to the sponsor. The sponsor responded the agency's request on February 7, 2013. The sponsor's response dated on March 12, 2013 to the Agency's March 7's information request was also reviewed.

1.2 Data Analyzed and Sources

There was no data submitted for review.

1.3 Sponsor's proposal and justification

1.3.1 Sponsor's proposal for sampling plan with 2-tier test of total 30 canisters (60 observations)

GSK proposes to apply the acceptance criteria for Content Uniformity of the Emitted Dose outlined in the FDA's October 25, 2005 Advisory Committee of Pharmaceutical Science Proposal for Parametric Tolerance Interval Test (PTI Test) Criteria.

The sponsor's acceptance criteria for sample size 20 at 1st tier and 60 at 2nd tier are following: at the first tier, the Content Uniformity of the Emitted Dose test is initially performed on 10 inhalers and yields 20 results. Each individual determination represents a single inhalation (the minimum clinical dose). The lot passes the PTIT test and is then released to the market if the sample mean falls within (b) (4) label claim and the 2 1-sided tolerance interval with (1-$\alpha 1$)*100% confidence level and at least 87.5% coverage, ($X_1 - K_1 * S_1$, $X_1 + K_1 * S_1$), falls within (80, 120)% Label claim. X_1, S_1 are the sample mean of 20 determinations and sample standard deviation of 20 determinations, respectively, where K_1 = (b) (4). If not, the sponsor continues to the second stage. At the second tier testing, an additional 20 inhalers are sampled to provide 60 results in total. The lot passes the PTIT test and is then released to the market if the sample mean falls within (b) (4) label claim and the 2 1-sided tolerance interval with (1-$\alpha 2$)*100% confidence level and at least 87.5% coverage, ($X_2 - K_2 * S_2$, $X_2 + K_2 * S_2$), falls within (80, 120)% Label claim. If not, the lot can't be released to the market. X_2, S_2 are the sample mean of 60 determinations and sample standard deviation of 60 determinations, respectively, where K_2 = (b) (4). $\alpha 1$ and $\alpha 2$ are determined (b) (4)

1.3.2 Sponsor's alternative sample size

In the original submission, the sponsor proposed an alternative sample size at post-approval and intended to commit to maintain a 1:3 ratio between the sample sizes for the first and second tier, and to use the Lan-DeMets implementation of the Pocock approach to calculate the

corresponding coefficients (Novick et al, 2009). But the sponsor did not pre-specify any sample size to be potentially used.

1.4 Reviewer's comments on the sponsor's PTIT

The statistical reviewer evaluated the sponsor PTIT method for 2-tier test with total 30 canisters (60 observations) and concluded the sponsor's PTIT method for sample size 60 was acceptable given that the sponsor added one of 2 observations per canister from beginning and the other of 2 observations per canister from the end.

Because the sponsor did not specify what alternative sample would be at post-approval, the agency sent the sponsor the information request regarding alternative sample size on January 7, 2013. In the information request, the agency asked the sponsor the following:
"Prespecify the alternative sample sizes and the corresponding k-values (the tolerance coefficient). You may propose extending the two, 1-sided PTIT procedure at sample size (b) (4) by intersecting with the OC curve of the (b) (4) for the two, 1-sided PTIT procedure at a pre-specified acceptance probability, e.g., 90%."

The sponsor proposed some possible samples and the corresponding k-values listed in Table 1 of the sponsor's February 7, 2013 response to FDA information request on January 7, 2013. From the characteristic operating curves, it can be easily seen that operating curves of PTIT for alternative sample sizes all intersect with operating curve of PTIT for total sample size 60 at (b) (4) passing probability. Hence after the statistician reviewed the sponsor's February 7, 2013 response to the information request of January 7, 2013, the agency sent the following information request to the sponsor on March 7, 2013.

> "Your response (dated 7-Feb-2013) to item 9 of the Agency's information requests regarding the alternative sample sizes and the corresponding k-values for emitted dose uniformity testing is not acceptable. As indicated by the Operating Characteristic (OC) curve you provided, your possible sampling approach (of an alternative sample size at a 1:3 ratio between the sample sizes for the first and second tier) allows increased passing probability of a given batch with sheer increase in sample size. This is not acceptable. To resolve the issue, you may confirm not to change the currently proposed sample size of 20 for Tier 1 and (b) (4) for Tier 2 testing without prior agreement or approval from the Agency.
>
> Alternatively, you may choose the sample sizes, but the alternative sample-size test should have a 90% probability to pass at the quality standard at which the test for the 20/60 sample size plan (20 at 1st tier, 60 at 2nd tier) has a 90% probability of at least 87.5% coverage with 95% confidence level tolerance interval for the total of 60 samples falling between 80% and 120% of label claim."

In the sponsor's response (dated on March 12, 2013) to above information request, the sponsor claimed that "The Two 1-Sided PTIT Procedure with 87.5% Coverage, 95% Confidence and a 1:3 Tier Ratio is designed (b) (4) ".

In order to comment on the sponsor's probability claim, the two one-sided hypotheses can be set up as:

H_0^U: Pr $(X \geq U) \geq P_U$ versus H_a^U: Pr $(X \geq U) < P_U$ (1)
H_0^L: Pr $(X \leq L) \geq P_L$ versus H_a^L: Pr $(X \leq L) < P_L$ (2)

Where X is the random variable for delivery dose through out the life of usage of the inhaler, L=80, U=120, and P_U=P_L= (b) (4)

To illustrate our reasoning, we will use tier 1 as an example. (b) (4)
(b) (4)

α1 is type I error rate, the probability of rejecting H_0^U and H_0^L under H_0^U or H_0^L. (1-α1)*100% is probability of not rejecting H_0^U and H_0^L under H_0^U or H_0^L. Clearly, the confidence level (1-α1)*100% is not the probability of rejecting H_0^U and H_0^L under H_a^U and H_a^L. (b) (4)

Hence the sponsor's k values (tolerance factors) for different sample sizes are derived on the basis of (b) (4). However, (b) (4) is not a concern from practical point of view (b) (4) Hence the k-values should be derived on the maintaining 90% power for different sample size such that the alternative sample-size test should have a 90% passing probability to pass at the quality standard at which the test for the 20/60 sample size plan (20 at 1st tier, 60 at 2nd tier) has a 90% probability of at least 87.5% coverage with 95% confidence level tolerance interval for the total of 60 samples falling between 80% and 120% of label claim.

1.5 Conclusions and Recommendation

The statistical reviewer evaluated the sponsor PTIT method for 2-tier test with total 30 canisters (60 observations) and concluded the sponsor's PTIT method for sample size 60 was acceptable given that the sponsor added one of 2 observations per canister from beginning and the other of 2 observations per canister from the end.

The alternative sample-size test should have a 90% passing probability to pass at the quality standard at which the test for the 20/60 sample size plan (20 at 1st tier, 60 at 2nd tier) has a 90% probability of at least 87.5% coverage with 95% confidence level tolerance interval for the total of 60 samples falling between 80% and 120% of label claim.

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

/s/

MEIYU SHEN
03/18/2013

YI TSONG
03/18/2013

STATISTICS FILING CHECKLIST FOR A NEW NDA/BLA

NDA Number: 204275 **Applicant: GSK** **Stamp Date: July 12, 2012**

Drug Name: Breo Ellipta (fluticasone furoate/vilanterol) Inhalation Powder **NDA/BLA Type: NDA**

On **initial** overview of the NDA/BLA application for RTF:

	Content Parameter	Yes	No	NA	Comments
1	Index is sufficient to locate necessary reports, tables, data, etc.	X			
2	ISS, ISE, and complete study reports are available (including original protocols, subsequent amendments, etc.)	X			
3	Safety and efficacy were investigated for gender, racial, and geriatric subgroups investigated (if applicable).	X			
4	Data sets in EDR are accessible and do they conform to applicable guidances (e.g., existence of define.pdf file for data sets).	X			

IS THE STATISTICAL SECTION OF THE APPLICATION FILEABLE? ____Yes____

If the NDA/BLA is not fileable from the statistical perspective, state the reasons and provide comments to be sent to the Applicant.

Please identify and list any potential review issues to be forwarded to the Applicant for the 74-day letter.

No comments to the Sponsor for the 74-day letter.

Content Parameter (possible review concerns for 74-day letter)	Yes	No	NA	Comment
Designs utilized are appropriate for the indications requested.	X			
Endpoints and methods of analysis are specified in the protocols/statistical analysis plans.	X			
Interim analyses (if present) were pre-specified in the protocol and appropriate adjustments in significance level made. DSMB meeting minutes and data are available.			X	
Appropriate references for novel statistical methodology (if present) are included.	X			
Safety data organized to permit analyses across clinical trials in the NDA/BLA.	X			
Investigation of effect of dropouts on statistical analyses as described by applicant appears adequate.	X			

File name: 5_Statistics Filing Checklist for a New NDA 204275

STATISTICS FILING CHECKLIST FOR A NEW NDA/BLA

David Hoberman, Ph.D.	9/5/2012
Reviewing Statistician	Date
Kiya Hamilton, Ph.D.	9/5/2012
Reviewing Statistician	Date
Joan Buenconsejo, Ph.D.	9/5/2012
Supervisor/Team Leader	Date

File name: 5_Statistics Filing Checklist for a New NDA 204275

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

/s/

KIYA HAMILTON
09/05/2012

JOAN K BUENCONSEJO
09/05/2012
I concur

4/10/12

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

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

IND:	**74696**
Drug Name:	**GW642444 inhalation powder (vitalterol)**
Indication:	**Maintenance treatment of chronic obstructive pulmonary disease and asthma**
Applicants:	**GlaxoSmithKline (GSK)** **7333 Mississauga Road North** **Mississauga, Ontario Canada** **CROs:** (b) (4) [redacted]
Date(s):	**To Reviewer: 7 November 2011** **Completed: 2 March 2012** **Revised: 10 April 2012**
Review Priority:	**Standard**
Biometrics Division:	**Division 6**
Statistical Reviewer:	**Steve Thomson**
Concurring Reviewers:	**Karl Lin, Ph.D.**
Medical Division:	**Pulmonary, Allergy and Rheumatology Products**
Toxicologist Team:	**Asoke Mukherjee, Ph.D** **Molly Topper, Ph.D.**
Project Manager:	**Angela Ramsey, Ph.D.**

Keywords: Carcinogenicity, Cox regresson, Kaplan-Meier product limit, Survival analysis, Trend test, Bayesian

Table of Contents

1. EXECUTIVE SUMMARY

Reports from two studies, in rats and mice, were provided. The rat study was conducted [redacted]. The mouse study was conducted [redacted] The **rat report states that:** "The objective of this study was to investigate the carcinogenic potential of GW642444 following daily nose-only **inhalation administration to rats for a minimum of 104 consecutive weeks."** (page 21 of rat report) **According to the mouse report:** "The objective of this study was to determine the possible effects of GW642444M on the incidence and morphology of tumors in a 104-week inhalation carcinogenicity **study in mice." (page 43 of mouse report)**

1.1. Conclusions and Recommendations

The product is described as compound "GW642444M is the triphenyl acetate salt of GW642444, a beta-2 agonist. All doses and concentrations (including analyte concentration in plasma and aerosol concentration) are expressed in terms of the parent compound, which for the purpose of this report is referred **to as GW642444." (page 21 of rat** report). The same applies to the mouse study. In both species, dosing was accomplished by placing the animal into an inhalation chamber with aerosoled GW642444 for up to one hour daily.

Gross aspects of the basic study designs for the main study animals are summarized below:

Table 1. Design of Rat Study (dosed at µg/kg/day)

Treatment Group	# Animals	Nominal Dosage (Males)	Estimated Dosage (Males)	Nominal Dosage (Females)	Estimated Dosage (Females)
1. Control	60	0	0	0	0
2. Low	60	10	10.5	10 /3[1]	10.5 /3.47[1]
3. Low-Mid	60	80	84.4	80 /25[1]	84.4/ 28.2[1]
4. High-Mid	60	220	223	220[2]	223[2]
5. High	60	659	657	650[2]	657[2]

[1] From Week 86 dose was reduced from the first value to the second.

[2] Dosing stopped at week 85 due to increased mortality.

Table 2. Design of Mouse Study (dosed at µg/kg/day)

Treatment Group	# Animals	Nominal Dosage	Estimated Dosage
1. Control	84	0	0
2. Low-Low	84	6	6.4
3. High-Low	84	60	62
4. Low-Mid	84	600	615
5. High-Mid	84	6000	6150
6. High	84	30000	29500

More detailed descriptions of the studies are provided in Sections 3.2.1 and 3.2.2 below. Simple summary life tables are presented in these sections of the report.

In Appendix 1, Figures A.1.1 and A.1.2 for rats, and Figures A.1.3 and A.1.4 for mice, display Kaplan-Meier estimated survival curves for each study group for each species and gender combination. The results of the tests of trend and differences in survival are displayed in Tables 3 and 4 below:

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

Hypothesis Tested	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Rat Homogeneity over Groups 1-5	0.1752	0.0770	< 0.0001	< 0.0001
Homogeneity over Groups 1-3	0.3221	0.5885	0.0001	< 0.0001
No trend over Groups 1-5	0.1169	0.0224	0.0001	0.0005
No trend over Groups 1-3	0.2256	0.4479	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 5	0.6075	0.1946	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 3	0.1377	0.3315	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 2	0.3196	0.4314	0.1220	0.0276

From Figure A.1.1 in Appendix 1, in male rats, for some time after day 400 there appears to be a clear separation of the rather intertwined High and High-mid dose groups (i.e. groups 4 and 5) from the remaining, also rather intertwined, dose groups. However, by the end of the study, survival is close in all dose groups. This is consistent with the results above, i.e., the decrease in survival between between groups 1, 2, and 3 versus groups 4 and 5, in the middle part of the study, with less difference at the end resulted in an equivocal test of trend (Logrank $p = 0.1169$, Wilcoxon $p = 0.0224$) and the test of overall homogeneity (Logrank $p = 0.1752$, Wilcoxon $p = 0.0770$). The three tests in groups 1-3 were primarily to match the same and rather more relevant tests in females, but none were statistically significant (all six Logrank and Wilcoxon $p \geq 0.1377$).

Results in female rats are quite different. In female rats, from Figure A.1.2, the Control dose group has the highest survival, with the Low dose group next, with the remaining dose groups largely intertwined. This is sufficient to result in the generally highly significant tests comparing survival in female rats. That is, the test of overall homogeneity and a comparison between the High dose and Control were all highly statistically significant (i..e., both Logrank and Wilcoxon $p < 0.0001$), with a highly statistically significant test of trend (Logrank $p = 0.0001$, Wilcoxon $p = 0.0005$). Dosing was stopped in Week 85 in the High-mid and High dose groups (i.e., Groups 4 and 5). Hence dosing in these groups is no longer proportional to the dosing in the remaining dose groups. And thus there may be some interest in the results of dropping these dose groups and thus restricting attention to dose groups 1-3 (i.e., Control thru Low-mid dose). The overall test of homogeneity among Groups 1-3 was also highly statistically significant (i.e., Logrank $p = 0.0001$ and Wilcoxon $p < 0.0001$), with a highly statistically significant test of trend (both Logrank p and Wilcoxon $p < 0.0001$). The comparisons between

Control and both the High dose (group 5) and the Low-mid group (group 3) were all highly statistically significant (i.e., all four Logrank and Wilcoxon $p < 0.0001$). Finally, in female rats, there may be some interest in comparing differences in survival between the Low dose and Control. These were equivocal (Logrank $p = 0.1220$ and Wilcoxon $p = 0.0276$) due to the large number of censored survival times and a different pattern of survival from Day 400 to Day 650, but with closer patterns of survival at the end of the study.

Statistical significance levels of tests on survival in mice are summarized as follows:

Table 4. Statistical Significances of Tests of Homogeneity and Trend in Survival in the Mouse Study

Hypothesis Tested	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Mice Homogeneity over Groups 1-6	0.7455	0.5696	0.3945	0.4618
No trend over Groups 1-6	0.4520	0.1321	0.8634	0.3034
No Difference Between Groups 1 vs 6	0.9896	0.4417	0.6140	0.8865

From Figure A.1.3, in male mice, survival curves were all fairly closely intertwined, although during weeks 40-90 the High dose group generally had the lowest survival rate. These slight differences were not sufficient to result in any statistically significant tests of overall homogeneity, trend, or pairwise differences between the High dose and Control (all six $p \geq 0.1321$). In female mice, from Figure A.1.4, it seems that the Low-low dose group tended to have the highest survival, with the other groups largely intertwined. As with male mice, these slight differences were not sufficient to result in any statistically significant tests of overall homogeneity, trend, or pairwise differences between the High dose and Control (all six $p \geq 0.3034$).

A Bayesian analysis of survival utilizing a piecewise proportional hazards model is presented in Appendix 2. This analysis is consistent with that above in that in male rats there is some evidence of a dose related trend, but no strong evidence of a treatment related difference from vehicle. In female rats there is strong evidence of a dose related trend in survival, with strong evidence of a pairwise differerence between each of the Low-mid, High-mid, and High dose groups with the Control group. In both genders in mice there is no strong evidence of any particular dose related trend over treatment groups, or pairwise differences between the vehicle and the other treatment groups.

