**Statistical Analysis Plan (SAP)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Title** | |  | |
| **CRU/Department/Division/Center** | |  | |
| **IRB Number** | |  | |
| **Investigators:** | |  | |
| **Lead Investigator** | |  | |
| **Mentors** | |  | |
| **Co-authors (if known)** | |  | |
| **Biostatistician(s)** | | Jackson Dial | |
| **Supervising Biostatistician** | | Elizabeth Hauser, PhD | |
| **Original Creation Date** | | April 16, 2022 | |
| **Version Date** | |  | |
| **Project Folder Location** | |  | |
| **Project Goal(s)** | |  | |
| **Submission Deadline(s)** | |  | |
| **Effort Estimate (optional)** | |  | |
| **Investigator Agreement** | All statistical analyses included in an abstract or manuscript should reflect the work of the biostatistician(s) listed on this SAP. No changes or additional analyses should be made to the results or findings without discussing with the project biostatistician(s).  All biostatisticians on this SAP should be given sufficient time to review the full presentation, abstract, manuscript, or grant and be included as co-authors on any abstract or manuscript resulting from the analyses.  If substantial additional analysis is necessary or the aims of the project change, a new SAP will need to be developed.  Publications resulting from this SAP are supported in part by the Duke CTSA and must cite grant number UL1TR002553 and be submitted to PubMed Central.  I have reviewed the SAP and understand that any changes must be documented.  *Acknowledged by:* Click or tap here to enter text.  *Date:* Click or tap to enter a date. | | |
| **Activity Log** | *Add initial information…………………………………………………April 16, 2022*  *Write Plans……………………………………………………………….... April 19, 2022*  *Perform Analyses…………………………………………………………April 24, 2022* | | |
| **Acronyms** |  | |  |
|  |  | |  |
|  |  | |  |

# Study Overview

Background/Introduction: *Heart failure is a serious issue that is a form of heart disease. It has been shown that the drug digoxin found in the foxglove plant is useful in treating heart failure. It has also been shown that differing levels of blood potassium levels can change the effect of digoxin on heart failure.*

## Study Aims

**Aim 1**

Investigate the association between baseline subject characteristics (Baseline Form + derived variables age, BMI, NSYM, and TRTMT) and baseline potassium levels.

**Aim 2**

Investigate the association between baseline potassium levels (KLEVEL) and increased risk of hospitalization for worsened heart failure (WHF on Event Form).

**Aim 3**

Investigate the interaction between baseline potassium level and Digoxin treatment to change effect of treatment on hospitalization for worsened heart failure.

## Study Hypotheses

## Primary Hypotheses

There is a relationship between KLEVEL and WHF such that as KLEVEL decreases, risk for WHF increases.

The TRTMT variable does not have an effect on the effect of KLEVEL on WHF.

## Secondary Hypotheses

Baseline characteristics are associated with KLEVEL.

# Study Population

## Inclusion Criteria

* Data were excluded from the analysis if they had any NA values for any of the variables used in that aim. This means that there were different sample sizes for each aim. Aim 1 included 5385 observations, awhile aims 2 & 3 included 5998 observations. See table 1 for NA counts per variable.

## Exclusion Criteria

* Exclusion was applied to any observations with missing values in any variables used in the analysis.

## Data Acquisition

*Fill in all relevant information:*

|  |  |
| --- | --- |
| Study design | Placebo-Controlled Clinical Trial |
| Data source/how the data were collected | Followed enrolled patients from 302 centers across the US and Canada |
| Contact information for team member responsible for data collection/acquisition |  |
| Date or version (if downloaded, provide date) | April 16, 2022 |
| Data transfer method and date |  |
| Where dataset is stored |  |

# Outcomes, Exposures, and Additional Variables of Interest

## Primary Outcome(s)

|  |  |  |  |
| --- | --- | --- | --- |
| **Outcome** | **Description** | **Variables and Source** | **Specifications** |
| WHF | Flag for a patient being hospitalized due to worsened heart failure | Event form Question 02 | Binary variable where:  0 = No  1 = Yes |

