| | |
|---|---|
| **Official Title:** | A Phase III, Multicenter, Randomized, Double-Blind, Placebo-Controlled, Parallel-Group Study to Assess the Efficacy and Safety of Tocilizumab Versus Placebo in Patients With Systemic Sclerosis |
| **NCT Number:** | **NCT02453256** |
| **Document Dates:** | SAP version 2: 23-Jan-2018 |

# STATISTICAL ANALYSIS PLAN

| | |
|---|---|
| **TITLE:** | **A PHASE III, MULTICENTER, RANDOMIZED, DOUBLE BLIND, PLACEBO-CONTROLLED, PARALLEL-GROUP STUDY TO ASSESS THE EFFICACY AND SAFETY OF TOCILIZUMAB VERSUS PLACEBO IN PATIENTS WITH SYSTEMIC SCLEROSIS** |
| **PROTOCOL NUMBER:** | WA29767 |
| **STUDY DRUG:** | Tocilizumab (RO4877533) |
| **VERSION NUMBER:** | 2 |
| **IND NUMBER:** | 112406 |
| **EUDRACT NUMBER:** | 2015-000424-28 |
| **SPONSOR** | F. Hoffmann-La Roche Ltd |
| **PLAN PREPARED BY:** | [redacted] |
| **DATE FINAL:** | Version 1: 8 August 2017 |
| **DATE(S) AMENDED:** | Version 2: See electronic date stamp below. |

## STATISTICAL ANALYSIS PLAN AMENDMENT APPROVAL

| Name | Reason for Signing | Date and Time (UTC) |
|---|---|---|
| [redacted] | Company Signatory | 23-Jan-2018 09:28:29 |

# STATISTICAL ANALYSIS PLAN AMENDMENT RATIONALE

The Statistical Analysis Plan WA29767 Version 2 was amended as follows:

- Section 2.2.2 (Secondary Efficacy Outcome Measures): Change from baseline in mRSS and FVC at Week 24 were listed as exploratory endpoints in the protocol. However, the analysis of these endpoints is part of the multiplicity testing procedure and the resulting P-values will not be considered exploratory. The mRSS and FVC at Week 48 are key endpoints, and showing a treatment effect following 24 weeks of treatment is considered important.
- Section 3.2.1 (Definition of SSc-related Complications): For SSc-related complications clarified that these are adverse events that are adjudicated independently by committee before the study is unblinded.
- Section 4.6.1 (Controlling for Type 1 Error): Changed the order of secondary endpoints in the multiplicity hierarchy
- Section: 4.6.3.2 (Time to Treatment Failure): Clarified that time to an event will be relative to first dose of study drug, rather than date of randomisation as specified in the protocol since only treatment emergent adverse events are being adjudicated. Defined censoring for the analysis (previously omitted).
- For events deemed SSc-related complications only serious events will be included in the time to treatment failure endpoint. Although non serious Grade 3 and 4 events are also being adjudicated, the information made available on these events to the adjudication committee is limited to eCRF data as detailed patient narratives are not available for non-serious events.
- Included time to event analysis of the components of time to treatment failure.
- Defined a sensitivity analysis of time to treatment failure where discontinuation from study for the following reasons will also be classed as a treatment failure: death, lack of efficacy, lost to follow up, withdrawal by subject, physician decision (whereas withdrawal from study for the following reasons will not be counted as an event: adverse event, pregnancy, protocol violation, non-compliance, study terminated by Sponsor, other).
- Clarified that Serious SSc-related complications will be summarised by system organ class (SOC), and will be further subcategorised descriptively for assessment of internal organ involvement (including, but not limited to, cardiac disease, renal disease, and GI disease).
- Section 4.6.3.3 (Binary Endpoints): For the binary endpoints of ≥20%, 40%, and 60% improvement in mRSS at Week 48, there will be an additional analysis whereby data from patients post initiation of immuno-modulation therapy (e.g., DMARDs and/or Biologics) will be censored so that these patients will become non responders in the analysis.

- Section 4.6.5.1(Sensitivity to Primary Estimand): Added details of a tipping point analysis for the primary estimand, and expanded on details of the pattern mixture model to allow programming of this analysis to be clear from the SAP text. As a sensitivity to the primary estimand, a linear regression analysis (with Huber–White sandwich errors) of change from baseline in mRSS at Week 48 is added.
- Section 4.6.5.3 (Sensitivity to Secondary Endpoints): Added tipping point analyses and pattern mixture models to the secondary endpoints included in the multiplicity testing procedure.
- Section 4.6.5.3.2 (Subgroup Analyses): Added further subgroup analyses for additional disease characteristics and baseline demography; namely baseline mRSS, sex, age, race, and region.
- Removed Appendix 3: Details of date imputations will be included in the programming documentation
- Appendix 10: Added details of the Cochran–Mantel–Haenszel test for difference in proportions.

Additional minor changes have been made to improve clarity and consistency.

## TABLE OF CONTENTS

## LIST OF TABLES



## LIST OF FIGURES



## LIST OF APPENDICES

## 1. BACKGROUND

This document describes the statistical methods to be used for the analysis of the clinical efficacy, clinical safety, pharmacokinetic (PK), and pharmacodynamic (PD) data from the Phase III Study WA29767 (FocuSSced). The study drug tocilizumab (TCZ, RO4877533) is a recombinant humanized monoclonal antibody directed against the human interleukin 6 (IL-6) receptor. Rheumatoid arthritis (RA) and systemic sclerosis (SSc) have been linked to the effects of IL-6 on cell proliferation.

Study WA29767 is a multicenter, randomized, double-blind, placebo-controlled, two-arm, parallel-group trial in patients with SSc. There is a 48-week blinded period followed by a 48-week open-label period.

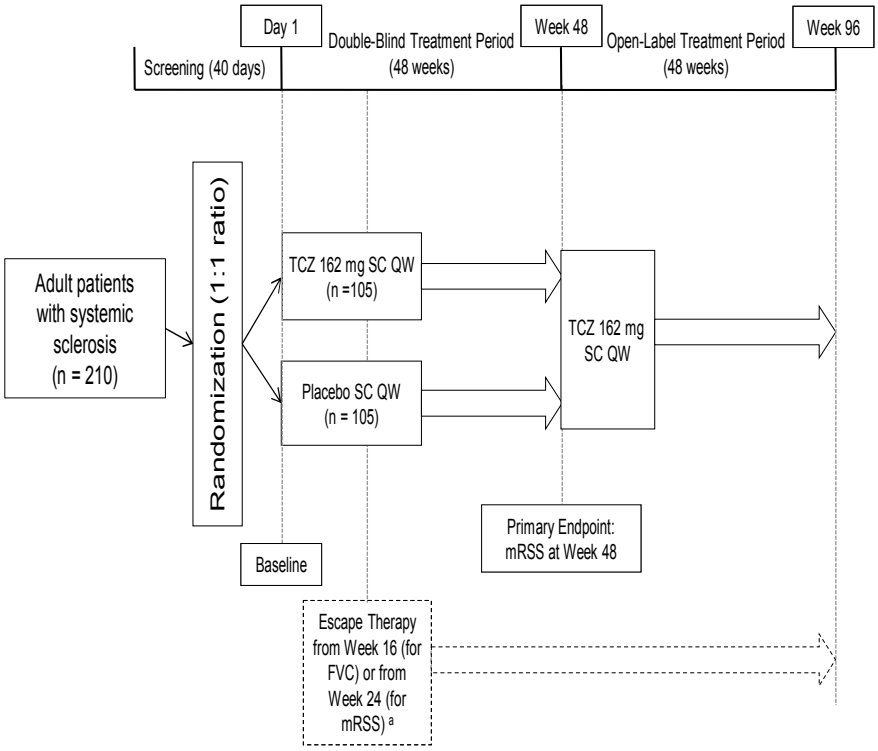
The primary endpoint is the difference in the change in modified Rodnan skin score (mRSS) from baseline at Week 48.

This document only describes the statistical methods and analyses used in the reporting of the double-blind phase of the study, up to Week 48. The analysis of biomarkers, including skin biopsies, will be covered in a separate document. The exploratory data analysis up to Week 96 will be described in a separate analysis plan.

## 2. STUDY DESIGN

This Phase III, multicenter, randomized, double-blind, placebo-controlled, two-arm, parallel-group study is designed to assess the efficacy and safety of TCZ in patients with SSc. The study consists of two periods: a 48-week, double-blind, placebo-controlled period, followed by a 48-week open-label treatment period. Patients will be randomized in a 1:1 ratio to receive SC injections of 162 mg of TCZ QW (once weekly) or placebo QW for 48 weeks during the double-blind treatment period. During the open-label treatment period, all patients will receive SC injections of 162 mg of TCZ QW for up to 48 weeks. Patients receive their first dose of open-label treatment at Week 48. Approximately 210 patients will be enrolled at approximately 75 global sites. The study design is presented schematically in Figure 1.

**Figure 1 Study Schema**




FVC=forced vital capacity; mRSS=modified Rodnan Skin Score; n=number; QW=once weekly; SC=subcutaneous; TCZ=tocilizumab.
[a]Also includes escape therapy for worsening of systemic sclerosis.

## 2.1 PROTOCOL SYNOPSIS

The Protocol Synopsis is in Appendix 1 and includes the study objectives, inclusion and exclusion criteria, outcome measures, and statistical methods as stated in the Protocol.

## 2.2 OUTCOME MEASURES

All assessments and procedures are detailed in the Protocol.

### 2.2.1 Primary Efficacy Outcome Measures

Change in mRSS from baseline at Week 48.

