SUPPLEMENT 3

Statistical Analysis Plan

Oral Iron Repletion Effects on Oxygen Uptake in Heart Failure

IRONOUT-HF

A randomized, double-blind, placebo-controlled study of chronic NYHA class II-IV heart failure with reduced ejection fraction and iron deficiency (with or without anemia).

Protocol Date
December 22, 2014 – Amendment 1

Sponsor
National Heart, Lung and Blood Institute

Protocol Principal Investigators
Greg Lewis, MD
# Table of Contents

1. Overview ................................................. 3
   1.1 Synopsis .............................................. 3
   1.2 Study Treatments ...................................... 3

2. Study Design ........................................... 3
   2.1 Overview .............................................. 3
   2.2 Randomization ......................................... 3
   2.3 Data Sources ........................................... 3

3. Analysis Population and Missing Data ........................................... 3

4. General Methodology ....................................... 4

5. Primary Endpoint ........................................ 4

6. Secondary Endpoints ........................................ 4

7. Exploratory Endpoints ...................................... 5

8. Primary Endpoint Description ................................... 5
   8.1 Primary Endpoint ....................................... 5

9. Secondary Endpoint Descriptions ................................... 6
   9.1 Secondary Endpoint #1 ................................... 6
   9.2 Secondary Endpoint #2 ................................... 6
   9.3 Secondary Endpoint #3 ................................... 7
   9.4 Secondary Endpoint #4 ................................... 8
   9.5 Secondary Endpoint #5 ................................... 8

10. Exploratory Endpoint Descriptions .................................. 9
    10.1 Exploratory Endpoint #1 .................................. 9
    10.2 Exploratory Endpoint #2 ................................ 10
    10.3 Exploratory Endpoint #3 ................................ 10
    10.4 Exploratory Endpoint #4 ................................ 11

11. Safety Analyses ........................................ 11
    11.1 Safety Endpoint #1 ..................................... 11
    11.2 Safety Endpoint #2 ..................................... 12

12. Interim Analysis .......................................... 12

13. Subgroup of Interest ....................................... 12
1. Overview

1.1 Synopsis

The Oral Iron Repletion Effects on Oxygen Uptake in Heart Failure (IRONOUT-HF) trial is testing the efficacy and safety of therapy with twice daily Oral Polysaccharide Iron Complex in the reduced ejection fraction population to assess functional capacity and symptoms in heart failure over a 16 week treatment and follow-up period.

1.2 Study Treatments

Placebo or Oral Polysaccharide Iron (150 mg) will be taken twice daily and administered for sixteen weeks. Study treatment is administered separately from meals (1 hour prior or 2 or more hours after).

2. Study Design

2.1 Overview

The IRONOUT-HF study is a randomized, double-blind, placebo-controlled study in chronic NYHA class II-IV heart failure patients with reduced ejection fraction and iron deficiency. A total of 220 patients will be enrolled in the trial.

The treatments in this study are blinded. Treatment is expected to last for 16 weeks.

The over-arching hypothesis is that, compared to placebo, therapy with the oral polysaccharide iron complex will be associated with improvement in functional capacity as assessed by CPET at week 16.

2.2 Randomization

Patients are randomized in a 1:1 ratio to either oral polysaccharide iron or placebo. The randomization scheme consists of a permuted block design with stratification by clinical site and presence/absence of anemia. Per protocol section 8.7, anemia is defined as hemoglobin < 12 g/dL when determining the appropriate strata. For analysis purposes, anemia will be defined as in Section 13 of this Statistical Analysis Plan.

2.3 Data Sources

A database of case report form and biomarker core lab data will be created in Inform, and the data then transferred to SAS for analysis. The randomized treatment assignment will be provided through data provided by the Axcess system, an Almac Clinical Services web-based randomization system.

3. Analysis Population and Missing Data

All randomized patients will be included in the analysis population for assessing the primary, secondary and exploratory endpoints. However, as described in subsequent sections of this document, some patients may be excluded from certain analyses if key data elements are missing. With the extensive efforts being made
in connection with the clinical sites to ensure data quality and completeness, it is expected that exclusion of patients for any endpoint analysis will be minimal.

Placement of a left ventricular assist device (LVAD) or receipt of a heart transplant (HTP) prior to the end of the study follow-up is a possibility in this population. Due to the affect these procedures would have on the heart failure symptoms and quality of life, any data collected post-LVAD or HTP will not be used in analyses.

The specific endpoint descriptions in Sections 8 through 10 describe the circumstances that would lead to a patient being excluded from a specific analysis.