Of course in a carcinogenicity study, primary interest is on the occurrence of cancers. Statistical analysis compares tumor incidence over dose groups. Tables 5 – 7, below, display those organ tumor combinations that had at least one test of trend or pairwise difference from Control that was statistically significant at the usual 0.05 level. For each species by gender by organ the number of animals analyzed and used in the statistical tests is presented first. The tumor incidence for each organ is presented next, with the significance levels of the tests of trend, and the results of pairwise tests between each dose group and Control. These statistical tests are conditional upon the animals actually evaluated, ignoring those not analyzed.

Complete tumor incidence tables for each organ are provided in Tables A.3.4 through A.3.7 in Appendix 3.

To adjust for the multiplicity of tests the so-called Haseman-Lin-Rahman rules discussed in Section 1.3.1.5, below, are often applied. That is, when testing for trend over dose and the difference between the highest dose group with a control group, to control the overall Type I error rate to roughly 10% for a standard two species, two sex study, one compares the unadjusted significance level of the trend test to 0.005 for common tumors (incidence > 1%) and 0.025 for rare tumors, and the pairwise test to 0.01 for common tumors and 0.05 for rare tumors. As also discussed in section 1.3.1.4, using these adjustments for other tests, like the trend over the Control, Low, and Low-mid dose groups in female rats, or for other pairwise comparisons than the simple comparison between the high dose and the control dose, can be expected to increase the overall type I error rate to some value above the nominal rough 10% level, possibly considerably higher than the nominal 10% rate. **The period '.' in these tables denotes the p-values of tests of dose groups with no tumors in any group. In rats, the treatment groups 1-5 are denoted by "VC" for vehicle control, "Low", "LM" for Low-mid, "HM" for High-mid, and "Hi"** for High dose, respectively.

Table 5. Potentially Statistically Significant Neoplasms in Male Rats

Organ/ Tumor	Incidence VC	Low	LM	HM	Hi	Significance Level trend	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
SUBCUTANEOUS TISSUE										
# Evaluated	13	8	10	13	7					
Fibroma	7	2	8	5	7	.0292	.9866	.3328	.8701	.1765

Note that since the vehicle control incidence is greater than 1% fibromas would be classified as common tumors. After applying the Haseman-Lin adjustment for multiplicity the test of trend in fibromas is not statistically significant (p = 0.0292 > 0.005).

Table 6. Potentially Statistically Significant Neoplasms in Female Rats

Organ/ Tumor	Incidence VC	Low	LM	HM	Hi	Significance Level trend	trend 1-3	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
LIVER											
# Evaluated	60	60	60	60	60						
Adenoma: hepatocellular	0	0	1	0	2	.0327	.2544	.	.3816	.	.1425
MESOVARIAN LIGAMENT											
# Evaluated	60	60	60	60	60						
Leiomyoma	0	0	5	4	4	.0220	.0009	.	.0072	.0203	.0203
PITUITARY											
# Evaluated	60	60	60	60	60						
Adenoma/Carc. pars dist.	54	55	54	57	60	.0131	.2309	.1157	.2413	.1072	.0274
Adenoma: pars distalis	44	47	48	51	53	.0137	.1387	.1222	.1138	.0306	.0119
THYROID											
# Evaluated	60	60	60	60	60						
Adenoma/Carcinoma C-Cell	4	2	3	2	7	.0142	.3654	.8355	.5323	.7522	.0745
Adenoma: C-cell	2	2	2	2	7	.0029	.3511	.5989	.4856	.4998	.0163

From the vehicle control, for the listed tumors, tumor incidence in the liver and mesovarian ligament tumors in female rats would be classified as rare, while the remaining tumors above would be catergorized as common. Again, after applying the Haseman-Lin adjustment for multiplicity the test of trend in hepatocellular adenomas of the liver is close to statistical significance ($p = 0.0327 \approx 0.025$). For leiomyoma in the mesovarian ligament the tests of trend over all five treatment groups and over groups 1-3 are both statistically significant ($p = 0.0220, 0.0009$, both < 0.025). The pairwise tests between the Low-mid, High-mid, and High dose groups and the Control in this leiomyoma were also all statistically significant ($p = 0.0072, 0.0203, 0.0203 < 0.05$, respectively). Note however that including these tests between the Low-mid and High-mid dose groups with the Control can be expected inflate the significance level to some value above the nominal approximate 10% level. The test of trend over all five dose groups in C-cell adenoma of the thyroid was statistically significant ($p = 0.0029 < 0.005$) while the comparison between the High dose and Control was fairly close to statistical significance ($p = 0.0163 \approx 0.01$). The comparison between the High dose group and Control in pars distalis adenoma of the pituitary was close to statistical significance ($p = 0.0119 \approx 0.01$). No other comparison achieved the multiplicity adjusted level of significance, even when including the pairwise comparisons of the non-high dose groups with control.

Table 7 provides similar results in mice. In this table, mouse treatment groups 1-6 are **labeled "Veh" or "VC" for vehicle control, "LL" for Low-low, "HL" for High-Low, "LM" for Low-Mid, "HM" for High-mid, and, again "Hi'** for High dose, respectively. Again, the organ tumor combinations in mice that involved at least one test that was nominally statistically significant at a 0.05 level are reproduced below:

Table 7 Potentially Statistically Significant Neoplasms in Mice

	Incidence						Significance Level					
Organ/ Tumor	Veh	LL	HL	LM	HM	Hi	trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
Male Mice												
GALLBLADDER												
TESTES												
# Evaluated	84	84	84	84	84	84						
BENIGN INTERSTITIAL CELL	0	0	0	0	0	2	.0275	.	.	.	.	.2485
Female Mice												
OVARIES												
# Evaluated	84	83	84	84	84	83						
TUBULOSTROMAL ADENOMA	0	0	1	0	2	6	.0001	.	.5000	.	.2485	.0137
UTERUS W/ CERVIX												
# Evaluated	84	84	84	84	84	84						
LEIOMYOSARCOMA	0	1	2	4	6	4	.1134	.5000	.2485	.0603	.0142	.0603
Leiomyoma/Leiomysarcoma	2	3	7	9	7	6	.3547	.5000	.0839	.0284	.0839	.1385

Adjusting for multiplicity, the test of trend in benign interstitial cell tumor of the testes (in male mice) would be arguably statistically significant ($p = 0.0275 \approx 0.025$). In tubulostromal adenoma of the ovaries in female mice both the test of trend and the pairwise comparison to

Control were highly statistically significant ($p = 0.0001 < 0.025$, $p = 0.0137 < 0.05$, respectively). Accepting the inflation in overall Type I error from using the other pairwise comparisons, the comparison with the High-mid group in leiomysarcoma of the uterus would be statistically significant ($p = 0.0142 < 0.05$). No other tests achieved the multiplicity adjusted nominal rough 10% level.

Complete incidence tables in male and female mice are presented in Tables A.3.6 and A.3.7 in Appendix 3.

1.2. Brief Overview of the Studies

This submission had a rat study:

Glaxo Study 79808: GW642444M: Inhalation Carcinogenicity Study in Rats,

and a similar, mouse study:

(b) (4) **Study 07-6304: GW642444M: Inhalation Carcinogenicity Study in Mice.**

1.3. Statistical Issues and Findings

1.3.1. Statistical Issues

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

1.3.1.1. Multiple Housing and Dosing of Animals:

In the rat study animals were housed 2-3 per cage while mice were housed in groups of 5, at least initially. Social interaction is important for the welfare of the animals. However, an argument could be made that housing animals together should be treated as part of the treatment of the experiment, and thus the appropriate unit of analysis would be the cage of the animals, not the individual animal. This would reduce the study sample size. Carcinogenicity tendencies that could be communicated across animals, competition for food, fighting, or other within cage effects could cause positive or negative correlations in response. Several animals were dosed together in a chamber, apparently from the same group and possibly the same cage. Variations in dosing across occasions might also induce positive or negative correlations. Thus, it is possible that within treatment estimated variances may be too large or too small, resulting in conservative or anti-conservative tests (in terms of Type I error). Unless it has been clearly shown that tumor incidence is independent of cage or group in chamber, from a purely statistical analysis point of view, this reviewer would generally recommend single housing and dosing of animals. However, with this method of dosing, it would seem that the effect of multiple housing would likely be rather small, as opposed, say, to dietary dosing.

Dose values in the FDA analysis are based on the Sponsor supplied estimated average dose value in the treatment groups. It is expected that these values will be close to the actual dose experienced by each animal. Apparently the Sponsor's statistical analysis used the nominal target dose values. This can be expected to usually be further from the actual dose and thus to add a bias to results, but since the difference between the estimated dose and the nominal dose is small this bias should be small.

1.3.1.2. Survival Analysis:

The survival analyses presented here are based on both the log rank test and the Wilcoxon test comparing survival curves. Log rank tests tend to put higher weight on later events, while the Wilcoxon test tends to weight events more equally, and thus is more sensitive to earlier differences in survival. The logrank test is most powerful when the survival curves track each other, and thus the hazards, i.e., the conditional probability of the event in the next infinitesimal interval, would be roughly proportional. Both of these tests were used to test both homogeneity of survival among the treatment groups and the effect of dose on trend in survival. Appendix 1 reviews the specific animal survival analyses in more detail. The log rank test for mortality is the test reported by the Sponsor in the mouse study. In the rat study the Sponsor reports results using Tarone's test, which weights events between the weights used in the log rank and Wilcoxon tests. The results of the Sponsor's analysis are summarized in Sections 3.2.1.1 and 3.2.2.1. An experimental Bayesian analysis of survival is given in Appendix 2.

1.3.1.3. Multiplicity of Tests on Survival:

Using the logrank and Wilcoxon tests, for each gender in rats there are 14 tests of survival differences. In mice there are six similar tests of survival in each gender. If we were to assume that any set of tests are independent across comparisons, which clearly they are not, and assume that there is absolutely no difference in survival, the probability of at least one statistically significant result in each gender, at the usual 0.05 level, is about 0.512 in rats. In mice the probability of at least one statistically significant result in each gender, at the usual 0.05 level, is about 0.265, and about 0.46 of at least one statistically significant result in at least one gender. Such is the possible price paid for the multiplicity of hypothesis tests in the frequentist paradigm.

1.3.1.4. Tests on Neoplasms:

The Sponsor uses Peto type analysis of neoplasms. The analyses in the FDA report are based on poly-k analysis of tumor incidence. The poly-k test is a modification of the original Cochran-Armitage test of trend in response to dose, adjusted for differences in mortality (please see Bailer & Portier, 1988, Bieler & Williams, 1993). It was noted in the report of the Society of Toxicological Pathology "town hall" meeting in June 2001 that the poly-k modification of the Cochran-Armitage tests of trend has been recommended over the corresponding Peto tests.

1.3.1.5. Multiplicity of Tests on Neoplasms:

Frequentist hypothesis testing involves accepting or rejecting hypotheses about the parameters of interest on the basis of the values of some statistic. If one does not provide some sort of multiplicity adjustment to the significance level, the chances of rejecting one or more

true null hypothesis increases as the number of such tests increases. To avoid this it is common to adjust for multiplicity in hypothesis testing resulting in an adjustment in experiment-wise Type I error (i.e., the probability of rejecting a true null hypothesis). Based on his extensive experience with such carcinogenicity analyses in standard laboratory rodents, for pairwise tests between the high dose groups and control in two species, Haseman (1983) claimed that for a roughly 0.10 (10%) overall false positive error rate, rare tumors should be tested at a 0.05 (5%) level, and common tumors (with a historical control incidence greater than 1%) at a 0.01 level. Similarly, Lin and Rahman (1998) showed that tests of trend should be tested at a 0.025 level for rare tumors and 0.005 for common tumors. This approach is intended to balance both Type I error and Type II error (i.e., the error of concluding there is no evidence of a relation to tumorgenicity when there actually is such a relation).

Significance levels of the pairwise tests between the Control group and the non-high dose groups are also provided in the FDA analysis. However, including these tests can be expected to increase the overall type I error rate to some level above the rough 10% level. Even following the Haseman-Lin-Rahman rules above, the overall type I error associated with including these tests (plus possibly others) may be considerably larger than the rough 10% appropriate when these rules are restricted to the test of trend and pairwise difference between the high dose and the vehicle.

1.3.1.6. Validity of the Designs:

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

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

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

Lin and Ali (2006), quoting work by Haseman, have suggested that in standard laboratory rodent species, a survival rate of about 25 animals out of 50 or more animals, between weeks 80-90 of a two-year study may be considered a sufficient number of survivors as well as one measure of adequate exposure. Note that as a percentage of animals that survived to week 91, this criterion is met or exceeded in the High dose group in either gender in either species. (Please see Tables 16 and 17 on page 18, and Tables 23 and 24 on page 25). Like the other comments in this section, this requires the expertise of the toxicologist, but may suggest that the MTD was met or not achieved, but not exceeded.

The mean weight values and derived differences and ratios in the following table were **taken directly from the Sponsor's reports (Rat** Tables 7 and 8, pages 167-194, Mouse Tables 10 and 11, pages 323-355). The change from baseline in the table below is the simple difference between the means at the specified dates, and thus animals that die are only counted at the study initiation, not at the end of the study.

Chu, Ceuto, and Ward (1981), citing earlier work by Sontag *et al* (1976) recommend that the MTD "is taken as 'the highest dose that causes no more than a 10% weight decrement as compared to the appropriate control groups, and does not produce mortality, clinical signs of toxicity, or pathologic lesions (other than those that may be related to a neoplastic response) that would be predicted to shorten the animal's natural life span' " From Tables 7 and 8 below, the weight decrement criterion was met in rats. In female mice there seems to be an actual weight gain in females in the treated dose groups. However, the criterion is exceded in male mice. Again, although this requires the expertise of the toxicologist, this may be evidence that the MTD was not exceeded in rats and female mice, but have been exceded in male mice.

Table 7. Mean Weights and Changes (in g) in Male Rats

Dose Group	Est. Dose µg/kg/day	Week		Change from baseline	% change relative to Control
		-1	99		
1. Control	0	203.3	729.3	526.0	
2. Low	10.5	202.6	737.5	535.2	101.7%
3. Low-Mid	84.4	199.5	690.1	490.6	99.3%
4. High-Mid	223	193.3	706.1	512.8	97.5%
5. High	6570	191.9	664.0	472.1	89.8%

Table 8. Mean Weights and Changes (in g) in Female Rats

Dose Group	Est. Dose µg/kg/day	Week		Change from baseline	% change relative to Control	Week	Change from baseline	% change relative to Control
		-1	87			91		
1. Control	0	158.2	459.8	301.6		462.5	304.3	
2. Low	10.5/3.47[1]	158.0	462.4	304.4	100.9%	469.1	311.1	102.2%
3. Low-Mid	84.4/28.2[1]	154.1	448.7	294.6	97.7%	420.0	265.9	87.4%
4. High-Mid	223[2]	153.4	469.7	316.3	104.9%	483.5	330.1	108.5%
5. High	6570[2]	154.4	449.0	294.6	97.7%	433.3	278.9	91.7%

[1] From Week 86 dose was reduced from the first value to the second.

[2] Dosing in groups 4 and 5 was stopped at week 85 due to increased mortality.

Table 9. Mean Weights and Changes (in g) in Mice

Dose Group	Est. Dose µg/kg/ day	Males				Females			
		Week		Change from baseline	% change relative to Control	Week		Change from baseline	% change relative to Control
		-1	101			-1	104		
1. Control	0	29.7	37.8	8.1		22.3	29.3	7.0	
2. Low-Low	6.4	30.0	37.3	7.3	90.1%	22.2	31.3	9.1	130.0%
3. High-Low	62	29.4	36.9	7.5	92.6%	21.3	31.1	9.8	140.0%
4. Low-Mid	615	30.4	37.2	6.8	84.0%	22.9	30.9	8.0	114.3%
5. High-Mid	6150	30.7	37.6	6.9	85.2%	22.8	29.6	6.8	97.1%
6. High	29500	30.3	36.2	5.9	72.8%	22.1	29.8	7.7	110.0%

The Sponsor summarizes food consumption during the rat study as follows: "Increases in overall food consumption were observed in both sexes given ≥ 223 µg/kg/day during the first 8

weeks of treatment (up to 1.12X control). This initial mild increase continued on through to Week 85 in males given ≥ 223 µg/kg/day (up to 1.05X control), whilst in females, mild increases in food consumption were seen in all treated groups between weeks 8 and 85 (up to 1.13X control)." (page 61 of rat report)

In mice: "During the first half of the study in males and for slightly longer in females, higher than control group mean food intake was apparent in all treated groups, but there was no evidence of dose relationship. The effect became less obvious as the study progressed and by the end of the study there was no clear difference between control and test article treated animals." (page 64 of mouse report)

Again from 2) above, excess mortality not associated with any tumor or sacrifice in the higher dose groups might suggest that the MTD was exceeded. This suggests that a useful way to assess whether or not the MTD was achieved is to measure early mortality not associated with any identified tumor. If this is high in the higher dose groups, it suggests that animals tend to die before having time to develop tumors. Tables 10 and 11, below, displays the number of animals in each dose group that died of a natural death or moribund sacrifice, but did not show any tumors (i.e., the "Event"):

Table 10. Natural Death with No Identified Tumor in Rats (Male/Female)

	1. Vehicle	2.Low	3. Low-mid	4.High-mid	5. High
Males Event	3	5	6	5	3
No event	57	55	54	55	57
Females Event	0	0	1	0	0
No event	60	60	59	60	60

So in both genders there is no particular evidence of an early death not associated with any neoplasm.

