CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

205874Orig1s000

<u>STATISTICAL REVIEW(S)</u>

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 205-874
Supplement #:	
Drug Name:	Ferric Citrate
Indication(s):	lower serum phosphorus levels in subjects with End Stage Renal Disease on hemodialysis
Applicant:	Keryx Biopharmaceuticals, Inc.
Date(s):	8/7/2013
Review Priority:	Standard
Biometrics Division:	DBI
Statistical Reviewer:	John Lawrence, Ph D
Concurring Reviewers:	Jim Hung, Ph D
Medical Division:	Cardiorenal.
Clinical Team:	Nancy Xu, MD
Project Manager:	Russell Fortney
Keywords:	multiple comparisons, dose-response

Table of Contents

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

The submission includes the results of three randomized phase 3 clinical trials. One trial (PBB00101) included 3 dose groups and a placebo group. Because there was no plan to control the error rate for the testing of the 3 doses on the primary endpoint nor the numerous secondary endpoints, no efficacy claim can be made based on the trial and the sponsor is not requesting any efficacy results from this trial be shown in the label. The second trial (trial 304) was open label and compared one arm of the study drug to placebo. This study showed that the study drug was effective at lowering phosphorous levels. The third trial (305) was open label and compared three doses of the study drug (no placebo). This trial showed that that the study drug was effective at lowering serum phosphorous levels; the two higher doses were significantly better than the lowest dose. In all the trials, either no multiple comparison procedure was used or a sequential strategy was used. In all cases, no efficacy claims can be made for any secondary endpoints because the higher ranking hypothesis tests failed (for cases where the sequential testing was used).

INTRODUCTION

1.1 Overview

In patients with end stage renal disease (ESRD), there is a marked decrease in phosphorus excretion in the urine leading to hyperphosphatemia. If not treated, this may cause adverse clinical outcomes. The National Kidney Foundation Kidney Disease Outcome Quality Initiative (K/DOQI) Clinical Practice Guidelines recommends maintaining serum phosphorus levels between 3.5 and 5.5 mg/dL.

The submission includes two Phase 3 trials in subjects with ESRD on hemodialysis.

Table: List of all studies included in analysis

	Phase and Design	Treatment Period	Follow-up Period	# of Subjects per Arm	Study Population
Study KRX-0502-304	*Phase 3,open label, placebo controlled efficacy period*	*58 Weeks*	*58 Weeks total. 52 weeks safety, 2 weeks washout, 4 weeks efficacy*	*96 (efficacy period)*	*ESRD*
Study KRX-0502-305	*Phase 3,open label, three doses*	*28 Days*	*28 Days*	*52 low dose, 52 medium, 50 high dose*	*ESRD*
Study PBB00101	*Phase 3, double-blind, three doses vs. placebo*	*4 weeks*	*4 weeks*	*16 placebo, 33 low dose, 34 medium, 33 high dose*	*ESRD*

1.2 Data Sources

Electronic datasets and Study Reports:

\\cdsesub1\evsprod\NDA205874\\205874.enx

\\cdsesub1\evsprod\NDA205874\0000\m5\datasets

STATISTICAL EVALUATION

1.3 Data and Analysis Quality

There were many issues with the data and analysis quality for all three trials.

First of all, the ITT population was not used for any efficacy analyses. In these trials, there were subjects in the ITT population that were not included in the efficacy analyses. In trial 305, the SAP stated " For all efficacy analyses, only on-treatment data will be analyzed. On-treatment values are defined as any efficacy values evaluated one day after taking study drug." Study 305 claimed to be using the ITT population, but made up its own (incorrect) interpretation of what ITT means. The other trial used what was called the Full Analysis population, which again was not ITT. Fortunately, there were few subjects not included in the ITT population and the results where they were significant for the remaining subjects were overwhelming. In all trials, subjects who choose to stop taking the study drug should still have measurements and those measurements should be included in the primary efficacy analysis. The next issue is a problem with the design and not the analysis. Studies 304 and 305 used an open label study design. Even for endpoints that are objectively measured laboratory values, there can be bias created by this kind of design and it should be avoided whenever possible. The analysis of trial PBB00101 did not include an appropriate method to control the error rate for the 3 comparisons. Because of that, it is not possible to interpret the p-values for individual pairwise comparisons of each dose to placebo. A placebo-controlled phase 3 trial with multiple doses should have a plan to test overall the question of whether any dose is superior to placebo and to test specifically which dose or doses are superior to placebo.

On the plus side, all the datasets had a reasonable size and I could open them on my computer using JMP software. Using the datasets for trial PBB00101, I could not replicate the results in the Study Report. The number of subjects in the dataset was the same as in Table 11-1 of the Study Report, but the range of values as well as the median or mean differed from the values in the table in some cases. For trial 305, there were several subjects that were randomized, but did not appear at all in the lab datasets. Those subjects should have had lab values from screening and randomization visits and those should be in the lab datasets. The subject level dataset should include important information such as baseline lab values and important patient characteristics such as whether the subject was on peritoneal dialysis. It did not. For Study 304, I could not replicate the results shown in the study report during the 52 week safety period. There should be a column in the dataset that tells me whether each value is used in the efficacy analysis. There was, but the problem is that when I used those values that were flagged as being used, I found very different results from what was in the Study Report. This tells me that whatever analysis was done by the sponsor, they did not use all of those values flagged as being used. For example, the Study Report would say there were 253 and 137 subjects in the two arms, but only 194 and 114 of those had values of serum ferritin at Week 52. But, when I looked at the adphse.xpt (lab analysis efficacy) dataset, I found all 253 and 137 subjects had non-missing values of CHG(=change from baseline) in serum ferritin at Week 52. It was obvious to me that the sponsor only used a portion of those values, but I did not understand why or which ones they used. That was not the primary efficacy analysis period, but it still is a problem with the datasets.

1.4 Evaluation of Efficacy

1.4.1 Study Design and Endpoints

Trial PBB00101 was a randomized, double-blind, placebo-controlled, dose-ranging study to assess the effect of ferric citrate on serum phosphate concentrations in patients with ESRD who were undergoing hemodialysis. Patients who met the inclusion/exclusion criteria underwent a one to two week washout from all phosphate-binding agents and were randomized to one of four treatment groups in a ratio of 1:2:2:2:
Placebo arm
Ferric citrate 2 g per day (g/day) arm
Ferric citrate 4 g/day arm
Ferric citrate 6 g/day arm

All patients were to receive 4 capsules TID, with meals, for 28 days (starting on Day 0).

The primary efficacy variable was change in PO_4 concentration from baseline (Day 0) to Day 14 and Day 28. The secondary efficacy variables were change in Ca x PO_4 product, Ca concentration, serum iron concentration, ferritin concentration, transferrin saturation percentage, and total IBC from baseline to Day 14 and Day 28.

According to the study report (Section 9.7.2): "Assume that the array of treatment response, evaluated as serum PO_4, decreased from the post-washout baseline is 0.8 mg/dL following ferric citrate 2 g/day, 1.2 mg/dL following ferric citrate 4 g/day, 1.6 mg/dL following ferric citrate 6 g/day, and 0 following placebo treatment. Also, assume that at least 70 patients (83% of 84 randomized) have baseline and follow-up evaluations of serum PO4 and the standard deviation in each treatment group is 1.4 mg/dL. Then, the study had greater than 80% power to detect a statistically significant difference between ferric citrate 6 g/day and placebo, based on a two-sided, least significant difference test evaluated at the α= 0.05 level. Statistical power for the 6 g versus placebo contrast is approximately 83%."

Trial KRX-0502-304 was an open label trial comparing ferric citrate to placebo. The primary endpoint was change in serum phosphorous from the beginning of the efficacy period (the end of the initial 52 week safety period) to the end of the 4 week efficacy period (week 56 of the entire trial). The dose was not fixed; rather, subjects started with 6 g/day during the 52 week safety period and the dose was titrated to achieve the target goal for serum phosphorus of 3.5 to 5.5 mg/dL. I could not find any summary in the Study Report of the actual dose used in the efficacy period. There were four secondary endpoints.

- change from baseline in ferritin at Week 52
- change from baseline in TSAT at Week 52
- cumulative use of IV iron over 52 weeks
- cumulative use of EPO (ESA) over 52 weeks.

Trial KRX-0502-305 was an open label trial comparing three doses of ferric citrate (1 g, 6 g, 8 g). The primary endpoint was change in serum phosphorous over 28 days. There were numerous secondary endpoints:

- Changes from baseline in serum phosphorus, serum calcium, and calcium times phosphorus product, ferritin, TSAT, and bicarbonate at all post-baseline assessment time points, where the baseline is the last respective assessment prior to receiving the first dose of the study drug. Since there were 4 time points and 3 doses, this represents 72 different analyses across these 6 endpoints.
- proportion of treatment failures (defined as serum phosphorus ≥ 9.0 mg/dL) at the end of treatment
- proportion of patients with serum phosphorus < 5.5 mg/dL at Visit 8 (Day 28)

1.4.2 Statistical Methodologies

For trial PBB00101, the 3 ferric citrate treatment groups were compared to the placebo treatment group using a one-way analysis of variance model with a fixed effect for treatment group. Model significance from the linear regression were reported using the F-test P-value. The secondary efficacy variables (change in Ca x PO_4 product, Ca concentration, serum iron concentration, ferritin concentration, transferrin saturation percentage, and total IBC from baseline to Day 14 and Day 28) were analyzed in the same manner as the primary efficacy variable. According to the study report (Section 9.7.1): "No adjustments for multiple comparisons were applied, since the results of this study provide a first review for evidence of a dose effect in the ferric citrate treatment."

