

STATISTICAL REVIEW AND EVALUATION

NDA #: 21-272
Applicant: United Therapeutics Corporation
Name of Drug: Uniprost™ (treprostinol sodium)
Indication: Treatment for pulmonary arterial hypertension
Document reviewed: Volumes 2.1, 2.24, and 2.27-2.50
Date of submission: October 16, 2000
Statistical Reviewer: John Lawrence, Ph.D. (HFD–710)
Medical Reviewer: Abraham Karkowsky, M.D. (HFD–110)

1. Introduction

Uniprost™, or UT-15, is a structural analog of epoprostenol (Flolan®) with a similar pharmacological profile. Flolan has been approved for the chronic treatment of patients with primary pulmonary hypertension and has been used to treat patients with pulmonary hypertension associated with other conditions. Unlike Flolan, Uniprost is chemically stable at room temperature and it has a longer half-life than Flolan. For these reasons, the sponsor believes that Uniprost would improve risks associated with treatment and should be considered as an alternative therapy for pulmonary arterial hypertension (PAH). There were two Phase III studies conducted by the sponsor to support the safety and efficacy of the treatment- Studies P01:04 and P01:05.

2. Study Design

The design of Studies P01:04 and P01:05 were identical. Each study was a multicenter, double-blind, parallel-group study. Patients between the ages of 8 and 75 were eligible for each study if they had a current documented diagnosis of PAH. On Day 1 of the Screening Period, routine baseline assessments were performed. On Day 2, the baseline Six-Minute Walk Test was administered. Patients whose baseline exercise capacity was less than 50 m or greater than 450 m were excluded from entering the Treatment Phase. Patients were randomized within strata determined by dichotomous levels of etiology of the disease (primary PH/ secondary PH) and baseline exercise capacity (low = 50-150 m/ high = 151-450 m). Randomization among patients with secondary PH was further stratified by use of vasodilators. The 12-Week Treatment Phase began immediately after baseline assessments and randomization on Day 2. Six-Minute Walk Tests were scheduled at Day 9, Day 44, and Day 87.

In order to select the sample size, an estimate of the expected treatment effect was made using data from a study using the active treatment Flolan. The treatment effect in the Flolan study was an improvement of 45 m in change from baseline compared to placebo. Assuming a treatment effect for Uniprost of 55 m over placebo, it was expected that a sample size of 210 in a single study would provide a 95% chance of rejecting the null hypothesis at $\alpha=0.05$. So, the actual sample sizes of 224 in Study P01:04 and 246 in P01:05 should have been adequate if the estimate of the treatment effect was reasonable.

Of the 470 patients randomized in both studies, 233 were assigned to receive the active treatment and 237 received the placebo. One patient assigned to the placebo group never received treatment. The remaining 469 patients constitute the modified Intent-To-Treat population (*mITT*). In the *mITT* population, the average age was 44.5, there were 382 females and 87 males, 396 Caucasians, 21 Blacks, 13 Asians, 33 Hispanics, 2 Native Americans, and 4 from a race other than those listed.

Patients received an initial dose of Uniprost or placebo of 1.25 ng/kg/min. This was the maximum allowable dose at the end of Week 1, but could be decreased to a tolerated dose. Following Week 1, patients were contacted weekly to assess whether changes in dosage were warranted. The dose was increased if symptoms did not improve and was reduced at the onset of any adverse experience that was judged to be related to study drug or there were changes in hemodynamics, vital signs, or clinical signs or symptoms that warranted reductions.

3. Primary Efficacy Variable

The primary endpoint of the two studies was change in exercise capacity at Week 12 as measured by distance walked in six minutes.

4. Secondary Efficacy Variables

Three principal reinforcing endpoints were prospectively identified: signs and symptoms of PAH, Dyspnea-Fatigue Rating, and an assessment of the occurrence of death, transplantation, or discontinuation from study drug due to clinical deterioration. Hemodynamics and Borg Dyspnea Score were defined as secondary endpoints.

5. Protocol Specified Planned Statistical Analysis

The primary analysis was a nonparametric analysis of covariance using the *mITT* population and the pooled data from the two studies. There is no provision for analyzing patients in the *mITT* population with no post-baseline walking distances. First, separate least squares regression models were fit to the Week 1, Week 6, and Week 12 distance walked as a

function of baseline distance walked, center, etiology of PH (primary or secondary), and vasodilator use at baseline. On p. 30 of the Final Analysis Plan [Vol. 2.33] an additional covariate for use of steroids to treat PHT at baseline is included. However, this covariate is not listed on p. 90 of the Study Report [Vol 2.27]. Standardized mid-ranks (also known as modified ridit scores), defined as $\text{rank}/(\# \text{ observations} + 1)$, were determined from the residuals from the ordinary least squares regression. Missing values were imputed by carrying forward the standardized midrank from the last valid observation. The lowest standardized rank (0) was assigned to deaths, transplants, or clinical deterioration. Standardized mid-ranks were then recalculated and compared between treatment groups using the Cochran-Mantel-Haenzsel procedure mean score statistic with table scores stratified by the stratification factors used during randomization [*Source: Vol. 2.27 pages 88-92*].

According to a letter from the sponsor dated March 23, 2000, the analysis plan was modified slightly: if an exercise test is missing because “patient was too critically ill”, the lowest standardized rank will be used for the nonparametric analysis.