Table 11. Natural Death with No Identified Tumor in Mice (Male/Female)

	1. Vehicle	2.Low-Low	3. High-Low	4.Low-Mid	5. High-Mid	6. High
Males Event	39	35	45	39	39	46
No event	45	49	39	45	45	38
Females Event	38	37	37	41	45	46
No event	46	47	47	43	39	38

In mice there are a large number of such early deaths, but it seems to be unrelated to dose. This is clearly apparent in the table, but if one needed a statistical test, chi-square tests of no differences across treatments give large p-values (Males $p = 0.5249$, Females $p = 0.5673$). Strictly speaking, such tests do not show there is no dose related effect, only that there is no strong evidence of such an effect. Still this is evidence of no treatment differences in early deaths, and suggests that early deaths are not associated with the experimental process, especially the method of dose administration.

Again, the actual determination of whether the MTD was achieved or exceded requires the expertise of the toxicologist.

1.3.2. Statistical Findings

Please see Section 1.1 above.

2. INTRODUCTION

2.1. Overview

This submission summarizes the results of two year rat and mouse inhalation studies to assess the carcinogenic potential carcinogenic potential of aerosoled compound GW642444 when dosed daily for about 104 consecutive weeks. The rat study was conducted (b) (4) . The mouse study was conducted (b) (4)

2.2. Data Sources

The Sponsor provided two SAS transport files, both labeled tumor.xpt, each containing a SAS tumor data set named tumor.sas7bdat.

3. STATISTICAL EVALUATION

3.1. Evaluation of Efficacy

NA

3.2. Evaluation of Safety

3.2.1. Glaxo Study 79808: GW642444M: Inhalation Carcinogenicity Study in Rats,

STUDY DURATION: 104 Weeks (planned)
EXPERIMENTAL (DOSING) START DATE: 29 January 2008
FINAL DOSING DATE: 31 January 2010
TREATMENT DURATION: Males 99 to 104 weeks depending upon dose group
Females 97 to 104 weeks depending upon dose group
RAT STRAIN: Sprague Dawley Crl:CD(SD) Rats
ROUTE: Daily nose-only inhalation

Study conduct was summarized as follows: "GW642444 was given to rats (60/sex/group) at estimated achieved doses of 0, 10.5, 84.4, 223 and 657 µg/kg/day for 60 minutes once daily for 85 weeks by nose-only inhalation. Due to increased mortality, dosing was stopped for females given 223 and 657 µg/kg/day at Week 85 (26 and 23 animals surviving in these groups, respectively). These females remained on study without further treatment until group survival fell to 15 (Weeks 95 or 96 respectively) at which time they were electively killed. From Week 86, the doses for the remaining females were reduced to 3.47 (from 10.5) and 28.2 (from 84.4) µg/kg/day by decreasing the daily exposure duration from 60 to 20 minutes for the remainder of the study. Females at 84.4/28.2

µg/kg/day were terminated in Week 95 due to survival reaching 15. Control females and females given 10.5/3.47 µg/kg/day were killed in Week 104. All males were electively killed in Week 101 when the number of survivors in the control group reached less than 20." (page 17 of rat report)

The basic study design is summarized in tabular form in Table 12, below, actually a repeat of Table 1:

Table 12. Design of Rat Study (dosed at µg/kg/day)

Treatment Group	# Animals	Nominal Dosage (Males)	Estimated Dosage (Males)	Nominal Dosage (Females)	Estimated Dosage (Females)
1. Control	60	0	0	0	0
2. Low	60	10	10.5	10 /3[1]	10.5 /3.47[1]
3. Low-Mid	60	80	84.4	80 /25[1]	84.4/ 28.2[1]
4. High-Mid	60	220	223	220[2]	223[2]
5. High	60	659	657	650[2]	657[2]

[1] From Week 86 dose was reduced from the first value to the second.
[2] Dosing stopped at week 85 due to increased mortality.

The Sponsor indicates that an additional 9 animals/sex/group were included within each dosing level for toxicokinetic evaluation in Weeks 4 and 26.

"The lactose vehicle control group [1] was exposed to lactose only. The concentration of lactose in the aerosol given to this group was targeted to be the same as the concentration of lactose given to Group 5 with its dosing duration handling being the same for test article groups. The dosing duration of females in the control was reduced in order to provide target for Groups 2 and 3 after Day 595. The lactose concentration in aerosols was determined by both gravimetric and chemical methods." (page 43 of rat report)

Further, "GW642444 . . . was administered as a dry power formulation blended in lactose at a nominal concentration of 0.4% w/w (used to dose animals given 10.5/3.47 or 84.4/28.2 µg/kg/day) or 4% w/w (used to dose animals given 223 or 657 µg/kg/day).

"The inhalation exposure system consisted of nose-only flow through inhalation chambers. The animals were restrained in (b) (4) restraint tubes which were inserted onto the chambers. The test atmospheres were generated into the top section of the inhalation chambers (b) (4)

"The various test article concentrations were achieved by altering the rate of test article introduction into the chamber and/or the exposure duration." (page 21 of rat report)

Dosing levels were justified as follows: "The target dose of 650 µg/kg/day was based on unacceptable respiratory tract irritancy seen at higher doses (10000 µg/kg/day and above) during

a previous 13-week inhaled study. Therefore, 650 µg/kg/day was considered to be a suitable high dose for 2 years of dosing." (page 22 of rat report)

Animals were housed 2-3 together with water and apparently food available ad libitum. Although it is probably kinder to the animals, as discussed in Section 1.3.1.1, multiple housing may cause problems with the analysis of study results, particularly with other forms of dosing.

3.2.1.1. Sponsor's Results and Conclusions

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

Sponsor's Survival analysis:

The Sponsor provided the following summary of survival at the end of the 104-week treatment period:

"Preterminal deaths occurred in approximately two-thirds of the rats in the control group or treated groups. The total group mortality of animals dying or killed during the study was as follows:" (page 55 of report)

The following table was copied from the Sponsor's report:

Table 13 (Sponsor Un-numbered Table on page 55 of rat report) Summary of Mortality

Sex	Males					Females				
Estimated Achieved Dose (µg/kg/day)	0	10.5	84.4	223	657	0	10.5/ 3.47	84.4/ 28.2	223	657
Mortality	42/60	36/60	32/60	43/60	40/60	37/60	42/60	45/60	45/60	45/60
%Survival (at terminal kill)	30	40	47	28	33	40	33*	25**	25**	25**

Statistical significance: *= P ≤0.05 ** = P ≤0.001 (Peto) (presumably Tarone test, not Peto trest)

"Overall, a statistically significant increase in mortality was noted in female animals from all treated groups (P≤0.05 at 10.5/3.47 µg/kg/day and P≤0.001 at ≥84.4/28.2 µg/kg/day).

"When the distribution of mortality over time was assessed, a dose related increase in mortality was apparent in males given ≥223 µg/kg/day GW642444 by Week 65 (i.e. between Week 1 and the end of Week 64) and in females given ≥84.4/28.2 µg/kg/day by Week 52 of the study. This trend to increased mortality, particularly in treated females, continued in these groups as the study progressed and, as a consequence, in Week 85 treatment was withdrawn for females in the 223 and 657 µg/kg/day dose groups and the dose for the two lower dose group females (initial dose of 10.5 or 84.4 µg/kg/day) was reduced to 3.47 or 28.2 µg/kg/day as of Week 86, respectively. Surviving females from the 84.4/28.2 µg/kg/day dose group were eventually electively killed during Week 95 of the study when survival reached 15 animals in the group. Control females and females at 10.5/3.47 µg/kg/day were killed in Week 104. All males were killed in Week 101 when

the number of survivors in the control reached less than 20." (page 56 of rat report)

Note that the data set used in the FDA analysis indicates that one more male animal in dose group 4 and one more female in group 2 were counted in the terminal sacrifice than is indicated by the dose group totals above.

Sponsor's Tumorigenicity analysis:

Under the heading **"Test article Related Neoplastic Lesions / Pituitary Tumors"** on page 64 the Sponsor states the following:

"Application of the Peto one-sided test for pairwise group comparisons indicated a statistically significant increase in pituitary adenomas but not carcinomas when compared to the control group for female animals given ≥ 84.4/28.2 µg/kg/day ($p \leq 0.01$) but no effect on those given 10.5/3.47 µg/kg/day:" The following table was copied from the Sponsor's report:

Table 14 (Sponsor Un-numbered Table on page 64 of rat report)

	Males					Females				
Group number	1	2	3	4	5	1	2	3	4	5
Inhaled Dose µg/kg/day	0	10.5	84.4	223	657	0	10.5/3.47	84.4/28.2	223	657
No of animals examined	60	60	60	60	60	60	60	60	60	60
Adenoma - pars distalis	42	37	42	39	45	44	47	48$	51$	53$
Carcinoma - pars distalis	3	4	1	1	1	10	8	6	6	7
Total - pars distalis Tumors	45	41	43	40	46	54	55	54$	57$	60$

$ Statistically significant p≤0.01 for pairwise comparison, common tumor, using Peto's one side trend test following recommendations of Lin and Rahman

"Test article-related earlier mortality due to pars distalis tumors of the pituitary gland was seen in males given ≥ 223 µg/kg/day and females given ≥84.4/28.2 µg/kg/day. These effects were apparent by week 65 in males given ≥ 223 µg/kg/day and by Week 52 in females given ≥ 84.4/28.2 µg/kg/day. This was supported by the observation that Peto's survival-adjusted one tailed trend test indicated, a statistically significant increasing trend ($p = 0.0002$) for benign pituitary tumors in females." (page 65 of rat report).

"Mesovarian Ligaments

Test-article-related smooth muscle hyperplasia/hypertrophy and occasionally leiomyomata of the mesovarian ligaments were observed in animals given ≥84.4/28.2 µg/kg/day. The findings were present in decedent females and those surviving to terminal kill and were not present in control females or those given 10.5/3.47 µg /kg/day." (page 67 of rat report) The incidence of these changes is shown in the following table:

Table 15 (Sponsor Un-numbered Table on page 68 of rat report)

Sex	Female				
Dose (µg/kg/day)	0	10.5/3.47	84.4/28.2	223	657
No of animals in group	60	60	60	60	60
Leiomyomata	0	0	5$	4$	4
Smooth muscle hyperplasia/hypertrophy					
Minimal	0	0	1	6	6
Slight	0	0	1	1	6
Total	0	0	2	7	12

$ Statistically significant $p \leq 0.05$ for pairwise comparison, common tumor, using Peto's one side trend test following recommendations of Lin and Rahman 1998

The Sponsor claims that: **"There were no test article-related** increased incidences of any other neoplastic findings in **any tissues." (page 68 of rat report)** However the Sponsor notes further that **for thyroid glands there was "an increased incidence of thyroid C-cell adenomas in** female animals given 657 µg/kg/day was statistically significant by Peto's survival-adjusted one tailed trend test but not significant by pairwise group comparison. This observation was not considered to be test article-related in light of the similarity to background spontaneous incidence **levels in this species." (page 68 of rat report)**

3.2.1.2. FDA Reviewer's Results

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

Survival analysis:

The following tables (Table 16 for male rats, Table 17 for females) summarize the mortality results for the study groups. The data were grouped for the specified time period, and present the number of deaths during the time interval over the number at risk at the beginning of the interval. The percentage cited is the percent that survived at the end of the interval. In these tables the terminal period only includes those animals were sacrificed. Animals that died of other causes during the terminal period are included in the preceding, but overlapping time period. The Kaplan-Meier survival plots in Appendix 1 provide a more detailed picture of the profile of mortality losses.

Table 16. Summary of Male Rats Survival (estimated dosed in µg/kg/day)

Period (Weeks)	Control 0	Low 10.5	Low-Mid 84.4	High-Mid 223	High 657
1-52	3/60[1] 95%[2]	0/60 100%	4/60 93.3%	4/60 93.3%	2/60 96.7%
53-78	7/57 83.3%	8/60 86.7%	8/56 80%	16/56 66.7%	20/58 63.3%
79-91	15/50 58.3%	14/52 63.3%	10/48 63.3%	9/40 51.7%	9/38 48.3%
92-101	17/35 30%	14/38 40%	10/38 46.7%	13/31 30%	9/29 33.3%
Terminal[3] 101	18	24	28	18	20

[1] number of deaths / number at risk
[2] overall per cent survival to end of period.
[3] number of animals that survived to terminal sacrifice

Table 17. Summary of Female Rats Survival (estimated dosed in µg/kg/day)

Period (Weeks)	Control 0	Low 0.5/3.47[4]	Low-Mid 84.4/28.2[4]	High-Mid 223/0[5]	High 657/0[5]
1-52	0/60[1] 100%[2]	4/60 93.3%	4/60 93.3%	5/60 91.7%	4/60 93.3%
53-78	7/60 88.3%	15/56 68.3%	22/56 56.7%	22/55 55%	24/56 53.3%
79-91	12/53 68.3%	14/41 45%	13/34 35%	16/33 28.3%	14/32 30%
92-104	18/41 38.3%	8/27 31.7%	6/21 25%	2/17 25%	3/18 25%
Terminal[3] 105	23	19	15	15	15

[1] number of deaths / number at risk
[2] overall per cent survival to end of period.
[3] number of animals that survived to terminal sacrifice
[4] from Week 86 dose was reduced from the first value to the second.
[5] dosing stopped at week 85 due to increased mortality.

Note that the terminal sacrifice totals in the High-Mid dose group in male rats and the Low dose group in females differ by one animal from those reported by the Sponsor slightly from those reported in the Sponsor reproduced in tables immediately above. This may be due to slightly different ways of aggregating groups.

Table 18 below provides the significance levels of the tests of homogeneity and trend over dose groups as proposed in Section 1.3.1.1, above.

Table 18. Statistical Significances of Tests of Homogeneity and Trend in Survival in Rats

Hypothesis Tested	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Rat Homogeneity over Groups 1-5	0.1752	0.0770	< 0.0001	< 0.0001
Homogeneity over Groups 1-3	0.3221	0.5885	0.0001	< 0.0001
No trend over Groups 1-5	0.1169	0.0224	0.0005	0.0006
No trend over Groups 1-3	0.2256	0.4479	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 5	0.6075	0.1946	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 3	0.1377	0.3315	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 2	0.3196	0.4314	0.1220	0.0276

From Figure A.1.1 in Appendix 1, in male rats, for some time after day 400 there appears to be a clear separation of the rather intertwined High and High-mid dose groups (i.e., groups 4 and 5) from the remaining, also rather intertwined, dose groups. However, by the end of the study, survival is close in all dose groups. This is consistent with the results above, i.e. the decrease in survival between between groups 1, 2, and 3 versus 4 and 5, in the middle part of the study, with less difference at the end resulted in an equivocal test of trend (Logrank p = 0.1169, Wilcoxon p = 0.0224) and test of overall homogeneity (Logrank p = 0.1752, Wilcoxon p = 0.0770). The comparisons in groups 1-3 were primarily to match the same, somewhat more relevant tests in females, but none were statistically significant (all six Logrank and Wilcoxon p $\geq$ 0.1377).

Results in female rats are quite different. In female rats from Figure A.1.2, in female rats the Control dose group has the highest survival, with the Low dose group next, and the remaining dose groups largely intertwined. This is sufficient to result in the generally highly significant tests comparing survival in female rats. That is, the test of overall homogeneity and a comparison between the High dose and Control were all highly statistically significant (i.e., both Logrank and Wilcoxon p < 0.0001), with a highly statistically significant test of trend (Logrank p = 0.0005, Wilcoxon p = 0.0006). Dosing was stopped in Week 85 in the High-mid and High dose groups (i.e., Groups 4 and 5). Hence dosing in these groups is no longer directly comparable to the dosing in the remaining dose groups. Thus there may be some interest in the results comparing dose groups 1-3. The overall test of homogeneity among Groups 1-3 was highly statistically significant (i.e., Logrank p = 0.0001, Wilcoxon p < 0.0001), with a highly statistically significant test of trend (both Logrank and Wilcoxon p < 0.0001), and comparison between Control and the Low-mid dose group (i.e., both Logrank and Wilcoxon p < 0.0001). There may be some interest in comparing differences in survival between the Low dose and Control. These were equivocal (Logrank p = 0.1220 and p = 0.0276) due to the large number of censored survival times and a different pattern of survival form Day 400 to Day 650, but closer survival patterns at the end of the study.

Results from a supporting experimental Bayesian nonparametric analysis of survival are provided in Appendix 2. This analysis is consistent with that above in that in male rats there is some evidence of a dose related trend, but no strong evidence of a treatment related difference from vehicle. In female rats there is strong evidence of a dose related trend in survival, with

strong evidence of a pairwise differerence between each of the Low-mid, High-mid, and High dose groups with the Control group.