### 2.2.2 Secondary Efficacy Outcome Measures

- Proportions of patients with ≥20%, ≥40%, and ≥60% improvement in mRSS at Week 48 compared with baseline
- Change in forced vital capacity (FVC) from baseline to Week 48
- Change in Health Assessment Questionnaire Disability Index (HAQ-DI) from baseline to Week 48
- Change in Patient's Global Assessment from baseline to Week 48

- Change in Physician's Global Assessment from baseline to Week 48
- Time to treatment failure at Week 48

Analyses of the following endpoints, which were listed as exploratory endpoints in the protocol, have been included in the multiplicity testing procedure and the resulting p-values will not be considered exploratory (Section 4.6.1):

- Change in FVC (forced vital capacity) from baseline to Week 24
- Change in mRSS from baseline to Week 24

### 2.2.3 Exploratory Efficacy Outcome Measures

- Proportions of patients who achieve a response, as determined by the investigator using Combined Response Index for Systemic Sclerosis (CRISS), at Week 48
- Change in the Visual Analogue Scale (VAS) component of the Scleroderma Health Assessment Questionnaire (SHAQ) from baseline to Week 24 and baseline to Week 48
- Change in Work Productivity and Activity Impairment Questionnaire: General Health (WPAI-GH) score from baseline to Week 24 and baseline to Week 48
- Change in EuroQol 5-Dimension Questionnaire with three levels of severity (EQ-5D-3L) score from baseline to Week 24 and baseline to Week 48
- Change in total score and subscores of Saint George's Respiratory Questionnaire (SGRQ) from baseline to Week 48
- Change in total and domain scores of the Scleroderma Skin Patient-Reported Outcome (SkinPRO) questionnaire from baseline to Week 48 (administered to patients in North America only)
- Change in Functional Assessment of Chronic Illness Therapy (FACIT)-Fatigue score from baseline to Week 48.
- Change in high-resolution computed tomography (HRCT) fibrosis score from baseline (based on HRCT scan performed within 12 months prior to screening) to Week 48
- Change in diffusion capacity of the lung for carbon monoxide ($DL_{co}$) from baseline to Week 48
- Proportion of patients with ≥15% decline in observed $DL_{CO}$ at Week 48 relative to baseline
- Proportion of patients with ≥15% decline in percentage of predicted $DL_{CO}$ at Week 48 relative to baseline
- Proportion of patients with ≥10% decline in observed FVC at both Week 24 and Week 48 relative to baseline
- Proportion of patients with ≥10% decline in percentage of predicted FVC at Week 24 and at Week 48 relative to baseline

- Proportion of patients achieving MCID (change ≥0.22 in the HAQ-DI ) in the HAQ-DI at Week 48

### 2.2.4 Pharmacokinetic Outcome Measures

The PK and exposure-effect outcome measures for this study are as follows:

- Predose observed serum TCZ concentration at baseline and at specified timepoints thereafter ($C_{trough}$).
- Correlation between PK parameters (e.g., $C_{trough}$) for TCZ and efficacy (e.g., mRSS or any positively trending efficacy outcome noted in Sections 2.2.1 and 2.2.2, safety (CTC grades for neutrophils, platelets, ALT, and AST, or any negatively trending safety outcome noted in Section 2.2.7), or immunogenicity outcome measures, as relevant. $C_{trough}$ may be categorized into exposure bins (e.g., high, medium, and low) to assist in identifying trends.

### 2.2.5 Pharmacodynamic Outcome Measures

The PD outcome measures for this study are as follows:

- Serum sIL-6R, IL-6, erythrocyte sedimentation rate (ESR) and C reactive protein (CRP) levels predose at baseline and at subsequent time points after initiation of study drug.

### 2.2.6 Immunogenicity Outcome Measures

The immunogenicity outcome measures for this study are as follows:

- Incidence of anti-drug (TCZ) antibodies (ADA) during the study relative to the prevalence of anti-TCZ antibodies at baseline
- Correlation between ADA status and efficacy, safety or PK measures

### 2.2.7 Safety Outcome Measures

The safety evaluations will consist of the following:

- Nature, frequency, and severity of adverse events (AEs) and serious AEs
- Frequency of deaths
- Incidence of specific laboratory abnormalities
- Change from baseline in digital ulcer count
- Frequency of SSc-related complications as determined by the Clinical Adjudication Committee

## 2.3 DETERMINATION OF SAMPLE SIZE

A sample size of approximately 105 patients in the TCZ group and 105 patients in the placebo group (a total of 210 patients) will give power in the range of >75% to 80%, (allowing for an estimated patient dropout rate of approximately 15% to 20%) to detect a between-group difference of 3.55 units (common standard deviation of 8.43) in mean change in mRSS from baseline to Week 48 using a two-group t-test, with a 5% two-sided significance level. The minimal detectable difference (smallest treatment

difference that would give a statistically significant result) under these assumptions, and with a patient dropout rate of 20%, is approximately 2.6.

The loss of power at Week 48 due to performing a futility analysis based on mRSS at Week 24 (see Section 4.12) on a subset of patients was minimal, at approximately 3% based on simulations.

The following power calculations for key secondary endpoints use the maximum available alpha according to the multiplicity procedure (Section 4.6.1).

The required sample size of 210 patients based on the primary endpoint, gives 94% power for change from baseline in percent predicted FVC at week 48 at the 4% two-sided significance level, assuming 15% dropout (mean difference 0.043, common standard deviation 0.075, Wilcoxon–Mann–Whitney two sample test).

Assuming 70% of patients on TCZ are event free at 48 weeks of follow up, 50% of patients event free on placebo (hazard ratio [HR] 0.52) gives a power of 80% at the two-sided 4.5% significance level for the time to treatment failure endpoint based on 210 patients with 15% dropout at 48 weeks.

Under the same assumptions for number of patients and dropouts the power for the change from baseline in HAQ-DI at Week 48 is 65% using a two-group t-test at the two-sided 5% significance level (mean difference 0.181, common standard deviation 0.517).

## 2.4 ANALYSIS TIMING

The primary study analysis will occur when the last patient has either withdrawn or completed his or her Week 48 visit, and will be based on cleaned data for all patients up to and inclusive of their Week 48 assessment date, as well as safety follow-up data for patients that withdrew prior to Week 48. A Week 48 Clinical Study Report (CSR) based on these analyses will be produced.

There will be a final analysis and CSR when all patients have reached the end of the open-label period (Week 96) of the study and/or have completed their safety follow-up visits. The analyses for Week 96 data will be covered in a separate Statistical Analysis Plan.

# 3. STUDY CONDUCT

## 3.1 RANDOMIZATION ISSUES

At baseline, patients will be randomized (stratified by IL-6 level ($<$10; $\geq$10 pg/mL) at screening) to study drug assignment and assigned randomization and study drug assignment numbers through an interactive voice/web response system (IxRS). Patients will be randomized in a 1:1 ratio to receive either 162 mg of subcutaneous (SC)

TCZ weekly (QW) (Group A) or SC placebo QW (Group B) for 48 weeks. To ensure randomization is carried out correctly, Roche Internal User Acceptance Testing of the IxRS system will be conducted prior to the first patient being enrolled into the study. In addition, the independent data coordinating center (iDCC) will review IxRS data as part of the independent data monitoring committee (iDMC) preparation.

## 3.2 DATA MONITORING

An external iDMC will review unblinded safety data and will convene at least two times per year. Analyses required for the iDMC's data review will be performed as described in the iDMC Charter.

A review for safety will be performed at the same time as the futility analysis when approximately 76 patients have completed 24 weeks of treatment (see Protocol Section 6.9).

### 3.2.1 Definition of SSc-Related Complications

A Clinical Adjudication Committee (AC), an independent and blinded expert clinician panel, adjudicated serious AEs and Grade 3 and 4 AEs with regard to their classification as SSc-related complications. Serious AEs only (not Grade 3 and 4 events that are without patient narratives) positively adjudicated as SSC-related complications will be used for the analysis of time to treatment failure.

These adjudication decisions are captured in the electronic Case Report Form (eCRF), and the review of these adverse events will be completed before study unblinding. Full details of the adjudication committee process have been documented in a charter which will be made available.

# 4. STATISTICAL METHODS

## 4.1 ANALYSIS POPULATIONS

Disposition summaries will be based on the All Patients population. Analysis of safety data will be based on the safety population. The primary analysis population for efficacy will be the intention to treat (ITT) population.

### 4.1.1 All Patients

The All Patients population will include all subjects randomized in the study.

### 4.1.2 Intent-to-Treat Population

The ITT (intent-to-treat) population will be a modified ITT and include all subjects randomized in the study who received any study drug. The treatment group for this population will be defined according to the treatment assigned at randomization by IxRS.

### 4.1.3 Pharmacokinetic-Evaluable Population

The PK population will include all patients who received at least one TCZ injection and had at least one PK sample with detectable results.

### 4.1.4 Safety Population

The safety population will include all patients who received any study drug and provided at least one post-dose safety assessment (withdrawal, AE, death, laboratory assessment, vital signs). Patients will be included in the treatment arm for the treatment most frequently received at the time of the analysis.

## 4.2 ANALYSIS OF STUDY CONDUCT

The patients excluded from the safety and ITT populations will be summarized, including the reason for exclusion, by treatment group. A summary of enrolment by country and investigator name will be produced.

The number of patients that completed or discontinued from the study by Week 48, including a reason for discontinuation, will be summarized by treatment group. A listing of withdrawals for the Week 48 analysis will be produced.

Total duration of patient participation in the study will be summarized for the safety population at Week 48. Total duration on study will be calculated as the date of the last safety assessment (including study drug exposure) up to and including the Week 48 cut date, minus the date of first injection of study drug, plus 1 day.

Duration of study treatment will be summarized for the safety population at Week 48, calculated as the date of the last injection up to and including the Week 48 cut date, minus the date of first injection of study drug, plus 1 day.

Compliance will be summarized by dose intensity (the number of doses actually received divided by the expected number of doses), the cumulative dose, and the number of patients with missed doses.

The number and frequency of patients first receiving immuno-modulating treatments (e.g., disease-modifying anti-rheumatic drugs [DMARDS]) will be summarized by visit up to Week 48 (refer to Section 4.6.5.3).

A listing of the number of patients in each treatment group that were randomized according to an incorrect stratification value given at baseline will be produced at Week 48 (IxRS data will be compared to the clinical database).

A listing of patients with major protocol deviations including those that entered the study without meeting all inclusion/exclusion entry criteria will be produced.

## 4.3 ANALYSIS OF TREATMENT GROUP COMPARABILITY

To descriptively assess the comparability of treatment arms (placebo versus TCZ 162 mg SC) at baseline, summary tables will be produced for the safety population for clinically important baseline demographic and disease characteristics by treatment received. If it is found that there is a large difference in the number of patients in the safety population compared to the ITT population, additional summaries will be produced as randomized.