The null hypothesis of no treatment difference was to be rejected if the two-sided p-value from the pooled analysis was less than 0.049 and both of the p-values from the individual studies were less than 0.049. This is the traditional standard for two confirmatory studies with an adjustment because the sponsor wanted to test the null hypothesis within the subgroup of PPH patients at $\alpha=0.001$. If the global null hypothesis was not rejected, then the protocol states the null hypothesis would be rejected if the p-value from the pooled analysis was less than 0.01 and at least one of the analyses from a single study had a p-value less than 0.049. This gives the sponsor a second chance to reject the null hypothesis. This issue is discussed more thoroughly in Section 7.

6. Characteristics of Patients at Baseline and Dropouts

The baseline characteristics of the patients in the two treatment arms for the two studies are in Table 6.1. There was no significant difference between the two treatment arms with respect to any of these characteristics.

Table 6.1 Characteristics of the patients in the two groups at baseline. For continuous variables, this table shows the group mean \pm standard error of mean. [Source: Vol. 2.27, Tables 11.2.1, 11.2.2.1, and 11.2.2.4]

Characteristic	Uniprost Group	Placebo Group
N	233	237
Age (years)	44.6 \pm 1	44.4 \pm 1
Male (%)	15.5	21.6
Caucasian (%)	85	84
Years with PAH	4.3 \pm 0.5	3.3 \pm 0.4
NYHA Class II (%)	11	12
NYHA Class III (%)	82	82
NYHA Class IV (%)	8	7
Primary PH (%)	41	41
PAH associated with Scleroderma (%)	5	5
"" Limited Scleroderma (%)	6	3
"" Mixed Connective Tissue Disease (%)	3	4
"" Systemic Lupus Erythematosus (%)	3	8
"" Overlap Syndrome (%)	0.4	0.8
"" congenital systemic-to-pulmonary shunts (%)	25	22
Distance walked at baseline (m)	326 \pm 5.5	327 \pm 5.7

In the Uniprost group, 200 patients completed the 12 weeks of treatment. 6 patients discontinued due to clinical deterioration, 18 withdrew for adverse experiences, 7 died on study drug, and 2 withdrew consent. In addition to the 7 patients who died on Study Drug, 2 more patients died within 12 weeks from being randomized after they had withdrawn from the study. A total of 13 patients withdrew for death, transplantation, or clinical deterioration [Source: Vol. 2.27 Tables 10.1A, 11.4.1.2.3 and 12.5.5.].

In the placebo group, 221 patients completed the 12 weeks of treatment, 6 patients deteriorated, 1 withdrew for adverse experiences, 7 died on study drug, 1 patient had a transplant, and 1 withdrew consent. In addition to the 7 patients who died on Study Drug, 3 more patients died within 12 weeks from being randomized after they had withdrawn from the study. A total of 16 patients withdrew for death, transplantation, or clinical deterioration [Source: Vol. 2.27 Tables 10.1A, 11.4.1.2.3 and 12.5.5.].

In the *mITT* population, one patient did not have any exercise tolerance measurements post baseline, 455 patients had a Six-Minute Walk Test at Week 1, 468 patients had a Six-Minute Walk Test at Week 6, and 419 patients had a Six-Minute Walk Test at Week 12 [Source: Vol. 2.27 Tables 11.4.1.1.2B, 11.4.1.1.4G, and 11.4.1.1.4H].

7. Statistical Comments About the Analysis Plan

The decision to impute a worst possible score for those patients who died or discontinued for transplantation or clinical deterioration is reasonable. A nonparametric analysis is suitable because we can then assign a worst score, or a rank of 0, for these patients. It might be more appropriate to rank all the patients who died below those who discontinued for clinical deterioration and those patients, in turn, below all those who completed the study. The relative ranks among those patients who died and among those patients who discontinued for clinical deterioration can be determined by length of time in the study. However, there were roughly the same number of patients in each arm who died and discontinued for clinical deterioration, so this will not likely have an impact here.

However, there was a substantial imbalance in the number of patients who discontinued the study due to serious adverse experiences (18 versus 1). These patients all had their last rank carried forward in the analysis, rather than a worst rank assigned. When it is not entirely clear whether serious adverse experiences can also be associated with clinical deterioration or vice versa, assigning these patients a worst rank may be needed. As a supportive analysis, it may be illustrative to see the impact of using the last rank carried forward for these patients by assigning a rank of 0 for these patients also.

A more important issue is the overall Type I error rate for the proposed analysis in this submission. First, consider the traditional standard for approval at the FDA based on two confirmatory trials. Even if the efficacy of a treatment is shown convincingly in one study, the agency likes to see replication in a second study because we will then be in a better position to infer that the results generalize to the entire population of patients with the disease. The overall Type I error rate (or false positive rate) is the chance that both studies will have a p-value less than 0.05 and the results of both studies are in the same direction. If the treatment effects in the two studies are identically 0, then the chance that both p-values will be less than 0.05 and both treatment effects are in the same direction is 0.00125^1 . For this reason, the Division of Cardio-Renal Drugs has often advised sponsors that one study with a p-value less than 0.00125 may be sufficient for approval. When there is no between trial variability in the treatment effect, these two standards are indeed equivalent.

Now, consider the approach that is used in this submission. We will reject the null hypothesis of no treatment effect under either of these two circumstances:

- 1) both studies have p-values <0.049 and the pooled data has a p-value <0.049
- 2) either study has a p-value <0.049 and the pooled data has a p-value <0.01

¹ P[first p-value <0.05 and second p-value <0.05 and direction is the same]
= P[first p-value <0.05] * P[second p-value <0.05] * P[direction is the same] = $0.05*0.05/2 = 0.00125$

Furthermore, if neither 1) nor 2) occurs, we will reject the null hypothesis of no treatment effect in the subgroup of PPH patients under the following condition:

- 3) the data on PPH patients pooled from both studies has a p-value <0.001 .