Tumorigenicity analysis:

As discussed in Section 1.3.1.5, the Haseman-Lin-Rahman rules for adjusting for multiplicity in a two species study specify that for a very rough 0.10 (10%) overall false positive error rate, overall trend should be tested at a 0.05 (5%) level in rare tumors (background incidence 1% or less) and at 0.01 (1%) level in common tumors. The comparison between the High dose and Control should be tested at a 0.025 (2.5%) level in rare tumors (background incidence 1% or less) and at 0.005 (0.5%) level in common tumors. The following tables 19 and 20 below, are a repeat of tables 5 and 6, above, and display those organ-tumor combinations that were associated with at least one test with a unadjusted significance level of 0.05 or less. In this **table the treatment groups are denoted by "VC" for vehicle control, "Low", "LM" for Low-mid, "HM" for High-mid , "Hi" for High dose. In female rats dosing was stopped early in the two highest dose groups, so there may be interest in the test of trend over the the remaining three dose groups. The statistical significance level of this test is denoted "trend/1-3."**

Table 19. Potentially Statistically Significant Neoplasms in Male Rats

Organ/ Tumor	Incidence VC	Low	LM	HM	Hi	Significance Level trend	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
SUBCUTANEOUS TISSUE										
# Evaluated	13	8	10	13	7					
Fibroma	7	2	8	5	7	.0292	.9866	.3328	.8701	.1765

Note that since the vehicle control incidence is greater than 1% fibromas would be classified as common tumors. After applying the Haseman-Lin-Rahman adjustment for multiplicity the test of trend in fibromas is not statistically significant ($p = 0.0292 > 0.005$).

Table 20. Potentially Statistically Significant Neoplasms in Female Rats

Organ/ Tumor	Incidence VC	Low	LM	HM	Hi	Significance Level trend	trend 1-3	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
LIVER											
# Evaluated	60	60	60	60	60						
Adenoma: hepatocellular	0	0	1	0	2	.0327	.2544	.	.3816	.	.1425
MESOVARIAN LIGAMENT											
# Evaluated	60	60	60	60	60						
Leiomyoma	0	0	5	4	4	.0220	.0009	.	.0072	.0203	.0203
PITUITARY											
# Evaluated	60	60	60	60	60						
Adenoma/Carc. pars dist.	54	55	54	57	60	.0131	.2309	.1157	.2413	.1072	.0274
Adenoma: pars distalis	44	47	48	51	53	.0137	.1387	.1222	.1138	.0306	.0119
THYROID											
# Evaluated	60	60	60	60	60						
Adenoma/Carcinoma C-Cell	4	2	3	2	7	.0142	.3654	.8355	.5323	.7522	.0745
Adenoma: C-cell	2	2	2	2	7	.0029	.3511	.5989	.4856	.4998	.0163

From the vehicle control incidence the liver and mesovarian ligament tumors in female rats would be classified as rare, while the remaining tumors would be catergorized as common. Again, after applying the Haseman-Lin-Rahman adjustment for multiplicity the test of trend in hepatocellular adenomas of the liver is close to statistical significance ($p = 0.0327 \approx 0.025$). The tests of trend over all five treatment groups and over groups 1-3 are both statistically significant ($p = 0.0220, 0.0009$, both < 0.025). The pairwise tests between the Low-mid, High-mid, and High dose groups and the Control were all statistically significant ($p = 0.0072, 0.0203, 0.0203 < 0.05$, respectively). Note however that including the tests between the Low-mid and High-mid dose groups can be expected inflate the significance level to some value above the nominal approximate 10% level. The test of trend over all five dose groups in C-cell adenoma of the thyroid was statistically significant ($p = 0.0029 < 0.005$) while the comparison between the High dose and Control was close to statistical significance ($p = 0.0163 \approx 0.01$). The comparison between the High dose group and Control in pars distalis adenoma of the pituitary was close to statistical significance ($p = 0.0119 \approx 0.01$). No other comparison achieved the multiplicity adjusted level of significance, even when including the pairwise comparisons of the non-high dose groups with control.

Complete incidence table in male rats and female rats are presented in Tables A.3.4 and A.3.5 in Appendix 3.

3.2.2. (b) (4) Study 07-6304: GW642444M: Inhalation Carcinogenicity Study in Mice.

STUDY DURATION: 104 Weeks (planned)
EXPERIMENTAL START DATE (INITIATING DOSING): 13 December 2007
END OF TREATMENT: Males 15 November - 6 December 2009 (Days 704-724)
Females Group 5 2- 6 December 2009 (Days 721-725)
1-4, 6 13-16 December 2009 (Days 732-735)
MOUSE STRAIN: (b) (4) Crl:CD-1® Mice
ROUTE: Daily nose only inhalation for 60 minutes

The Sponsor summarized study conduct as follows: " GW642444M was given to CD-1 Mice at estimated achieved doses of 0 (vehicle), 6.4, 62, 615, 6150 or 29500 µg/kg/day [once] daily for 60 minutes by nose-only inhalation for 101 weeks to males and 103 or 104 weeks to females. Sixty mice/sex were originally assigned to each group in the main study. An additional 66 animals/sex were added at each dose level for toxicokinetic evaluation. Due to the high mortality that occurred across all groups during the early stages of the study, when compared with historical control data, 24 animals/sex/group previously designated as toxicokinetic animals were reassigned to main study; these animals had not previously been subject to any blood sampling. All data related to these animals have been combined with the main study animals and is reported together. The total group **size was therefore 84 animals/sex/group."** (page 13 of mouse report)

The study design is summarized as follows (actually a repeat of Table 2):

Table 21. Design of Mouse Study (dosed at µg/kg/day)

Treatment Group	# Animals	Nominal Dosage	Estimated Dosage
1. Control	84	0	0
2. Low-Low	84	6	6.4
3. High-Low	84	60	62
4. Low-Mid	84	600	615
5. High-Mid	84	6000	6150
6. High	84	30000	29500

"GW642444M is the triphenyl acetate salt form of GW642444, a long-acting beta-2 agonist (LABA). All doses and concentrations (including analyte concentration in aerosol, powders and plasma) are expressed in terms of the parent compound, which for the purpose of this report is referred to as GW642444. . . .

"GW642444M . . . was administered as a dry power formulation blended in Lactose at a nominal concentration of 0.4% or 20% (Groups 2 and 3: 0.4% w/w; Group 4: 0.4% w/w Days1 and 2 (main study), Day 1 (TK study) and 20% w/w thereafter; Groups 5 and 6: 20% w/w) to achieve the target concentrations.

"The inhalation exposure system consisted of nose-only flow through inhalation chambers. The animals were restrained in (b) (4) restraint tubes which were inserted onto the chambers. The test atmospheres were generated into the top section of the inhalation chambers (b) (4)

"The various test article concentrations were achieved by altering the rate of test article introduction into the chamber." (page 17 of report)

Dosing was justified as follows: "In a previous 13-week study in CD-1 mice, an achieved dose of 63600 µg/kg/day caused clinical signs of irregular and/or labored breathing, subdued behavior, half-closed eyes and 10 deaths by Day 8. This dose was reduced to an achieved dose of 38200 µg/kg/day on Day 9, and the clinical signs persisted through Week 4. Minimal degeneration/regeneration of the nasal turbinate respiratory and olfactory epithelium, slight laryngeal squamous metaplasia, decreased hepatocellular cytoplasmic rarefaction and slight uterine myometrial hypertrophy were also noted at this dose. Since there were no further clinical signs after Week 4 and no more deaths attributable to GW642444, 38200 µg/kg/day was considered the maximum tolerated dose (MTD) in this study. . . .

"Based on a MTD of 38200 µg/kg/day in the 13-week study, a target high dose of 30000 µg/kg/day was selected. A target low dose of 6 µg/kg/day was predicted to provide a small multiple over the predicted maximum human repeat dose AUC. Target doses of 6000, 600 and 60 µg/kg/day were selected in order to evaluate dose-response relationships." (page 44 of mouse report)

3.2.1.1. Sponsor's Results and Conclusions

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

Survival analyisis:

The Sponsor summarizes results as follows: "In the early stages of the study, there was a higher than expected incidence of deaths when compared with historical control data, frequently associated with abdominal distension and/or abnormal breathing across all groups including control. [redacted] (b) (4) Alterations [redacted] (b) (4) were made in an attempt to counteract the abdominal distension and lower the mortality rate. The mortality rate decreased and stabilized from about Week 34 when tubes without screens were used suggesting a procedural element [redacted] (b) (4) in the cause of this finding.

"The distribution in the number of unscheduled deaths, scheduled deaths and survival relative to the group size of 84 mice per sex, per group is summarised below:" (page 61 of report)

Table 22. Sponsor Provided Table of Incidence of Deaths (pages 61 and 62)

Males	**Weeks**	**Dose (µg/kg/day)**					
		0	**6.4**	**62**	**615**	**6150**	**29500**
Dead/Killed	1-8	0	0	1	1	2	3
	9-25	10	6	9	6	9	10
	26-34	10	13	12	8	6	15
	35-53	3	2	7	8	4	7
	54-73	11	12	5	9	9	7
	73-101	30	26	25	24	25	17
Total		64	59	59	56	55	59
Terminal kill	101	20	25	25	28	29	25
Survival (%)		24	30	30	33	35	30

Females	**Weeks**	**Dose (µg/kg/day)**					
		0	**6.4**	**62**	**615**	**6150**	**29500**
Dead/Killed	1-8	0	0	0	0	1	1
	9-25	9	4	6	14	7	18
	26-34	14	16	19	11	17	8
	35-53	10	6	2	7	4	5
	54-73	5	2	10	7	8	6
	74-104	20	28	29	21	31	19
Total		58	56	66	60	68[a]	57
Terminal kill	105	26	28	18	24	16[a]	27
Survival (%)		31	33	21	29	19[a]	32

[a] Terminal kill in Week 104

"All surviving males were killed during Week 101 of treatment due to control survival reaching 20 males; females given 6150 µg/kg/day were killed from the end of Week 103 when survival approached 15. Remaining female groups were killed after 104 completed weeks of treatment. For statistical analysis, all females which died, were found dead or were killed as part of the scheduled kill after the end of Week 104 were considered as terminal kill animals. Overall, mortality occurred across all groups with no test article-related increase in death rate. The trend test

was not statistically significant when all groups were included in the analysis (males: p=0.685 and females: p=0.368). None of the pairwise comparisons were statistically significant." (page 62 of mouse report)

Tumorigenicity analysis:

The Sponsor summarizes statistical tumorigenicity results:

"Males
None of the comparisons were statistically significant.

"Females
Ovaries
For benign tubulostromal adenoma, the trend test was statistically significant when all groups were included in the analysis ($p<0.001$). Upon exclusion of the 30000 µg/kg/day treated group the trend test was no longer significant (p=0.050). The pairwise comparison of the control group with the 30000 µg/kg/day treated group was statistically significant (p=0.011).

"Ovarian findings
For total incidence of tubulostromal hyperplasia, the 30000 µg/kg/day treated group had significantly higher incidence than the control group (p=0.031).

For total animals with tubulostromal hyperplasia and/or adenoma, the 6000 and 30000 µg/kg/day treated groups had significantly higher incidence than the control group (p = 0.032 and p = 0.001 respectively).

For total animals with sex cord stromal hyperplasia and/or adenoma, the 60 and 6000 µg/kg/day treated groups had significantly higher incidence than the control group (p = 0.005 and p = 0.017 respectively).

For total animals with tubulostromal and/or sex cord stromal hyperplasia and/or adenoma, the 60, 6000 and 30000 µg/kg/day treated groups had significantly higher incidence than the control group (p=0.001, p=0.003 and p=0.005 respectively)." (page 2411-2412 of mouse report, page 8-9 of statistics report).

3.2.1.2. FDA Reviewer's Results

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

Survival analysis:

The following tables (Table 23 for male mice, Table 24 for females) summarize the mortality results for the study groups. The data were grouped for the specified time period, and present the number of deaths during the time interval over the number at risk at the beginning of the interval. The percentage cited is the percent that survived at the end of the interval. In these tables the terminal period only includes those animals were sacrificed. Animals that died of

other causes during the terminal period are included in the preceding, but overlapping time period. The Kaplan-Meier survival plots in Appendix 1 provide a more detailed picture of the profile of mortality losses. Note that the four animals excluded from the carcinogenicity analysis are included here (please see Section 2.2 for details.)

Table 23. Summary of Male Mice Survival (estimated dosed in µg/kg/day)

(Weeks)	Control 0	Low-Low 6.4	High-Low 62	Low-Mid 615	High-Mid 6150	High 29500
1-52	23/84[1] 72.6%[2]	21/84 75%	29/84 65.5%	23/84 72.6%	19/84 77.4%	34/84 59.5%
53-78	15/61 54.8%	20/63 51.2%	6/55 58.3%	11/61 59.5%	16/65 58.3%	9/50 48.8%
79-91	14/46 38.1%	7/43 42.9%	12/49 44.0%	11/50 46.4.%	15/49 40.5%	9/41 38.1%
92- 104	12/32 23.8%	11/36 29.8%	12/37 29.8%	11/39 33.3%	5/34 34.5%	7/32 29.8%
Terminal[3] 105	20	25	25	28	29	25

[1] number of deaths / number at risk
[2] overall per cent survival to end of period.
[3] number of animals that survived to terminal sacrifice

Table 24. Summary of Female Mice Survival (estimated dosed in µg/kg/day)

(Weeks)	Control 0	Low-Low 6.4	High-Low 62	Low-Mid 615	High-Mid 6150	High 29500
1-52	33/84[1] 60.7%[2]	26/84 69.0%	27/84 67.9%	32/84 61.9%	29/84 65.5%	32/84 61.9%
53-78	10/51 48.8%	6/58 61.9%	11/57 54.8%	10/52 50%	16/55 46.4%	13/52 46.4%
79-91	5/41 42.9%	13/52 46.4%	14/46 38.1%	5/42 44.0%	13/39 31.0%	7/39 38.1%
92- 104	10/36 31.0%	11/39 33.3%	14/32 21.4%	14/37 27.4%	10/26 19.0%	5/32 32.1%
Terminal[3] 105	26	28	18	23	16	27

[1] number of deaths / number at risk
[2] overall per cent survival to end of period.
[3] number of animals that survived to terminal sacrifice

Note that in Table 24, in the female Low-mid dose group there is a discrepancy of one female terminal sacrifice compared to the corresponding totals reported in the Sponsor provided table 21. The following table, Table 25, summarizes the results from tests comparing survival profiles across study groups in the tumorigenicity data sets:

Table 25. Statistical Significances of Tests of Homogeneity and Trend in Survival in Mice

Hypothesis Tested	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Mice Homogeneity over Groups 1-6	0.7455	0.5696	0.3945	0.4618
No trend over Groups 1-6	0.4520	0.1321	0.8634	0.3034
No Difference Between Groups 1 vs 6	0.9896	0.4417	0.6140	0.8865

From Figure A.1.3, in male mice, survival curves were all fairly closely intertwined, although from weeks 40-90 the High dose group generally had the lowest survival rats. These slight differences were not sufficient to result in any statistically significant tests of overall homogeneity, trend, or parwise differences between the High dose and Control (all six $p \geq 0.1321$). In female mice, from Figure A.1.4, it seems that the Low-low dose group tended to have the highest survival, with the other groups largely intertwined. Again these slight differences were not sufficient to result in any statistically significant tests of overall homogeneity, trend, or parwise differences between the High dose and Control (all six $p \geq 0.3034$).

Results from a supporting experimental Bayesian nonparametric analysis of survival are provided in Appendix 2.

Tumorigenicity analysis:

As discussed in Section 1.3.1.5, for common tumors, the Haseman-Lin-Rahman rules adjusting for multiplicity in a standard two species study specify that for a very rough 0.10 (10%) overall false positive error rate, overall trend should be tested at a 0.025 (2.5%) level in rare tumors and at 0.005 (0.5%) level in common tumors. The comparison between Control and the High dose should be tested at a 0.05 (5%) level in rare tumors and at 0.01 (1%) level in common tumors. Those organ-tumor combinations with at least nominally statistically significant result ($p \leq 0.05$) in mice are summarized below:

Table 26. Potentially Statistically Significant Neoplasms in Mice

	Incidence						Significance Level					
Organ/ Tumor	Veh	LL	HL	LM	HM	Hi	trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
Male Mice												
TESTES												
# Evaluated	84	84	84	84	84	84						
BENIGN INTERSTITIAL CELL	0	0	0	0	0	2	.0275	.	.	.	.	.2485
Female Mice												
OVARIES												
# Evaluated	84	83	84	84	84	83						
TUBULOSTROMAL ADENOMA	0	0	1	0	2	6	.0001	.	.5000	.	.2485	.0137
UTERUS W/ CERVIX												
# Evaluated	84	84	84	84	84	84						
LEIOMYOSARCOMA	0	1	2	4	6	4	.1134	.5000	.2485	.0603	.0142	.0603
Leiomyoma/Leiomysarcoma	2	3	7	9	7	6	.3547	.5000	.0839	.0284	.0839	.1385

Adjusting for multiplicity the test of trend in benign interstitial cell tumor of the testes (in male mice) would be close to statistically significance ($p = 0.0275 \approx 0.025$, but recall that overall type I error is roughly at least 10%). In tubulostromal adenoma of the ovaries in female mice both the test of trend and the pairwise comparison to the control were statistically significant ($p = 0.0001 < 0.025$, $p = 0.0137 < 0.05$, repectively).. Accepting the inflation in overall Type I error from using the other pairwise comparisons, the comparison with the High-mid group in leiomysarcoma would be statistically significant ($p = 0.0142 < 0.05$). No other tests achieved the multiplicity adjusted nominal 10% level.

