A Phase 2 Randomized, Single-Blind Study of a Single Dose of Peginterferon lambda-1a Compared with Placebo in Outpatients with Mild COVID-19

Statistical Analysis Plan

NCT04331899

September 14, 2020

Stanford MEDICINE | Quantitative Sciences Unit

Statistical Analysis Plan

Study Name *A Phase 2 Randomized, Single-Blind Study of a Single Dose of Peginterferon lambda-1a Compared with Placebo in Outpatients with Mild COVID-19*

Version Number 2.0

Date 14 September 2020

Approved by:

Manisha Desai
Electronically signed by: Manisha Desai
Reason: I am approving this document.
Date: Sep 14, 2020 11:18 PDT

Manisha Desai, PhD
Co-Investigator & Faculty Biostatistician

Haley Hedlin
Electronically signed by: Haley Hedlin
Reason: I am approving this document.
Date: Sep 14, 2020 10:57 PDT

Haley Hedlin, PhD
Lead Biostatistician

Upinder Singh
Electronically signed by: Upinder Singh
Reason: I have reviewed this document.
Date: Sep 14, 2020 11:32 PDT

Upinder Singh, MD
Principal Investigator

Revision history

Revision	Date	Section/Page	Changes Made -- Reasons for the Change
2.0	9/14/20	Throughout	Updated to match Protocol version 3
2.0	9/14/20	Sections 2 & 3	Details added about statistical methods and missing data handling

TABLE OF CONTENTS

INTRODUCTION

This statistical analysis plan (SAP) is a comprehensive and detailed description of the strategy, rationale, and statistical techniques that will be used in the study.

1 STUDY DETAILS

1.1 Study design

This phase 2 randomized, single-blind study intends to evaluate the efficacy of a single dose of peginterferon lamba-1a in outpatients with mild COVID-19 compared with placebo in reducing the duration of shedding of SARS-CoV-2 virus.

One hundred and twenty (60 per arm) enrolled patients will be randomized in a 1:1 fashion. Thus, there will be two arms: (1) the control arm, comprised of placebo in addition to standard supportive care and (2) the treatment arm, comprised of a single dose of peginterferon lamba-1a in addition to standard supportive care. An interim analysis for safety and overwhelming efficacy will be performed once 50% of patients have 24 hours of follow-up complete.

The primary analysis is based on intent-to-treat principles in that all patients who are randomized to study arm are included in the analysis, where cessation of shedding is assessed according to randomized study arm. The secondary analysis of adverse events (AE) are assessed according to dose received.

1.2 Primary endpoint

The primary endpoint is time until cessation of oral shedding of SARS-CoV-2 virus, defined as the time in days from randomization to the first of two negative oropharyngeal tests. Patients who do not experience the endpoint will be censored at 28 days after randomization (end of study) or at the time of premature withdrawal.

1.3 Secondary endpoints

Secondary endpoints include the following:

- Time until resolution of all symptoms;
- SARS-CoV-2 viral load over time in oropharyngeal tests;
- SARS-CoV-2 viral load area under the curve in oropharyngeal tests through Day 14;
- Number of emergency department visits or hospitalizations within 28 days of treatment.

1.4 Exploratory endpoints

Exploratory endpoints include the following:

- Time until sustained symptom resolution;
- Time until resolution of respiratory symptoms (cough, sore throat, runny nose, chest pain, and shortness of breath);

- Time until resolution of systemic and respiratory symptoms (headache, fever/chills, muscle/joint pain, fatigue and respiratory symptoms described above);
- Time until disease progression;
- SARS-CoV-2 infection-free by negative RT PCR testing results from nasal swabs on Days 5, 7, 10, and 14;
- SARS-CoV-2 infection-free by negative RT PCR testing results from oropharyngeal swabs on Days 5, 7, 10, and 14;
- SARS-CoV-2 viral load over time in nasal tests;
- SARS-CoV-2 viral load area under the curve in nasal tests through Day 14;
- Seroconversion by Day 28 (presence of IgG antibodies to SARS-CoV-2 Spike Receptor Binding Domain, as measured by ELISA);
- Cellular immune responses;
- Host genetics.