According to this reviewer's simulation, if 40% of the patients have PPH then the overall Type I error rate for the criteria used in this submission is 0.01. However, it is widely recognized that even when the designs are identical, the treatment effect may vary from study to study. If there is any between trial variability in the treatment effect, the chance that any of the three conditions will hold is inflated. The appendix of this review illustrates this in more detail.

An overall Type I error rate of 0.01 is already more liberal than the error rate of 0.00125 for the traditional FDA approach. Now, if we include other conditions that were not pre-specified under which the sponsor can claim that efficacy was demonstrated, the Type I error rate will be inflated even further. For instance, suppose one p-value from an individual study had been 0.009 and the second had been 0.10 and the p-value from the pooled data was 0.015. Someone might look at this and argue that the drug should be approved because Condition 2 was almost satisfied since the p-value from one study was significantly less than 0.049 and the second was in the right direction and the p-value from the pooled data was really close to 0.01. However, if we allow this to happen, then it is possible that our minds cannot stretch wide enough to imagine all of the possible scenarios that are "close enough" and therefore, we have no hope of calculating, much less controlling, the real Type I error rate.

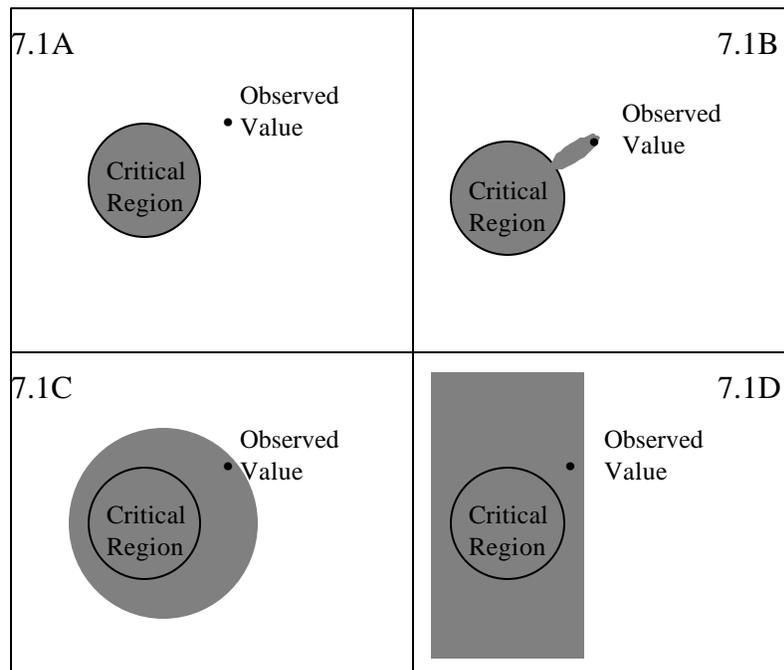
There are many possible ways to calculate an overall p-value from this experiment and therefore, there is no correct way to do this. In order to make things simple, assume that the statistic is univariate and has a standard normal distribution under the null hypothesis. We create a test by prospectively specifying a critical region, which defines the set of values for the statistic for which the null hypothesis will be rejected. If the significance level is 0.05, then the probability of observing a value in the critical region is 0.05 if the null hypothesis is true. Now, suppose we prospectively define the critical region to be all numbers greater than 1.96 in absolute value, but when we actually do the experiment, we observe a value of 1.7. The p-value is the probability of observing something as extreme or more extreme than 1.7. In this case, nobody would argue that any value greater than 1.7 in absolute value is more extreme, so the p-value is $2 \Phi(-1.7) = 0.089$.

The situation here is more complex because the outcome is not univariate. There are outcomes from two studies and the outcome of the data pooled together and the outcome from the analysis of the PPH subgroup. When the outcome is not univariate, it is harder to see what is more extreme than what was actually observed. Clearly, if the observed value is not in the critical region, then anything in the critical region would have to be considered more extreme. The approach that would give the smallest p-value is to assume that only the exact outcome that was observed or anything in the critical region is counted in computing the p-value. Figure 7.1 illustrates in two dimensions several possible regions that could be used to calculate the p-value.

In these figures, the gray area represents the region that is as extreme or more extreme in calculating the p-value. Figure 7.1A corresponds to the region where only the critical region and the actual observed value are considered to be as extreme or more extreme. Figure 7.1B corresponds to the region where only the critical region and a very small set of values that connect the critical region to the actual observed value are considered as extreme or more extreme. The other two figures allow more scenarios that were not actually observed to be considered as extreme or more extreme than what was actually observed. Nobody knows the right way to calculate the p-value and that is why we have to prospectively specify what outcomes we might observe in this experiment that would convince us that the null hypothesis does not adequately explain the data.

The goal of the agency is not only controlling the Type I error rate, i.e. making sure that ineffective drugs are not approved. It is also important to make sure that effective drugs do get approved. Is the bar set too high in the protocol? Assume that the real average treatment effect across studies is 45 m. This represents a 14% increase from baseline assuming that the placebo group is unchanged and is equal to the observed effect in the Flolan study and is a smaller effect than the sponsor expected for this drug. The probability that Conditions 1, 2, or 3 would be satisfied is 0.999. Using the FDA traditional standard (similar to Condition 1 alone), the probability of two positive trials is 96%. So, the bar is not set too high by either the traditional FDA criteria or the actual criteria stated in the protocol. To put it simply, a drug that allows patients in this

Figure 7.1 Different regions that could define values as extreme or more extreme than the observed value.



population to improve walking distance by an average of 45 m more than placebo should have no trouble demonstrating this in these two studies. The reader is again referred to the appendix for an illustration of the power when there is between study variability in the treatment effect.