Further details on these tests and complete incidence tables in both genders are provided in Appendix 3.

4. FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

NA

5. SUMMARY AND CONCLUSIONS

5.1. Statistical Issues and Collective Evidence

Please see Section 1.3 above.

5.2. Conclusions and Recommendations

Please see Section 1.1 above.

Appendices

Appendix 1. FDA Survival Analysis

Simple summary life tables in mortality are presented in the report (Tables 16, 17, 23, and 24, above). Kaplan-Meier estimated survival curves across study groups for each gender are displayed below in Figures A.1.1 and A.1.2 for rats and Figures A.1.3 and A.1.4 for mice. These plots include 95% confidence intervals around each survival curve (colored area around each curve). These plots are also supported by tests of homogeneity and trend in survival over the five (in rats) and six (in mice) different treatment groups, as well as tests comparing the highest dose to the control. Note that the tests of trend over dose used the estimated dose, not the nominal dose apparently used ine the Sponsor's statistical analysis. Due to differences in dosing in female rats the rat study includes several other comparisons that may be of interest. The statistical significance levels (i.e., p-values) are provided in Tables A.1.1. and A.1.2., below. One might note that the log rank tests places greater weight on later events, while the Wilcoxon test tends to weight them more equally, and thus places more weight on earlier differences than does the log rank test.

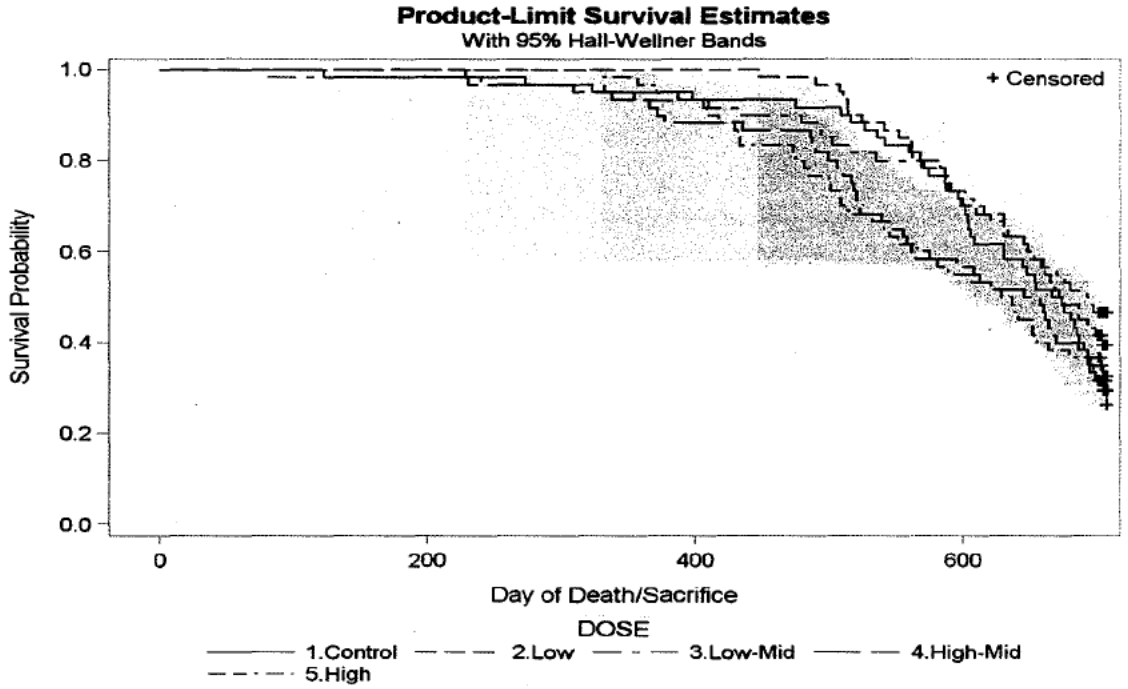
Table A.1.1 Statistical Significances of Tests of Homogeneity and Trend in Survival in the Rat Study

Hypothesis Tested	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Rat Homogeneity over Groups 1-5	0.1752	0.0770	< 0.0001	< 0.0001
Homogeneity over Groups 1-3	0.3221	0.5885	0.0001	< 0.0001
No trend over Groups 1-5	0.1169	0.0224	0.0005	0.0006
No trend over Groups 1-3	0.2256	0.4479	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 5	0.6075	0.1946	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 3	0.1377	0.3315	< 0.0001	< 0.0001
No Difference Between Groups 1 vs 2	0.3196	0.4314	0.1220	0.0276

From Figure A.1.1, in male rats, for some time after day 400 there appears to be a clear separation of the rather intertwined High and High-mid dose groups (i.e. groups 4 and 5) from the remaining, also rather intertwined, dose groups. However, by the end of the study, survival is close in all dose groups. This is consistent with the results above, i.e. the decrease in survival between between groups 1, 2, and 3 versus 4 and 5, in the middle part of the study, with less difference at the end resulted in an equivocal test of trend (Logrank $p = 0.1169$, Wilcoxon $p = 0.0224$) and test of overall homogeneity (Logrank $p = 0.1752$, Wilcoxon $p = 0.0770$). The comparisons in groups 1-3 and 1-2 were primarily to match the same and somewhat more relevant tests in females, but none were statistically significant (all six Logrank and Wilcoxon $p \geq 0.1377$).

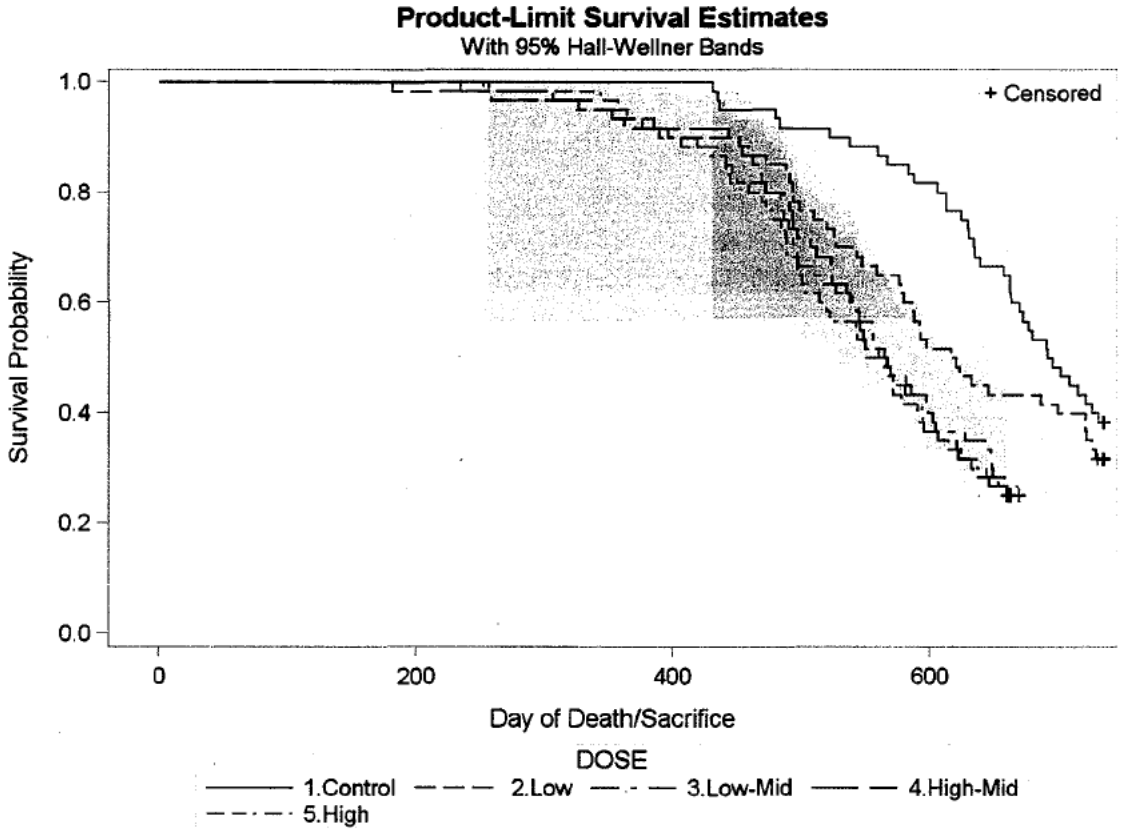
Recall if the time of natural death is greater than the last noted time the survival time is described as "censored." Since we would expect an animal to live past the time it was sacrificed, the time of sacrifice is treated as a time of censoring.

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



Results in female rats are quite different. In female rats, from Figure A.1.2, the Control dose group has the highest survival, with the Low dose group next, and the remaining dose groups largely intertwined. This is sufficient to result in the generally highly significant tests comparing survival in female rats. That is, the test of overall homogeneity and a comparison between the High dose and Control were all highly statistically significant (i..e., both Logrank and Wilcoxon $p < 0.0001$), with a highly statistically significant test of trend (Logrank $p = 0.0001$, Wilcoxon $p = 0.0005$). Dosing was stopped in Week 85 in the High-mid and High dose groups (i.e. Groups 4 and 5). Hence dosing in these groups is no longer proportional to the dosing in the remaining dose groups. And thus there may be some interest in the results of dropping these dose groups and thus restricting attention to dose groups 1-3. The overall test of homogeneity among Groups 1-3 was also highly statistically significant (i.e., Logrank $p = 0.0001$ and Wilcoxon $p < 0.0001$), with a highly statistically significant test of trend (both Logrank p and Wilcoxon $p < 0.0001$). The comparisons between Control and both the High dose (group 5) and the High-low group (group 3) were all highly statistically significant (i.e., all four Logrank and Wilcoxon $p < 0.0001$). Finally, in female rats, dose groups 3-5 were sacrificed early, so there may be some interest in comparing differences in survival between the Low dose and Control. These were equivocal (Logrank $p = 0.1220$ and $p = 0.0276$) due to the large number of censored survival times and a different pattern of survival from Day 400 to Day 650, but with closer patterns of survival at the end of the study.

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



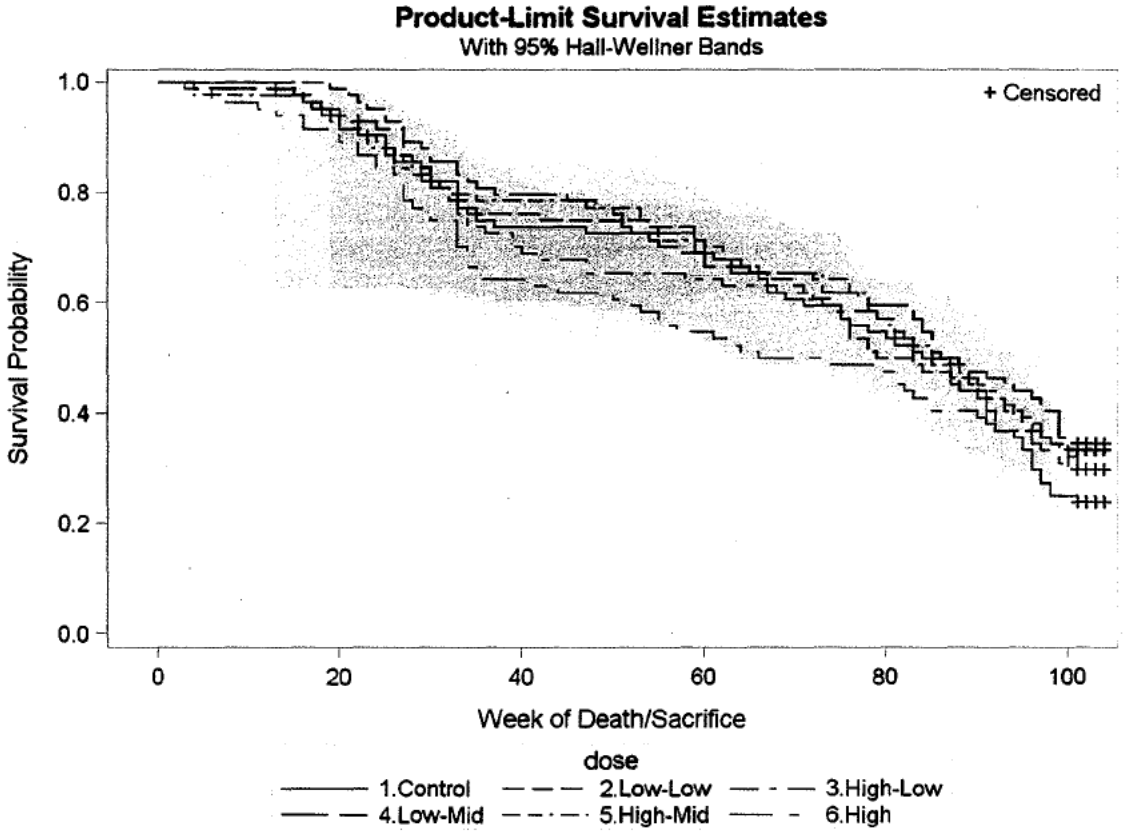
Figures A.1.3 through A.1.4, below, provide similar survival curves for each mouse gender, while Table A.1.2 provides a similat tabulation of p-values for the mouse study.

Table A.1.2. Statistical Significances of Tests of Homogeneity and Trend in Survival in the Mouse Study

Hypothesis Tested	Males		Females	
	Log rank	Wilcoxon	Log rank	Wilcoxon
Mice Homogeneity over Groups 1-6	0.7455	0.5696	0.3945	0.4618
No trend over Groups 1-6	0.4520	0.1321	0.8634	0.3034
No Difference Between Groups 1 vs 6	0.9896	0.4417	0.6140	0.8865

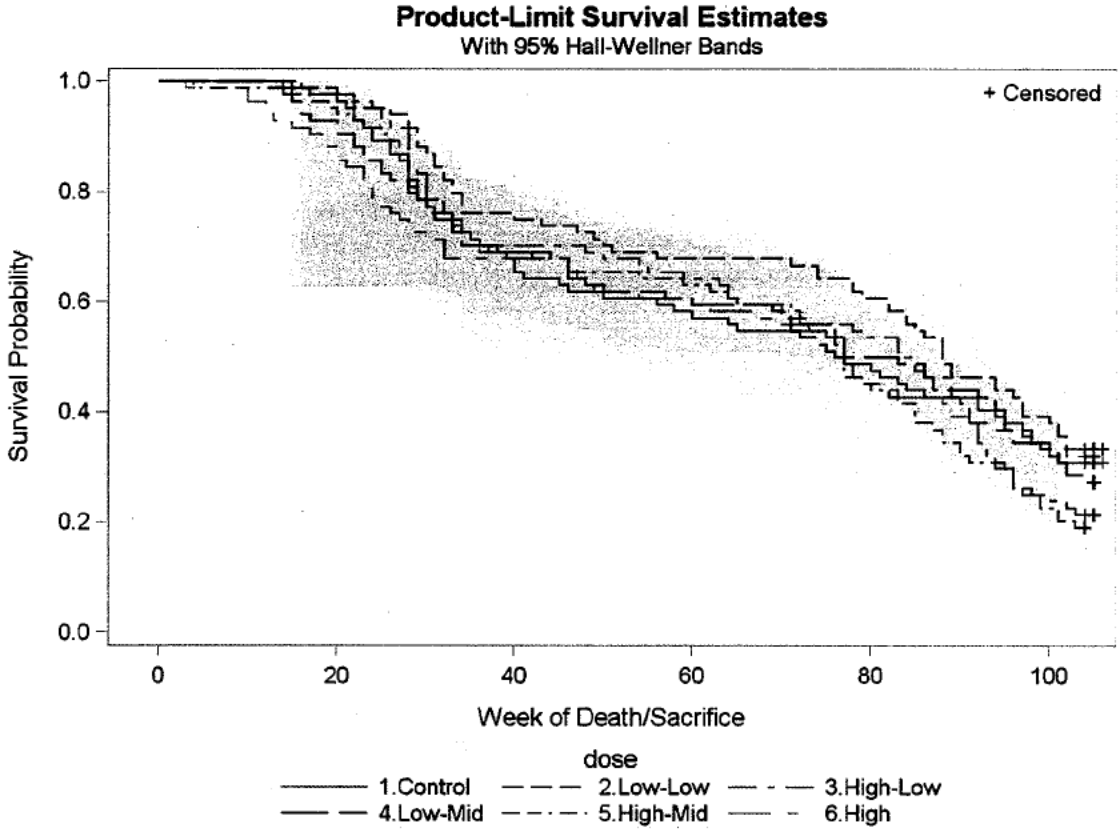
From Figure A.1.3, in male mice, survival curves were all fairly closely intertwined, although from weeks 40-90 the High dose group generally had the lowest survival rats. These slight differences were not sufficient to result in any statistically significant tests of overall homogeneity, trend, or parwise differences between the High dose and Control (i.e., all six $p \geq 0.1321$).