When there is no plan to control the error rate for multiple comparisons, the trial cannot be interpreted the same way as in a confirmatory trial where the error rate was controlled. This trial can be interpreted as an exploratory trial.

The study report states that under certain assumptions, with 70 subjects, the trial would have greater than 80% power based on a two-sided, least significant difference test. The least significant difference test is well-known to not control the familywise error rate when there are more than 3 groups. Here, there are 4 groups and the familywise error rate can be more than double the targeted error rate (see Hayter, Anthony J. "The maximum familywise error rate of Fisher's least significant difference test." *Journal of the American Statistical Association* 81.396 (1986): 1000-1004.). Furthermore, there is no way to correctly adjust any p-values after the fact to adjust for this. In addition, I calculated that under the stated assumptions, the trial would have only approximately 70% power. See Appendix for further information including a formula for the familywise error rate of the least significant difference test, the power of the F-test and a comparison of the power of four different closed test procedures that all would correctly control the familywise error rate (closure of the least significant difference test, Holm's procedure, closure of Dunnett's test, closure of test based on pooling doses of test drug, closure of test for linear dose response). Although it is too late to go back and change anything in this trial, the results in the appendix may be of interest for future trials.

Trial KRX-0502-304 was an open label trial comparing ferric citrate to placebo. The primary endpoint was change in serum phosphorous from the beginning of the efficacy period (the end of the initial 52 week safety period) to the end of the 4 week efficacy period (week 56 of the the entire trial). Change in serum phosphorus from the Week-52-baseline to the end of the Efficacy Assessment Period (Week 56) was analyzed using an analysis of covariance (ANCOVA) model with treatment as the fixed class effect and Study-baseline (Week 52) as the covariate. Missing data were imputed using last observation carried forward (LOCF).

If the primary endpoint was successful, then the following four secondary endpoints would be tested in this order using a sequential testing strategy:
change from baseline in ferritin at Week 52.
change from baseline in TSAT at Week 52.
cumulative use of IV iron over 52 weeks.
cumulative use of EPO (ESA) over 52 weeks.

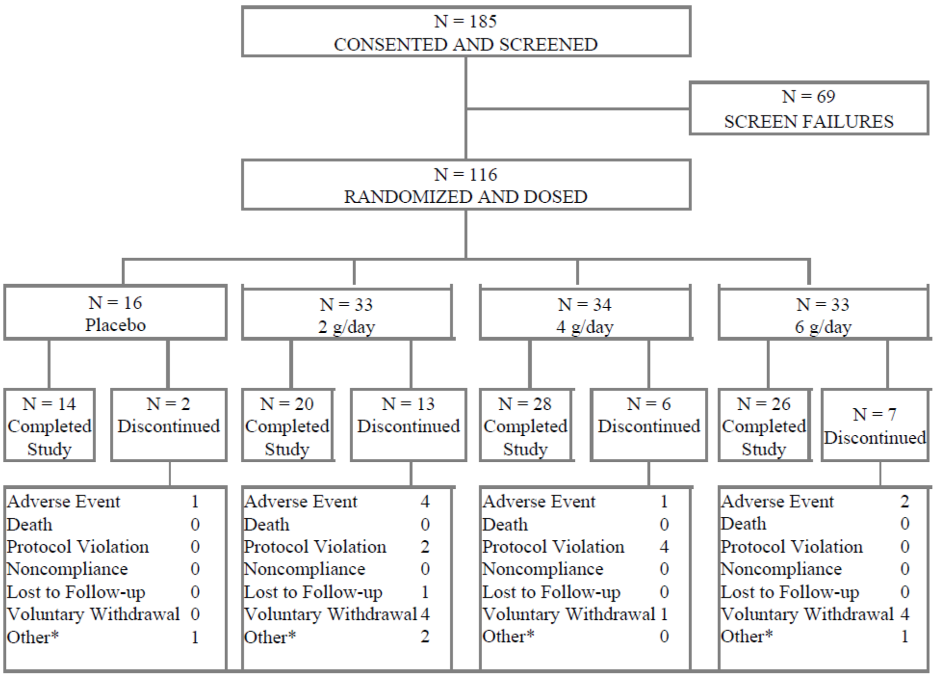
The change in ferritin and TSAT from Study-baseline (Visit 4) to Week 52 (Visit 21) was analyzed in the same way as the primary endpoint (i.e. ANCOVA with LOCF for missing values). The cumulative IV iron and ESA administration from randomization to Week 52 was compared between treatment groups using ANCOVA methods with no imputation for missing values.

Trial KRX-0502-305 was an open label trial comparing three doses of ferric citrate. The primary endpoint was change in serum phosphorous at Day 28 or last value on treatment. The primary analysis used a simple linear regression model with dose effect. Positive dose ranging will be confirmed if the null hypothesis of slope =0 is rejected at a significance level of 0.05. If that hypothesis was rejected, then each pairwise comparison for the primary endpoint would be done sequentially starting with 8 g vs. 1 g, then 6 g vs 1 g, followed by 8 g vs. 6 g. No multiple testing strategy was used for the secondary endpoints. There was a section in the SAP title "ADJUSTMENT FOR MULTIPLICITY" that stated simply: "There will be only one primary efficacy assessment (see Section 9.1)."

1.4.3 Patient Disposition, Demographic and Baseline Characteristics

For Study PBB00101, the patient disposition and demographics are shown in Figure 1 and Table 1. The Table does not include all subjects in the ITT population, just those in the sponsor's analysis.

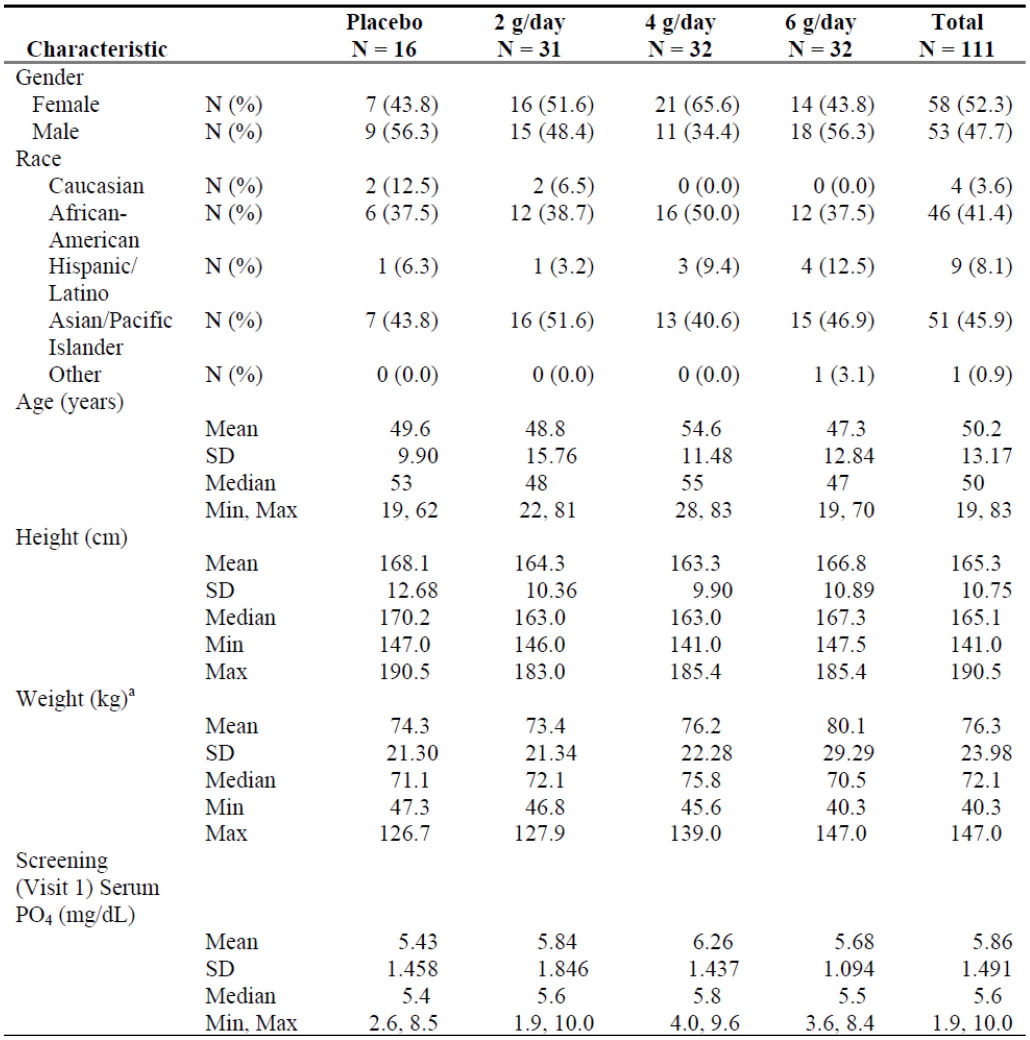
Figure 1 Disposition of subjects in Study PBB00101.