8. Primary Analysis

Using the pre-specified analysis the study report indicates that the p-values from the primary analysis for the pooled studies, Study P01:04 alone, and Study P01:05 alone were 0.0064, 0.0607, and 0.0550 respectively. The median change from baseline in the treatment group using the pooled data was 10 m and in the individual studies, the median changes were 3 m and 16 m. The median change from baseline in the placebo group using the pooled data was 0 m and in the individual studies the median changes were 1 m and -3 m [*Source: Vol. 2.27 Table 11.4.1.1.1A*]. The results of the sponsor's analysis are summarized in Table 8.1.

Table 8.1. Results from sponsor's primary analysis. Baseline and Week 12 walking distance and change from baseline are summarized by median and the first and third quartiles. [*Source: Vol. 2.27 Tables 11.2.2.4 and 11.4.1.1.1A except where noted*].

Study	Group	Baseline	Week 12^H	Change	P-value
P01:04	Placebo (n=111)	349 m (272, 407)	346 m (275, 400)	1.0 m (-53.0, 30.8)	0.0607
	Treatment (n=113)	341 m (264, 390)	340 m (306, 400)	3.0 m (-27.4, 36.6)	
P01:05	Placebo (n=125)	338 m (272, 377)	348 m (293, 400)	-3.0 m (-37.0, 35.0)	0.0550
	Treatment (n=119)	348 m (268, 396)	357 m (304, 404)	16.0 m (-22.0, 50.0)	
Pooled	Placebo (n=236)	342 m (272, 396)	333 m (277, 400)	0.0 m (-44.5, 32.5)	0.0064
	Treatment (n=232)	345 m (264, 395)	351 m (304, 402)	10.0 m (-24.5, 47.5)	

^H This column was produced by the FDA reviewer from all the observed data at Week 12 for completeness of the table (no imputation was done for missing values). The reviewer could not find this information in the sponsor's report.

The FDA's interpretation of the primary analysis differs from the sponsor's in a few minor ways. These differences arise from issues that were not prospectively defined in the protocol.

Patient number 7004: This patient was assigned to treatment and had a baseline walking distance of 345 m. This patient had a Week 1 walking distance of 393 m and a Week 12 walking distance of 398 m. No Week 6 walking distance was measured because the patient was too critically ill. The sponsor uses the Week 12 walking distance to calculate a score for this patient while the FDA analysis imputes a worst score for this patient. The letter dated March 23, 2000 states: *In addition to the descriptions of the handling of missing data in Table 8.3.1 on page 14 of the final analysis plan, if an exercise test is missing because “patient was too critically ill”, the lowest standardized rank will be used for the nonparametric analysis and a distance of 0 meters will be used for the parametric analysis. Data missing for any other reason will have last standardized ranks carried forward for the nonparametric analyses and last observations carried forward for the parametric analyses.* The literal interpretation of this is that if any ETT is missing, the patient gets a worst score, not only if the Week 12 ETT is missing. This is not just a technical semantic argument- it is difficult to understand why patients who were too ill to walk at Week 12 should be analyzed differently than those who were too ill to walk at Week 6 because there was already a method defined prospectively for imputing a score for patients with no walking distance measured at Week 12.

Patient number 10507: This patient was assigned to the active treatment arm and had a baseline walking distance of 183 m but no subsequent walking distances were measured. The patient withdrew on day 9 for an adverse event. The last day of follow-up on the patient was 39 days after randomization. There are several ways to handle this patient including: a) analyze the data without this patient b) fit a regression of baseline vs. the remaining covariates and carry forward the standardized rank for this patient c) carry forward a worst rank. The sponsor uses the first approach. Since this patient is included in the *mITT* population, it does not seem reasonable to ignore this patient. There is a strong argument for imputing a worst possible score because of the circumstances. Approach b) is in the same spirit as the planned analysis. Patients who do not have complete followup are imputed by carrying forward the last value after adjusting for several covariates. This approach is not perfect because patients with lower baseline tended to show greater improvement. Therefore, this approach will tend to carry forward a smaller rank than that which would be used if post-baseline walking distances were observed. In this case, approach b) would carry forward a standardized rank of 0.138 for this patient. This is the approach used in the FDA analysis.

Patient number 52006: This patient was assigned to placebo and had only the first walking distance measured post-baseline. The patient died within 100 days of randomization. Since the assessment window for all measurements at week 12 extends to Study Day 100, this patient is assigned a worst possible score in the FDA analysis. The last observed standardized rank at Week 1 is used by the sponsor.

Patient number 61008: This patient was assigned to placebo and had a baseline walking distance of 357 m. This patient had a Week 1 walking distance of 338 m and a Week 12 walking distance of 256 m. No Week 6 walking distance was measured because the patient

was too critically ill. The sponsor uses the Week 12 walking distance to calculate a score for this patient while the FDA analysis imputes a worst score for this patient.

Patient number 18501: This patient was assigned to the placebo group and had a baseline walking distance of 362 m. Subsequent walking distances were measured 35, 55, and 71 days after randomization. The first two of these fell within the window that would be counted in the Week 6 visit, but the last did not fall within the Week 6 or the Week 12 window. The idea of the imputation used in the primary analysis is to compare measurements between individuals at the same time in the study (using residuals from the linear regression) and to carry the ranks forward. There was no other patient that had a measurement between the windows for Week 6 and Week 12. Hence, it is not possible to calculate a rank for the measurement on day 71 for this patient. So, two alternatives are i) carry the actual observation at day 71 to Week 12 and do the entire analysis as if it were a Week 12 observation or ii) find the rank of the residual for the day 55 observation and carry this rank forward (in other words, ignore the unscheduled measurement at day 71 entirely). The sponsor uses alternative i) and the FDA uses alternative ii).