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



In female mice, from Figure A.1.4, it seems that the Low-low dose group tended to have the highest survival, with the other groups largely intertwined. Again these slight differences were not sufficient to result in any statistically significant tests of overall homogeneity, trend, or parwise differences between the High dose and Control (i.e., all six $p \geq 0.3034$).

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



Appendix 2. Bayesian Survival Analysis

Let $S(t)$ be the survival function, i.e., with T denoting the survival time,

$$S(t) = Pr(T > t),$$

and $f(t)$ the density of T. The instantaneous hazard function is $h(t) = f(t)/S(t)$ with cumulative hazard:

$$H(t_i) = \int_0^{t_i} h(u)du$$

So $f(t) = h(t)\, S(t)$. Also $\log(S(t)) = -H(t)$, so $S(t) = e^{-H(t)}$. Then $f(t) = h(t)\, e^{-H(t)}$.

The standard Cox regression form of the proportional hazards model for survival specifies the hazard function:

$$h(t \mid x) = h_0(t)\, exp(x^t\beta).$$

The term $h_0(t)$ is called the baseline hazard, defined by holding all covariates at 0, while the term $exp(x^t\beta)$ is called the partial likelihood. Note if the the only covariates in the partial likelihoood are treatment indicators, this implies that the logarithm of the survival curves should be approximately parallel. From the survival curves in Appendix 1, this seems to be a reasonable approximation for female rats and possibly male rats and is even less certain for male mice, but is clearly not quite appropriate for female mice. Thus results for mice, particularly female mice, should treated with extreme caution.

If we assume the baseline hazard and the partial likelihood are functionally independent, one can clearly optimize these terms separately. A typical frequentist analysis of the effect of covariates is based solely on the partial likelihood, using asymptotics to analyze the linear predictor, ignoring the baseline hazard $h_0(t)$. Such a frequentist analysis takes parameters as fixed and assesses the likelihood of the observed data, and results in the log rank tests of Appendix 1. A Bayesian analysis starts by noting that parameters are not known, and assumes that so-called prior probability distributions are natural measures of this lack of exact knowledge about the parameters. Then the Bayesian analysis conditions on the observed data, and assesses the effect of the observed data on the estimated probabilities of the parameters.

In each gender in rats there were five distinguishable treatment groups, including the control, while in mice there were six treatment groups. We specify the estimated dose levels in rats as 0, 10.5, 84.4, 223, and 657 µg/kg/day and 0, 6.4, 62, 615, 6150, and 29500 µg/kg/day in mice. In the formulation above, the baseline hazard is confounded with any intercept term in the partial likelihood, and hence the partial likelihood can not have such an intercept. Thus there are only three degrees of freedom for testing differences among the four treatment groups.

When parameterizing each treatment group separately, using so called dummy coding, we can define, for each treatment group i, except the control dose:

δ_i = 1 for the ith treatment group,
0 otherwise.

With this parameterization each labeled effect actually represents the differential effect of the specified treatment over the effect of the vehicle control group.

At least two possible models are suggested:
(1) Parameterization of a differential effect over the control treatment group,
i.e.: $x_i^t\beta = \beta_1*\delta_1 + \beta_2*\delta_2 + \ldots + \beta_r*\delta_r$, for r = 4,5 (for rats and mice).
(2) Parameterization of a linear, slope effect of dose, $x_i^t\beta = \delta_1*$ dose $_1 + \ldots + \delta_r*$ dose $_r$.

In model (1) above, β_k, denoted beta1-beta4 or beta5 in the tables below, measures the differences between the k^{th} dose in the model and the control group (k=1 to 4,5).

Tables A.2.1 and A.2.2, below, summarize the estimated posterior distributions of the treatment group parameters in rats. The three right most columns provide the lower endpoints of an estimated 95% credible interval, the medians, and the upper endpoints of an estimated 95% credible interval . That is, the posterior probability that the parameter is in the interval is 0.95 is between the lower and upper percentiles. One way to translate this to a hypothesis testing framework is to suggest that if 0 is in the posterior interval we would conclude that the parameter could be zero. If it is near the center of the interval there is little to no evidence it is not zero. When interpreting the slope parameter it should be noted that, for numerical computational reasons, the actual dose has been divided by 100. Further, distances should be assessed in units of the standard deviation.

Table A.2.1 Posterior Summaries of Treatment Parameters in Rats

		Standard	Percentiles		
Parameter	Mean	Deviation	2.5%	50%	97.5%
Male Rats					
beta[1]	-0.2626	0.2264	-0.7044	-0.2619	0.1817
beta[2]	-0.3716	0.236	-0.8363	-0.3715	0.0864
beta[3]	0.09696	0.2194	-0.3355	0.09698	0.5304
beta[4]	0.07678	0.223	-0.3566	0.0779	0.5149
slope	3.743E-4	0.003863	2.25E-5	3.997E-4	9.502E-4
Female Rats					
beta[1]	0.3838	0.2521	-0.111	0.3839	0.8788
beta[2]	0.9341	0.2377	0.4721	0.9341	1.404
beta[3]	0.9288	0.2389	0.4621	0.9281	1.397
beta[4]	0.9488	0.2381	0.4852	0.9459	1.419
slope	8.924E-4	2.684E-4	3.585E-4	8.948E-4	0.001412

In male rats there is some evidence of a positive trend, i.e. decreasing survival over increasing dose, consistent with the frequentist Wilcoxon test. Note however that the lower limit is quite close to zero in terms of the standard deviation. From the credible intervals, in male rats there is no strong evidence of dose differences with vehicle, although there is weak evidence that the overall effect of the Low-low and High-low dose groups (i.e., beta[1] and beta[2]) is

negative, i.e., higher survival than the vehicle control. Note, however, in male rats, unlike female rats, the proportional hazards assumption of the basic model is somewhat questionable.

In female rats, there is weak evidence that survival in the Low-low group is somewhat less than the vehicle (since 0 is near the lower limit of the 95% credible interval for beta[1]), while there is strong evidence of decreased survival in the other actual dose groups when compared to control (since each of the credible intervals for beta[2] to beta[4]) are bounded away from zero relative to the standard deviation. Similarly, the slope term provides strong evidence of a decrease in survival over increasing dose.

Table A.2.2 Posterior Summaries of Treatment Parameters in Mice

		Standard	Percentiles		
Parameter	Mean	Deviation	2.5%	50%	97.5%
Male Mice					
beta[1]	-0.1648	0.1789	-0.514	-0.1655	0.1906
beta[2]	-0.1455	0.1794	-0.4994	-0.1454	0.2046
beta[3]	-0.2601	0.1827	-0.6208	-0.2589	0.09567
beta[4]	-0.2314	0.1819	-0.5838	-0.2328	0.1297
beta[5]	-0.03302	0.1802	-0.3825	-0.03383	0.3245
slope	3.446E-4	5.121E-4	-0.0006612	3.46E-4	0.001373
Female Mice					
beta[1]	-0.1804	0.1853	-0.5457	-0.1802	0.1819
beta[2]	0.09257	0.177	-0.2576	0.09149	0.4401
beta[3]	0.02106	0.1816	-0.3371	0.02113	0.3763
beta[4]	0.2113	0.1769	-0.1384	0.2114	0.5586
beta[5]	-0.005923	0.1837	-0.3664	-0.004618	0.354
slope	2.475E-5	4.944E-4	-9.994E-4	2.246E-5	9.715E-4

In mice there is no evidence of any difference in any dose effect from control or slope, indicating no strong evidence of any differences in survival.

Appendix 3. FDA Poly-k Tumorigenicity Analysis

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

For each species by gender by organ combination the number of animals analyzed and used in the statistical tests is presented first. Note that indicating an organ was not examined requires a specification in the data so this analysis assumes that unless there is such a record the organ was examined. The tumor incidence for each tumor in the organ is presented next, with the significance levels of the tests of trend, and the results of pairwise tests between the Control and other dose groups. These statistical tests are conditioned on the animals actually evaluated, ignoring those not analyzed. In rats the treatment groups 1-5 are denoted by "VC" for vehicle control, "Low" or "Lw", "LM" for Low-mid, "HM' for High-mid, and "Hi" for High dose, respectively. In mice, treatment groups 1-6 are labeled "Veh" for control, "LL" for Low-low, "HL" for High-Low, "LM" for Low-mid or Low-medium, "HM" for High-mid, and, again "Hi' for High dose, respectively. In female rats dosing was stopped early in the two highest dose groups, so there may be interest in the test of trend over the the remaining three dose groups. The significance level of this test is denoted "trend/1-3."

To adjust for the multiplicity of tests the so-called Haseman-Lin-Rahman rules discussed in Section 1.3.1.4 are often applied. That is, when testing for trend over dose and the difference between the highest dose group with a Control group, to control the overall Type I error rate to roughly 10% for a standard two species, two sex study, one compares the unadjusted significance level of the trend test to 0.005 for common tumors and 0.025 for rare tumors, and the pairwise test to 0.01 for common tumors and 0.05 for rare tumors. Incidence in the vehicle group is used to assess background tumr incidence, and thus whether a tumor is classified as rare (background incidence <1%) or common. As also discussed in section 1.3.1.4, using these adjustments for other tests, like the trend over the vehicle, Low, and Low-mid dose groups in rats, or pairwise comparisons between the vehicle and any other dose group than the High dose group (i.e. Group 5 in rats and 6 in mice) can be expected to increase the overall type I error rate to some value above the nominal rough 10% level, quite possibly considerably higher than 10% .rate.

Tables A.3.1 and A.3.2 in rats and Table A.3.3 in mice show the tumors that had at least one mortality adjusted test whose nominal statistical significance was at least 0.05. Note that when one adjusts for multiplicity these nominally significant comparisons may not be statistically significant. Tables A.3.4 and A.3.5 display all incidences and statistical test results for male and female rats, respectively, while Tables A.3.6 and A.3.7 present similar results in male and female mice. The p-values of the poly-k test are based on exact tests from StatXact as discussed above. As also noted above, the period '.' **denotes the p-values** of tests of dose groups with no tumors in any group.

Table A.3.1 Potentially Statistically Significant Neoplasms in Male Rats

	Incidence					Significance Level				
Organ/ Tumor	VC	Low	LM	HM	Hi	trend	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
SUBCUTANEOUS TISSUE										
# Evaluated	13	8	10	13	7					
Fibroma	7	2	8	5	7	.0292	.9866	.3328	.8701	.1765

Note that since the vehicle control incidence is greater than 1% fibromas would be classified as common tumors. After applying the Haseman-Lin-Rahman adjustment for multiplicity the test of trend in fibromas is not statistically significant ($p = 0.0292 > 0.005$).

Table A.3.2 Potentially Statistically Significant Neoplasms in Female Rats

	Incidence					Significance Level					
Organ/ Tumor	VC	Low	LM	HM	Hi	trend	trend 1-3	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
LIVER											
# Evaluated	60	60	60	60	60						
Adenoma: hepatocellular	0	0	1	0	2	.0327	.2544	.	.3816	.	.1425
MESOVARIAN LIGAMENT											
# Evaluated	60	60	60	60	60						
Leiomyoma	0	0	5	4	4	.0220	.0009	.	.0072	.0203	.0203
PITUITARY											
# Evaluated	60	60	60	60	60						
Adenoma/Carc. pars dist.	54	55	54	57	60	.0131	.2309	.1157	.2413	.1072	.0274
Adenoma: pars distalis	44	47	48	51	53	.0137	.1387	.1222	.1138	.0306	.0119
THYROID											
# Evaluated	60	60	60	60	60						
Adenoma/Carcinoma C-Cell	4	2	3	2	7	.0142	.3654	.8355	.5323	.7522	.0745
Adenoma: C-cell	2	2	2	2	7	.0029	.3511	.5989	.4856	.4998	.0163

From the vehicle control incidence the liver and mesovarian ligament tumors in female rats would be classified as rare, while the remaining tumors would be catergorized as common. Again, after applying the Haseman-Lin-Rahman adjustment for multiplicity the test of trend in hepatocellular adenomas of the liver is close to statistical significance ($p = 0.0327 \approx 0.025$). The tests of trend over all five treatment groups and over groups 1-3 are both statistically significant ($p = 0.0220, 0.0009$, both < 0.025). The pairwise tests between the Low-mid. High-mid, and High dose groups and the Control were all statistically significant ($p = 0.0072, 0.0203, 0.0203 < 0.05$, respectively). Note however that including the tests between the Low-mid and High-mid

dose groups can be expected inflate the significance level to some value above the nominal approximate 10% level. The test of trend overall five dose groups in C-cell adenoma of the thyroid was statistically significant ($p = 0.0029 < 0.005$) while the comparison between the High dose and Control was close to statistical significance ($p = 0.0163 \approx 0.01$). The comparison between the High dose group and Control in pars distalis adenoma of the pituitary was close to statistical significance ($p = 0.0119 \approx 0.01$). No other comparison achieved the multiplicity adjusted level of significance, even when including the pairwise comparisons of the non-high dose groups with control.

Table A.3.3 Potentially Statistically Significant Neoplasms in Mice

	Incidence						Significance Level					
Organ/ Tumor	Veh	LL	HL	LM	HM	Hi	trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
Male Mice												
GALLBLADDER												
TESTES												
# Evaluated	84	84	84	84	84	84						
BENIGN INTERSTITIAL CELL	0	0	0	0	0	2	.0275	.	.	.	.	.2485
Female Mice												
OVARIES												
# Evaluated	84	83	84	84	84	83						
TUBULOSTROMAL ADENOMA	0	0	1	0	2	6	.0001	.	.5000	.	.2485	.0137
UTERUS W/ CERVIX												
# Evaluated	84	84	84	84	84	84						
LEIOMYOSARCOMA	0	1	2	4	6	4	.1134	.5000	.2485	.0603	.0142	.0603
Leiomyoma/Leiomysarcoma	2	3	7	9	7	6	.3547	.5000	.0839	.0284	.0839	.1385

Adjusting for multiplicity the test of trend in benign interstitial cell tumor of the testes (in male mice) would be statistically significant ($p = 0.0275 \approx 0.025$). In tubulostromal adenoma of the ovaries in female mice both the test of trend and the pairwise comparison to Control were statistically significant ($p = 0.0001 < 0.025$, $p = 0.0137 < 0.05$, repectively). Accepting the inflation in overall Type I error from using the other pairwise comparisons, the comparison with the High-mid group in leiomysarcoma would be statistically significant ($p = 0.0142 < 0.05$). No other tests achieved the multiplicity adjusted nominal 10% level. $p = 0.0163 \approx 0.01$). The comparison between the High dose group and Control in pars distalis adenoma of the pituitary was close to statistical significance ($p = 0.0119 \approx 0.01$). No other comparison achieved the multiplicity adjusted level of significance, even when including the pairwise comparisons of the non-high dose groups with Control.