## Additional Variables of Interest

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Variables and Source** | **Specifications** |
| SEX | Identified sex of patient | Baseline Form Question 4 | Categorical variable where:  1 = Male  2 = Female |
| RACE | Race of patient | Baseline Form Question 5 | Categorical variable where:  1 = White  2 = Non-white |
| TRTMT | Treatment Group Digoxin/Placebo | Derived | Binary variable where:  0 = Placebo  1 = Treatment |
| CREAT | Serum Creatinine Level, measures in mol/L | Baseline Form Question 9. | Continuous variable with valid range 0-3.761 |
| KLEVEL | Serum Potassium Level, measured in mmol/l | Baseline Form Question 9a. | Continuous variable with valid range 0-6.3 |
| SYSBP | Systolic Blood Pressure | Baseline Form Question 22 | Continuous variable with valid range 74-220 |
| DIABP | Diastolic Blood Pressure | Baseline Form Question 22 | Continuous variable with valid range 25-184 |
| FUNCTCLS | Current NYHA Functional Class | Baseline Form Question 23 | Categorical variable where:  1 = Class I  2 = Class II  3 = Class III  4 = Class IV  \* See DIG Documentation for description of classes |
| CHFETIOL | Etiology of CHF | Baseline Form Question 24 | Categorical variable where:  1 = Ischemic  2 = Hypertensive  3 = Valvular  4 = Idiopathic  5 = Alcohol Related  6 = Other |
| PREVMI | Previous Myocardial infarction | Baseline Form Question 25 | Binary variable where:  0 = No  1 = Yes |
| DIABETES | History of Diabetes | Baseline Form Question 27 | Binary variable where:  0 = No  1 = Yes |
| HYPERTEN | History of Hypertension | Baseline Form Question 28 | Binary variable where:  0 = No  1 = Yes |
| DIURETK | Potassium-Sparing Diuretics | Baseline Form Question 30 | Binary variable where:  0 = No  1 = Yes |
| KSUPP | Potassium Supplement | Baseline Form Question 31A | Binary variable where:  0 = No  1 = Yes |
| AGE | Age of Patient | Derived as INT(Date of Randomization - Date of Birth)/365.25 | Continuous variable with valid range 21 - 94 |
| BMI | BMI of Patient | Derived as Weight (kg) / Height (meters)^2 | Continuous variable with valid range 14.445 – 62.664 |
| NSYM | Number of symptoms of CHF | Derived as the sum of Baseline Form questions 13-20 | Count variable with valid range 0 - 4 |