Adverse Event	1	Adverse Event	4	Adverse Event	1	Adverse Event	2
Death	0	Death	0	Death	0	Death	0
Protocol Violation	0	Protocol Violation	2	Protocol Violation	4	Protocol Violation	0
Noncompliance	0	Noncompliance	0	Noncompliance	0	Noncompliance	0
Lost to Follow-up	0	Lost to Follow-up	1	Lost to Follow-up	0	Lost to Follow-up	0
Voluntary Withdrawal	0	Voluntary Withdrawal	4	Voluntary Withdrawal	1	Voluntary Withdrawal	4
Other*	1	Other*	2	Other*	0	Other*	1

Source: Figure 10-1 of Study Report.

Table 1 Demographics and baseline characteristics for Study PBB00101.

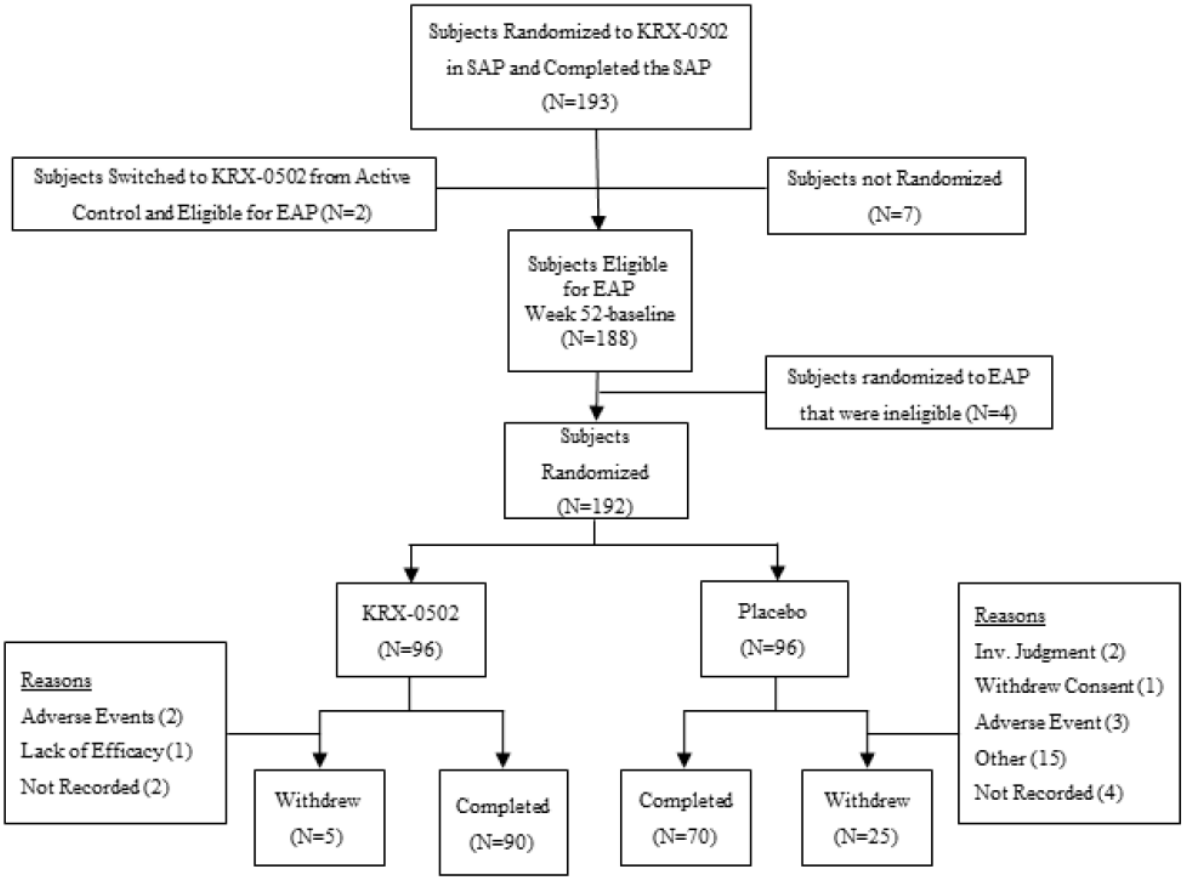


Characteristic		Placebo N = 16	2 g/day N = 31	4 g/day N = 32	6 g/day N = 32	Total N = 111
Gender						
Female	N (%)	7 (43.8)	16 (51.6)	21 (65.6)	14 (43.8)	58 (52.3)
Male	N (%)	9 (56.3)	15 (48.4)	11 (34.4)	18 (56.3)	53 (47.7)
Race						
Caucasian	N (%)	2 (12.5)	2 (6.5)	0 (0.0)	0 (0.0)	4 (3.6)
African-American	N (%)	6 (37.5)	12 (38.7)	16 (50.0)	12 (37.5)	46 (41.4)
Hispanic/ Latino	N (%)	1 (6.3)	1 (3.2)	3 (9.4)	4 (12.5)	9 (8.1)
Asian/Pacific Islander	N (%)	7 (43.8)	16 (51.6)	13 (40.6)	15 (46.9)	51 (45.9)
Other	N (%)	0 (0.0)	0 (0.0)	0 (0.0)	1 (3.1)	1 (0.9)
Age (years)						
	Mean	49.6	48.8	54.6	47.3	50.2
	SD	9.90	15.76	11.48	12.84	13.17
	Median	53	48	55	47	50
	Min, Max	19, 62	22, 81	28, 83	19, 70	19, 83
Height (cm)						
	Mean	168.1	164.3	163.3	166.8	165.3
	SD	12.68	10.36	9.90	10.89	10.75
	Median	170.2	163.0	163.0	167.3	165.1
	Min	147.0	146.0	141.0	147.5	141.0
	Max	190.5	183.0	185.4	185.4	190.5
Weight (kg)[a]						
	Mean	74.3	73.4	76.2	80.1	76.3
	SD	21.30	21.34	22.28	29.29	23.98
	Median	71.1	72.1	75.8	70.5	72.1
	Min	47.3	46.8	45.6	40.3	40.3
	Max	126.7	127.9	139.0	147.0	147.0
Screening (Visit 1) Serum PO_4 (mg/dL)						
	Mean	5.43	5.84	6.26	5.68	5.86
	SD	1.458	1.846	1.437	1.094	1.491
	Median	5.4	5.6	5.8	5.5	5.6
	Min, Max	2.6, 8.5	1.9, 10.0	4.0, 9.6	3.6, 8.4	1.9, 10.0

Source: Table 10-1 of Study report.

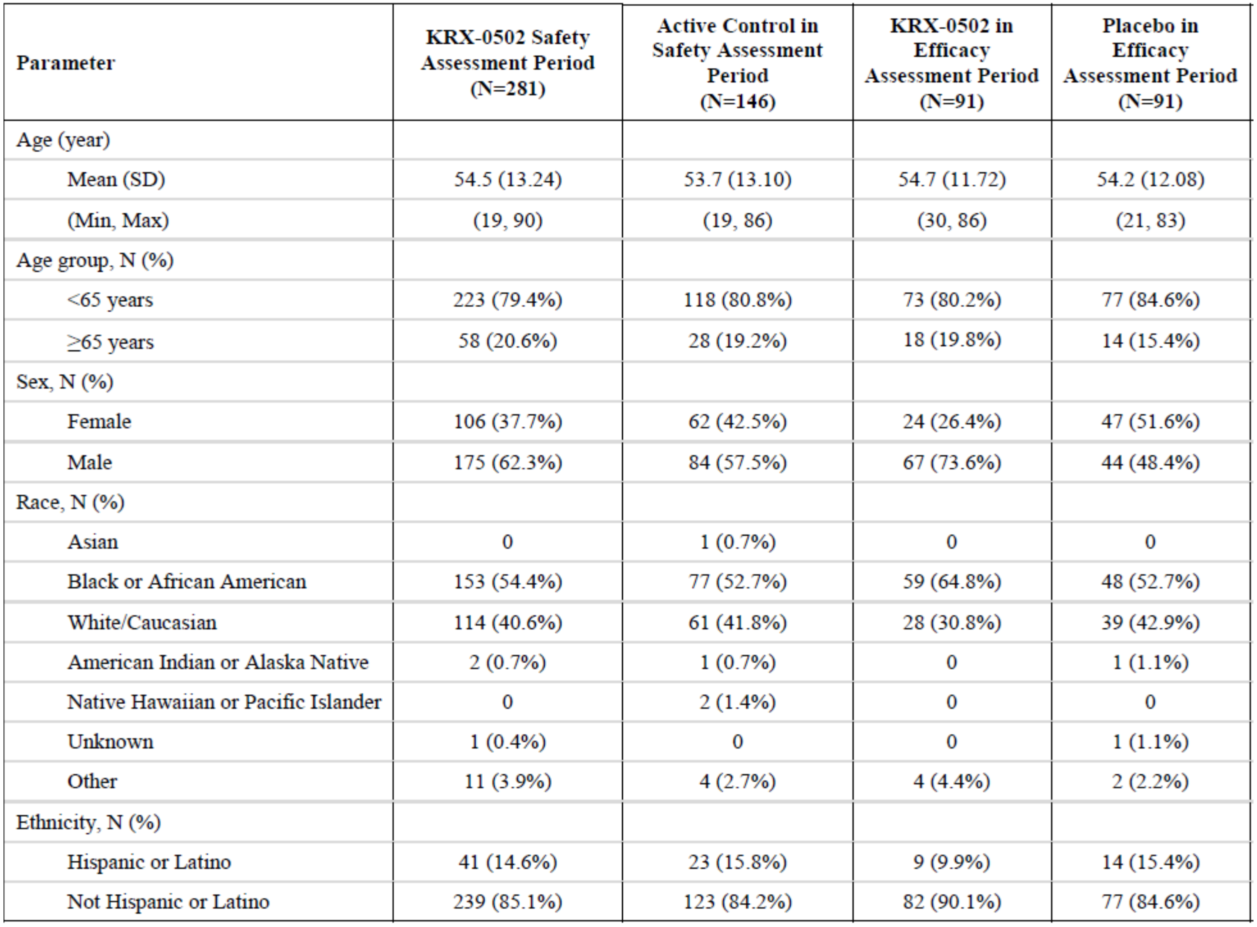
For Study 304, the patient disposition and demographics are shown in Figure 2 and Table 2. There were 192 subjects randomized and in the ITT population. The demographics are only for the subjects that were included in the sponsor's analysis.