Complete incidence table in male rats and female rats are presented in Tables A.3.4 and A.3.5 below:

Table A.3.4 Overall Tumorigenicity Results in Male Rats

Organ/ Tumor	Incidence VC	Lw	LM	HM	Hi	Significance Level trend	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
ABDOMEN										
# Evaluated	1	2	0	2	3					
Fibroma	0	0	0	0	1	.5000	.	.	.	.
Fibroma/Fibrosarcoma	0	0	0	1	1	.5000	.	.	.	.
Fibrosarcoma	0	0	0	1	0	.6667	.	.	.	.
ADRENAL										
# Evaluated	60	60	60	60	60					
Adenoma: cortical	0	1	0	1	0	.5207	.5181	.	.4737	.
Benign pheochromocytoma	2	3	1	1	4	.1107	.5229	.8840	.8543	.2771
Malignant pheochromocytoma	0	1	0	0	0	.7949	.5181	.	.	.
Pheochromocytoma [B]&[M]	2	4	1	1	4	.1561	.3725	.8840	.8543	.2771
BRAIN										
# Evaluated	60	60	60	60	60					
Benign granular cell tumor	1	1	0	0	0	.9588	.7708	1	1	1
Malignant astrocytoma	1	0	1	1	0	.6413	1	.7531	.7198	1
Medulloblastoma	0	0	0	1	0	.3641	.	.	.4737	.
CAVITY NASAL/SINUS										
# Evaluated	60	60	60	60	60					
Papilloma: squamous cell	0	0	0	0	1	.1795	.	.	.	.4667
Sq. Cell Carcinoma/Papilloma	0	0	0	0	1	.1795	.	.	.	.4667
CAVITY ORAL										
# Evaluated	0	0	1	0	0					
Carcinoma: squamous cell	0	0	1	0	0	1	.	.	.	.
Sq. Cell Carcinoma/Papilloma	0	0	1	0	0	1	.	.	.	.
HEAD										
# Evaluated	0	0	2	0	0					
Squamous cell carcinoma	0	0	1	0	0	1	.	.	.	.
HEMOLYM. TISSUE										
# Evaluated	60	60	60	60	60					
Histiocytic sarcoma	1	3	0	1	0	.8936	.3262	1	.7198	1
Malignant lymphoma	3	2	1	1	4	.1279	.8273	.9420	.9255	.4433
JEJUNUM										
# Evaluated	60	60	60	60	60					
Adenocarcinoma	0	0	0	1	0	.3641	.	.	.4737	.
Adenoma	1	0	0	0	0	1	1	1	1	1
Leiomyoma	0	1	0	0	0	.7949	.5181	.	.	.
Sarcoma (not otherwise specified	0	0	0	0	1	.1795	.	.	.	.4667
KIDNEY										
# Evaluated	60	60	60	60	60					
Liposarcoma	2	0	0	0	0	1	1	1	1	1
L. NODE MANDIBULAR										
# Evaluated	59	60	58	59	60					
Carcinoma: metastasis	0	0	0	0	1	.1823	.	.	.	.4667
L.NODE MESENTERIC										
# Evaluated	60	59	60	60	60					
Hemangioma	2	1	0	1	1	.5070	.8840	1	.8543	.8543
LIVER										
# Evaluated	60	60	60	60	60					
Adenoma/Carcinoma hepato.	4	5	3	3	2	.7769	.5456	.7840	.7396	.8595
Adenoma: hepatocellular	2	0	2	2	0	.7697	1	.6922	.6416	1
Carcinoma: hepatocellular	2	5	1	1	2	.6227	.2468	.8796	.8599	.6303

Table A.3.4 (cont.) Overall Tumorigenicity Results in Male Rats

Organ/ Tumor	Incidence VC	Lw	LM	HM	Hi	Significance Level trend	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
LUNG										
# Evaluated	60	60	60	60	60					
Carcinoma: alveolar/bronchiolar	0	0	1	0	0	.5744	.	.5062	.	.
Carcinoma: metastasis	0	0	1	0	0	.5765	.	.5122	.	.
MAMMARY GLAND										
# Evaluated	60	60	60	59	59					
Adenoma	0	0	1	1	1	.1482	.	.5062	.4667	.4737
Adenoma/Adenocarcinoma	0	0	1	1	1	.1482	.	.5062	.4667	.4737
Fibroadenoma	2	0	0	1	0	.7476	1	1	.8484	1
Fibroma/Fibroadenoma	2	0	0	1	0	.7476	1	1	.8484	1
MUSCLE SKELETAL MI										
# Evaluated	2	2	1	0	2					
Lipoma	0	0	0	0	1	.5000	.	.	.	.
PANCREAS										
# Evaluated	60	60	60	60	60					
Adenoma/Carcinoma islet cell	10	7	4	9	2	.9517	.8832	.9812	.5820	.9962
Adenoma: islet cell	6	5	2	9	2	.7480	.7663	.9690	.1983	.9535
Carcinoma: islet cell	4	2	2	0	0	.9946	.9097	.9047	1	1
PARATHYROID GLAND										
# Evaluated	60	60	58	60	60					
Adenoma	0	1	0	1	0	.5223	.5181	.	.4737	.
PITUITARY										
# Evaluated	60	60	60	60	60					
Adenoma/Carcinoma Any	46	41	43	41	47	.2300	.9282	.7933	.8186	.5907
Adenoma: pars distalis	42	37	42	39	45	.1531	.9371	.5980	.7225	.4635
Adenoma: pars intermedia	1	0	0	1	1	.2352	1	1	.7269	.7198
Carcinoma: pars distalis	3	4	1	1	1	.8633	.5400	.9449	.9252	.9252
PROSTATE										
# Evaluated	60	60	60	60	60					
Adenocarcinoma	0	0	1	0	0	.5765	.	.5122	.	.
RECTUM										
# Evaluated	60	60	60	60	60					
Adenoma	0	1	0	0	0	.7949	.5181	.	.	.
SALIV.GL. MANDIBUL										
# Evaluated	60	60	59	59	60					
Fibroma	0	0	0	1	0	.3641	.	.	.4737	.
SKIN MISCELLANEOUS										
# Evaluated	15	17	11	14	19					
Adenoma/Carcinoma Basal cell	0	2	1	0	1	.5157	.2609	.4500	.	.5217
Adenoma: basal cell	0	0	0	0	1	.2264	.	.	.	.5217
Adenoma: sebaceous	0	1	0	0	0	.7925	.5217	.	.	.
Carcinoma: basal cell	0	2	1	0	0	.8607	.2609	.4500	.	.
Keratoacanthoma	4	4	0	4	2	.7754	.7140	1	.6111	.9366
Papilloma: squamous cell	0	0	1	1	0	.4927	.	.4500	.4762	.
SPINAL CORD LUMBAR										
# Evaluated	60	60	60	60	60					
Chordoma	0	1	0	0	0	.7949	.5181	.	.	.
SPLEEN										
# Evaluated	60	60	60	60	60					
Fibrosarcoma	0	1	0	0	0	.7949	.5181	.	.	.
STOMACH										
# Evaluated	60	60	60	60	60					
Benign neuroendocrine cell tumor	1	0	0	0	0	1	1	1	1	1
Sarcoma (not otherwise specified	0	0	0	0	1	.1795	.	.	.	.4667

Table A.3.4 (cont.) Overall Tumorigenicity Results in Male Rats

Organ/ Tumor	Incidence VC	Lw	LM	HM	Hi	Significance Level trend	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
SUBCUTANEOUS TISSUE										
# Evaluated	13	8	10	13	7					
Fibroma	7	2	8	5	7	.0292	.9866	.3328	.8701	.1765
Fibroma/fibrosarcoma	7	2	8	5	7	.0292	.9866	.3328	.8701	.1765
Fibrosarcoma	0	0	0	0	1	.1622	.	.	.	.4286
Hemangiosarcoma	0	0	0	1	0	.3947	.	.	.5294	.
Lipoma	1	3	1	1	0	.9285	.2308	.7667	.7941	1
Malignant schwannoma	1	0	0	0	0	1	1	1	1	1
Sarcoma (not otherwise specified	0	0	0	1	0	.3947	.	.	.5294	.
Squamous cell Carcinoma	1	0	0	0	0	1	1	1	1	1
TAIL										
# Evaluated	1	1	0	1	2					
Fibrosarcoma	0	0	0	1	0	1	.	.	.	.
TESTIS										
# Evaluated	60	60	60	60	60					
Adenoma: interstitial cell	4	3	2	2	0	.9714	.8029	.8994	.8672	1
THORAX										
# Evaluated	1	0	0	2	0					
Hibernoma	0	0	0	1	0	1	.	.	.	.
THYROID										
# Evaluated	60	60	60	60	60					
Adenoma/Carcinoma C-Cell	6	8	7	10	4	.6996	.4412	.5000	.1438	.7739
Adenoma/Carcinoma foll. cell	2	1	2	3	1	.5378	.8881	.6922	.4507	.8484
Adenoma: C-cell	4	7	4	10	3	.6209	.3172	.6574	.0508	.7270
Adenoma: follicular cell	1	1	2	2	1	.4504	.7648	.5000	.4611	.7123
Carcinoma: C-cell	2	1	3	0	1	.6892	.8881	.5116	1	.8484
Carcinoma: follicular cell	1	0	0	1	0	.5968	1	1	.7263	1
TONGUE										
# Evaluated	60	60	60	60	60					
Carcinoma: squamous cell	0	1	1	0	0	.7340	.5181	.5122	.	.
Papilloma: squamous cell	0	1	0	0	0	.7949	.5181	.	.	.
Sq. Cell Carcinoma/Papilloma	0	2	1	0	0	.8407	.2654	.5122	.	.

Table A.3.5 Overall Tumorigenicity Results in Female Rats

Organ/ Tumor	Incidence VC	Low	LM	HM	Hi	Significance Level trend	trend 1-3	Lw vs VC	LM vs VC	HM vs VC	Hi vs VC
ADRENAL											
# Evaluated	60	60	60	60	60						
Adenoma: cortical	2	0	1	0	1	.4480	.5893	1	.7693	1	.7693
Benign pheochromocytoma	1	0	0	1	0	.5595	1	1	1	.6207	1
BRAIN											
# Evaluated	60	60	60	60	60						
Malignant astrocytoma	2	0	0	0	1	.4255	1	1	1	1	.7599
Malignant oligodendroglioma	0	0	0	0	1	.1706	.	.	.	.	.3816
CAVITY NASAL/SINUS											
# Evaluated	60	60	60	60	60						
Adenocarcinoma	0	0	1	0	1	.1417	.2544	.	.3816	.	.3816
CECUM											
# Evaluated	60	60	60	60	60						
Leiomyoma	1	0	0	0	0	1	1	1	1	1	1
HEAD											
# Evaluated	1	0	0	0	0						
Squamous cell carcinoma	1	0	0	0	0	1	1	.	.	.	.

Table A.3.5 (cont.) Overall Tumorigenicity Results in Female Rats

Organ/ Tumor	Incidence VC	Low	LM	HM	Hi	Significance Level trend	trend 1-3	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
HEAD											
# Evaluated	1	0	0	0	0						
Squamous cell carcinoma	1	0	0	0	0	1	1	.	.	.	.
HEMOLYM. TISSUE											
# Evaluated	60	60	60	60	60						
Histiocytic sarcoma	1	0	0	0	1	.3129	1	1	1	1	.6207
Malignant lymphoma	0	1	0	0	0	.7235	.5877	.4535	.	.	.
JEJUNUM											
# Evaluated	60	60	60	60	60						
Adenocarcinoma	0	0	1	0	0	.5000	.2544	.	.3816	.	.
Leiomyoma	0	0	1	0	0	.5000	.2544	.	.3816	.	.
KIDNEY											
# Evaluated	60	60	60	60	60						
Adenoma: tubular cell	1	0	0	0	1	.3129	1	1	1	1	.6207
LIVER											
# Evaluated	60	60	60	60	60						
Adenoma: hepatocellular	0	0	1	0	2	.0327	.2544	.	.3816	.	.1425
LUNG											
# Evaluated	60	60	60	60	60						
Adenoma: alveolar/bronchiolar	0	1	0	0	0	.7219	.5841	.4471	.	.	.
MAMMARY GLAND											
# Evaluated	59	60	60	60	60						
Adenocarcinoma	9	14	11	11	10	.2794	.1543	.1271	.1250	.1250	.1737
Adenoma	13	7	9	5	10	.2092	.3546	.8989	.5682	.9153	.4147
Adenoma/Adenocarcinoma	19	18	19	13	17	.2141	.1201	.5380	.1818	.6195	.2272
Fibroadenoma	27	27	19	13	17	.7958	.6383	.3515	.6299	.9607	.7200
Fibroma	0	0	0	1	0	.3373	.	.	.	.3867	.
Fibroma/Fibroadenoma	27	27	19	14	17	.7865	.6383	.3515	.6299	.9324	.7200
MESOVARIAN LIGAMEN											
# Evaluated	60	60	60	60	60						
Leiomyoma	0	0	5	4	4	.0220	.0009	.	.0072	.0203	.0203
OVARY											
# Evaluated	60	60	60	60	60						
Benign granulosa-theca cell tumor	0	0	0	0	1	.1657	.	.	.	.	.3733
PANCREAS											
# Evaluated	60	59	60	60	60						
Adenoma/Carcinoma islet cell	0	1	2	0	2	.1176	.0798	.4405	.1425	.	.1362
Adenoma: islet cell	0	1	2	0	1	.3356	.0798	.4405	.1425	.	.3733
Carcinoma: islet cell	0	0	0	0	1	.1667	.	.	.	.	.3733
PARATHYROID GLAND											
# Evaluated	57	59	59	60	59						
Adenocarcinoma	0	0	1	0	0	.5030	.2545	.	.3889	.	.
PITUITARY											
# Evaluated	60	60	60	60	60						
Adenoma/Carc. pars distalis	54	55	54	57	60	.0131	.2309	.1157	.2413	.1072	.0274
Adenoma: pars distalis	44	47	48	51	53	.0137	.1387	.1222	.1138	.0306	.0119
Carcinoma: pars distalis	10	8	6	6	7	.3753	.5497	.6425	.6513	.6762	.5145
SALIV. GLAND PAROT											
# Evaluated	60	60	59	57	57						
Adenoma	0	0	1	0	0	.4880	.2478	.	.3733	.	.

Table A.3.5 (cont.) Overall Tumorigenicity Results in Female Rats

Organ/ Tumor	Incidence VC	Low	LM	HM	Hi	Significance Level trend	trend 1-3	Low vs VC	LM vs VC	HM vs VC	Hi vs VC
SKIN MISCELLANEOUS											
# Evaluated	5	8	3	4	6						
Carcinoma: squamous cell	0	0	0	0	1	.2667	.	.	.	.	.5714
Fibrosarcoma	0	1	0	0	0	.8000	.6667	.6250	.	.	.
STOMACH											
# Evaluated	60	60	60	60	60						
Adenocarcinoma	0	0	0	1	0	.3353	.	.	.	.3816	.
Benign neuroendocrine cell tmr	0	0	0	0	1	.1657	.	.	.	.	.3733
SUBCUTANEOUS TISSU											
# Evaluated	2	4	3	3	2						
Fibroma	2	1	0	0	0	1	1	1	.	1	.
Fibroma/fibrosarcoma	2	2	1	0	0	.8929	.7143	1	.	1	.
Fibrosarcoma	0	1	1	0	0	.3929	.1905	.6667	.3333	.	.
Lipoma	0	2	0	1	0	.1500	.3000	.3000	.	.3333	.
Malignant schwannoma	0	0	0	0	1	.1429	.	.	.	.	.3333
THORAX											
# Evaluated	0	1	0	0	0						
Hibernoma	0	1	0	0	0	1	1	.	.	.	.
THYROID											
# Evaluated	60	60	60	60	60						
Adenoma/Carcinoma C-Cell	4	2	3	2	7	.0142	.3654	.8355	.5323	.7522	.0745
Adenoma: C-cell	2	2	2	2	7	.0029	.3511	.5989	.4856	.4998	.0163
Carcinoma: C-cell	2	0	1	0	0	.8772	.5893	1	.7693	1	1
TONGUE											
# Evaluated	60	60	60	60	60						
Carcinoma: squamous cell	0	0	0	0	1	.1706	.	.	.	.	.3816
UTERUS											
# Evaluated	60	60	60	60	60						
Adenocarcinoma: endometrial	0	1	0	0	0	.7219	.5841	.4471	.	.	.
Benign granular cell tumor	2	0	2	0	0	.8544	.2673	1	.4945	1	1
Leiomyoma	0	0	0	1	0	.3353	.	.	.	.3816	.
Polyp: endometrial stromal	9	4	6	9	2	.7930	.3615	.9224	.5843	.2614	.9641
Sarcoma: endometrial stromal	0	1	0	0	0	.7235	.5877	.4535	.	.	.
Stromal polyp/sarcoma	9	5	6	9	2	.8270	.4052	.8513	.5843	.2614	.9641
VAGINA											
# Evaluated	60	59	60	60	60						
Benign granular cell tumor	0	0	0	0	1	.1657	.	.	.	.	.3733

Complete incidence table in male and female mice are presented in Tables A.3.6 and A.3.7 below:

Table A.3.6 Overall Tumorigenicity Results in Male Mice

Organ/ Tumor	Incidence Veh	LL	HL	LM	HM	Hi	Significance Level trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
Male Mice												
ADRENAL GLANDS												
# Evaluated	83	84	84	84	84	84						
CORTEX: ADENOMA	1	0	2	1	2	1	.4059	1	.5045	.7545	.5045	.7545
SUBCAPSULAR ADENOMA	1	2	0	0	0	0	.9821	.5045	1	1	1	1
BONE												
# Evaluated	13	9	7	10	9	5						
OSTEOGENIC SARCOMA	0	0	1	0	0	0	.5849	.	.3500	.	.	.
OSTEOMA	2	0	1	0	0	0	.9343	1	.7491	1	1	1
Bone												
# Evaluated	84	84	84	84	84	84						
Osteoma/osteosarcoma	0	0	1	0	0	0	.6667	.	.5000	.	.	.
COLON												
# Evaluated	84	83	84	84	84	83						
ADENOMA	1	0	0	0	0	0	1	1	1	1	1	1
DUODENUM												
# Evaluated	81	82	80	78	79	82						
ADENOCARCINOMA	0	0	1	0	0	0	.6618	.	.4969	.	.	.
EXTREMITY												
# Evaluated	14	14	14	21	18	15						
SQUAMOUS PAPILLOMA	1	1	0	0	0	2	.1110	.7593	1	1	1	.5268
FEMORAL MARROW												
# Evaluated	84	84	84	84	84	84						
HEMANGIOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
HEMANGIOSARCOMA	1	0	0	0	0	0	1	1	1	1	1	1
GALLBLADDER												
# Evaluated	82	83	83	82	77	81						
ADENOCARCINOMA	0	1	0	0	0	0	.8320	.5030	.	.	.	.
ADENOMA/PAPILLARY ADENOMA	0	1	0	0	0	0	.8320	.5030	.	.	.	.
HARDERIAN GL												
# Evaluated	84	84	84	83	84	84						
ADENOMA	3	1	4	4	2	1	.8629	.9397	.5000	.4933	.8163	.9397
CARCINOMA	0	0	0	0	1	0	.3340	.	.	.	.5000	.
HEAD												
# Evaluated	0	1	1	1	0	0						
SQUAMOUS CELL CARCINOMA	0	0	0	1	0	0	.3333	.	.	.	.	.
LIVER												
# Evaluated	84	84	84	84	84	84						
Adenoma/Carcinoma hepato.	12	13	6	9	7	5	.9537	.5000	.9608	.8245	.9288	.9810
HEMANGIOMA	0	0	0	1	1	0	.3890	.	.	.5000	.5000	.
HEMANGIOSARCOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
HEPATOCELLULAR ADENOMA	8	8	4	7	4	1	.9954	.6032	.9343	.7051	.9343	.9984
HEPATOCELLULAR CARCINOMA	5	5	2	2	4	4	.4018	.6269	.9414	.9414	.7522	.7522
LUMBAR SC												
# Evaluated	84	84	84	83	84	84						
MALIGNANT MENINGIOMA	0	0	0	0	1	0	.3340	.	.	.	.5000	.
LUNGS												
# Evaluated	84	84	84	84	84	84						
Adenoma/Carcinoma bronch. Alv.	13	16	12	14	12	6	.9879	.3418	.6674	.5000	.6674	.9755
BRONCHIOLO/ALV. ADENOMA	9	9	8	9	7	3	.9818	.5981	.6950	.5981	.7843	.9840
BRONCHIOLO/ALV. CARCINOMA	4	7	4	5	5	3	.7948	.2674	.6401	.5000	.5000	.7784