These summary tables will include number of patients, mean, standard deviation, and median for continuous demographic/disease characteristics and number and percentage of patients for categorical characteristics.

Demographic and baseline disease characteristics will be summarized as described in the following sections.

### 4.3.1 Demographics

- Sex
- Age
- Height
- Weight
- Race
- Ethnicity
- Country
- Reproductive status
- Smoking history

### 4.3.2 Disease Characteristics

- IL-6
- IL-6 level ($<10$; $\geq 10$ pg/mL)
- Duration of SSc
- mRSS
- CRP
- ESR
- Platelets
- HAQ-DI
- Physician's Global VAS
- Patient's Global VAS
- % predicted FVC

- % predicted DLCO (hemoglobin corrected)
- Anti-topoisomerase (positive ≥20 U/mL)
- Anti-RNA polymerase (positive ≥20 U/mL)
- Anti-centromere (positive ≥1:40 dilution)
- Anti-nuclear antibodies (ANA) (positive ≥1:40 dilution).

The number of patients positive for hepatitis B surface antigen (HBsAg) and hepatitis C antibody (anti-HCV) at screening will be summarized for the safety population.

Medical history data, including surgery and procedures and baseline conditions, will be summarized descriptively by treatment group using the safety population. Descriptive summaries of any previous and concomitant treatment will be produced by treatment group. A glossary showing the mapping of investigator verbatim terms to diseases will be produced for the medical history data included in the analysis for Week 48.

## 4.4 DATA CUT FOR ANALYSES

For each patient, the day of first injection of study treatment will be designated study day 1. Each subsequent assessment point will be assigned a study day calculated as (date of assessment−date of first injection of study treatment) + 1.

Efficacy datasets will be cut at the Week 48 exposure date plus 28 days. The exposure date will be the date of the Week 48 study drug injection, identified as the injection that occurs within the protocol time window of 337 (+/−7 days). Typically, this will be the first open label dose. If more than one injection occurs in the time window then the study day of the earliest open label drug will be used, or, if there is no open label dose in the time window, the day of the latest injection in the window will be used. If the Week 48 study drug injection is missing the data sets will be cut at study day 337 plus 28 days.

The efficacy data within the cut datasets that is included for analysis, in particular for the Week 48 visit, is described in Section 4.5.

For safety, the data will be cut at the day prior to the Week 48 exposure date or the day prior to the first dose of open label treatment, whichever is later. If the Week 48 exposure date is missing, the data will be cut at study day 337. All data including data collected in safety follow-up visits or spontaneously reported AEs will be included in safety summaries.

There will also be a final analysis of all available data including all visits up to Week 96 and the 8 weeks of safety follow-up, which will be described in a separate analysis plan.

## 4.5 VISIT WINDOWS

All patient assessments within the cut efficacy datasets that are collected by scheduled visits will be assigned to a study week using the actual study day of the assessment; this

includes withdrawal visits and any unscheduled visits. Time windows will be continuous from the midpoint between two consecutive study visits to the next midpoint, and will be dependent on the schedule of assessments (as given in Appendix 2, Appendix 3, and Appendix 4) for each variable independently. An example table of time windowing for the mRSS and FVC is shown below (Table 1). Mapping of other variables will similarly be based on the scheduled visits. Data will never be mapped to visits for which the assessment was not scheduled in the Protocol.

**Table 1 Time Windows for Assigning Assessment Study Days to Study Visits (Weeks) for Efficacy (mRSS and FVC)**

| Study Visit | Scheduled Study Day +7days | Efficacy Time Window |
|---|---|---|
| Baseline | 0 | ≤1 |
| Week 8 | 56 | >1 to≤84 |
| Week 16 | 112 | >84 to≤140 |
| Week 24 | 168 | >140 to≤210 |
| Week 36 | 252 | >210 to≤294 |
| Week 48 | 336 | >294 *to study day of first OL dose |

*Except for FVC, HRCT and DLco, where the upper bound of week 48 is the cut date, i.e. week 48 exposure + 28 days regardless of open label dosing.

For FVC assessments of insufficient quality, repeats were requested during study conduct within 4 weeks of the initial assessment, therefore for FVC (and HRCT, DLco) all data within the cut (Week 48 exposure plus 28 days) will be mapped to Week 48. Any efficacy assessments that occurred after the study day of first open label dose that is used in the analysis or summary tables will be flagged in the relevant listing.

For all efficacy assessments except for mRSS, if more than one particular efficacy assessment occurs within the same time window, then the nearest non-missing assessment to the nominal time point will be assigned to that visit. If there are two (or more) efficacy assessments that occur the same time away from the nominal time point, then the latest assessment will be assigned to that visit. For multiple mRSS assessments that occur within the same time window, in order to account for repeated assessments to allow assessor consistency the latest assessment within a time window will be used.

For calculations of percent predicted $DL_{CO}$, the hemoglobin value nearest to the $DL_{CO}$ assessment will be used.

If more than one particular safety assessment (e.g., laboratory result or vital sign) occurs within the same time window then the worst value will be assigned to the visit. The last

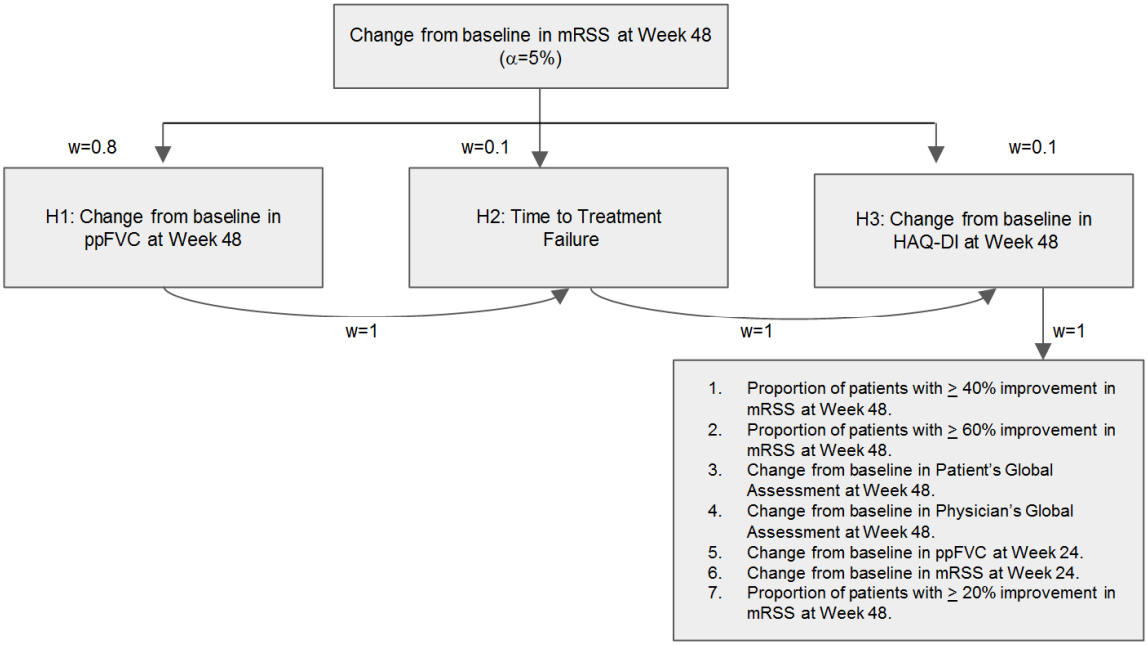
value from screening will be used for baseline assessments of safety if there is no baseline (study day 1) value.

For summaries of data not collected by visit, such as AEs, medical history, and concomitant medications, all data in the Week 48 safety data cut will be included.

## 4.6 EFFICACY ANALYSIS

### 4.6.1 Controlling for Type 1 Error

The following diagram shows the order and weighting applied to the testing of the key clinical endpoints, including primary, secondary and some exploratory endpoints. This roll-back approach will preserve the overall type I error of 5%. Please note that no adjustment to the alpha level for mRSS at Week 24 will be made to account for the futility analysis (Section 4.12).




Assuming the primary efficacy endpoint is significant at the 5% level, then percent predicted FVC (ppFVC) will be weighted at 0.8 and will hence be tested at a 4% alpha. Both time to treatment failure and HAQ-DI will be weighted at 0.1 so that if ppFVC fails to meet significance at the 4% level, each can still be tested at a minimum of 0.5% significance level ($\alpha$=0.005). If ppFVC is significant, then the alpha will be transferred to the time to treatment failure endpoint which can then be tested at a 4.5% alpha. If ppFVC and time to treatment failure are significant at the 4% and 4.5% alpha respectively, then the alpha will be passed to the HAQ-DI hypothesis and HAQ-DI will be tested at an alpha of 5%. If ppFVC is not significant, but time to treatment failure is, the 0.5% alpha can be transferred to HAQ-DI which can then be tested at the 1% level. If the time to treatment failure endpoint is not significant then HAQ-DI will be tested at an alpha level of 0.5%.

If HAQ-DI is significant at the acquired significance level then the other secondary endpoints listed below HAQ-DI will be tested in order at the alpha level passed down by HAQ-DI. If the endpoint being tested is non-significant at the acquired significance level then that end point and those below in the hierarchy will be non-significant, although p-values will still be produced. If HAQ-DI is non-significant, none of the subsequent secondary endpoints will be claimed for statistical significance and instead descriptive p-values will be presented.

If the primary endpoint is not significant at the 5% level the testing will stop.

### 4.6.2 Primary Efficacy Endpoint

The estimand of interest for the primary analysis is the difference between treatment arms in the mean change in the mRSS at Week 48 for the ITT population in the regimens as actually taken. The study has been designed to continue to capture efficacy data on patients who discontinue study drug prematurely and may receive additional immuno-modulating treatments (e.g., DMARDs) and in patients who receive escape therapies (i.e., immuno-modulating treatment in addition to study drug) during the double-blind treatment period. All data will be included in the primary analysis. The primary analysis specified above assumes a missing-at-random missing-data mechanism whereby patients who are lost to follow-up from the TCZ arm will tend to have similar efficacy to that of patients on TCZ who remained in the study.