4. General Methodology

Medians, 25th and 75th percentiles will be presented for continuous variables; the number and percentage of patients in each category will be presented for categorical variables. For all endpoints a p-value ≤0.05 will be considered statistically significant. Analyses will be performed using validated SAS software (SAS Institute, Inc, Cary, NC). Appropriate statistical models will be used to examine the effect of treatment with oral polysaccharide iron on both the primary, secondary, and exploratory outcomes in the study.

For the rank-based endpoints, a non-parametric testing strategy will be employed. For continuous endpoint variables, general linear models will be used. For endpoints where the response is dichotomous (binary), the logistic regression model will be used. For time-to-event endpoints, the Cox regression model will be used.

In those instances where multiple imputation is used to address incomplete data, 100 imputations will be used when generating the multiple imputation datasets.

5. Primary Endpoint

Primary Endpoint

#1: Change in Peak VO2 (ml/min) by CPET from baseline to week 16

See Section 8 for a detailed description of the primary endpoint, including rules that will be followed for handling incomplete data.

6. Secondary Endpoints

Secondary Endpoints

#1: Change in O2 uptake kinetics as defined by the mean response time as measured by CPET from baseline to week 16
#2: Change in ventilatory efficiency as defined by Ve/VCO2 slope by CPET from baseline to week 16
#3: Change in 6 minute walk from baseline to week 8 and week 16
#4: Change in core lab NT Pro BNP from baseline to week 16
#5: Change in Kansas City Cardiomyopathy Questionnaire (KCCQ) symptom scores from baseline to week 8 and week 16

See Section 9 for a detailed description of each secondary endpoint, including rules that will be followed for handling incomplete data.
7. Exploratory Endpoints

Exploratory Endpoints

#1: Change in renal function from baseline to 16 weeks
#2: Time to death or heart failure hospitalization
#3: Change in VO2 at anaerobic threshold from baseline to week 16
#4: Change in iron bioavailability markers from baseline to week 16

See Section 10 for a detailed description of each exploratory endpoint, including rules that will be followed for handling incomplete data.

8. Endpoint Descriptions

8.1 – Primary Endpoint

Endpoint Description: Change in peak VO2 (ml/min) by CPET from baseline to week 16.

Response Variable Definition: Change is defined as the post-baseline value – baseline value. The CPET is performed at baseline and week 16.

Additional Covariates: Baseline peak VO2

Handling of Dropouts and Missing Data:

If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing peak VO2 data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline peak VO2, and week 16 peak VO2.

Three sensitivity analyses will be performed that accounts for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

2) Repeated Measures: A repeated measures model will be used with baseline peak VO2 as a response instead of a covariate.

3) Worst rank analysis: The observed change in peak VO2 will be ranked from smallest to largest. Any patient that died or had a LVAD prior to the assessment point will be assigned the worst rank possible based on the observed data.

Statistical Tests:

A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups. A mixed model (PROC MIXED) will be used to generate the repeated measures model results.

Interpretation of Results: An increase in peak VO2 indicates a better outcome. Lower rank scores indicate worse outcomes in the worst rank analysis.
9. Secondary Endpoint Descriptions

9.1 – Secondary Endpoint #1

**Endpoint Description:** Change in \(O_2\) uptake kinetics as defined as mean response time by CPET at week 16.

**Response Variable Definition:** Change for response is defined as the post-baseline value – baseline value. The CPET is performed at baseline and week 16.

**Additional Covariates:** Baseline value of the response variable

**Handling of Dropouts and Missing Data:**
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing mean response data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline mean response time and Week 16 mean response time.

Two sensitivity analyses will be performed that accounts for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

2) Worst rank analysis: The observed change in mean response time will be ranked from smallest to largest. Any patient that died or had a LVAD prior to the assessment point will be assigned the worst rank possible based on the observed data.

**Statistical Tests:**
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups. The non-parametric Wilcoxon Rank Sum test (PROC NPAR1WAY) will be used for the worst rank score analysis.

**Interpretation of Results:** For the general linear models, an increase in mean response time indicates worse outcomes. Lower rank scores indicate worse outcomes in the worst rank analysis.

9.2 – Secondary Endpoint #2

**Endpoint Description:** Change in ventilatory efficiency defined by \(V_E/VCO_2\) slope by CPET from baseline to 16 weeks.

**Response Variable Definition:** Change is defined as the post-baseline value – baseline value. The CPET is performed at baseline and Week 16.

**Additional Covariates:** Baseline \(V_E/VCO_2\) slope

**Handling of Dropouts and Missing Data:**
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing \(V_E/VCO_2\) slope data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline \(V_E/VCO_2\) slope and Week 16 \(V_E/VCO_2\) slope.
Two sensitivity analyses will be performed that accounts for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

2) Worst rank analysis: The observed change in $V_E/VCO_2$ slope will be ranked from smallest to largest. Any patient that died or had a LVAD prior to the assessment point will be assigned the worst rank possible based on the observed data.