Figure 2 Disposition for Study 304


Subjects Randomized to KRX-0502 in SAP and Completed the SAP (N=193)
Subjects Switched to KRX-0502 from Active Control and Eligible for EAP (N=2)
Subjects not Randomized (N=7)
Subjects Eligible for EAP Week 52-baseline (N=188)
Subjects randomized to EAP that were ineligible (N=4)
Subjects Randomized (N=192)
KRX-0502 (N=96)
Placebo (N=96)
Reasons
Adverse Events (2)
Lack of Efficacy (1)
Not Recorded (2)
Withdrew (N=5)
Completed (N=90)
Completed (N=70)
Withdrew (N=25)
Reasons
Inv. Judgment (2)
Withdrew Consent (1)
Adverse Event (3)
Other (15)
Not Recorded (4)

Source: Figure 3 of Study Report.

Table 2 Demographics for Study 304.

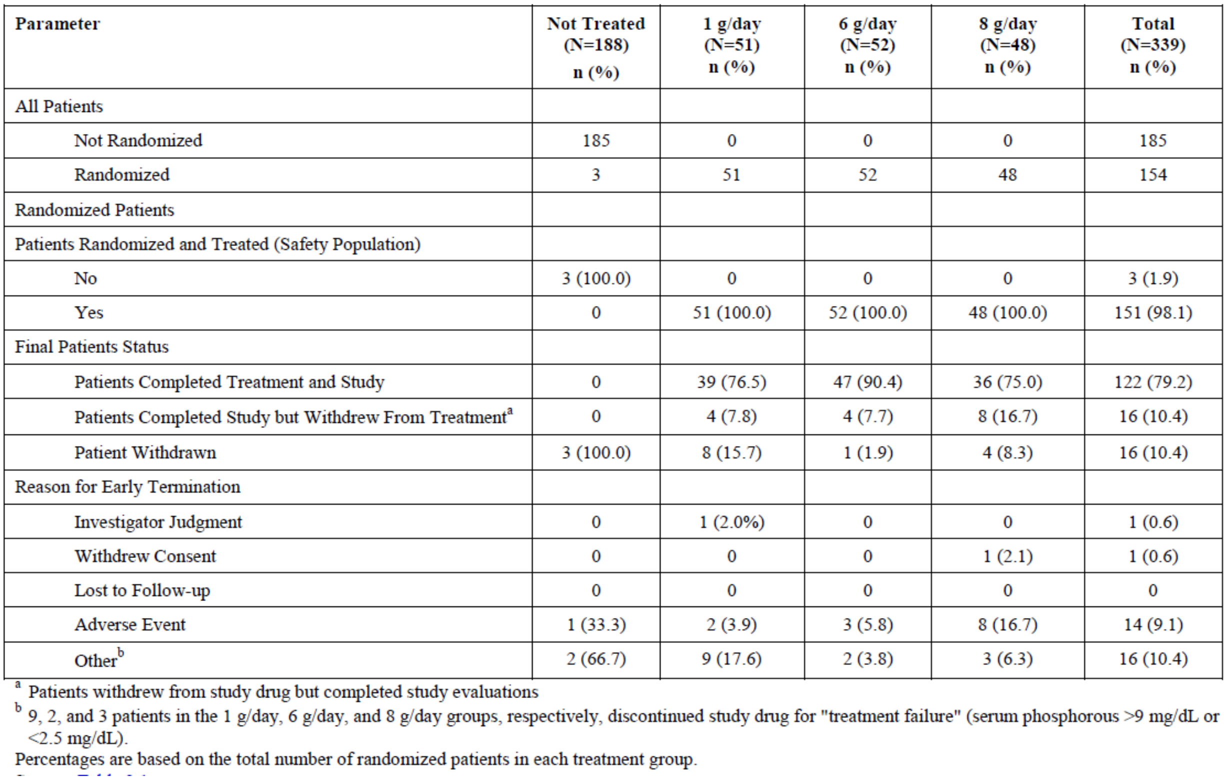


Parameter	KRX-0502 Safety Assessment Period (N=281)	Active Control in Safety Assessment Period (N=146)	KRX-0502 in Efficacy Assessment Period (N=91)	Placebo in Efficacy Assessment Period (N=91)
Age (year)				
Mean (SD)	54.5 (13.24)	53.7 (13.10)	54.7 (11.72)	54.2 (12.08)
(Min, Max)	(19, 90)	(19, 86)	(30, 86)	(21, 83)
Age group, N (%)				
<65 years	223 (79.4%)	118 (80.8%)	73 (80.2%)	77 (84.6%)
≥65 years	58 (20.6%)	28 (19.2%)	18 (19.8%)	14 (15.4%)
Sex, N (%)				
Female	106 (37.7%)	62 (42.5%)	24 (26.4%)	47 (51.6%)
Male	175 (62.3%)	84 (57.5%)	67 (73.6%)	44 (48.4%)
Race, N (%)				
Asian	0	1 (0.7%)	0	0
Black or African American	153 (54.4%)	77 (52.7%)	59 (64.8%)	48 (52.7%)
White/Caucasian	114 (40.6%)	61 (41.8%)	28 (30.8%)	39 (42.9%)
American Indian or Alaska Native	2 (0.7%)	1 (0.7%)	0	1 (1.1%)
Native Hawaiian or Pacific Islander	0	2 (1.4%)	0	0
Unknown	1 (0.4%)	0	0	1 (1.1%)
Other	11 (3.9%)	4 (2.7%)	4 (4.4%)	2 (2.2%)
Ethnicity, N (%)				
Hispanic or Latino	41 (14.6%)	23 (15.8%)	9 (9.9%)	14 (15.4%)
Not Hispanic or Latino	239 (85.1%)	123 (84.2%)	82 (90.1%)	77 (84.6%)

Source: Table 6 of Study Report.

The disposition and demographics for Study 305 are shown in Table 3 and Table 4. Not all the subjects in the ITT population are included in the demographics.

Table 3 Patient Disposition Study 305.



Parameter	Not Treated (N=188) n (%)	1 g/day (N=51) n (%)	6 g/day (N=52) n (%)	8 g/day (N=48) n (%)	Total (N=339) n (%)
All Patients					
Not Randomized	185	0	0	0	185
Randomized	3	51	52	48	154
Randomized Patients					
Patients Randomized and Treated (Safety Population)					
No	3 (100.0)	0	0	0	3 (1.9)
Yes	0	51 (100.0)	52 (100.0)	48 (100.0)	151 (98.1)
Final Patients Status					
Patients Completed Treatment and Study	0	39 (76.5)	47 (90.4)	36 (75.0)	122 (79.2)
Patients Completed Study but Withdrew From Treatment[a]	0	4 (7.8)	4 (7.7)	8 (16.7)	16 (10.4)
Patient Withdrawn	3 (100.0)	8 (15.7)	1 (1.9)	4 (8.3)	16 (10.4)
Reason for Early Termination					
Investigator Judgment	0	1 (2.0%)	0	0	1 (0.6)
Withdrew Consent	0	0	0	1 (2.1)	1 (0.6)
Lost to Follow-up	0	0	0	0	0
Adverse Event	1 (33.3)	2 (3.9)	3 (5.8)	8 (16.7)	14 (9.1)
Other[b]	2 (66.7)	9 (17.6)	2 (3.8)	3 (6.3)	16 (10.4)