# Statistical Analysis Plan

## Demographic and Clinical Characteristics (“Table 1”)

|  | Digoxin (N=3397) | Placebo (N=3403) | Overall (N=6800) |
| --- | --- | --- | --- |
| **Sex** |  |  |  |
| Female | 755 (22.2%) | 764 (22.5%) | 1519 (22.3%) |
| Male | 2642 (77.8%) | 2639 (77.5%) | 5281 (77.7%) |
| **Race** |  |  |  |
| Non-White | 487 (14.3%) | 504 (14.8%) | 991 (14.6%) |
| White | 2910 (85.7%) | 2899 (85.2%) | 5809 (85.4%) |
| **Serum Creatinine** |  |  |  |
| Mean (SD) | 1.28 (0.366) | 1.29 (0.372) | 1.29 (0.369) |
| Median [Min, Max] | 1.20 [0.500, 3.76] | 1.21 [0.100, 3.05] | 1.20 [0.100, 3.76] |
| **Serum Potassium** |  |  |  |
| Mean (SD) | 4.33 (0.511) | 4.46 (7.87) | 4.40 (5.57) |
| Median [Min, Max] | 4.30 [0, 6.30] | 4.30 [0, 434] | 4.30 [0, 434] |
| Missing | 391 (11.5%) | 410 (12.0%) | 801 (11.8%) |
| **SystolicBP** |  |  |  |
| Mean (SD) | 126 (19.9) | 126 (19.9) | 126 (19.9) |
| Median [Min, Max] | 122 [78.0, 220] | 124 [74.0, 202] | 123 [74.0, 220] |
| Missing | 1 (0.0%) | 2 (0.1%) | 3 (0.0%) |
| **DiastolicBP** |  |  |  |
| Mean (SD) | 74.9 (11.5) | 74.9 (11.1) | 74.9 (11.3) |
| Median [Min, Max] | 75.0 [25.0, 184] | 75.0 [38.0, 140] | 75.0 [25.0, 184] |
| Missing | 2 (0.1%) | 3 (0.1%) | 5 (0.1%) |
| **NYHA Class** |  |  |  |
| Class I | 465 (13.7%) | 442 (13.0%) | 907 (13.3%) |
| Class II | 1810 (53.3%) | 1854 (54.5%) | 3664 (53.9%) |
| Class III | 1042 (30.7%) | 1039 (30.5%) | 2081 (30.6%) |
| Class IV | 76 (2.2%) | 66 (1.9%) | 142 (2.1%) |
| Missing | 4 (0.1%) | 2 (0.1%) | 6 (0.1%) |
| **CHF Etiology** |  |  |  |
| Alcohol Related | 92 (2.7%) | 130 (3.8%) | 222 (3.3%) |
| Hypertensive | 272 (8.0%) | 311 (9.1%) | 583 (8.6%) |
| Idiopathic | 525 (15.5%) | 482 (14.2%) | 1007 (14.8%) |
| Ischemic | 2405 (70.8%) | 2398 (70.5%) | 4803 (70.6%) |
| Other | 46 (1.4%) | 22 (0.6%) | 68 (1.0%) |
| Valvular | 48 (1.4%) | 51 (1.5%) | 99 (1.5%) |
| Missing | 9 (0.3%) | 9 (0.3%) | 18 (0.3%) |
| **Previous Myocardial Infarction** |  |  |  |
| No | 1198 (35.3%) | 1182 (34.7%) | 2380 (35.0%) |
| Yes | 2199 (64.7%) | 2221 (65.3%) | 4420 (65.0%) |
| **History of Diabetes** |  |  |  |
| No | 2436 (71.7%) | 2431 (71.4%) | 4867 (71.6%) |
| Yes | 961 (28.3%) | 972 (28.6%) | 1933 (28.4%) |
| **History of Hypertension** |  |  |  |
| No | 1869 (55.0%) | 1846 (54.2%) | 3715 (54.6%) |
| Yes | 1528 (45.0%) | 1557 (45.8%) | 3085 (45.4%) |
| **Potassium-Sparing Diuretics** |  |  |  |
| No | 3159 (93.0%) | 3124 (91.8%) | 6283 (92.4%) |
| Yes | 238 (7.0%) | 279 (8.2%) | 517 (7.6%) |
| **Potassium Supplement** |  |  |  |
| No | 2121 (62.4%) | 2087 (61.3%) | 4208 (61.9%) |
| Yes | 1276 (37.6%) | 1316 (38.7%) | 2592 (38.1%) |
| **Age of Patient** |  |  |  |
| Mean (SD) | 63.4 (11.0) | 63.5 (10.8) | 63.5 (10.9) |
| Median [Min, Max] | 64.0 [21.0, 94.0] | 65.0 [22.0, 92.0] | 65.0 [21.0, 94.0] |
| **BMI of Patient** |  |  |  |
| Mean (SD) | 27.0 (5.19) | 27.2 (5.19) | 27.1 (5.19) |
| Median [Min, Max] | 26.4 [15.2, 58.3] | 26.6 [14.4, 62.7] | 26.5 [14.4, 62.7] |
| Missing | 0 (0%) | 1 (0.0%) | 1 (0.0%) |
| **Number of CHF Symptoms** |  |  |  |
| Mean (SD) | 3.65 (0.805) | 3.67 (0.783) | 3.66 (0.794) |
| Median [Min, Max] | 4.00 [0, 4.00] | 4.00 [0, 4.00] | 4.00 [0, 4.00] |
| **Worsened Heart Failure** |  |  |  |
| No | 2487 (73.2%) | 2223 (65.3%) | 4710 (69.3%) |
| Yes | 910 (26.8%) | 1180 (34.7%) | 2090 (30.7%) |