Patient 60005: Assigned to active treatment, dropped informed consent after 46 days. The patient was followed-up after withdrawal and had a 12 week walking distance measured. The sponsor's analysis uses the measurement at week 12 while the FDA carries the standard rank from week 6 (the last observation before the patient withdrew).

Patients 2004, 52003 and 52004: All were assigned to placebo and correctly received placebo treatment for the first 6 weeks on study. However, they were inadvertently switched to active treatment for the last 6 weeks of the study. The sponsor carries forward the standardized rank from week 6 for these patients, while the FDA uses the week 12 walking distance.

Both the FDA and the sponsor's analysis begin by finding the standardized ranks of the residuals from linear regression models at Weeks 1, 6, and 12. These regression models included main effects for etiology, baseline distance walked, vasodilator use, and center. The residuals from these linear regression models were ranked and the last observed rank was carried forward to Week 12 but a value of 0 (worst case) was assigned for patients who died or discontinued for clinical deterioration or were too ill to take the ETT. The pre-specified analysis is the CMH (mean score) statistic adjusted for the stratification variables used at randomization. The Final Study Report indicates that because of the low number of patients with low baseline walking distance (defined as less than 150 m), the primary analysis was modified to not include baseline as a covariate. The FDA analysis uses baseline distance as a covariate and finds the significance of the mean score statistic from the asymptotic chi-square approximation except in the case of the P01:05 study where the permutation distribution was used. The reason for the use of the permutation distribution to find the p-value is that in one stratum, there was only one patient and this causes one term in the asymptotic formula to have a zero denominator. The p-value from the FDA analysis for the data from both studies pooled together is 0.0153 and the p-values from the individual studies are 0.104 and 0.081.

The analysis that uses data only from those patients with PPH did not convincingly show a benefit in this subgroup ($p=0.0433$ for both studies pooled together [*Source: Study Report Table 11.4.1.1.5, not verified by the FDA*]).

Whether one uses the sponsor's or the FDA's primary analysis, it is clear that the pre-specified criteria was technically not met, but there appears to be some evidence of efficacy in these two studies. In Sections 9 and 10, some supportive analyses are presented that may be helpful in making a decision about approval.

9. Sponsor's Supportive Analysis of Primary Efficacy Variable

The report contains several planned and unplanned supportive analyses of the primary endpoint. This review will discuss two of these supportive analyses. For the first supportive analysis, the primary analysis was repeated using the per-protocol population. All patients who did not follow the protocol, using pre-specified criteria, were removed in this analysis. The p-values from the individual studies are 0.103 and 0.086 and the p-value for the pooled data is 0.015 [*Source: Vol. 2.27 Table 11.4.1.1.2B*].

For the second supportive analysis, the *mITT* population was used but the method of imputing missing values was modified. Recall that for the primary analysis, worst possible ranks were imputed for discontinuations due to death, transplants, or clinical deterioration while the last rank was carried forward for discontinuations due to other reasons. In this supportive analysis, the last rank was carried forward for all patients without a measurement at Week 12, regardless of the reason. Using this approach, the p-values for the individual studies were 0.083 and 0.075 and the p-value from the pooled data is 0.011 [*Source: Vol. 2.27 Table 11.4.1.1.4B*].

In summary, both of these supportive analyses tend to show the same thing as the primary analysis by the sponsor. That is, both studies taken individually show that the drug was numerically, but not significantly, better than placebo. Since the results of the two studies are consistent, when the data from both studies are combined, the p-value from the pooled analysis is smaller than either p-value from the individual studies.

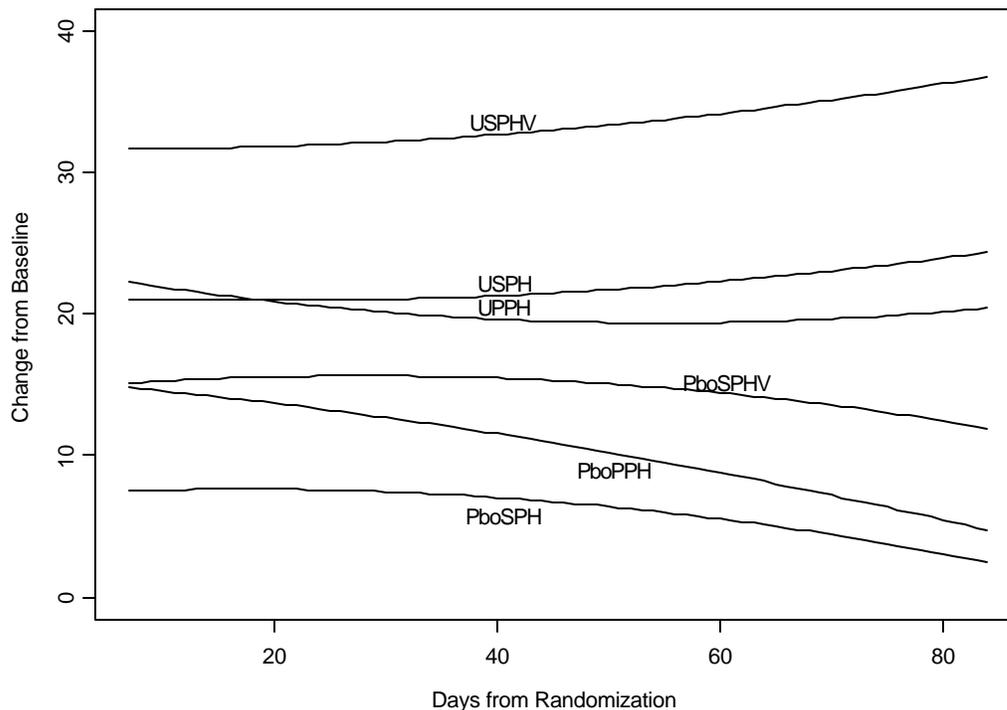
10. FDA's Supportive Analysis of Primary Efficacy Variable