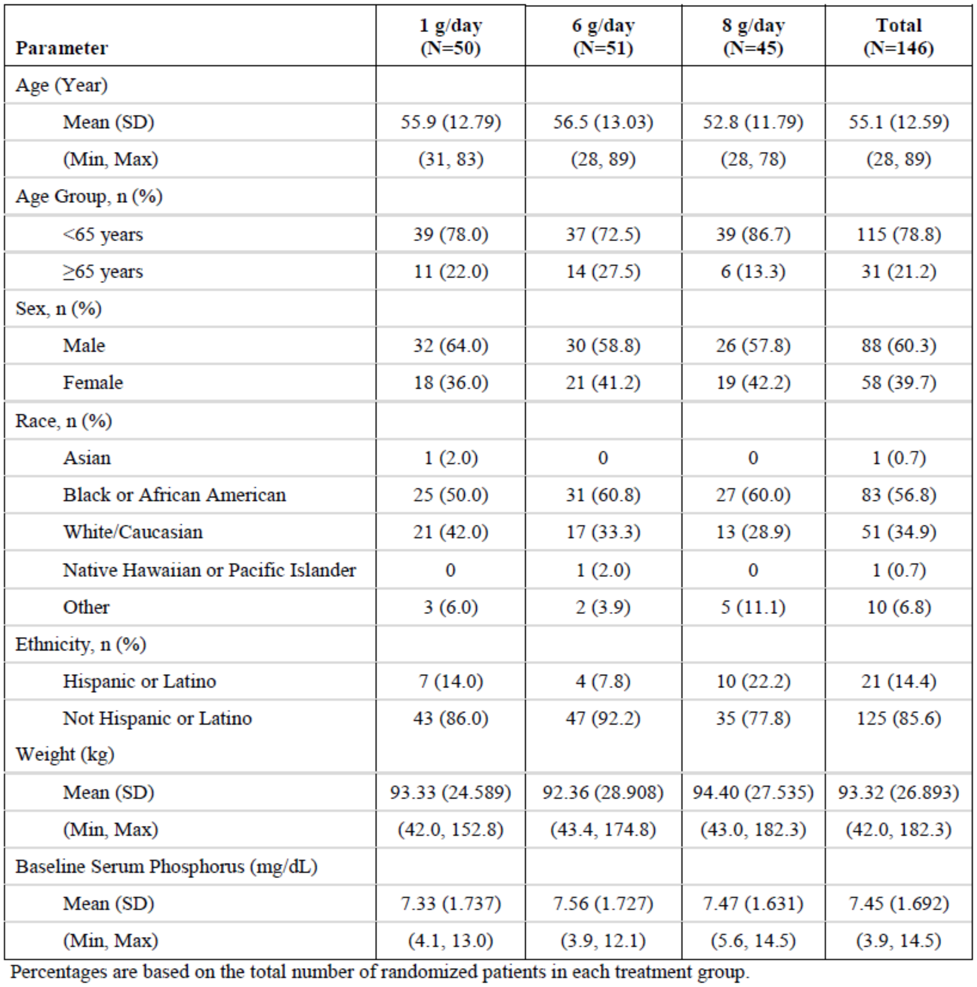
[a] Patients withdrew from study drug but completed study evaluations

[b] 9, 2, and 3 patients in the 1 g/day, 6 g/day, and 8 g/day groups, respectively, discontinued study drug for "treatment failure" (serum phosphorous >9 mg/dL or <2.5 mg/dL).

Percentages are based on the total number of randomized patients in each treatment group.

Source: Table 3 of Study Report

Table 4 Demographics Study 305.



Parameter	1 g/day (N=50)	6 g/day (N=51)	8 g/day (N=45)	Total (N=146)
Age (Year)				
Mean (SD)	55.9 (12.79)	56.5 (13.03)	52.8 (11.79)	55.1 (12.59)
(Min, Max)	(31, 83)	(28, 89)	(28, 78)	(28, 89)
Age Group, n (%)				
<65 years	39 (78.0)	37 (72.5)	39 (86.7)	115 (78.8)
≥65 years	11 (22.0)	14 (27.5)	6 (13.3)	31 (21.2)
Sex, n (%)				
Male	32 (64.0)	30 (58.8)	26 (57.8)	88 (60.3)
Female	18 (36.0)	21 (41.2)	19 (42.2)	58 (39.7)
Race, n (%)				
Asian	1 (2.0)	0	0	1 (0.7)
Black or African American	25 (50.0)	31 (60.8)	27 (60.0)	83 (56.8)
White/Caucasian	21 (42.0)	17 (33.3)	13 (28.9)	51 (34.9)
Native Hawaiian or Pacific Islander	0	1 (2.0)	0	1 (0.7)
Other	3 (6.0)	2 (3.9)	5 (11.1)	10 (6.8)
Ethnicity, n (%)				
Hispanic or Latino	7 (14.0)	4 (7.8)	10 (22.2)	21 (14.4)
Not Hispanic or Latino	43 (86.0)	47 (92.2)	35 (77.8)	125 (85.6)
Weight (kg)				
Mean (SD)	93.33 (24.589)	92.36 (28.908)	94.40 (27.535)	93.32 (26.893)
(Min, Max)	(42.0, 152.8)	(43.4, 174.8)	(43.0, 182.3)	(42.0, 182.3)
Baseline Serum Phosphorus (mg/dL)				
Mean (SD)	7.33 (1.737)	7.56 (1.727)	7.47 (1.631)	7.45 (1.692)
(Min, Max)	(4.1, 13.0)	(3.9, 12.1)	(5.6, 14.5)	(3.9, 14.5)

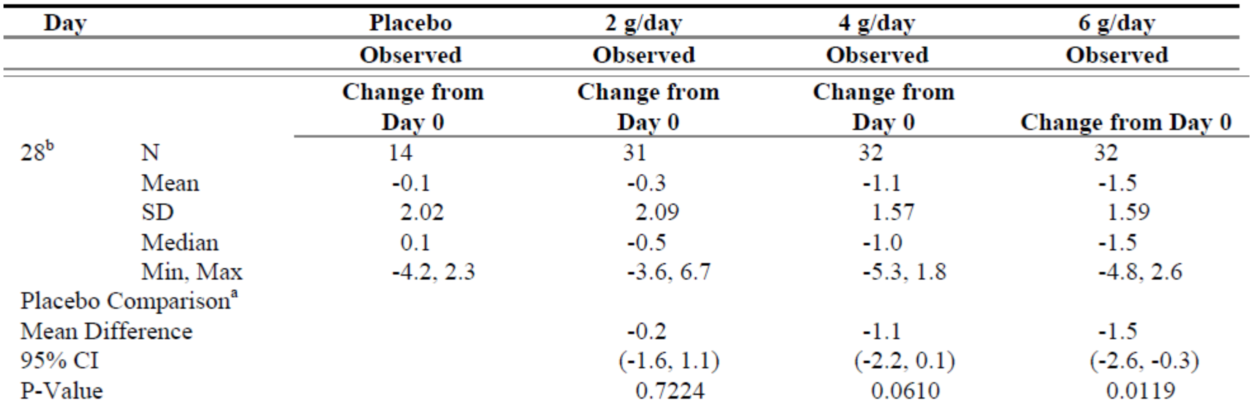
Percentages are based on the total number of randomized patients in each treatment group.

Source Table 4 of Study Report.

1.4.4 Results and Conclusions

For Study PBB00101, the sponsor's results for the primary endpoint are shown in Table 5. The p-value for the F-test does not appear here, so we don't know whether stage 1 was passed for the least significant difference test. Even if that happened, the procedure does not control the familywise error rate, so I cannot say that there was a statistically significant difference between the 6g/day dose and placebo. I found the same number of subjects in the dataset as shown in the table (the other subjects in the ITT population were not in the dataset). However, 4 of these subjects had no baseline value and therefore no change from baseline. That left me with N=13, 31, 30, and 31 respectively. Also, the mean, median, min and max were the same as in the table except for the 2 g/day arm. In that arm, the mean I found is 0.8, the median is -0.4, the min was -3.6 and the max was 33.3. The subject with a change of 33.3 had a baseline of 6.5 and a final value of 39.8. The p-value from the F-test was 0.07, which means the procedure stops with no further comparisons. Despite that, none of the pairwise were significant either in my analysis.

Table 5 Sponsor's results for primary endpoint Study PBB00101.