## Analyses Plan for Aim 1

A linear regression model will be fit to the data with KLEVEL as the response and all the other variables of interest as predictor variables, using stepwise variable selection. The level of significance for entrance and removal from the model will be 0.05. This model will be further investigated based on subset selection using forward and backward selection, independently. If these 3 methods all provide the same variables, then it can be concluded that those variables are associated with the response. If a variable is only included in one of the 3 models, it will be concluded that there is a possible association, but further investigation is necessary.

To further solidify this conclusion, the same process will be performed using stepwise, forward, and backward selection but will now be based on AIC values instead of p-values. Again, if variables are contained in both the models selected by p-value and models selected by the AIC, then it will be concluded that the variables in both models are associated with the outcome variable. Variables included in the selection but not in all will be pointed out for further analysis.

Assumptions for Linear Regression (See plots in Appendix):

1. Linearity between X and Y
2. Homoscedasticity, or constant variance
3. Independence of observations
4. Normality

## Analyses Plan for Aim 2

A simple logistic regression model will be fit to the data with WHF as the binary outcome and KLEVEL as the continuous predictor and will use the following hypotheses:

Once the logistic regression model is fit to the data, conclusions will be drawn from the model based on the hypotheses and coefficients of the model, if the assumptions have been met.

Assumptions for Logistic Regression:

1. Outcome variable is binary
2. Independent observations
3. No multicollinearity
4. Linearity of independent variables and log-odds
5. Large sample size

## Analyses Plan for Aim 3

A multiple logistic regression model will be fit to the data with the same binary outcome and continuous variable as in Aim 2, but will add in a binary predictor variable TRTMT (digoxin treatment group) to assess the effect that TRTMT has on the effect of KLEVEL on WHF.

The following hypotheses will be tested:

After assumptions have been checked, if they are met then conclusions will be drawn from the model. Otherwise, the model will not be used to make conclusions.

# Analysis

## Analysis for Aim 1

The 3 algorithms of variable selection using p-value as entrance and removal criteria all selected the same variables as significant predictors of the outcome variable, KLEVEL. These variables were DIABETES and CREAT.

When the entrance and removal criteria was changed to be AIC, a different subset of variables were selected. Each of these 3 algorithms (stepwise, forward, and backward) selected the same subset of variables, which included DIABETES, CREAT, PREVMI, and SEX.

It can be concluded that there is a statistically significant relationship between DIABETES and CREAT with KLEVEL. It can be concluded that higher levels of creatinine (CREAT) are associated with higher levels of serum potassium (KLEVEL), specifically for a 1 unit increase in creatinine, there is a 0.03787 average increase serum potassium. Having Diabetes (DIABETES) also increases the average level of serum potassium, with an individual having diabetes with an average of 0.03158 mmol/L higher than an individual without diabetes.

Previous Myocardial Infarction and Sex were also identified when using AIC as the criteria, and these variables should be further investigated before association conclusions can be made.

As can be seen in the diagnostic plots in the appendix, there is a big issue with the variance, as there is a clear pattern in the scale-location plot.

See appendix for outputs from RStudio.

## Analysis for Aim 2

The fitted logistic regression model with WHF as the outcome variable and KLEVEL as the predictor variable yielded a p-value of 0.65839 in relation to the hypotheses. This means failure to reject the null hypothesis and it can be concluded that the coefficient for KLEVEL is equal to 0. In this situation, there is no logistic relationship between KLEVEL and WHF. See appendix for output.

Assumptions:

1. The outcome variable, WHF, was defined to be binary.
2. Independence of observations is automatically met for the data given that each observation is an independent person.
3. No multicollinearity can exist with only 1 predictor variable.
4. See appendix for plot showing linearity.
5. The total sample size and group sample sizes were deemed large enough, see table 1 for values.