The primary analysis is a nonparametric analysis. One of the main arguments for a nonparametric analysis is that a patient who dies or discontinues for clinical deterioration should be counted as having a worse outcome than any patient who completed the study. If we do not use ranks, then we would have to answer the question of what walking distance at Week 12 should we assign to these patients. The use of ranks takes some of the subjectivity out of the

process. One of the drawbacks of this nonparametric analysis is that it does not yield an easily interpretable estimate of the treatment effect.

A linear mixed effect model can be used here as an exploratory analysis in order to see the treatment effect over time. The model that we will use makes the assumption that those patients who discontinue early- regardless of the reason- would have walking distances similar to those patients who completed the study. In other words, if a patient in the placebo group had a Week 6 walking distance but no Week 12 measurement, then the model can be used to predict a Week 12 observation for this patient by using the data from the other patients that have similar characteristics to this one. Since each patient would theoretically have three measurements post-baseline, the change from baseline was modeled as a quadratic function of time. The specific linear model that was used includes fixed effects for treatment group, baseline distance walked, etiology, vasodilator use among secondary PH patients, and time as a quadratic function. In addition, all two-way interactions between treatment group and the other variables as well as the two-way interactions between stratification (etiology/ vasodilator use) and time were included in the model. There were random effects for the intercept, slope, and the quadratic term for time. The strategy was to specify a complex model and let the data decide which terms were important. The curves for each stratification level at the average baseline walking distance are shown in Figure 10.

Figure 10.1 Fitted curves from linear mixed effects model at the average baseline value. USPHV=Uniprost, secondary PH, vasodilator use; PboPPH=Placebo, PPH, etc.



From Figure 10.1, it appears that at Week 1, patients in all strata in the placebo group improved walking distance by an average of about 10 m, but over the course of the trial, the improvement from baseline decreased slightly. In the Uniprost group, the change at Week 1 was about 30 m in the SPH vasodilator subgroup and about 20 m in the other two subgroups, but over the course of the trial, the improvement was maintained or increased slightly.

Although the large change from baseline in all subgroups at day 7 appears to be unusual because of the low starting dose and the short amount of time involved, this is not just an artifact of the model- the data support this change. There were 451 patients who had a walking distance measured at Week 1. For these patients, the median change from baseline at Week 1 was 10 m and the p-value from the Wilcoxon test has eight zeroes after the decimal.

Another issue that arises from this graph is that a large portion of the treatment effect at Week 12 is already present at Week 1. For example, in the subgroup of patients with secondary PH who were not using vasodilators, the difference between the curves at Week 12 is about 17 m, while the difference at Week 1 is about 13 m. In order to investigate the treatment effect at Week 1, this reviewer repeated the primary efficacy analysis ignoring all data observed after Week 1. The p-value from this analysis is 0.12. Although not significant, this supports what is shown in Figure 10.1. That is, there is already a fairly large difference between

the treatment groups at Week 1. Over the next 11 weeks of the trial, this difference increases slightly so that at the end of the trial the difference reaches statistical significance.

A test for a treatment effect at Week 12 can be obtained by using the likelihood ratio test where the null model is defined by forcing the curves to pass through the same point at Week 12. In order to increase the power of the test, the full model was made less complex by eliminating the interaction terms between treatment and the other covariates. The p-value from this likelihood ratio test with one degree of freedom is 0.0105. This p-value must be interpreted with some caution because this analysis was not pre-specified and there was some model selection involved. The estimate of the treatment effect (improvement from baseline in walking distance relative to placebo) at Week 12 is 15.5 m and the confidence interval for the treatment effect at Week 12 is (3.7, 27.3).

At the request of the medical officer, a second supportive analysis was done to investigate the robustness of the results. This was done mainly because of the disparity between the number of patients who withdrew for adverse events in the two arms in consideration of the way that the primary analysis carried forward data for these patients. Some, but not all, patients who withdrew for adverse events were assigned a worst score in this analysis. This analysis was identical to the primary analysis by the FDA with the following exceptions:

Patients 52008, 54012, 54018, 2001, 2006, 2020, 19502: all used Flolan within 96 days of randomization after discontinuing due to adverse events. This suggests that their clinical status had in fact deteriorated. Therefore, in this analysis these patients will be assigned the worst possible outcome.

Patient 2016: discontinued after 47 days, but subsequent follow-up indicates status after 12 weeks was worse. Therefore, in this analysis these patients will be assigned the worst possible outcome.

Patient 14012: had increased SOB upon discontinuation of study. Therefore, in this analysis these patients will be assigned the worst possible outcome.

Patient 19008: lost to follow-up 45 days after randomization. We will assume a worst case outcome.

The p-value from this analysis of the data from both studies pooled together is 0.134. In this analysis, these selected patients who were classified as having dropped out for adverse events are treated the same way as those patients who were classified as withdrawing for clinical deterioration. So, the non-significant p-value suggests that the observed significance of the primary analysis is not robust to this reclassification.

11. Quality of Life and Secondary Endpoints

At Week 1, the median change from baseline in distance walked for the patients in the treatment group (pooled studies) was 11 m and the median change for the placebo group was 7.6 m. At Week 6, the median change was 13 m for the treatment group and 4.5 m for the placebo group [Source: Vol. 2.27, Tables 11.4.1.1.4G & H].