**Statistical Tests:**
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups. The non-parametric Wilcoxon Rank Sum test (PROC NPAR1WAY) will be used for the worst rank score analysis.

**Interpretation of Results:** For the general linear models, an increase in $V_E/VCO_2$ slope indicates a worse outcome. Lower rank scores indicate worse outcomes in the worst rank analysis.

### 9.3 – Secondary Endpoint #3

**Endpoint Description:** Change in functional status as assessed by six minute walk distance at 8 weeks and 16 weeks after randomization.

**Response Variable Definition:** Change for each response is defined as the post-baseline value – baseline value. The six minute walk is performed at baseline, week 8 and week 16.

**Additional Covariates:** Baseline walk distance

**Handling of Dropouts and Missing Data:**
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing walk distance data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline walk distance, week 8 walk distance, and week 16 walk distance.

Two sensitivity analyses will be performed that account for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

2) Worst rank analysis: The observed change in walk distance will be ranked from smallest to largest. Any patient that died or had a LVAD prior to the assessment point will be assigned the worst rank possible based on the observed data.

**Statistical Tests:**
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups. The non-parametric Wilcoxon Rank Sum test (PROC NPAR1WAY) will be used for the worst rank score analysis.

**Interpretation of Results:** For the general linear models, a decrease in walk distance indicates worse outcome. Lower rank scores indicate worse outcomes in the worst rank analysis.
9.4 – Secondary Endpoint #4

**Endpoint Description:** Change in core lab NT Pro BNP from baseline to week 16.

**Response Variable Definition:** Change is defined as the post-baseline value – baseline value. The core lab biomarkers are collected at baseline and week 16.

**Additional Covariates:** Baseline NT Pro BNP

**Handling of Dropouts and Missing Data:**
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing biomarker data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline NT Pro BNP, and week 16 NT Pro BNP.

Two sensitivity analyses will be performed that accounts for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

2) Worst rank analysis: The observed change in NT Pro BNP score will be ranked from smallest to largest. Any patient that died or had a LVAD prior to the assessment point will be assigned the worst rank possible based on the observed data.

**Statistical Tests:**
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups. The non-parametric Wilcoxon Rank Sum test (PROC NPAR1WAY) will be used for the worst rank score analysis.

**Interpretation of Results:** For the general linear models, an increase in NT Pro BNP indicates worse outcome. Lower rank scores indicate worse outcomes in the worst rank analysis.

9.5 – Secondary Endpoint #5

**Endpoint Description:** Change in symptoms as assessed by the Kansas City Cardiomyopathy Questionnaire (Overall and Clinical Summary Scores) at 8 weeks and 16 weeks after randomization.

**Response Variable Definition:** Change for each response is defined as the post-baseline value – baseline value. The KCCQ is completed at baseline, week 8 and week 16.

**Additional Covariates:** Baseline value of the response variable

**Handling of Dropouts and Missing Data:**
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing KCCQ data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline overall summary score, baseline clinical summary score, week 8 overall...
Two sensitivity analyses will be performed that account for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

2) Worst rank analysis: The observed change in KCCQ score will be ranked from smallest to largest. Any patient that died or had a LVAD prior to the assessment point will be assigned the worst rank possible based on the observed data.

Statistical Tests:
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups. The non-parametric Wilcoxon Rank Sum test (PROC NPAR1WAY) will be used for the worst rank score analysis.

Interpretation of Results: For the general linear models, a decrease in KCCQ score indicates worse outcome. Lower rank scores indicate worse outcomes in the worst rank analysis.

10. Exploratory Endpoint Descriptions

10.1 – Exploratory Endpoint #1

Endpoint Description: Change in renal function from baseline to 16 weeks. The variables will include core lab cystatin C and site reported creatinine.

Response Variable Definition: Change for each response is defined as the post-baseline value – baseline value. The lab values are collected at baseline and week 16.

Additional Covariates: Baseline value of the response variable

Handling of Dropouts and Missing Data:
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing biomarker data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline creatinine, baseline cystatin C, week 16 creatinine and week 16 cystatin C.

One sensitivity analysis will be performed that accounts for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

Statistical Tests:
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups.

Interpretation of Results: For the general linear models, an increase in cystatin C or creatinine indicates a worse outcome.
10.2 – Exploratory Endpoint #2

**Endpoint Description:** Time to death or heart failure hospitalization.

**Response Variable Definition:** The time from randomization to patient death or the first heart failure hospitalization (days).

**Additional Covariates:** None

**Handling of Dropouts and Missing Data:**
All patients should have some information regarding re-hospitalization. Death and heart transplant will be censoring variables in these analyses.