#### 4.6.2.1 Definition of Modified Rodnan Skin Score

The mRSS is a measure of skin thickness across 17 different body sites and has a range of 0 to 51 in whole units (see Protocol Section 4.5.9 for further details). There will be no imputation of missing data prior to the analysis detailed in Section 4.6.2.2 and the total scores as collected in the site electronic device will be used, rather than derived total scores. All missing assessments will be adjusted for within the model fitting process.

#### 4.6.2.2 Statistical Analysis

The primary efficacy endpoint is the change in mRSS from baseline at Week 48. The mean change from baseline will be analyzed using a restricted maximum likelihood-based repeated measures approach. The analysis will include the fixed, categorical effects for treatment, visit, the stratification factor IL-6 level ($<$10; $\geq$10 pg/mL) at screening, IL-6 level at screening-by-visit interaction, and treatment-by-visit interaction, as well as the continuous covariates of baseline mRSS score and baseline mRSS score–by–visit interaction. Analyses will be implemented in SAS® using PROC MIXED. An unstructured covariance structure will be used to model the within-patient errors. The Kenward–Roger approximation (Kenward and Roger 1997) will be used to estimate the denominator degrees of freedom.

The primary treatment comparison will be the contrast between treatments at the Week 48 timepoint. Least square means, along with a difference in means, a 95% CI for

the difference in means, and the p-value will be reported. The significance test will be based on a two-sided $\alpha$=0.05.

#### 4.6.2.3 Stratification Factors

Efficacy analyses will adjust for the stratification factor at randomization, which is the categorical variable IL-6 level (<10; ≥10 pg/mL) at screening. Data from the IxRS used for the randomization, not the actual value of IL-6 at screening, will be used.

### 4.6.3 Secondary Efficacy Endpoints

Further details on the definition of secondary endpoints are provided in Section 6.4.2 of the Protocol.

#### 4.6.3.1 Continuous Endpoints

For all continuous secondary endpoints, except for FVC, the primary method of analysis for change from baseline will be a repeated measures model to compare the treatment effect at Week 48, as described above in Section 4.6.2.2.

Least square means, along with a difference in means, a 95% CI for the difference in means, and the p-value will be reported at Week 48.

The least square means of the change from baseline by treatment group in all primary and secondary continuous endpoints, along with 95% CIs will be plotted and tabulated by visit up to Week 48.

No imputation will be made for missing assessments prior to analyses. All missing data will be adjusted for within the model fitting process where applicable.

The primary method of analysis for change from baseline in FVC endpoints will be a non-parametric analysis method based on the expectation that the distribution of data will be non-normal. Change from baseline to Week 48 in percent predicted FVC (ppFVC) will be analyzed using a Van Elteren test stratified by screening IL-6 level (<10; ≥10 pg/mL). The median change from baseline for each treatment group and the corresponding 95% CI for the median will be presented for ppFVC, along with the Van Elteren p-value. A cumulative distribution plot will be produced for both FVC endpoints (absolute and percent predicted).

The median change from baseline, along with 95% CIs will be summarized and plotted by visit up to Week 48.

Additional parametric analyses consistent with the method used for other secondary endpoints will also be conducted for the FVC.

**Forced Vital Capacity (FVC):**

Percent predicted FVC and change from baseline in percent predicted FVC will be calculated by visit. The predicted FVC is derived using height (ht) in (cm) at screening,

age at screening, sex, and race (Caucasian, Mexican American, and African American). An example calculation is given below (Hankinson et al. 1999).

(Caucasian men, ≥20 years of age)

$$FVC = -0.1933 + (0.00064 \cdot age) - (0.000269 \cdot age \cdot age) + (0.00018642 \cdot ht \cdot ht)$$

(Caucasian women, ≥ 18 years of age)

$$FVC = -0.3560 + (0.01870 \cdot age) - (0.000382 \cdot age \cdot age) + (0.00014815 \cdot ht \cdot ht)$$

For Asian and Asian-American patients, a correction factor of 0.88 is applied (Hankinson et al. 2010). A plot of the mean change from baseline by treatment group by visit for the percent predicted FVC will be produced up to Week 48.

### HAQ-DI:

The HAQ-DI is a 20-item, validated questionnaire used to assess difficulty in performing activities of daily living. The HAQ-DI assesses eight domains of physical functioning: Dressing and Grooming (2 items), Hygiene (3 items), Arising (2 items), Reach (2 items), Eating (3 items), Grip (3 items), Walking (2 items), Common Daily Activities (3 items). The questions assess usual abilities ranging from 0 "without any difficulty" to 3 "unable to do." A negative change from baseline indicates improvement. The Minimal Clinically Important Difference (MCID) for improvement ranges from 0.14–0.22 (Khanna et al. 2006; Pope 2011).

The total HAQ-DI score will be calculated according to the developer's scoring algorithm (see Appendix 5 for details).

Patients in Japan were administered the Japanese version of the HAQ-DI, known at the J-HAQ. The two questionnaires contain the same evaluation components and are comparable. The pain VAS aspect of the HAQ-DI (which is not part of J-HAQ) was not implemented for this study and therefore HAQ-DI and J-HAQ will be combined for analysis.

HAQ-DI will also be summarized descriptively as a binary endpoint based on the proportion of patients achieving an improvement from baseline ≥ 0.22.

### Patient Global Assessment:

The Patient's Global Assessment represents the patient's overall assessment of his or her current SSc status on a 100 mm horizontal VAS. The extreme left end of the scale indicates "has no effect at all" (symptom free), and the extreme right end indicates "worst possible effect". A negative change from baseline indicates improvement. The MCID has been reported by Sekhon and Pope 2010 to be –6.70.

### Physician Global Assessment:

The Physician Global Assessment represents the clinician's overall assessment of the patient's current SSc status on a 100 mm horizontal VAS. The extreme left end of the scale indicates "has no effect at all" (symptom free), and the extreme right end indicates

"worst possible effect". Higher scores indicate worse disease in terms of severity, damage, or overall disease, but there is no standardization for the scale (Pope 2011). A negative change from baseline indicates improvement. Expert consensus has suggested a range of 8–13 units for the MCID (Gazi et al. 2007).

#### 4.6.3.2 Time to Treatment Failure

Time to treatment failure was previously defined in the protocol as the time from randomization to the time of first:

- death
- decline in percent predicted FVC >10% relative to baseline
- increase in mRSS >20% and an increase in mRSS of ≥5 points relative to baseline
- occurrence of a predefined SSc-related complication as adjudicated by the Clinical Adjudication Committee

  Whichever occurs first, during the Week 48 double-blind treatment period.

However, specifically, time to an event will be relative to first dose of study drug, rather than date of randomisation as only treatment emergent AEs were adjudicated, and as described in Section 3.2.1, for events deemed SSc-related complications only serious events will be included in the time to treatment failure endpoint.

Time to treatment failure will be summarized descriptively by Kaplan-Meier curves and the median, 25th, and 75th percentiles (where possible) and 95% CI for the median. The treatment groups will be compared using a Cox proportional hazards model adjusting for the stratification factor applied at randomization. A 95% CI for the HR and a p-value will be produced. Data from patients who discontinue from the study prior to Week 48 will be censored from the study day after the discontinuation date. Data post initiation of immuno-modulation therapy (e.g., DMARDs and/or Biologics) and/or data collected after stopping study drug will not be censored.

These analyses will be repeated, where appropriate, for the individual components of time to treatment failure; namely death, decline in percent predicted FVC >10% relative to baseline, relative increase in mRSS >20%, and an increase in mRSS of ≥5 points, or occurrence of a predefined serious SSc-related complication.

An analysis of time to treatment failure excluding serious infections from the adjudicated serious SSc–related complications will also be conducted.

A sensitivity analysis of time to treatment failure will be performed where discontinuation from study for the following reasons will also be classed as a treatment failure: death, lack of efficacy, lost to follow-up, withdrawal by subject, physician decision (whereas withdrawal from study for the following reasons will not be counted as an event: AE, pregnancy, protocol violation, non-compliance, study terminated by sponsor, other).

The proportion of patients with each of the components of time to treatment failure will be summarized descriptively. Serious SSc-related complications will be summarised by system organ class (SOC), and will be further subcategorised descriptively for assessment of internal organ involvement (including, but not limited to, cardiac disease, renal disease, and GI disease). All SSc-related complications (including adjudicated non-serious CTC grade 3 and 4 AEs) will be listed.

#### 4.6.3.3 Binary Endpoints

For binary secondary endpoints, such as proportion of patients with ≥40% improvement in mRSS at Week 48 compared to baseline, the weighted difference in proportion will be presented, together with the 95% CI using the extended Mantel-Haenszel method and the p-value calculated using the Cochran–Mantel–Haenszel test, adjusting for the stratification factor, IL-6 level (<10; ≥10 pg/mL) at screening (see Appendix 11 for statistical formulas) (Koch et al. 1989). Patients who have a missing Week 48 assessment will be considered non-responders in the analysis.

For mRSS binary endpoints only there will be an additional analysis whereby data from patients post initiation of immuno-modulation therapy (e.g., DMARDs and/or Biologics) will be censored so that these patients will become non-responders in the analysis.

### 4.6.4 Exploratory Efficacy Endpoints

All exploratory endpoints will be summarized and where applicable exploratory p-values will be produced using the same methodology as specified for the secondary endpoints in Section 4.6.3 unless stated otherwise.