Quality of Life was measured by the Heart Failure Questionnaire. This instrument evaluates three QoL dimensions: physical, emotional, and global. Out of the 470 patients randomized in both studies, 330 had measurements at baseline and at 12 Weeks. In the treatment group, there was an average change of -6.6 for global QoL, -4.5 for physical, and -1.3 for emotional. In the placebo group, there was an average change of -1.9 for global, -1.8 for physical, and -0.3 for emotional. The pairwise comparisons using the Wilcoxon rank sum test appeared to show a significant difference only in the physical dimension (global $p=0.175$, physical $p=0.0064$, emotional $p=0.3678$) [Source: Vol 2.27, Table 11.4.1.4]. These p-values are nominal and are not adjusted for multiplicity.

Table 11.1 contains a list of various symptoms of pulmonary hypertension. The number of patients in each treatment arm who had this symptom at baseline, but did not have this symptom at the end of the study and vice versa are included in this table. If an unusually large number of patients in the active treatment arm had the symptoms at baseline, but not at the end of the study, then this would indicate a treatment benefit. The p-value in this table is from Fisher's exact test. This is a two-sided p-value and is not adjusted for multiple testing.

Table 11.1 Symptoms of pulmonary hypertension and the number of patients in each treatment arm who had the symptom present at baseline, but absent at the end of the study (labelled as P→A), or absent at baseline, but present at the end of study (labelled as A→P), and the number of patients with no change in the symptom.

Symptom	Uniprost		Placebo		P-value
	A→P	P→A	A→P	P→A	
Dyspnea	0	8	1	4	.385
Fatigue	5	14	12	12	.13
Orthopnea	17	29	30	14	.004
Palpitations	27	46	22	25	.34
Chest Pain	8	48	30	37	.0004
Syncope	1	15	7	10	.04
Dizziness	27	55	33	35	.07
3 rd heart sound	12	7	12	15	.24
4 th heart sound	19	14	26	24	.66
Ventricular heave	25	20	25	24	.68
Loud P2	7	7	8	5	.70
Systolic murmur	19	10	19	15	.45
Diastolic murmur	4	5	10	8	.70
Distention	19	33	30	21	.03
Edema	18	36	29	23	.03
Hepatomegaly	7	19	7	8	.31

The p-values in Table 11.1 corresponding to the symptoms Orthopnea, Chest Pain, Syncope, Distention, and Edema are less than 0.05. This suggests that there were significantly more patients than would be expected to occur by chance in the Uniprost group who had these symptoms present at baseline, but absent at the end of the study.

12. Adverse Events

In the treatment group, 43% had at least one dose reduction due to adverse events compared to 6% in the placebo group. In the treatment group, the most common adverse event which required dose reduction was infusion site pain (64 patients) followed by infusion site reaction (31), nausea (11), and pain, headache, and vasodilation (9 each). In comparison, in the placebo group, there was either 0 or 1 patients with each of these adverse events requiring dose reduction [Source: Study Report Vol 2.27 Table 12.1.2B]. Twenty patients in the treatment group and two in the placebo group prematurely discontinued due to adverse events or withdrew consent [Source: Study Report Vol 2.27 Section 12.1.3].

Restricting attention to those adverse events that were possibly or reasonably attributable to study drug, there were 228 events overall in the treatment group and 154 in the placebo group. The most common of these were infusion site pain (200 in treatment group/ 58 in placebo group), infusion site reaction (196/ 51), headache (55/ 27), diarrhea (51/ 23), nausea (44/ 25), jaw pain (30/ 9), rash (27/ 16), pain (23/ 14), vasodilation (23/ 9), edema (18/ 2) [Source: Study Report Vol 2.27 Table 12.2.3B].

13. Conclusions

The criteria for demonstrating efficacy specified by the sponsor were not met. This criteria included a p-value from the pooled data less than 0.01 and at least one p-value from an individual study less than 0.049. The p-values from the sponsor's analysis for the individual studies were 0.0607 and 0.0550 and the p-value from the pooled data is 0.0064 [see Section 8]. The criteria itself was very liberal and should have been easily met by a drug with similar efficacy as that demonstrated by Flolan. Therefore, there is no justification for relaxing the criteria that was specified in the protocol. In addition to the primary analysis, both the sponsor and the FDA performed various exploratory analyses that are described in this review. The results of these analyses are summarized in Table 13.1.

Table 13.1 Results of different analyses [See sections 8, 9, 10 of this review for more details].

Characteristic	P-value
Sponsor's primary analysis (Nonparametric ANCOVA, <i>mITT</i> population)	0.0064 (pooled) 0.0607 (P01:04) 0.0550 (P01:05)
FDA's primary analysis (Nonparametric ANCOVA)	0.0153 (pooled) 0.104 (P01:04) 0.081 (P01:05)
Secondary analysis by sponsor (same as primary analysis but per-protocol population used)	0.015 (pooled) 0.103 (P01:04) 0.086 (P01:05)
Secondary analysis by sponsor (same as primary analysis but last rank carried forward for all patients regardless of reason for discontinuation)	0.011 (pooled) 0.083 (P01:04) 0.075 (P01:05)
Secondary analysis by FDA (linear mixed effects model using pooled data)	0.0105 (pooled) est. treatment effect = 15.5 m
Secondary analysis by FDA (same as primary analysis but some patients classified as discontinuing for a.e. reclassified as treatment failures)	0.134 (pooled)

John Lawrence, Ph.D.
Mathematical Statistician

This review consists of 20 pages of text, tables, and figures.