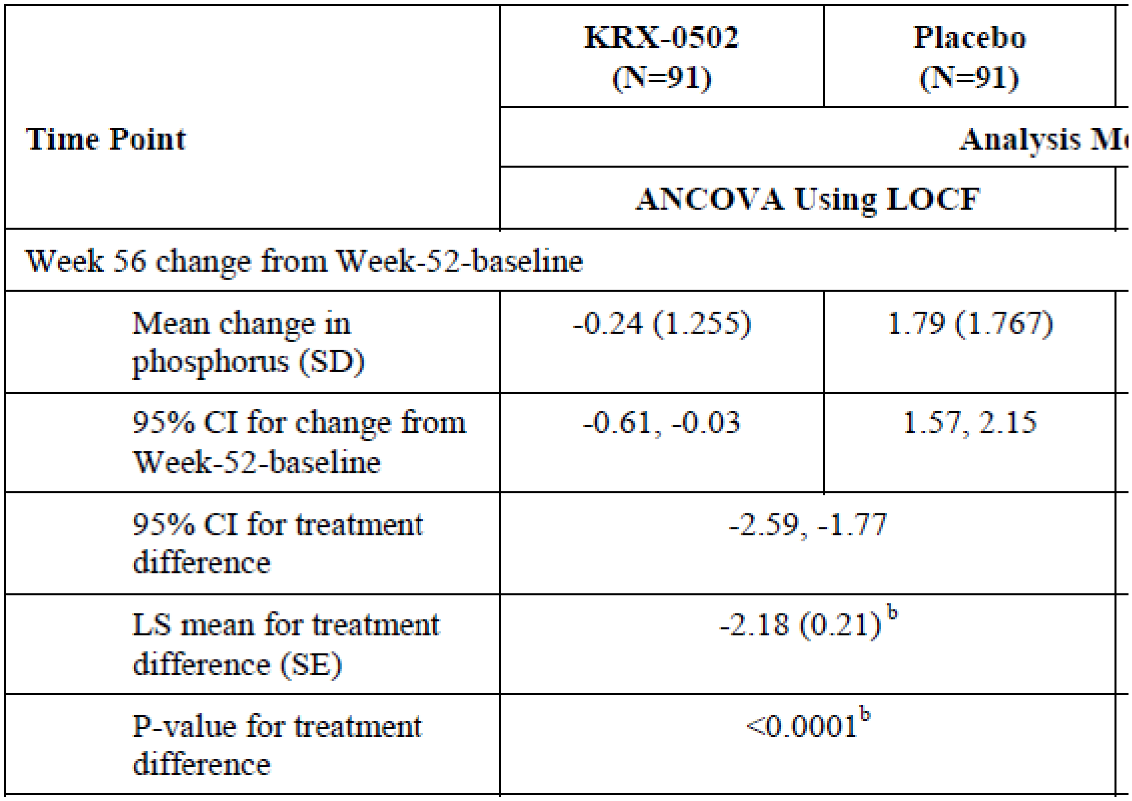


Day		Placebo	2 g/day	4 g/day	6 g/day
		Observed	**Observed**	**Observed**	**Observed**
		Change from Day 0	**Change from Day 0**	**Change from Day 0**	**Change from Day 0**
28[b]	N	14	31	32	32
	Mean	-0.1	-0.3	-1.1	-1.5
	SD	2.02	2.09	1.57	1.59
	Median	0.1	-0.5	-1.0	-1.5
	Min, Max	-4.2, 2.3	-3.6, 6.7	-5.3, 1.8	-4.8, 2.6
Placebo Comparison[a]					
Mean Difference			-0.2	-1.1	-1.5
95% CI			(-1.6, 1.1)	(-2.2, 0.1)	(-2.6, -0.3)
P-Value			0.7224	0.0610	0.0119

Source: Table 11-1 of Study Report.

For Study 304, the sponsor's results for the primary endpoint are shown in Table 6. I could not confirm these results. When I used the sponsors dataset, I found 92 subjects in each group with a change from baseline (efficacy period) serum phosphorous value at Week 56. My point estimate of the mean change from baseline is -2.15 and my 95% confidence interval is (-2.55, -1.74). Also, Figure 3 has the empirical distribution functions. Figure 4 shows the normal probability plot for the residuals. This shows no evidence that the residuals are not normally distributed. Moreover, the Shapiro-Wilks test for normality also does not reject the normality assumption (p=0.73).

Table 6 Sponsor's Results for Study 304



Time Point	KRX-0502 (N=91)	Placebo (N=91)
	Analysis M	
	ANCOVA Using LOCF	
Week 56 change from Week-52-baseline		
Mean change in phosphorus (SD)	-0.24 (1.255)	1.79 (1.767)
95% CI for change from Week-52-baseline	-0.61, -0.03	1.57, 2.15
95% CI for treatment difference	-2.59, -1.77	
LS mean for treatment difference (SE)	-2.18 (0.21) [b]	
P-value for treatment difference	<0.0001[b]	

[b] The LS mean treatment difference and P-value for the change in mean serum phosphorus were calculated via an ANCOVA model with treatment as the fixed effect and Week-52-baseline as the covariate.

Source: Table 10 of Study Report.

Figure 3 Empirical cumulative distribution functions (blue=placebo, red=ferric citrate). (FDA analysis)

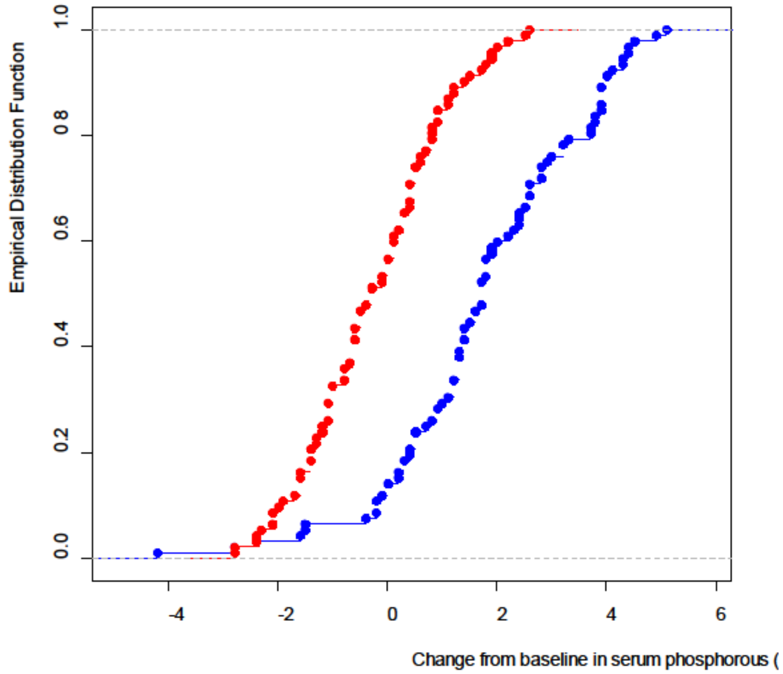
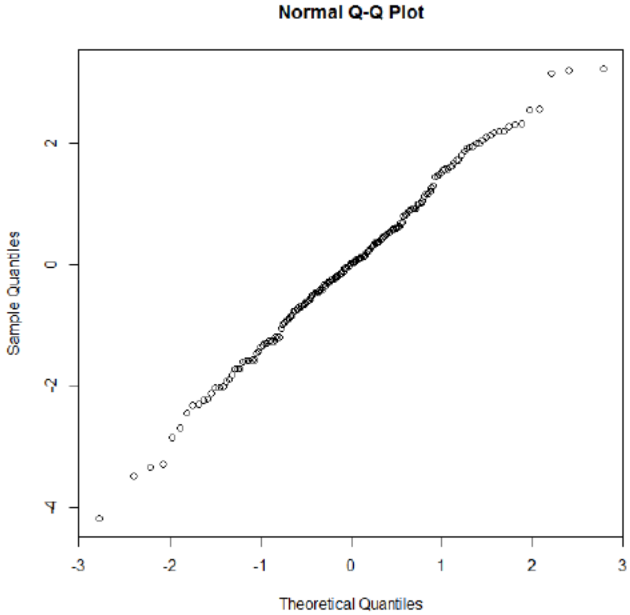
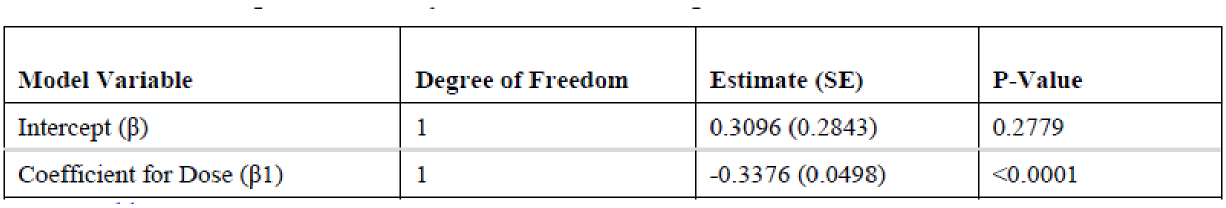


Figure 4 Normal probability plot of residuals from the primary efficacy analysis ANCOVA model. (FDA analysis).



For Study 305, the sponsor's results are shown in Table 7. I did confirm these estimates and p-values using the sponsor's dataset. This analysis uses last observation carried forward for subjects with no final value. That had the greatest impact in the 1 g arm. In the 1 g arm, 12 subjects had LOCF and their mean was +0.5 while the completers had a mean of -0.1. In the two higher dose arms, there were fewer dropouts and the means for the LOCF values were similar to the mean for the completers (-2.4 for dropouts and -1.9 for completers in the 6 g arm; -2.5 and -2.1 for the 8 g arm). The residuals were not normally distributed. The normal probability plot is shown in Figure 5 and the p-value from the Shapiro-Wilks test was 0.0018. The cumulative distribution function for the change from baseline in each arm is shown in Figure 6. According to the Study Report p, 48: "In the ANCOVA analysis, the mean differences in the change from baseline values between the 1 g/day group and the 6 g/day and 8 g/day groups were statistically significant (p=<0.0001), while the difference between the 6 g/day and 8 g/day groups was not (p=0.4864)."