#### Combined Response Index for Systemic Sclerosis

The Combined Response Index for Systemic Sclerosis (CRISS) will be summarized as both a binary and continuous endpoint. There will be no imputation of missing data. The calculation for the CRISS score is a two-step process and is as follows (Khanna et al. 2016):

Step 1: if a patient develops any of the following they will be assigned a probability of improving equal to 0.0

- new scleroderma renal crisis
- decline in ppFVC ≥15% relative to baseline confirmed by a second FVC within one month, HRCT to confirm ILD (if previous HRCT did not show ILD) and FVC <80% of predicted attributable to systemic sclerosis
- new onset of left ventricular failure (defined as left ventricular ejection fraction ≤45%) requiring treatment, attributable to systemic sclerosis
- new onset of pulmonary arterial hypertension on right–sided heart catheterization requiring treatment, attributable to systemic sclerosis

Step 2: for the remaining patients, compute the predicted probability of improving for each subject using the following equation (equation to derive predicted probabilities from a logistic regression model):

$$\frac{exp\left[-5.54 - 0.81 * \Delta_{MRSS} + 0.21 * \Delta_{FVC\%} - 0.40 * \Delta_{Pt-glob} - 0.44 * \Delta_{MD-glob} - 3.41 * \Delta_{HAQ-DI}\right]}{1 + exp\left[-5.54 - 0.81 * \Delta_{MRSS} + 0.21 * \Delta_{FVC\%} - 0.40 * \Delta_{Pt-glob} - 0.44 * \Delta_{MD-glob} - 3.41 * \Delta_{HAQ-DI}\right]}$$

where, $\Delta_{MRSS}$ indicates the change in MRSS from baseline to follow-up, $\Delta_{FVC}$ denotes the change in ppFVC from baseline to follow-up, $\Delta_{Pt\text{-}glob}$ indicates the change in patient global assessment, $\Delta_{MD\text{-}glob}$ denotes the change in physician global assessment, and $\Delta_{HAQ\text{-}DI}$ is the change in HAQ-DI. All changes are absolute change ($Time_2$-$Time_{baseline}$).

Subjects for which the predicted probability is greater or equal to 0.60 are considered improved, while subjects for which the predicted probability is below 0.60 are considered non-improved. Continuous analysis will be performed using non-parametric methods.

### Scleroderma Health Assessment Questionnaire

The SHAQ is composed of the HAQ-DI scale (or J-HAQ scale for Japanese patients) with the addition of five scleroderma-specific VAS scales to assess additional elements of SSc disease. Each VAS item is rated separately, with higher scores indicating more severe disease (range 0–100 mm). The five VAS scales are: 1) intestinal disease, 2) breathing problems, 3) Raynaud syndrome, 4) digital ulcers, and 5) overall disease. These VAS scales will all be analyzed independently. A negative change from baseline indicates improvement.

### Saint George's Respiratory Questionnaire

Saint George's Respiratory Questionnaire contains 50 items distributed over three scales: respiratory symptom severity (symptoms); impairment in patient activity as a result of respiratory symptoms (activity); and impact of respiratory symptoms on overall function and wellbeing (impact). Each scale is scored from 0 to 100, with the total score representing the weighted average of these three subscores.

Each item in the SGRQ has specific weight (lowest possible weight=0, highest=100). Scores for the three scales and total are calculated using the assigned weights. The numerator and denominator for the total and each scale are comprised of individual item weights, minus specific scale weights. Additional details pertaining to missing data and scoring are described in Appendix 6.

### FACIT Fatigue Scale

The FACIT fatigue scale is a 13-item measure of fatigue, with patients scoring each item on a 5-point scale. A positive change from baseline indicates improvement. FACIT-Fatigue total score will be calculated according to developer's algorithm (Appendix 7)

### SkinPRO Questionnaire

The SkinPRO is a 22-item, patient-completed questionnaire developed to measure skin health status as experienced by scleroderma patients with skin involvement. There are 4 separate domains: skin symptoms, physical function, social function, and emotional response. Although the 22-item was administered, it will be scored in accordance with the most recent validation work from the literature, which uses 18 items as described in Appendix 8 (Man A et al. 2017). There are 4 separate sub-scales: physical effects, physical limitations, emotional effects, and social effects. The response set for all items is a 7-point ordinal scale, ranging from "Not at all" to "Very much". Total and subscale scores will be calculated as described in Appendix 8. The questionnaire was only administered in patients in North America so results will be summarized for that subgroup only.

### EuroQol 5-Dimension Questionnaire

The EQ-5D-3L is a generic, preference-based health utility measure with questions about mobility, self-care, usual activities, pain/discomfort, and anxiety/depression that are used to build a composite of the patient's health status. It is a concise scale that has performed well for patients with SSc (Gualtierotti et al. 2016). The EQ-5D-3L will be utilized in this study for economic modeling. A positive change from baseline indicates improvement. The EQ-5D will be scored according to the developer's guidelines (EQ-5D-3L User Guide) using UK utility values. For further details of EQ-5D see Appendix 9.

### Work Productivity and Activity Impairment—General Health

The WPAI-GH questionnaire is a six-item scale, asking patients to estimate the amount of time that their work and daily activities were affected by their health over the previous 7 days. The WPAI yields four types of scores:

1. Absenteeism (work time missed)
2. Presenteesism (impairment at work / reduced on-the-job effectiveness)
3. Work productivity loss (overall work impairment / absenteeism plus presenteeism)
4. Activity Impairment

WPAI scores will be calculated according to developer's algorithm described in Appendix 10.

### $DL_{CO}$

Predicted $DL_{CO}$ values are derived using height (ht) at screening (inches), age at screening, and patient sex. The formula for predicted values for females and males are presented below (Pesola et al. 2004):

$$\text{Females: } 2.2382 - (0.111 \cdot \text{age}) + (0.4068 \cdot \text{ht})$$

$$\text{Males: } 12.9113 - (0.229 \cdot \text{age}) + (0.418 \cdot \text{ht})$$

The predicted value is then adjusted for the hemoglobin level (g/dL) at each visit using the following formula to give predicted hemoglobin adjusted values (MacIntyre et al. 2005):

Females: $DL_{,CO,predicted}$ for Hb=$DL_{,CO,predicted}$×(1.7 Hb/(9.38+Hb))

Males: $DL_{,CO,predicted}$ for Hb=$DL_{,CO,predicted}$×(1.7 Hb/(10.22+Hb))

A plot and summary of the median change from baseline by treatment group by visit for the percent predicted $DL_{CO}$ will be produced up to Week 48.

**HRCT**

Change from baseline to Week 48 in Quantitative Lung Fibrosis Score of the lobe of most involvement (QLF-LM) as determined by HRCT scans will be summarized for the exploratory endpoint of change in HRCT.

The lobe of most involvement is determined as the lobe with the worst (highest) score at baseline out of the 5 lobes (left lower, left upper, right lower, right middle, right upper) for the specified parameter of interest (QLF in this case).

Other parameters collected from HRCT readings may also be summarized and change from baseline of QLF-LM and other HRCT parameters may be summarised separately by patients with and without ILD at baseline.

HRCT scans that were done prior to baseline may be permitted in place of a baseline scan in some cases.

### 4.6.5 Sensitivity Analyses

#### 4.6.5.1 Sensitivity to Primary Estimand

If found to be statistically significant ($p<0.05$), the primary endpoint will be re-analyzed assuming that missing data is missing not at random (MNAR), rather than MAR, via tipping-point sensitivity analyses implemented using multiple imputation.

The PROC MI procedure in SAS using the Markov chain Monte Carlo method will be used to partially impute data and convert the dataset into a monotone missing data pattern. One thousand monotone datasets will be imputed using the seed 1689, by treatment group, including the stratification factor as an indicator variable.

After imputing non-monotone data, the missing values remaining will be imputed using the PROC MI monotone regression procedure in SAS. This regression based approach imputes missing data in a sequential manner with the use of univariate models; for each time point, a regression model based on all patients who have observations (imputed or observed) available at this time point is fitted and used to impute observations for patients with missing values at this time point. The variables included in the imputation model will be treatment group, the stratification factor, IL-6 level ($<10$; $\geq 10$ pg/mL) at screening, and the mRSS scores by visit. The seed that will be used is 13820647. Only

one complete dataset will be imputed per monotone dataset, resulting in one thousand complete datasets.

Tipping-point analysis will be implemented by adding (or subtracting) a constant delta to the MAR imputed values at the Week 48 analysis time point in the direction of lack of efficacy for TCZ and in the direction of improvement for placebo. A range of evenly spaced deltas will be used to adjust imputed values in the TCZ and placebo arms independently in order to produce a grid of delta adjustments to missing values in both arms. The size of the deltas will be chosen pragmatically based on the primary analysis, so that adjustments that lead to non-significance are encompassed in the grid. In particular, assuming MAR on placebo, i.e., no delta adjustment, an adjustment on the TCZ arm that just tips the analysis into non-significance will be included.

For each pair of delta adjustments (TCZ and placebo) each dataset will be analyzed separately using the same analysis model specified for the primary endpoint and the results combined using the PROC MIANALYZE procedure in SAS. The resulting point estimates of treatment effect, 95% CIs and corresponding p-values under each pair of deltas will be tabulated.

As an additional sensitivity analysis to the primary estimand a pattern-mixture model will be implemented, using multiple imputations, whereby missing data in the placebo arm will be imputed using a missing-at-random assumption and missing data in the TCZ arm will be imputed in a stepwise fashion using multiple calls to PROC MI with a monotone regression statement using data from placebo-treated patients as the basis for the imputation (Ratitch and O'Kelly 2011). This imputation method assumes that patients who are lost to follow-up in the TCZ arm, that stop study drug and may revert to standard of care, will have a trajectory similar to that of patients in the placebo arm. This is a reasonable assumption as there is a lack of proven effective treatment for patients to transition to. It also assumes that these patients will not have an immediate worsening of outcome upon study drug discontinuation. Conservatively, missing data on the placebo arm is assumed missing at random, since withdrawals on placebo in the Phase II Fasscinate Study on average had worse outcomes at the point of withdrawal than patients that continued in the study.

Firstly, the data will be imputed to be monotone exactly as described for the tipping point analysis. Successive calls to PROC MI will then be made, using seed 298 for all calls, starting with TCZ patients with missing data for the first post-baseline visit. As a first step these patients will be selected out and pooled with the placebo patients. A regression model based on all placebo patients who have observations available at this time point will be fitted and used to impute observations for TCZ patients with missing values at this time point. The imputation model will include the stratification factor, IL-6 level ($<10$; $\geq 10$ pg/mL) at screening, in addition to the baseline mRSS score, and first post baseline mRSS score. In step two, TCZ patients with the second post-baseline time point missing will be selected out and pooled with the placebo group. As well as

the stratification factor, the model will include the second post-baseline visit mRSS scores, as well as prior observed or imputed mRSS scores from the previous step. Data across all the visits will be imputed in this way, to result in a thousand imputed datasets (there will only be one imputation per monotone dataset). Each dataset will be analyzed separately using the same analysis model specified for the primary endpoint and the results combined using the PROC MIANALYZE procedure in SAS. The resulting point estimate of treatment effect, 95% CI and corresponding p-value will be presented (Ratitch and O'Kelly 2011).