1.4 Safety endpoints

Safety endpoints include all adverse events by standardized DIADS AE grading criteria, vital signs, and clinical laboratory tests from time of randomization until the patient completes their study participation.

1.5 Randomization scheme

Patients will be randomized in a 1:1 ratio to the treatment and control arms. Randomization will be stratified by age (>=50 and <50 years old) and sex.

2 ANALYSIS PLAN

2.1 Analysis populations

The intent-to-treat (ITT) population will include all randomized patients. Patients will be analyzed according to their assigned treatment arm. All efficacy analyses will be completed in the ITT population.

The as-treated population will include all randomized patients. Patients will be analyzed according to the treatment actually received.

2.2 Descriptive analyses

Descriptive statistics (proportions for categorical variables, means, medians, standard deviations and interquartile ranges for continuous variables) will be reported for all key patient variables, including baseline and demographic characteristics, use of medications, compliance, and study completion status. Data that are missing on key patient characteristics and the outcome will be fully described, including any patterns of missingness (i.e., any relationships between missingness of a variable and patient characteristics).

A CONSORT diagram displaying the number of patients screened, eligible, and consented along with reasons for ineligibility will be provided. Graphical tools such as histograms, boxplots, and scatterplots will be created to assess quality of data and to display patterns over time.

All tests will be performed at the alpha = 0.05 level of significance unless otherwise noted.

2.3 Primary efficacy analysis

Time until shedding cessation will be compared between the two treatment arms using a two-sided Cox proportional hazards model adjusted for age and sex. The test will be performed at the alpha = 0.05 level of significance. The hazard ratio for shedding cessation will be estimated, along with its 95% confidence interval, from a Cox proportional hazards model. If the proportional hazards assumption is not met, we will consider an extended Cox model that relaxes the proportional hazards assumption.

The distribution of shedding cessation will be estimated using the Kaplan-Meier method, and Kaplan-Meier curves will be presented for each treatment arm. Time to shedding cessation at the end of the study period along with 95% confidence intervals will be presented for each treatment arm.

2.4 Secondary efficacy analyses

Time to resolution of all symptoms will be defined as the time from randomization until the first study day in which no symptoms are reported. The distribution of time to resolution of symptoms will be estimated using the Kaplan-Meier method, and Kaplan-Meier curves will be presented for each treatment arm. Time to symptom resolution at the end of the study period along with 95% confidence intervals will be presented for each treatment arm. The hazard ratio for resolution of symptoms will be estimated, along with its 95% confidence intervals, from a Cox proportional hazards model adjusted for age and sex.

We will model SARS-CoV-2 viral load in oropharyngeal tests over time using a generalized mixed effects linear regression model with a random effect for study participant to account for correlation within a participant. Viral load with cycle threshold (Ct) values above the limit of detection will be calculated by substituting a Ct value of 42 into the equation 10("#.%&&*()*++.##,) ∗ 120 to derive viral load from Ct. A logarithmic transformation on viral load will be fit to treatment arm, time, armXtime, and adjusted for age and sex to account for therandomization stratification variables. The armXtime interaction term will allow us to assess whether the trajectory of viral load differs by treatment arm. Time will be defined as days from randomization and, depending on model fit, will either be modeled continuously or with a flexible approach allowing a non-linear trajectory for the change in viral load over time. Model fit will be assessed based on visual inspection of the residuals. Alternative transformations or generalized linear models with a link function besides the identity link will be considered if the model does not fit the data well.

Area under the curve (AUC) will be defined by a single participant's logarithmic base 10 viral load from oropharyngeal tests collected on days 0, 1, 3, 5, 7, 10, and 14. AUC will be calculated by the trapezoidal rule, using exact times of collection of each OP swab. Multiple imputation using chained equations will be used to impute missing viral load data prior to AUC calculation. Five data sets will be imputed, and imputed values will be calculated using non-missing viral load on each of the 7 sample collection days, treatment arm, age, sex, and whether or not a participant was hospitalized. A pooled linear regression model will be fit to log-transformed AUC as a function of treatment, age, and sex using the 5 imputed data sets. Difference in AUC by treatment arm and corresponding 95% confidence intervals will be reported using Rubin's Rules. The difference in AUC by treatment arm will also be done after refitting the model without participants who were hospitalized to assess sensitivity of results.