Concur: James Hung, Ph.D.
Acting Team Leader, Biometrics I
George Chi, Ph.D.
Division Director, Biometrics I

cc: NDA # 21-272
HFD-110/Dr. Lipicky
HFD-110/Dr. Karkowsky
HFD-110/Mr. Fromme
HFD-700/Dr. Anello
HFD-710/Dr. Chi
HFD-710/Dr. Hung
HFD-710/chron

LAWRENCEJ/594-5375/report.doc/07/18/01

Appendix

In the case of the Flolan study that was used to estimate the sample size needed in these studies, the estimated treatment effect was an improvement in change from baseline of 45 m over placebo. The sponsor assumed that the treatment effect for Uniprost would be 55 m over placebo. Since the two drugs have a similar pharmacological profile, for the purpose of illustration we will assume that the treatment effect can vary from study to study by an average of between 0 and 10 m. In other words, even if a drug has no effect on walking distance, it can appear to show an effect in a given trial and this effect can have a standard deviation of up to 10. When the standard deviation is 10, the effect in an individual trial will be somewhere between -20 and +20 most of the time. Since the average effect across trials is 0, we can still call this drug ineffective. This is the essence of meta-analysis.

Suppose two studies are done with this ineffective drug and in these two studies the true treatment effects are τ_1 and τ_2 . These treatment effects are independent normal random variables with mean 0 and standard deviation 10. In one study, there are 224 patients and in the other study, there are 246 patients. The between patient standard deviation is about 85 (this was relatively consistent in the two studies and the Flolan study). The probability that both p-values from the individual studies will be less than 0.05 and the treatment effects will be in the same direction is about 0.011^2 . This includes those outcomes where both trials show that the treatment is significantly worse than placebo, so in reality the chance that both trials will show that the drug is significantly better than placebo is only half of this.

The previous calculations were all done under the assumption of the traditional approach used by the FDA. Now, consider the approach that is used in this submission. We will again assume that **the drug has no effect on average**, but can appear to show an effect in an individual study that varies from trial to trial. We will reject the null hypothesis of no treatment effect under either of these two circumstances:

- 1) both studies have p-values <0.049 and the pooled data has a p-value <0.049
- 2) either study has a p-value <0.049 and the pooled data has a p-value <0.01

Furthermore, if neither 1) nor 2) occurs, we will reject the null hypothesis of no treatment effect in the subgroup of PPH patients under the following condition:

- 3) the data on PPH patients pooled from both studies has a p-value <0.001 .

² The unconditional distribution of the estimate of the treatment effect in trial i is $N(0, \text{Var}(\tau_i) + 4 \times 85^2/n_i)$.

$$\text{Hence, } P[\text{p-value in trial } i < 0.05] = 2 \Phi \left(-1.96 \sqrt{\frac{4 \times 85^2 / n_i}{\text{Var}(\tau_i) + 4 \times 85^2 / n_i}} \right).$$

We will make the assumption that the treatment effect for the PPH patients is the same as it is for the entire study population and 40% of the patients have PPH. Table A.1 gives the estimated proportion of times that each of these occur with different values of between trial variability of the treatment effect.

The first row in the Table A.1 corresponds to the case where there is no between trial variability in the treatment effect. The overall Type I error rate is in the last column. So, in the case where there is no between trial variability, the overall Type I error rate is controlled at 0.01. If there is any between trial variability in the treatment effect, the chance that any of the three conditions will hold is inflated. When the standard deviation of the treatment effect between trials is 10 m, then the chance that at least one of the conditions will hold is 5%. Under those circumstances, recall that the chance of success using the traditional FDA approach is maintained at about 1%. The traditional FDA approach is very similar to Condition 1 except there is no adjustment for the subgroup analysis (the individual study p-values are compared to 0.05 rather than 0.049).

Table A.1 Proportion of times that each of the conditions will occur assuming no average treatment effect across trials and different between-trial standard deviations of the treatment effect (estimates based on 100,000 simulations).

Between-trial standard deviation	Condition 1	Condition 2	Condition 3	Condition 1, 2, or 3
0	0.001	0.009	0.001	0.010
1	0.001	0.010	0.001	0.011
2	0.001	0.011	0.001	0.012
3	0.002	0.012	0.001	0.013
4	0.002	0.014	0.001	0.015
5	0.003	0.017	0.002	0.018
6	0.003	0.022	0.002	0.024
7	0.004	0.028	0.002	0.028
8	0.006	0.035	0.003	0.036
9	0.008	0.044	0.003	0.045
10	0.011	0.053	0.004	0.054

In order to calculate the power, assume that the real average treatment effect across studies is 45 m. Moreover, assume the treatment effect can vary from study to study with a standard deviation of up to 10 m. The probability that Conditions 1, 2, or 3 would be satisfied is shown in Table A.2. Using the FDA traditional standard (similar to Condition 1), the probability of two positive trials is 88% if the between trial standard deviation is 10 m.

Table A.2 Proportion of times that each of the conditions will occur assuming an average treatment effect across trials of 45 m and different between-trial standard deviations of the treatment effect (estimates based on 100,000 simulations).

Between-trial standard deviation (m)	Condition 1	Condition 2	Condition 3	Condition 1, 2, or 3
0	0.962	0.999	0.652	0.999
1	0.961	0.999	0.648	0.999
2	0.961	0.999	0.652	0.999
3	0.957	0.999	0.646	0.999
4	0.951	0.998	0.646	0.998
5	0.943	0.998	0.644	0.998
6	0.934	0.997	0.645	0.997
7	0.923	0.996	0.642	0.996
8	0.910	0.994	0.636	0.994
9	0.900	0.992	0.635	0.992
10	0.882	0.989	0.631	0.989