Table A.3.6 (cont.) Overall Tumorigenicity Results in Male Mice

Organ/ Tumor	Incidence Veh	LL	HL	LM	HM	Hi	Significance Level trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
LYMPH/RETIC SYS												
# Evaluated	84	84	84	84	84	84						
GRANULOCYTIC LEUKEMIA	0	0	2	0	0	0	.7780	.	.2485	.	.	.
HISTIOCYTIC SARCOMA	1	1	1	0	1	0	.7698	.7515	.7515	1	.7515	1
MALIGNANT LYMPHOMA	10	6	2	4	5	9	.1038	.9062	.9976	.9763	.9490	.6864
MAST CELL TUMOR	0	0	1	0	0	1	.1945	.	.5000	.	.	.5000
PITUITARY												
# Evaluated	84	84	84	83	84	84						
Adenoma pars dit./inter.	1	1	1	1	0	0	.9041	.7515	.7515	.7485	1	1
PARS DISTALIS-ADENOMA	1	1	1	1	0	0	.9041	.7515	.7515	.7485	1	1
PARS INTERMEDIA: ADENOMA	1	0	0	0	0	0	1	1	1	1	1	1
PREPUT/CLIT GL												
# Evaluated	83	83	84	83	82	80						
ADENOMA	0	1	0	0	0	0	.8323	.5000	.	.	.	.
PROSTATE												
# Evaluated	84	83	84	84	84	82						
ADENOMA	0	0	1	0	0	0	.6667	.	.5000	.	.	.
SKIN												
# Evaluated	84	84	84	84	84	84						
FIBROSARCOMA	0	1	0	2	2	0	.6314	.5000	.	.2485	.2485	.
FIBROUS HISTIOCYTOMA	0	1	0	1	0	0	.6948	.5000	.	.5000	.	.
HEMANGIOMA	0	0	0	1	0	0	.5000	.	.	.5000	.	.
HEMANGIOSARCOMA	0	0	0	0	1	0	.3333	.	.	.	.5000	.
KERATOACANTHOMA	2	0	0	0	0	0	1	1	1	1	1	1
RHABDOMYOSARCOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
SQUAMOUS CELL PAPILLOMA	1	1	0	0	1	0	.6908	.7515	1	1	.7515	1
Sq. Cell Papilloma/kerato.	3	1	0	0	1	0	.8690	.9397	1	1	.9397	1
SOFT TISSUE												
# Evaluated	2	0	1	1	0	0						
NEUROENDOCRINE TUMOR	1	0	0	0	0	0	1	.	1	1	.	.
SPLEEN												
# Evaluated	84	84	84	84	84	84						
HEMANGIOSARCOMA	0	0	0	0	0	1	.1667	.	.	.	.	.5000
STOMACH												
# Evaluated	84	84	84	84	84	84						
FORESTOMACH: SQUAMOUS CELL PAP.	1	0	0	0	0	0	1	1	1	1	1	1
LEIOMYOSARCOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
SQUAMOUS CELL CARCINOMA	0	0	0	1	0	0	.5000	.	.	.5000	.	.
Systemic												
# Evaluated	84	84	84	84	84	84						
HEMANGIOMA	0	1	0	2	1	0	.6271	.5000	.	.2485	.5000	.
HEMANGIOSARCOMA	1	1	0	1	1	1	.3570	.7515	1	.7515	.7515	.7515
Hemangioma/-sarcoma	1	2	0	3	2	1	.5445	.5000	1	.3102	.5000	.7515
TAIL												
# Evaluated	13	10	10	8	12	6						
HEMANGIOSARCOMA	0	0	0	1	0	0	.4407	.	.	.3810	.	.
TESTES												
# Evaluated	84	84	84	84	84	84						
Adenoma/Interstitual Tumor	0	1	0	0	0	2	.0597	.5000	.	.	.	.2485
BENIGN INTERSTIT. CELL TMR	0	0	0	0	0	2	.0275	.	.	.	.	.2485
RETE TESTIS: ADENOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.

Table A.3.6 (cont.) Overall Tumorigenicity Results in Male Mice

Organ/ Tumor	Incidence Veh	LL	HL	LM	HM	Hi	Significance Level trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
THYROID												
# Evaluated	84	84	83	84	83	84						
C-CELL ADENOMA	0	0	0	1	0	0	.5000	.	.	.5000	.	.
FOLLICULAR CELL ADENOMA	0	0	0	1	0	0	.5000	.	.	.5000	.	.
VESSEL												
# Evaluated	84	84	84	84	84	84						
HEMANGIOMA	0	1	0	2	1	0	.6271	.5000	.	.2485	.5000	.
HEMANGIOSARCOMA	1	1	0	1	1	1	.3570	.7515	1	.7515	.7515	.7515

Table A.3.7 Overall Tumorigenicity Results in Female Mice

Organ/ Tumor	Incidence Veh	LL	HL	LM	HM	Hi	Significance Level trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
ADIPOSE TISSUE												
# Evaluated	1	1	0	1	2	1						
LIPOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
ADRENAL GLANDS												
# Evaluated	83	84	84	84	84	84						
Adenoma	1	1	0	1	0	1	.4276	.7545	1	.7545	1	.7545
CORTEX: ADENOMA	1	0	0	1	0	1	.2977	1	1	.7545	1	.7545
SUBCAPSULAR ADENOMA	0	1	0	0	0	0	.8350	.5030	.	.	.	.
BONE												
# Evaluated	5	8	7	9	9	9						
OSTEOGENIC SARCOMA	0	1	2	0	0	0	.9422	.6154	.3182	.	.	.
OSTEOMA	0	0	0	0	1	0	.3830	.	.	.	.6429	.
Bone												
# Evaluated	84	84	84	84	84	84						
Osteoma/osteosarcoma	0	2	2	0	0	0	.9496	.2485	.2485	.	.	.
DISTAL FEMUR												
# Evaluated	84	84	84	84	84	84						
CHONDROMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
EXTREMITY												
# Evaluated	4	6	5	4	9	5						
SQUAMOUS PAPILLOMA	0	0	1	1	0	1	.2667	.	.5556	.5000	.	.5556
FEMORAL MARROW												
# Evaluated	84	84	84	84	84	84						
HEMANGIOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
HARDERIAN GL												
# Evaluated	84	84	84	84	84	84						
ADENOMA	5	2	3	4	0	2	.7920	.9414	.8615	.7522	1	.9414
CARCINOMA	1	0	1	0	0	0	.8893	1	.7515	1	1	1
ILEUM												
# Evaluated	83	83	81	80	82	81						
ADENOCARCINOMA	0	0	0	0	0	1	.1653	.	.	.	.	.4939
LIVER												
# Evaluated	84	84	84	84	84	84						
HEPATOCELLULAR ADENOMA	1	0	3	2	0	0	.9308	1	.3102	.5000	1	1
LUMBAR SC												
# Evaluated	84	84	84	84	84	84						
BENIGN MENINGIOMA	0	0	0	0	1	0	.3333	.	.	.	.5000	.

Table A.3.7 (cont.) Overall Tumorigenicity Results in Female Mice

Organ/ Tumor	Incidence Veh	LL	HL	LM	HM	Hi	Significance Level trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
LUNGS												
# Evaluated	84	84	84	84	84	84						
Adenoma/Carcinoma bronch. alv.	8	7	5	8	6	5	.7560	.7051	.8762	.6032	.7983	.8762
BRONCHIOLO/ALV. ADENOMA	5	3	3	3	2	4	.3991	.8615	.8615	.8615	.9414	.7522
BRONCHIOLO/ALV. CARCINOMA	3	4	2	5	4	1	.8931	.5000	.8163	.3599	.5000	.9397
LYMPH/RETIC SYS												
# Evaluated	84	84	84	84	84	84						
GRANULOCYTIC LEUKEMIA	1	1	0	0	0	0	.9725	.7515	1	1	1	1
HISTIOCYTIC SARCOMA	7	3	8	4	6	1	.9805	.9506	.5000	.8946	.7174	.9967
MALIGNANT LYMPHOMA	11	7	13	9	5	8	.7355	.8942	.4130	.7622	.9685	.8350
MAST CELL TUMOR	0	1	0	0	0	0	.8333	.5000	.	.	.	.
MAMMARY AREAS												
# Evaluated	83	84	83	84	84	84						
ADENOCARCINOMA	0	2	0	2	0	0	.8427	.2515	.	.2515	.	.
MALIGNANT ADENOACANTHOMA	0	0	1	0	0	0	.6673	.	.5000	.	.	.
MESENTERY/PERITO												
# Evaluated	20	11	22	17	15	10						
MALIGNANT MESOTHELIOMA	0	0	1	0	0	0	.6737	.	.5238	.	.	.
OVARIES												
# Evaluated	84	83	84	84	84	83						
BENIGN GRANULOSA CELL TMR	0	0	1	1	2	0	.5428	.	.5000	.5000	.2485	.
CYSTADENOMA	0	1	3	2	1	0	.8585	.4970	.1228	.2485	.5000	.
Granulosa cell tumor B&M	0	0	1	1	3	0	.6030	.	.5000	.5000	.1228	.
HEMANGIOSARCOMA	0	0	0	0	0	1	.1653	.	.	.	.	.4970
LEIOMYOSARCOMA	0	0	0	0	1	0	.3327	.	.	.	.5000	.
LUTEOMA	0	0	1	1	0	0	.6389	.	.5000	.5000	.	.
MALIGNANT GRANULOSA CELL TUMOR	0	0	0	0	1	0	.3327	.	.	.	.5000	.
SEX CORD STROMA ADENOMA	0	1	2	1	0	2	.2048	.4970	.2485	.5000	.	.2455
TUBULOSTROMAL ADENOMA	0	0	1	0	2	6	.0001	.	.5000	.	.2485	.0137
PANCREAS												
# Evaluated	84	83	84	84	84	83						
ISLET CELL ADENOMA	1	0	0	0	0	1	.3036	1	1	1	1	.7485
PITUITARY												
# Evaluated	83	84	82	82	84	82						
Adenoma pars dit./inter.	1	0	1	0	0	0	.8875	1	.7485	1	1	1
PARS DISTALIS-ADENOMA	1	0	1	0	0	0	.8875	1	.7485	1	1	1
SKIN												
# Evaluated	84	84	84	84	84	84						
BASOSQUAMOUS TUMOR	0	0	0	0	1	0	.3333	.	.	.	.5000	.
CARCINOMA NOT OTHERWISE SPECIFIED	1	0	0	0	0	0	1	1	1	1	1	1
FIBROSARCOMA	0	0	0	0	1	1	.0832	.	.	.	.5000	.5000
MYXOSARCOMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
SQUAMOUS CELL CARCINOMA	1	0	0	0	0	0	1	1	1	1	1	1
SOFT TISSUE												
# Evaluated	0	0	0	1	3	2						
FIBROSARCOMA	0	0	0	0	1	0	.8333	.	.	.	.	.
LEIOMYOSCAROMA	0	0	0	0	0	1	.3333	.	.	.	.	.
MALIGNANT SCHWANNOMA	0	0	0	0	1	0	.8333	.	.	.	.	.
SPLEEN												
# Evaluated	83	84	84	84	84	84						
HEMANGIOSARCOMA	0	0	1	0	0	0	.6680	.	.5030	.	.	.

Table A.3.7 (cont.) Overall Tumorigenicity Results in Female Mice

Organ/ Tumor	Incidence Veh	LL	HL	LM	HM	Hi	Significance Level trend	LL vs VC	HL vs VC	LM vs VC	HM vs VC	Hi vs VC
STOMACH												
# Evaluated	84	84	84	84	83	83						
FORESTOMACH: SQUAMOUS CELL PAPILLOMA	1	1	0	0	0	0	.9723	.7515	1	1	1	1
Systemic												
# Evaluated	84	84	84	84	84	84						
HEMANGIOMA	1	3	2	0	0	0	.9912	.3102	.5000	1	1	1
HEMANGIOSARCOMA	0	1	1	0	2	1	.2451	.5000	.5000	.	.2485	.5000
Hemangioma/-sarcoma	1	4	3	0	2	1	.7518	.1837	.3102	1	.5000	.7515
THYROID												
# Evaluated	83	84	83	84	84	84						
FOLLICULAR CELL ADENOMA	0	1	0	0	0	2	.0603	.5030	.	.	.	.2515
UTERUS W/ CERVIX												
# Evaluated	84	84	84	84	84	84						
BENIGN GRANULAR CELL TUMOR	0	1	0	0	0	0	.8333	.5000	.	.	.	.
DECIDUOMA	0	0	0	0	0	1	.1667	.	.	.	.	.5000
ENDOMETRIAL ADENOCARCINOMA	1	1	1	1	1	1	.4544	.7515	.7515	.7515	.7515	.7515
ENDOMETRIAL STROMAL POLYP	2	2	5	2	5	3	.4251	.6898	.2216	.6898	.2216	.5000
ENDO. STROMAL SARCOMA	0	1	1	1	0	2	.1378	.5000	.5000	.5000	.	.2485
FIBROMA	0	1	0	0	0	0	.8333	.5000	.	.	.	.
HEMANGIOMA	1	2	2	0	0	0	.9808	.5000	.5000	1	1	1
HEMANGIOSARCOMA	0	1	0	0	2	0	.4679	.5000	.	.	.2485	.
LEIOMYOMA	2	2	5	5	1	2	.7760	.6898	.2216	.2216	.8772	.6898
LEIOMYOSARCOMA	0	1	2	4	6	4	.1134	.5000	.2485	.0603	.0142	.0603
Leiomyoma/Leiomysarcoma	2	3	7	9	7	6	.3547	.5000	.0839	.0284	.0839	.1385
Stromal polyp/sarcoma	2	3	6	3	5	5	.2361	.5000	.1385	.5000	.2216	.2216
VESSEL												
# Evaluated	84	84	84	84	84	84						
HEMANGIOMA	1	3	2	0	0	0	.9912	.3102	.5000	1	1	1
HEMANGIOSARCOMA	0	1	1	0	2	1	.2451	.5000	.5000	.	.2485	.5000

Appendix 4. References

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

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

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

De Iorio, M., Muller, P., Rosner, R.L., and MacEachern, S. (2004), "An ANOVA Model for Dependent Random Measures", Bayesian Nonparametric Nonproportional Hazards Survival Modeling", *Journal of the American Statistical Association.* **99**, 465, 205-215.

De Iorio, M., Johnson, W.O., Mueller, P., and Rosner, R.L. (2009), "Bayesian Nonparametric Nonproportional Hazards Survival Modeling", *Biometrics*, **65**, 3, 762-771.

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

Jara, A. (2007), "Applied Bayesian Non- and Semi-parametric Inference using DPpackage", *Rnews*, 7, 3, 17-26.

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

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

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

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

R Development Core Team (2009). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rahman, M.A. and Lin, K.K. (2008), "A Comparison of False Positive Rates of Peto and Poly-3 Methods for Long Term Carcinogenicity Data Analysis Using Multiple Comparison Adjustment Method Suggested by Lin and Rahman", *Journal of Biopharmaceutical Statistics*. 18, 949-958.

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

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

---

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

/s/

STEVEN F THOMSON
04/16/2012
Statistical Carcinogenicity Review

KARL K LIN
04/17/2012
Concur with review
Note: This is a revised version (with a correction) of the review report that was put into DARRTS 4/10/12.

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

/s/

STEVEN F THOMSON
04/16/2012
Statistical Carcinogenicity Review

KARL K LIN
04/17/2012
Concur with review
Note: This is a revised version (with a correction) of the review report that was put into DARRTS 4/10/12.