The incidence of hospitalizations and emergency department visits will be estimated for each arm, with 95% confidence intervals.

2.5 Exploratory efficacy analyses

An alternative definition of the secondary endpoint of symptom resolution will be considered in an exploratory analysis. We will evaluate time to sustained symptom resolution, defined as time from randomization until the first day in which no symptoms are reported and are not reported for the duration of the study. The analysis of time to sustained symptom resolution will follow the methods described for the secondary endpoint evaluation symptom resolution in the previous section. Time to symptom resolution based on symptom subsets (i.e. respiratory, systemic) will be similarly analyzed, additionally excluding participants who did not report symptoms from the specified subset at screening.

Time to disease progression will be defined as the time from randomization until the first day of a two point or more increase in either cough or shortness of breath persisting for at least 48 hours or an increase to a score of 4 or 5, or first day of admission to the emergency department, hospitalization, or death. Analysis and reporting of results will be done following methods detailed for time to symptom resolution.

Freedom from SARS-CoV-2 infection by oropharyngeal and nasal swabs at Day 5, 7, 10, and 14 will be modeled using a multivariable logistic regression model adjusted for age and sex. The odds ratio will be presented along with its 95% confidence interval.

Similar approaches to those used for evaluating SARS-CoV-2 viral load in oropharyngeal tests over time and AUC will be used to evaluate viral load in nasal tests.

Correlation between virology and symptomology will be characterized by evaluating the relationship between CT values obtained from the oropharyngeal tests and each symptom through two main approaches. For symptoms graded on an ordinal scale, we will calculate a Spearman correlation and assess significance with a Z-test. For variables that capture presence of a symptom, differences in the median CT values by presence or absence of the symptom will be presented and assessed via a Wilcoxon rank sum test.

2.6 Safety analyses

The frequency of adverse events and serious adverse events will be tabulated by type and by treatment arm. AEs will be compared by arm using the Chi-squared test or Fisher's exact test, as appropriate, in the safety analysis set.

2.7 Effect modification, subgroup, and sensitivity analyses

A subgroup analysis will be performed for the primary endpoint using only patients who present with symptoms at baseline.

A statistical interaction term between treatment arm and the following baseline characteristics will be added to the Cox proportional hazards model to test for effect modification.

1) CT value < 30 from the baseline OP swab,

2) presence of antibodies as defined by presence of IgG antibodies to SARS-CoV-2 Spike Receptor Binding Domain (RBD) at baseline, as measured by ELISA. Samples are considered seropositive against RBD if their absorbance value was greater than the mean of plus four standard deviation (SD) of all negative controls (n=130),

3) number of risk factors or predictors for severe disease present at baseline (temp 99.5+, cough, or shortness of breath present at randomization [symptoms count as a single risk factor], age ≥ 60, male sex, Black race, Hispanic ethnicity, BMI ≥ 30, and lab values of baseline lymphocyte counts <1000 and baseline ALT ≥ 94),

4) age ≥ 50,

5) sex

A Wald test performed at the alpha = 0.05 level of significance will be used to test whether the effect modification is statistically significant. The sensitivity and effect modification analyses will be performed in other key endpoints as well.

3 HANDLING OF MISSING AND CENSORED DATA

All efforts will be made to minimize instances of missing data. However, we expect some missing data will occur. Our analyses will assume data are missing at random. For the primary endpoint of time to shedding cessation, any participant who drops out prior to having two consecutive negative tests will be censored at their last positive test. Participants who have no post-randomization test results available or a single test result that is negative will be censored at day 1. Similarly, participants will be right censored for the time to symptom resolution or progression at the time when their last symptom questionnaire is completed. Participants who are asymptomatic at screening and never report symptoms thereafter will be considered to have met symptom resolution at study Day 1. For other missing data, we will use multiple imputation-based analyses. This approach will be applied to any analysis involving endpoints or key

variables where any missing data occurs in order to adhere to the ITT principle. Sensitivity to assumptions regarding missingness will be addressed through sensitivity analyses.