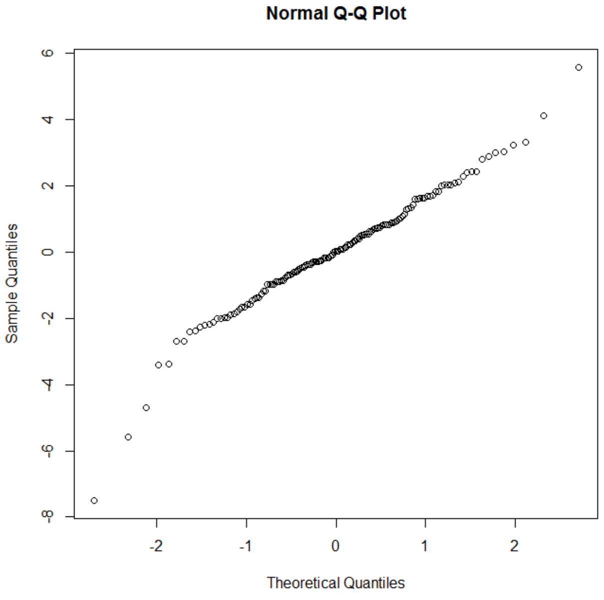
Table 7 Sponsor's results for primary endpoint Study 305.



Model Variable	Degree of Freedom	Estimate (SE)	P-Value
Intercept (β)	1	0.3096 (0.2843)	0.2779
Coefficient for Dose (β1)	1	-0.3376 (0.0498)	<0.0001

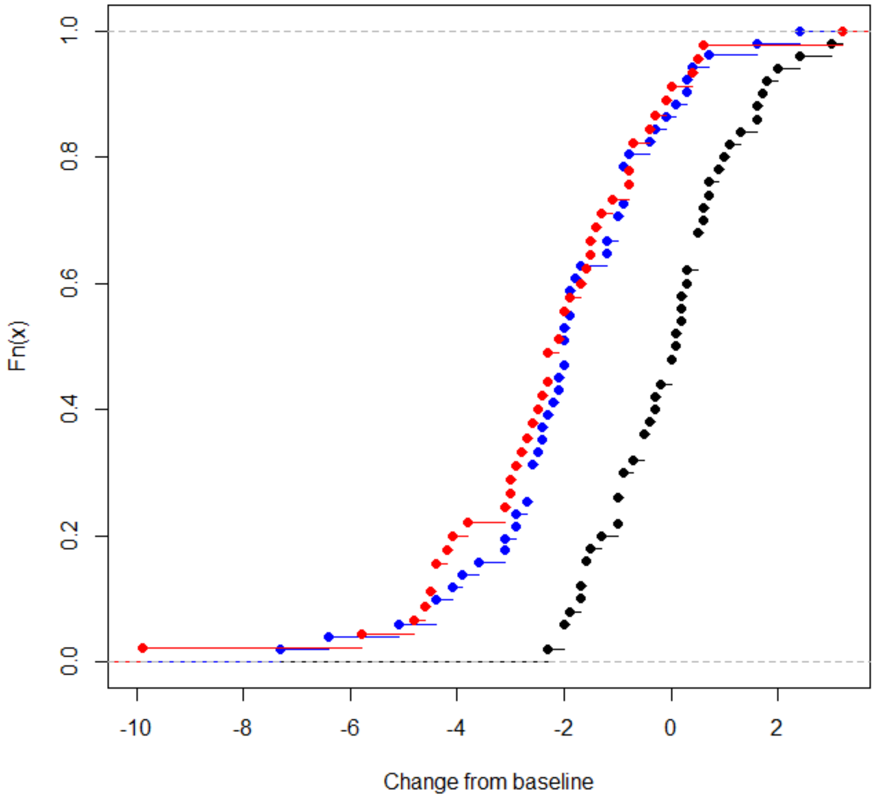
Source: Table 6 of Study Report and confirmed by the FDA.

Figure 5 Normal probability plot for primary analysis in Study 305.



Source: FDA analysis.

Figure 6 Empirical cumulative distribution function Study 305 (black= 1 g, blue = 6 g, red = 8 g)



Source:FDA analysis

1.5 Evaluation of Safety

See clinical review.

1.6 Benefit-Risk Assessment (Optional)

See clinical review.

FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

1.7 Gender, Race, Age, and Geographic Region

See clinical review.

1.8 Other Special/Subgroup Populations

See clinical review.

SUMMARY AND CONCLUSIONS

1.9 Statistical Issues

There were some problems with the design and the analysis of the three phase 3 trials. However, the 6 g per day dose used in the trials was effective at lowering serum phosphorous. Lower doses studied were not effective. There were secondary endpoints specified in the protocol, but no efficacy claims can be made for any of those endpoints.

1.10 Collective Evidence

The 6 g per day dose was effective at lower serum phosphorous in the patients studied.

1.11 Conclusions and Recommendations

Ferric citrate 6 g per day dose was effective at lowering serum phosphorous in the patients studied.

1.12 Labeling Recommendations (as applicable)

Secondary endpoints in the trials can be described with means and confidence intervals to give readers an idea of what to expect and for safety reasons. (b) (4) [redacted]

APPENDICES

The F-test for equality of means, power

The F-test is the ratio of the mean squares for treatments to the mean squared error. There is no guarantee that if the F-test is significant, then at least one pairwise comparison would be significant. For example, suppose the sample sizes for the 4 groups are 8, 8, 11, and 43, the MSE is 1.4^2 and the sample means are 0.69, 0.69, 0.482, and -0.38. In that case, the F-statistic will be ((8*0.69^2+8*0.69^2+11*0.482^2+43*0.38^2)/3)/1.4^2 =2.79 and the p-value will be 0.0475. However, the pairwise comparison p-values are 1 (group 1 vs. 2), 0.75 (1 vs. 3 and 2 vs. 3), 0.051 (1 vs. 4 and 2 vs. 4), and 0.073 (3 vs. 4).

Suppose there are *K* groups where the means in the groups are $\mu_1, \dots, \mu_K$ and the sample sizes are $n_1, \dots, n_K$. Let $\bar{X}_k$ denote the sample mean for group *k*. Then, we have $\bar{X}_k \sim N\left(\mu_k, \frac{\sigma^2}{n_k}\right)$. The grand mean is $\bar{\bar{X}} = \frac{\sum_{k=1}^K n_k \bar{X}_k}{\sum_{k=1}^K n_k} \sim N\left(\frac{\sum_{k=1}^K n_k \mu_k}{\sum_{k=1}^K n_k}, \frac{\sigma^2}{\sum_{k=1}^K n_k}\right)$.

Now,

$$\sum_{k=1}^{K} n_k \left(\bar{X}_k - \bar{\bar{X}}\right)^2$$

$$= \sum_{k=1}^{K} n_k {\bar{X}_k}^2 - 2\sum_{k=1}^{K} n_k \bar{X}_k \bar{\bar{X}} + \sum_{k=1}^{K} n_k \bar{\bar{X}}^2$$

$$= \sum_{k=1}^{K} n_k {\bar{X}_k}^2 - 2\bar{\bar{X}}\sum_{k=1}^{K} n_k \bar{X}_k + \bar{\bar{X}}^2 \sum_{k=1}^{K} n_k$$

$$= \sum_{k=1}^{K} n_k {\bar{X}_k}^2 - \bar{\bar{X}}^2 \sum_{k=1}^{K} n_k$$

Hence,

$$E\left\{\sum_{k=1}^{K} n_k {\bar{X}_k}^2 - \bar{\bar{X}}^2 \sum_{k=1}^{K} n_k\right\} = E\left\{\sum_{k=1}^{K} n_k {\bar{X}_k}^2\right\} - E\left\{\bar{\bar{X}}^2 \sum_{k=1}^{K} n_k\right\}$$

$$= \sum_{k=1}^{K} n_k E\left\{{\bar{X}_k}^2\right\} - E\{\bar{\bar{X}}^2\} \sum_{k=1}^{K} n_k$$

$$= \sum_{k=1}^{K} n_k \{Var\{\bar{X}_k\} + [E\bar{X}_k]^2\} - \left\{Var\{\bar{\bar{X}}\} + \left[E\bar{\bar{X}}\right]^2\right\} \sum_{k=1}^{K} n_k$$

$$= \sum_{k=1}^{K} n_k \left\{\frac{\sigma^2}{n_k} + {\mu_k}^2\right\} - \left\{\frac{\sigma^2}{\sum_{k=1}^{K} n_k} + \left[\frac{\sum_{k=1}^{K} n_k \mu_k}{\sum_{k=1}^{K} n_k}\right]^2\right\} \sum_{k=1}^{K} n_k$$

$$= \{K-1\}\sigma^2 + \sum_{k=1}^{K} n_k \{{\mu_k}^2\} - \frac{\{\sum_{k=1}^{K} n_k \mu_k\}^2}{\sum_{k=1}^{K} n_k}$$

The F-statistic has a noncentral F distribution with K-1 degress of freedom in the numerator, $\sum_{k=1}^{K} n_k - K$ degrees of freedom in the denominator and noncentrality parameter $\frac{1}{\sigma^2}\sum_{k=1}^{K} n_k\{{\mu_k}^2\} - \frac{\{\sum_{k=1}^{K} n_k \mu_k\}^2}{\sigma^2 \sum_{k=1}^{K} n_k}$.