A linear regression analysis (with Huber–White sandwich errors) of change from baseline in mRSS at week 48 will be performed. The model will include the stratification factor, baseline mRSS score, and treatment group.

#### 4.6.5.2 Alternative Estimand

A second estimand for the difference in change from baseline in mRSS will also be determined. Data post initiation of immuno-modulation therapy (e.g., DMARDs and/or Biologics) and/or data collected after stopping study drug will be censored and the mixed model repeated measurement (MMRM) model described in Section 4.6.2.2 will be implemented. The estimand of interest for this analysis is the difference between treatment arms in the mean change in the mRSS at Week 48 for the ITT population if the randomized treatments were taken as directed.

#### 4.6.5.3 Sensitivity to Secondary Endpoints

Similarly, tipping point analyses and pattern mixture models, using the same seeds and number of imputations will be performed for all the continuous variables in the multiplicity testing procedure shown in Section 4.6.1 with the exception of the time to treatment failure endpoint. For each endpoint the analysis method of the imputed datasets will be the same method as specified for the multiplicity testing procedure.

For the binary endpoints of the difference in proportions of patients with ≥20%, ≥40%, and ≥60% improvement in mRSS at Week 48 compared with baseline, a tipping point analysis will also be performed. Patients with missing change from baseline in mRSS at Week 48 are non-responders for the primary analysis method. These patients will be incrementally set to responders independently for each treatment arm. The number of patients with missing data set to responders per treatment arm will be presented in a grid, and the resulting difference in proportions, 95% CI and p-value will be tabulated.

#### 4.6.5.4 Immuno-Modulating Treatments

Introduction of an immuno-modulating treatment such as a DMARD or Biologic will be identified using preferred terms (e.g., mycophenolate mofetil).

A summary and listing of immuno-modulating treatments will be produced.

The number and percent of patients that initiated at least one additional immuno-modulating therapy will be summarized by the following categories (not mutually exclusive) and further sensitivity analyses may be performed on these subgroups as appropriate:

- Discontinued study drug
- Continued study drug
- Met mRSS protocol escape criterion
- Met FVC protocol escape criterion
- Met mRSS and FVC protocol escape criteria
- Did not meet either mRSS or FVC protocol escape criteria

#### 4.6.5.5 Subgroup Analyses

The analysis of the change from baseline in the mRSS will be conducted using a repeated measures model to compare the treatment effect at Week 48 for the subgroups IL-6 at screening ($<$10 pg/ml and $\geq$10 pg/ml) which will include the treatment-by-IL-6 at screening interaction in addition to the fixed and random effects specified for the primary analysis.

In addition, separate analyses of the change from baseline in mRSS will be conducted using a repeated measures model to compare the treatment effect at Week 48 for the subgroups listed below. For each subgroup analysis, the same model as used for the primary analysis will be used but will also include the parameter of interest as a class variable at baseline, as well as the baseline parameter of interest-by-treatment interaction. The treatment difference for each subgroup and a 95% CI will be presented.

- baseline CRP ($\leq$0.6, $>$0.6 mg/dL)
- baseline ESR ($<$28, $\geq$28 mm/hr)
- baseline platelets ($<$330, $\geq$330$\times$109/L)
- baseline disease duration ($<$2, $\geq$2 years)
- baseline mRSS score ($<$25, $\geq$ 25)
- Sex (Male, Female)
- baseline Age (<65,$\geq$65)
- An observed data summary of mean change from baseline in mRSS at Week 48 by treatment group for the following categories of race will be produced (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Other)
- An observed data summary of mean change from baseline in mRSS at Week 48 by treatment group for the following regions will be produced (North America, Western Europe, Central Europe, Latin America, Japan)

Specific safety and efficacy endpoints described in the PK section below will be produced by TCZ serum concentration quartiles. Other subgroup analyses may be performed as appropriate.

## 4.7 PHARMACOKINETIC ANALYSES

For the Week 48 analysis, TCZ serum concentrations ($C_{trough}$) will be summarized by visit using descriptive statistics, means, medians, standard deviations, minimum, maximum, and coefficients of variation (standard deviation/mean)*100). Plots of mean and standard deviation of TCZ ($C_{trough}$) by visit will also be presented. Tocilizumab concentration at Week 48 will be summarized based on tertiles of the values. Effect of ADA positivity on serum TCZ concentrations will be assessed by comparing the individual concentration-time profiles of ADA positive patients (definition in Section 4.10.6) with the individual profiles of the remainder of the population. Nonlinear mixed-effect modelling (NONMEM) will be used to estimate population and individual PK parameters and the influence of various covariates on these parameters will be investigated. Major baseline covariates will be identified and the dependence of PK parameters on each covariate will be described. Secondary PK parameters, such as AUC, $C_{max}$, and $C_{trough}$ at steady state, will be listed and summarized. Results of the popPK analysis will be reported separately from the CSR.

## 4.8 EXPOSURE EFFECT ANALYSES

In order to characterize exposure-efficacy relationships, continuous responses (e.g., mRSS and all positively trending secondary efficacy endpoints as noted in Section 2.2) will be summarized based on exposure tertiles (high, medium, and low exposures) in the active treatment group and compared to placebo patients. In addition, for binary secondary endpoints (e.g., patients who show a ≥20% improvement in mRSS at Week 48) a comparison of observed TCZ $C_{trough}$ at Week 48 grouped in tertiles, in responders will be compared to non-responders.

In order to characterize exposure-safety relationships, summary tables of AEs by tertiles of observed TCZ $C_{trough}$ at Week 48 will be produced. The proportion of patients by National Cancer Institute Common Terminology Criteria for Adverse Events (NCI CTCAE) grade (categorization based on worst post-baseline value) (NCI 2003) will be summarized by tertiles of observed TCZ $C_{trough}$ for neutrophils, platelets, ALT (alanine transaminase), and AST (aspartate aminotransferase) at Week 48.

An evaluation of exposure-response (PD, safety, efficacy) endpoints, where relevant, will additionally be conducted as part of the population based analysis and summarized in the popPK report.

## 4.9 PHARMACODYNAMIC ANALYSES

For the Week 48 analysis, sIL-6R, IL-6, CRP, and ESR levels and change from baseline will be summarized by visit using descriptive summary statistics, including the means,

medians, ranges, and standard deviations. For these PD parameters, plots of mean and standard deviation or standard errors by visit will also be presented.

## 4.10 SAFETY ANALYSES

Safety data will be summarized using descriptive statistics under the treatment received most frequently; no statistical analyses comparing safety data between two groups will be conducted. All safety data in the Week 48 safety data cut will be included for the primary analysis.

Since data entry for the study is ongoing, safety data that are not associated with a specific visit may be updated after the analysis time point for an individual patient. For example, an adverse event (AE) that starts on or prior to the Week 48 visit for a specific patient will be included in the Week 48 analysis. It is possible that the AE may be updated beyond Week 48 for the patient (e.g., an AE that was ongoing could be given an end date). Such updates would be reported in the Week 96 final CSR.

### 4.10.1 Adverse Events

Medical Dictionary for Regulatory Activities (MedDRA) will be used as the thesaurus for AEs and disease codes, and the GNE drug dictionary will be used for treatments. A glossary of these codes will be produced. Because all outputs produced for the Week 48 and Week 96, reporting will be produced using the latest versions of MedDRA and the GNE drug dictionary; there may be discrepancies between the different reporting events for the same AE/medication.

Only treatment-emergent AEs will be summarized. Treatment-emergent events are defined as those AEs with observed or imputed onset date on or after the start date of trial treatment. Only where the most extreme intensity is greater than the initial intensity (or if most extreme intensity is not missing and initial intensity is missing) will events with an onset date prior to the start of trial treatment be considered treatment-emergent. An AE with a completely missing start date will be assumed to be treatment-emergent unless the AE has a complete non-imputed end date that is prior to study day 1.

Adverse events will be coded and tabulated by system organ class (SOC) and/or preferred term. In tabulations, preferred terms and their associated SOC will be presented in order of descending frequency summed across the treatment arms. Adverse events will also be tabulated by severity and relationship to study medication as indicated by the investigator. Adverse events leading to withdrawal, AEs leading to death, and AEs leading to a dose modification or interruption will be summarized.

Adverse event rates per 100 patient–years exposure (defined as the number of events/total duration in the study multiplied by 100) will be calculated for each AE preferred term and SOC, along with the corresponding 95% CIs for the rates of the SOC (exact based on the $\chi^2$ distribution [Ulm et al. 1990]).

The following will also be summarized, and listings produced where required:

- Serious AEs
- AEs leading to withdrawal
- AEs leading to death
- AEs leading to a dose modification or interruption
- Non-serious AEs occurring in ≥5% of patients in at least one treatment group
- Injection-site reactions (ISRs) by severity
- Hypersensitivity AEs (AEs occurring immediately after or within 24 hours of the end of an injection that are not deemed "unrelated" to study treatment)
- Clinically significant Hypersensitivity AEs (AEs occurring immediately after or within 24 hours of the end of an injection that are not deemed "unrelated" to study treatment and that led to study treatment discontinuation)

Adverse events of special interest will be defined using published Standard MedDRA Queries (SMQs) or AE Grouped Terms (AEGTs) defined by Roche Drug Safety. The groupings of AEs will include but may not be limited to the following:

- All and Serious Infections (Infections and Infestations SOC)
- Opportunistic infections [OI] (Roche Standard AEGT Basket)
- Malignancies (Malignant or Unspecified tumors SMQ Narrow) confirmed by medical review
- Malignancies excluding NMSC (confirmed by medical review)
- All and/or Serious Hepatic events (Hepatic failure, Fibrosis, and Cirrhosis and Other Liver Damage-related Conditions SMQ Wide or Hepatitis, non-infectious SMQ Wide)
- Stroke (Ischemic Cerebrovascular Conditions SMQ Narrow or Hemorrhagic Cerebrovascular SMQ Narrow)
- Myocardial infarction [MI] (MI SMQ Narrow)
- Anaphylactic reaction events occurring immediately after or within 24 hours of injection of tocilizumab, summarized separately according to the following:
  - Roche Standard AEGT Basket according to Sampson's criteria (Sampson et al. 2006)
  - Anaphylactic Reaction SMQ Narrow
- Gastrointestinal perforations (Gastrointestinal perforation SMQ Wide) confirmed by medical review
- All and/or Serious Bleeding events (Hemorrhage terms [excluding laboratory terms] SMQ Wide)
- Demyelinating events (Demyelination SMQ Narrow)

A glossary showing the mapping of investigator verbatim terms to preferred terms will be produced for all AEs included in the analysis for Week 48. For each AE of special

interest table based on SMQs/AEGTs, a corresponding listing of the preferred terms that comprise the SMQ will be produced. A listing of deaths and SAEs that occur during the Week 48 double-blind period will be produced at Week 48.