Despite the p-value being insignificant, an interpretation of the model would be that for a 1 unit increase in KLEVEL (serum potassium), the odds of WHF increase by a factor of 0.976, or in other words they decrease by about 2.5%. The confidence interval for this value does include 1, supporting the hypothesis test conclusion, after converting from the beta coefficient to the log odds.

## Analysis for Aim 3

The fitted regression model with WHF as the outcome variable, KLEVEL as the continuous predictor variable and TRTMT as a binary predictor yielded p-values of 0.7321 and ~0 for , respectively. This results in failure to reject the null hypothesis that the coefficient for KLEVEL is equal to 0 but, rejection of the null hypothesis for the coefficient of TRTMT meaning this is not equal to 0.

Assumptions:

1. The outcome variable, WHF, was defined to be binary.
2. Independence of observations is automatically met for our data given that each observation is an independent person.
3. Multicollinearity does not exist because the computed variance inflation factors are both < 10.
4. See appendix for plots showing linearity.
5. The total sample size and group sample sizes were deemed large enough, see table 1 for values.

Since the coefficient for KLEVEL is not significant, it can be concluded that there is no interaction between KLEVEL and TRTMT on WHF.

# Appendix

Results for Aim 1:

Output for stepwise, forward, and backward selection based on p-value:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | * Estimate | * Std. Error | * t value | * Pr(>|t|) |
| * **(Intercept)** | * 4.269 | * 0.02571 | * 166 | * 0 |
| * **DIABETES** | * 0.03158 | * 0.01547 | * 2.042 | * 0.04121 |
| * **CREAT** | * 0.03787 | * 0.01898 | * 1.995 | * 0.04608 |

Output for stepwise, forward, and backward selection based on AIC:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | * Estimate | * Std. Error | * t value | * Pr(>|t|) |
| * **(Intercept)** | * 4.252 | * 0.0344 | * 123.6 | * 0 |
| * **DIABETES** | * 0.03166 | * 0.01546 | * 2.047 | * 0.04068 |
| * **CREAT** | * 0.03744 | * 0.01897 | * 1.973 | * 0.04854 |
| * **PREVMI** | * -0.02419 | * 0.01468 | * -1.648 | * 0.09944 |
| * **SEX** | * 0.02733 | * 0.0168 | * 1.627 | * 0.1039 |

Fitted regression model diagnostic plots:

*Chart, diagram

Description automatically generated*

Results for Aim 2:

Output for logistic regression model:

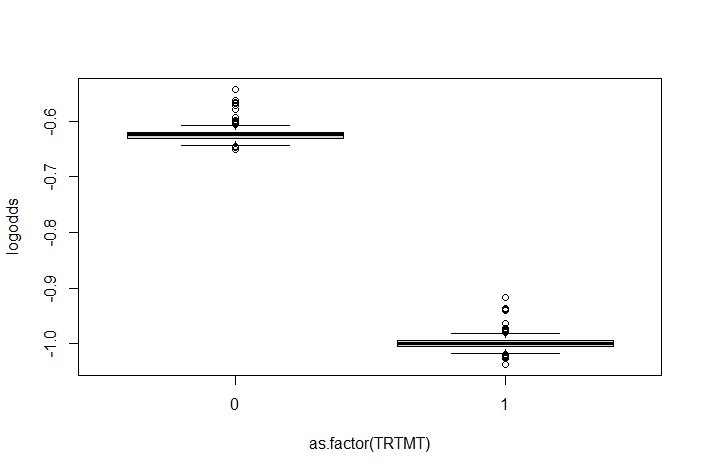
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| **(Intercept)** | -0.7008 | 0.2392 | -2.93 | 0.003386 |
| **KLEVEL** | -0.02429 | 0.05494 | -0.4421 | 0.6584 |

Aim 2:

Chart, scatter chart

Description automatically generatedLogistic regression assumption 4:

Aim 3:

Chart, scatter chart

Description automatically generatedLogistic regression assumption 4:

# References

* *RStudio*
* *tidyverse package*
* *table1 package*
* *olsrr package*
* *pander package*