For example, under the assumptions used to design the trial, $K = 4$, $n_1 = 10$, $n_2 = n_3 = n_4 = 20$, $\mu_1 = 0, \mu_2 = 0.8, \mu_3 = 1.2, \mu_4 = 1.6$, and $\sigma = 1.4$. The numerator degrees of freedom is 3, the denominator degrees of freedom is 66, and the noncentrality parameter is
20*(0.8^2+1.2^2+1.6^2)/1.4^2-((20*(0.8+1.2+1.6))^2)/(70*1.4^2)

The power for the first stage is 71.2%. The quantiles of the central F distribution and the upper tail probability of the noncentral F distribution can be found in R as follows:

(b) (4)

[redacted]

This can also be found by simulation as follows:

(b) (4)

[redacted]

On the other hand, if the allocation had been more balanced, say 16 in the placebo group and 18 in each dose group, then the power for this stage would have been greater than 80%:

(b) (4)

[redacted]

Fisher's least significant difference procedure, error rate

Fisher's least significant difference procedure has two stages. In stage one, we look at the F-test. If that is significant, then we look at the pairwise comparisons. It is well known that if there are more than 3 groups, the procedure does not control the familywise error rate. Table 1 of Hayter (Hayter, Anthony J. "The maximum familywise error rate of Fisher's least significant difference test." *Journal of the American Statistical Association* 81.396 (1986): 1000-1004.) shows that the upper bound for the error rate is 0.1222 with 4 groups and using a targeted type 1 error rate of 0.05. Note that this calculation includes all pairwise comparisons and not just the pairwise comparisons with the control group. The following modification of Hayter's Theorem 1 deals with this situation.

Theorem. Suppose group 1 is the control group and the remaining K-1 treatment groups will be compared with the control group by starting with an α-level F-test from a one-way analysis of variance at stage 1. If the equality of all means is rejected by the F-test, all pairwise comparisons with the control at stage 2 are done without adjustment. Suppose there are ν degrees of freedom for the error. Then, the maximum familywise error rate is

$$\alpha^*(K, \nu, \alpha) = P\left[\max_{i=1,..,K-2} |T_i| > t_{\alpha/2,\nu}\right]$$

where $(T_1, \dots, T_{K-2})'$ has the multivariate t-distribution with ν df and correlation matrix has elements $\left(1+\frac{n_1}{n_{i+1}}\right)^{-1}$ off the diagonal and $t_{\alpha/2,\nu}$ is the upper $\alpha/2$ point of the t-distribution with ν df.

The proof follows from looking at a limiting case where one of the treatment groups has a very large mean difference from the control group and all the remaining K-2 groups are the same as the control. The F-test in stage one is always large and we always go to stage 2.

For example, using the assumptions to design the trial, the maximum error rate is about 8.4% as found either using the formula in the theorem or by simulation.

(b) (4)

[redacted]

Fisher's least significant difference procedure, power for claiming high dose is superior

In order to find the power of the least significant difference test, we note that two things have to happen to claim the high dose is superior. First, the F-test would have to be significant. We know the chance that would happen is 71.2%. The second thing that has to happen is that the t-test for the pairwise comparison has to be significant. The power is about 69% and can be found by simulation.

(b) (4)

Fisher's least significant difference procedure, closed test

There are two simple ways to make Fisher's least significant difference test have the correct familywise error rate. One is based on the idea like Hayter used in his paper, i.e. use a different targeted α in order to make α^*=0.05. In the case of the assumptions used at the planning stage, this targeted α would be 0.0289. But, that may have to be revisited when the trial is over based on the actual sample sizes attained in each arm. I will call this procedure 1.

A second way is to use a closed test procedure to test each combination of equality of means using an F-test using the data only from the arms involved in the intersection. For example to test $\mu_1 = \mu_2 = \mu_4$ we can use the F-test from the one way analysis of variance using only the data from the control group, the low dose group and the high dose group. All the F-tests are done at level α=0.05. For the denominator in the F-test, we can use the MSE from all groups combined or the MSE from the groups used in the numerator- both are OK as long as it is decided in advance. In order to conclude the high dose is significantly different from the control group using this closed test procedure, we would need the following four F-tests to be significant for these combinations of groups: all 4 groups, all groups except low, all groups except medium, control and high alone (equivalent to the t-test). This is procedure 2.

Procedure 3 is Holm's procedure (closure of the Bonferroni test). Procedure 4 is Hochberg's step-up procedure. Procedure 5 is the closure of Dunnett's test (Dunnett's step down procedure). Procedure 6 is the closure of the test based on pooling all doses. Procedure 7 is the closure of the test that uses simple linear regression to test for a dose response. Based on the assumed means, the most powerful choice of levels to use for the x coordinates in the linear regression is not the actual numerical dose levels, but rather 0 for control, 2 for low dose, 3 for middle dose and 4 for high dose. That's what we will use. Procedure 8 tests each dose at level α sequentially starting from the high dose.

I considered 3 different scenarios. In all scenarios, the mean for the placebo arm is 0 and the standard deviation in all arms is 1.4 as postulated in the SAP. Scenario 1 is the scenario in the SAP where the 3 doses have means 0.8, 1.2 and 1.6. Scenario 2 is the scenario where all doses are equally effective and have mean 1.2. Scenario 3 is where only one dose is different from placebo and has mean 1.6.

```
#adjust means in x2  x3  x4 for other scenarios
```

(b) (4)

[redacted]

1 Page(s) has been Withheld in Full as b4 (CCI/TS) immediately following this page

(b) (4)

	Scenario 1	Scenario 2	Scenario 3
Procedure 1 (LSD with adjusted α)	0.613 (1.25)	0.419 (1.07)	0.748 (0.777)
Procedure 2 (closure of test using LSD)	0.695 (1.48)	0.511 (1.32)	0.817 (0.866)
Procedure 3 (Holm's)	0.724 (1.45)	0.605 (1.43)	0.689 (0.717)
Procedure 4 (Hochberg)	0.730 (1.48)	0.621 (1.51)	0.689 (0.719)
Procedure 5 (step down Dunnett)	0.733 (1.47)	0.616 (1.45)	0.698 (0.727)
Procedure 6 (closure of pooled doses)	0.655 (1.41)	0.662 (1.60)	0.193 (0.224)
Procedure 7 (closure of linear trend test)	0.815 (1.63)	0.507 (1.20)	0.736 (0.769)
Procedure 8 (sequential testing)	0.828 (1.66)	0.587 (1.436)	0.828 (0.861)

The table above shows the estimated probability of rejecting at least one hypothesis for each procedure and each scenario. The number in parentheses is the average number of rejected hypotheses. Procedures 1 and 2 are both ways of fixing the LSD procedure to make it control the FWER for this testing scenario. In these 3 scenarios, procedure 1 is not as good as procedure 2. Both of these procedures are better than any of the other six in scenario 3. That is the scenario where only one dose is effective; pooling or using a trend test makes it harder to find any significant result.

In scenario 1, there is a perfect linear relationship, so Procedure 7 should win hands down. But, I was surprised to see that the sequential testing procedure (Procedure 8) still manages a small victory over Procedure 7 in that scenario. The explanation for that is because using Procedure 7, in order to reject the elementary hypothesis for the high dose compared to placebo, it is required that four p-values are statistically significant (the trend test for all combinations of arms including at least placebo and the high dose). But, for Procedure 8, only one of those p-values needs to be significant (and that p-value is the same as one of the four for Procedure 7). Proc. 8 does not dominate Proc. 7 even in this scenario (i.e. it is possible for Proc. 7 to reject a hypothesis for the middle or low dose but not for the high dose). I think Procedure 7 might be more powerful than Procedure 8 if there were more subjects allocated to lower doses compared to the high dose or in certain alternatives where there was not a perfect linear relationship.

Scenario 2 is tailor made for Procedure 6 because all the doses are equally effective- pooling the doses improves the precision of the estimates. Likewise, Scenario 3 is tailor made for Procedure 8. Procedures 7 and 8 would seem to always be reasonably good procedures unless there was a U-shaped dose response. Holm's procedure is never the best, but it is never very far from the best. Head to head comparisons (with other procedures that control the FWER) like this shatter the misconception that the Bonferroni test is very conservative. Both the step down Dunnett procedure (Procedure 5) and Hochberg's procedure (Procedure 4) dominate Holm's procedure in a pointwise sense, i.e. for any set of data, if Holm's procedure rejects a hypothesis, then so will Dunnett's procedure and so will Hochberg's. However, in these scenarios there is not much gain in using Dunnett's or Hochberg's procedure. Also, note that all 8 of these procedures admit adjusted p-values. These adjusted p-values are very helpful and should be reported and used more often.

Lastly, the power can be improved by allocating the subjects differently. For example, under scenario 1, the power of Holm's procedure could be improved to 89.3% by allocating 22 subjects to placebo and the remainder equally among the 3 doses (16 per group).

(b) (4)

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

/s/

JOHN P LAWRENCE
04/22/2014

HSIEN MING J HUNG
04/22/2014

STATISTICS FILING CHECKLIST FOR A NEW NDA/BLA

NDA Number: 205874 **Applicant: Keryx** **Stamp Date: 8/7/2013**

Drug Name: KRX-0502 (ferric citrate) **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		I could not find any
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.

Content Parameter (possible review concerns for 74-day letter)	Yes	No	NA	Comment
Designs utilized are appropriate for the indications requested.				unknown
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.				unknown
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.				unknown

STATISTICS FILING CHECKLIST FOR A NEW NDA/BLA

Reviewing Statistician Date

Supervisor/Team Leader Date

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

/s/

JOHN P LAWRENCE
09/30/2013

HSIEN MING J HUNG
10/01/2013