The time to first non-infectious SAE (SAEs excluding infections), and time to first infection SAE will be presented as KM plots.

### 4.10.2 Deaths

Details of any deaths will be presented in the form of an individual patient listing and descriptive summaries, including a summary of the death rate per 100 patient-years of exposure (defined as the number of deaths/total duration in the study multiplied by 100) based on the safety population.

### 4.10.3 Laboratory Data

All laboratory data will be converted to SI units. The International Standard for the Handling and Reporting of Laboratory Data COG 3007 (Version 3.0) will be used to implement reference ranges and marked abnormalities for laboratory data where possible. Summary tables will detail the actual values and changes from baseline of the laboratory parameters over visits up to Week 48. Summaries of the number of patients by CTC grade for hematology and hepatic laboratory parameters will be produced (for summaries referring to NCI CTCAE (National Cancer Institute Common Terminology Criteria for Adverse Events) grading ([NCI 2003], Version 4.0 will be used) by Week 48. For liver laboratory tests, the number of patients will be summarized by category for baseline and worst post baseline result by Week 48.

The number of patients with marked abnormalities will be summarized by Week 48.

### 4.10.4 Lipid Data

Fasting lipids consist of total cholesterol, triglycerides, HDL, and LDL, and these will be summarized separately from the other laboratory parameters when referencing gradings or shifts.

Threshold for the presentation of lipid data are defined by the National Cholesterol Education Program Adult Treatment Panel III (NCEP 2001), and are shown in Table 2.

**Table 2 National Cholesterol Education Program (ATPIII) Thresholds**

| | | | | | |
|---|---|---|---|---|---|
| **LDL** (mg/dL) | <100 (optimal) | 100–129 (normal) | 130–159 (borderline high) | ≥160 (high) | |
| **HDL** (mg/dL) | <40 (low) | 40–59 (normal) | | ≥60 (high) | |
| **Total cholesterol** (mg/dL) | <200 (desirable) | | 200–239 (borderline high) | ≥240 (high) | |
| **Triglycerides** (mg/dL) | <150 (normal) | | | 150–499 (high) | ≥500 (very high) |

ATPIII=Adult Treatment Panel III; HDL=high density lipoprotein; LDL=low density lipoprotein.

### 4.10.5 Vital Signs

Summary tables will detail the actual values and changes from baseline of the laboratory parameters over visits up to Week 48 for pulse rate, systolic blood pressure, and diastolic blood pressure (after patient is supine for at least 5 minutes).

Blood pressure will also be summarized by visit using the following JNC 7 categories (as presented in Table 3):

**Table 3 Blood Pressure JNC 7 Category Criteria**

| Blood Pressure Category | Criteria |
|---|---|
| Normal | SBP<120 mmHg and DBP<80 mm Hg |
| Pre-hypertension | 120≤SBP<140 mmHg or 80≤DBP<90 mmHg |
| Stage I hypertension | 140≤SBP<160 mmHg or 90≤DBP<100 mmHg |
| Stage II hypertension | SBP≥160 mmHg or DBP≥100 mmHg |

SBP=systolic blood pressure

### 4.10.6 Immunogenicity

Samples for the detection of anti-TCZ antibodies (ADA, anti-drug antibodies, here refers to anti-TCZ antibodies) and analysis of immunogenicity will be taken at baseline and at Weeks 1, 8, 16, 24, 36, and 48 (or study withdrawal/completion). For patients experiencing anaphylaxis or hypersensitivity reactions that result in withdrawal from the study, a sample for detection of ADA will be taken at the time of the event and also at least 8 weeks after the last dose of TCZ/placebo.

Blood samples will be analyzed for the evaluation of immunogenicity of TCZ by a number of different assays. Samples that are positive for ADA in the screening assay will be further analyzed by a confirmation assay to confirm specificity. If the confirmation assay is positive, two additional tests will be performed: a neutralizing assay for the ability to inhibit the activity of TCZ and a test for ADA of the IgE isotype.

The immunogenicity analyses will include patients with at least one sample analyzed using immunogenicity assay. The numbers and proportions of ADA positive patients after TCZ exposure during the study (including both the treatment and follow-up periods)

will be summarized by treatment group. Patients are considered to be ADA positive if they are ADA negative at baseline but develop an ADA response following study drug administration (treatment-induced ADA response).

The relationship between ADA status and safety, efficacy, and PK (Section 4.7) endpoints will be analyzed and reported descriptively via subgroup analyses.

### 4.10.7 Digital Ulcers

The digital ulcer count at baseline and the change in digital ulcer category by visit will be summarized by treatment group. A shift table of the number of digital ulcers by visit from the previous visit by treatment arm will be produced.

## 4.11 SSC-SPECIFIC AUTOANTIBODY PANEL

The SSc-specific autoantibody panel will include tests for the following antibodies: anti-topoisomerase (positive ≥20 U/mL), anti-RNA polymerase (positive ≥20 U/mL), anti-PM/Scl (positive ≥20 U/mL), anti-U1 snRP (positive ≥20 U/mL), anti-histone (positive ≥1 U/mL), and anti-centromere (positive ≥1:40 dilution). Shift tables of the number of patients positive for each antibody at screening and Week 48 will be summarized for the safety population.

## 4.12 MISSING DATA

Methods for handling missing data for key efficacy variables are presented in Section 4.6. Partial dates for AEs, concomitant medications, laboratory assessments, and medical history will be imputed.

## 4.13 INTERIM ANALYSES

A futility analysis was conducted to which the sponsor remained blinded. The analysis was conducted after the first 76 patients reached the Week 24 visit or withdrew and was based on the change from baseline in mRSS at Week 24. The futility analysis was conducted by an external statistical group, iDCC (independent Data Co-ordinating Centre) and was reviewed by the iDMC. Details of the futility analysis, along with the rationale and timing are documented in the study iDMC charter. The beta spending function used and all other details of the futility analysis were exactly as specified in the charter; however the maximum information of the planned design was altered in the SEQ DESIGN procedure in SAS from the assumed 1.064 that had been specified in the charter upon instructions to the iDCC from the blinded Sponsor. The amount of information about an unknown parameter available from the data can be measured by the Fisher information. For a maximum likelihood statistic, the information level is the inverse of its variance. The maximum information is the information level at the final analysis stage of the group sequential trial, if the trial continued as planned. When designing the interim the predicted maximum information can be estimated from the sample size and an assumed standard deviation. The boundary and operating

characteristics of the interim were based on the information fraction at the interim being approximately 0.5.

The original design of this interim was based on the Phase II study data assuming a common standard deviation of 6.76 at Week 24. However the standard deviation of the pooled blinded week 24 data at the time of the interim snapshot was 5.43, much smaller than that originally assumed when designing the interim; thus the variability was lower in the Phase III study than had been expected. Assuming a standard deviation in both treatment arms of 5.43 and 98 patients per arm completing to Week 24 the standard error assumed for the final analysis at Week 24 was re-evaluated by the Sponsor; giving an expected maximum information of 1.66. This adjustment was made in order that the interim fraction would be approximately 0.5, in accordance with the planned boundary. The interim was conducted in December 2016, with a “continue” decision made by the iDMC on 16 December 2016. The Sponsor remained blinded throughout, with only a continue decision communicated. All documentation, including the closed meeting minutes and recommendations from the iDMC, will be made available after study unblinding.

## 5. REFERENCES


[NCEP] National Cholesterol Education Program (NCEP). Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). JAMA 2001;285:2486–97.

Common Terminology Criteria, Version 3.0, National Cancer Institute, USA, Published December 12, 2003.

EQ-5D-3L User Guide: Basic information on how to use the EQ-5D-3L instrument. (https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-3L_UserGuide_2015.pdf)

Gazi H, Pope JE, Clements P, et al. Outcome measurements in scleroderma: results from a Delphi exercise. J Rheumatol 2007;34:501–9.

Gualtierotti R, Ingegnoli F, Scalone L, et al., Feasibility, acceptability and construct validity of EQ-5D in systemic sclerosis. Swiss Med Wkly 2016;146: w14394.

Guideline on multiplicity issues in clinical trials EMA/CHMP/44762/2017.

Hankinson JL, Kawut SM, Shahar E, et al. Performance of American Thoracic Society-recommended spirometry reference values in a multiethnic sample of adults: the multi-ethnic study of atherosclerosis (MESA) lung study. Chest 2010;137(1):138-45.

Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med 1999;159:179–87.

Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 1997;53:983–97.

Khanna D, Beroccal VJ, Giannini EH, et al. The American College of Rheumatology Provisional Composite Response Index for Clinical Trials in Early Diffuse Cutaneous Systemic Sclerosis. Arthritis and Rheumatology 2016;68(2):299–311.

Khanna D, Furst DE, Hays RD, et al. Minimally important difference in diffuse systemic sclerosis: results from the D-penicillamine study. Ann Rheum Dis 2006;65(10):1325-9.

Koch GG, Car GJ, Amara A, et.al. Categorical data analysis. In StateBerry, D., A., Statistical Methodology in the Pharmaceutical Sciences, New York:Marcel Dekker, 1989:389-473.

MacIntyre N, Crapo R, Viegi G, et al. Standardisation of the single-breath determination of carbon monoxide uptake in the lung. Eur Respir J 2005;26:720–35.

Man A, Correa JK, Ziemek J, et al. Development and Validation of a Patient-reported Outcome instrumen t for skin involvement in patients with systemic sclerosis. Ann Rheum Dis 2017.

Pesola GR, Sunmonu Y, Huggins G, et al. Measured diffusion capacity versus prediction equation estimates in blacks without lung disease. Respiration 2004;71:484–92.