**Statistical Tests:**
The Cox regression model for survival data (PROC PHREG in SAS) will be used to test the statistical significance of differences in heart failure hospitalizations between the treatments. Kaplan-Meier curves will be generated to graphically display the event rates as a function of time from randomization in each treatment.

**Interpretation of Results:** Higher event rates indicate a worse outcome.

10.3 – Exploratory Endpoint #3

**Endpoint Description:** Change in VO₂ at anaerobic threshold (AT) by CPET at week 16.

**Response Variable Definition:** Change for response is defined as the post-baseline value – baseline value. The CPET is performed at baseline and week 16.

**Additional Covariates:** Baseline value of the response variable

**Handling of Dropouts and Missing Data:**
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing VO₂ at AT: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline VO₂ at AT, week16 VO₂ at AT.

One sensitivity analysis will be performed that accounts for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

**Statistical Tests:**
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups.

**Interpretation of Results:** For the general linear models, a decrease in VO₂ at AT indicates a worse outcome.
10.4 – Exploratory Endpoint #4

**Endpoint Description:** Change in iron bioavailability from baseline to 16 weeks. The variables will include core lab iron, ferritin, TIBC and transferrin saturation (defined as iron/TIBC).

**Response Variable Definition:** Change for each response is defined as the post-baseline value – baseline value. The lab values are collected at baseline and week 16.

**Additional Covariates:** Baseline value of the response variable

**Handling of Dropouts and Missing Data:**
If the patient is missing any of the required data such that a difference cannot be calculated, multiple imputation will be used to address the missing data. The following variables will be used in the calculation of missing biomarker data: treatment arm – (oral polysaccharide iron, placebo), stratification – (anemia, no anemia), age, sex, baseline lab value, and week 16 lab value.

One sensitivity analysis will be performed that accounts for missing data differently.

1) Complete Case Analysis: If the patient is missing any of the required data such that a difference cannot be calculated, the patient is excluded from that particular analysis.

**Statistical Tests:**
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in response variables between the two treatment groups.

**Interpretation of Results:** For the general linear models, the magnitude of the change from baseline in addition to the directionality of that change will help determine which treatment fared worse.

11. Safety analysis

11.1 – Safety Endpoint #1

**Safety Endpoint #1:** Comparison of adverse events through week 16

**Response Variable Definitions:** Each unique adverse event (AE) type based on preferred term within body system. If a patient has more than one event of the same type, the patient is only counted once. The overall rate of adverse events will also be compared by treatment arm.

**Additional Covariates:** None

**Handling of Dropouts and Missing Data:**
It is assumed that all patients will have some assessment of adverse events made through their individual follow-up durations. No data are expected to be missing.

**Statistical Tests:**
Statistical comparisons will be based on the Fisher’s mid-p test.

**Interpretation of Results:** Higher AE rates indicate a worse outcome.

11.2 – Safety Endpoint #2
Safety Endpoint #2: Change in local laboratory values at week 16.

Response Variable Definitions: Change in laboratory values will be calculated as post-baseline value – baseline value. The list of laboratory values includes sodium, potassium, chloride, bicarbonate, BUN, ALT, AST, alkaline phosphatase, total bilirubin, glucose, hemoglobin, hematocrit, red blood cell count, red cell distribution width, white blood cell count, and platelet count.

Additional Covariates: Baseline lab value

Handling of Dropouts and Missing Data:
The analyses will be conducted using only observed data with no adjustment for incomplete data.

Statistical Tests:
A general linear model (PROC GLM in SAS) will be used to estimate and statistically compare the differences in the change in local laboratory values measures between the two treatment groups.

Interpretation of Results: Since changes in many labs could be considered beneficial or troublesome, the interpretation of those results will be made in the context of the observed data.

12. Interim Analyses

Interim data analysis for efficacy will not be conducted due to the relatively small size and short duration of this clinical trial. Safety data will be periodically assessed by the Data and Safety Monitoring Board (DSMB) based on the reporting of adverse events and local laboratory changes. There are no pre-specified guidelines for determining stopping rules due to a safety concern. The clinical opinion from the DSMB deliberations will be sole determinant.

13. Subgroup of Interest

Selected analyses will be performed on the following subgroups: a) patients with and without anemia [Hgb < 12 for Females and <13.5 for Males based on standard definitions of anemia], b) patients with and without baseline venous congestion [defined as Extended/Distended JVP and/or Moderate/Severe Edema (≥ 2+)] and c) patients with and without an RER > 1.1 during maximum incremental exercise. In addition, the subgroup of gender will be included to satisfy the NHLBI Inclusion policy. Models will either adjust for presence of subgroup or will be performed in each subgroup, as appropriate.