If more than 5% of participants do not have the time of shedding cessation observed (i.e. they are right-censored), we will perform a sensitivity analysis to evaluate the robustness of our findings to missing data assumptions. In the sensitivity analysis, among those missing timing of shedding cessation, we will assume varying proportions (e.g. 25%, 50%) of participants in each arm who would have ceased shedding by Day 28. The proportions will differ by arm so we can evaluate whether our findings stand if it were true that people who dropped out stopped shedding earlier in the placebo arm vs the active arm.

4 SAMPLE SIZE CONSIDERATIONS

4.1 Statistical hypothesis

For the primary comparison, the following will be tested:

- Null hypothesis: time to cessation of oral shedding is equal in control and treatment;
- Alternative hypothesis: time to cessation of oral shedding differs between control and treatment.

Hypothesis tests will be two sided and conducted at an overall alpha = 0.05 level of significance.

4.2 Determination of sample size

We assume that the mean duration of virologic shedding in individuals receiving standard of care is 14 days, with a standard deviation of 4 days[1].

Assuming 1:1 randomization and the use of a two-sided log rank test at the alpha=0.04999 level of significance for the final analysis, 79 events will provide 80% power to detect a hazard ratio of 2.03. This leaves alpha=0.00001 to check for overwhelming efficacy after 50% of participants have completed 24 hours of follow-up. Assuming the control and treatment arm median cessation of shedding is 14 and 7 days, respectively, a two-month accrual period, a two-week follow-up period after randomization of the last patient, and a drop out of 10% in the control arm, it is estimated that the total sample size required to achieve 79 events is 120 (60 patients in each arm).

5 INTERIM ANALYSIS

An interim analysis for safety and overwhelming efficacy will be performed once 50% of patients have 24 hours of follow-up complete. Enrollment will pause once 50% of patients have received treatment and will remain paused until the DSMC makes their recommendation after their safety review. The DSMC will meet one or two days after enrollment is paused to review the safety data collected within the first 24 hours of follow-up on all enrolled patients. We

additionally expect blood count labs collected at day 5 to be available in approximately 25% of patients at the time of the DSMC review.

The DSMC will also review the efficacy data on all randomized participants at this meeting. The interim efficacy analysis will use the same methods as are planned for the final analysis using the ITT analysis. Based on the results of the interim analysis, the DSMC will either recommend to the sponsor to terminate the study for overwhelming efficacy ($p<0.00001$ at the interim analysis), terminate the study for safety concerns, modify the study, or continue the study as planned. No formal stopping rules for futility are planned.

6 POTENTIAL LIMITATIONS AND MITIGATION STRATEGIES

Given the fast-changing landscape during the COVID-19 pandemic, we provide the following power calculations showing how potential study alterations would impact the study's operating characteristics. Potential alterations considered are designed to mitigate the impact of newly identified therapies that change the standard of care and/or slower than planned enrollment.

The study team plans to evaluate the rate of enrollment two weeks after enrollment opens. If fewer than 10 patients have been enrolled within the first two weeks, a potential solution to increase enrollment is to change the randomization ratio to be 2:1 on treatment:control. We have selected a 1:1 ratio under the assumption of equipoise; however, we anticipate that participants may not want to have a 50% chance of being randomized to placebo. To address this concern, we could increase the ratio in order to reduce the chance of being randomized to placebo to 33%. Under the assumptions described above, the change in the randomization ratio will not require a larger sample size.

We also consider how a change in the standard of care could impact our study's power. If the standard of care changes and our expected effect size decreases, our expected power would decrease to 61%, 41%, or 22% under an assumed hazard ratio of 1.8, 1.6, or 1.4, respectively. The impact of the change to the standard of care will depend on what proportion of patients have been enrolled. If the standard of care changes before enrollment is 90% complete, we will estimate the conditional power based on the current enrollment and the expected decrease in the effect size; we would not look at the blinded study data or outcome data to estimate the conditional power. If the conditional power is < 70%, the study will be modified by increasing the sample size such that the conditional power reaches 80%. If the total required sample size to reach 80% conditional power is greater than 180, the study will be stopped for futility.

7 REFERENCES

1. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395(10229):1054-1062. doi:10.1016/S0140-6736(20)30566-3