Pope J. Measures of systemic sclerosis (scleroderma). Arthritis Care Res 2011;63(Suppl 11):S98–111.

Ratitch B, O'Kelly M. Implementation of pattern-mixture models using standard SAS/STAT procedures. Proceedings of the PharmaSUG conference: 2011 May 8–11.

Sampson HA, Muñoz-Furlong A, Campbell RL, et al. Second symposium on the definition and management of anaphylaxis: summary report – Second National Institute of Allergy and Infectious Disease/Food Allergy and Anaphylaxis Network symposium. J Allergy Clin Immunol 2006;117:391-7.

Sekhon S, Pope J. The minimally important difference in clinical practice for patient-centered outcomes including Health Assessment Questionnaire, fatigue, pain, sleep, Global Visual Analog Scale, and SF-36 in scleroderma. J Rheumatol 2010;37:591–8.

Ulm K. A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). Am J Epidemiol 1990;131:373−5.

# Appendix 1
## PROTOCOL SYNOPSIS

| | |
|---|---|
| **TITLE:** | **A PHASE III, MULTICENTER, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED, PARALLEL-GROUP STUDY TO ASSESS THE EFFICACY AND SAFETY OF TOCILIZUMAB VERSUS PLACEBO IN PATIENTS WITH SYSTEMIC SCLEROSIS** |
| **PROTOCOL NUMBER:** | WA29767 |
| **VERSION NUMBER:** | 6 |
| **EUDRACT NUMBER:** | 2015-000424-28 |
| **IND NUMBER:** | 112406 |
| **TEST PRODUCT:** | Tocilizumab (RO4877533) |
| **PHASE:** | III |
| **INDICATION:** | Systemic sclerosis |
| **SPONSOR:** | F. Hoffmann-La Roche Ltd |

**Objectives**

**Efficacy Objectives**

The primary efficacy objective for this study is as follows:

- To evaluate the efficacy of tocilizumab (TCZ) compared with placebo on skin sclerosis, as measured by modified Rodnan Skin Score (mRSS) at Week 48

The secondary efficacy objectives for this study are as follows:

- To evaluate the efficacy of TCZ compared with placebo on pulmonary function, as measured by forced vital capacity (FVC) at Week 48
- To evaluate the efficacy of TCZ compared with placebo on patient-reported outcomes (PROs), as measured by the Health Assessment Questionnaire Disability Index (HAQ-DI) and Patient's Global Assessment at Week 48
- To evaluate the efficacy of TCZ compared with placebo as measured by the Physician's Global Assessment at Week 48
- To evaluate the efficacy of TCZ compared with placebo by assessment of time to treatment failure (death, worsening of mRSS and/or FVC, or clinically significant systemic sclerosis [SSc] complication) up to Week 48

**Safety Objectives**

The safety objectives for this study are as follows:

- To evaluate the safety of TCZ compared with placebo, focusing on the nature, frequency, and severity of serious and non-serious adverse events, the frequency of SSc-related complications, and effects on vital signs, physical findings, and clinical laboratory results
- To evaluate the safety of TCZ compared with placebo by assessing the number of digital ulcers
- To assess the long-term safety of TCZ

# Appendix 1
## PROTOCOL SYNOPSIS (cont.)

**Immunogenicity Objectives**

The immunogenicity objectives for this study are as follows:

- To characterize the immunogenic potential of TCZ by measuring anti-TCZ antibodies
- To assess the potential relationship between development of anti-TCZ antibodies and efficacy, safety, or pharmacokinetic (PK) outcome measures

**Pharmacodynamic Objectives**

The pharmacodynamic (PD) objectives for this study are as follows:

- To compare changes in levels of PD biomarkers following treatment with TCZ versus placebo

**Pharmacokinetic Objectives**

The PK objectives for this study are as follows:

- To characterize the pharmacokinetics of TCZ
- To evaluate potential relationships between PK parameters for TCZ and efficacy, safety, or immunogenicity outcome measures

**Exploratory Objectives**

The exploratory objectives for this study are as follows:

- To evaluate the efficacy of TCZ compared with placebo on skin sclerosis, as measured by mRSS at Week 24
- To evaluate the efficacy of TCZ versus placebo measured by the proportion of responders as defined by the Combined Response Index for Systemic Sclerosis (CRISS) at Week 48
- To evaluate the efficacy of TCZ compared with placebo, as measured by the visual analog scale (VAS) component of the Scleroderma Health Assessment Questionnaire (SHAQ) at Weeks 24 and 48
- To evaluate the efficacy of TCZ compared with placebo, as measured by the Work Productivity and Activity Impairment—General Health (WPAI-GH) questionnaire at Weeks 24 and 48
- To evaluate the efficacy of TCZ compared with placebo, as measured by the EuroQol 5-Dimension Questionnaire with three levels of severity (EQ-5D-3L) at Weeks 24 and 48
- To evaluate the efficacy of TCZ compared with placebo, as measured by the Saint George's Respiratory Questionnaire (SGRQ) at Week 48
- To evaluate the effect of TCZ compared with placebo on fatigue as measured by Functional Assessment of Chronic Illness Therapy (FACIT)–Fatigue score at Week 48.
- To evaluate the efficacy of TCZ compared with placebo, as measured by the Scleroderma Skin Patient-Reported Outcome (SkinPRO) questionnaire at Week 48 (for North America only)
- To evaluate the efficacy of TCZ compared with placebo on the basis of change in pulmonary fibrosis, as determined using high-resolution computed tomography (HRCT) scans at Week 48
- To evaluate the efficacy of TCZ compared with placebo, as measured by diffusion capacity of the lung for carbon monoxide ($DL_{CO}$) at Week 48, and FVC at Week 24
- To evaluate the maintenance of efficacy of TCZ, as measured by mRSS and FVC at Week 96
- To assess whether non-inherited biomarkers are predictive of response to TCZ (i.e., predictive biomarkers), susceptibility to developing adverse events, or progression to a more severe disease state (i.e., prognostic biomarkers), can provide evidence of TCZ activity, or can increase the knowledge and understanding of disease biology

# Appendix 1
## PROTOCOL SYNOPSIS (cont.)

**Study Design**

**Description of Study**

This is a Phase III, multicenter, randomized, double-blind, placebo-controlled, two-arm, parallel-group study designed to assess the efficacy and safety of TCZ in patients with SSc.

**Number of Patients**

Approximately *212* patients with diffuse cutaneous systemic sclerosis (dcSSc) will be enrolled at approximately *75* global sites.

**Target Population**

Inclusion Criteria

Patients must meet the following criteria for study entry:

- Signed Informed Consent Form
- Age $\geq 18$ years at baseline (Day 1)
- Able to comply with the study protocol, in the investigator's judgment
- Diagnosis of SSc, as defined using the American College of Rheumatology [ACR]/European League Against Rheumatism [EULAR] criteria
- SSc disease duration of $\leq 60$ months (defined as time from the first non–Raynaud phenomenon manifestation)
- mRSS of $\geq 10$ and $\leq 35$ units at screening
- Active disease that meets at least one of the following criteria at screening:
  - Disease duration of $\leq 18$ months defined as time from the first non–Raynaud phenomenon manifestation
  - Increase in mRSS of $\geq 3$ units compared with the most recent assessment performed within the previous 6 months
  - Involvement of one new body area and an increase in mRSS of $\geq 2$ units compared with the most recent assessment performed within the previous 6 months
  - Involvement of two new body areas within the previous 6 months
  - Presence of at least one tendon friction rub
- Presence of at least one of the following at screening:
  - C-reactive protein (CRP) $\geq 0.6$ mg/dL ($\geq 6$ mg/L)
  - Erythrocyte sedimentation rate (ESR) $\geq 28$ mm/hr
  - Platelet count $\geq 330 \times 10^9$/L (330,000/$\mu$L)
- Uninvolved or mildly thickened skin at one of the following possible injection-site locations:
  - Front, middle region of the thigh
  - Abdomen, except for the 2-inch area directly around the navel
  - Outer area of the upper arm (if a patient caregiver is giving the injection)

- For women who are not postmenopausal (≥ 12 months of non-therapy-induced amenorrhea) or surgically sterile (absence of ovaries and/or uterus): agreement to remain abstinent or use single or combined contraceptive methods that result in a failure rate of < 1% per year during the treatment period and for up to 3 months after the last dose of study drug

  Abstinence is only acceptable if it is in line with the preferred and usual lifestyle of the patient. Periodic abstinence (e.g., calendar, ovulation, symptothermal, or postovulation methods) and withdrawal are not acceptable methods of contraception.

  Examples of contraceptive methods with a failure rate of < 1% per year include tubal ligation, male sterilization, hormonal implants, established, proper use of combined oral or injected hormonal contraceptives, and certain intrauterine devices.

  Alternatively, it is acceptable to combine the use of two methods (e.g., two barrier methods such as a condom and a cervical cap). Barrier methods must always be supplemented with the use of a spermicide.
- For men: agreement to remain abstinent or use contraceptive measures and agreement to refrain from donating sperm, as defined below:

  With female partners of childbearing potential or pregnant female partners, men must remain abstinent or use a condom during the treatment period and for at least 8 weeks after the last dose of study drug.

  Abstinence is acceptable only if it is in line with the preferred and usual lifestyle of the patient. Periodic abstinence (e.g., calendar, ovulation, symptothermal, or postovulation methods) and withdrawal are not acceptable methods of contraception.

  Men must refrain from donating sperm during the treatment period and for at least 8 weeks after the last dose of study drug.

Exclusion Criteria

Patients who meet any of the following criteria will be excluded from study entry:

- Pregnant or lactating, or intending to become pregnant during the study
- Women who are not postmenopausal (≥ 12 months of non-therapy-induced amenorrhea) or surgically sterile must have a negative serum pregnancy test at screening and a negative urine pregnancy test at baseline.
- Major surgery (including joint surgery) within 8 weeks prior to screening or planned major surgery within 12 months following randomization
- Skin thickening (scleroderma) limited to the face or areas distal to the elbows or knees at screening
- Rheumatic autoimmune disease other than SSc, including but not limited to rheumatoid arthritis (RA) (diagnosed using ACR/EULAR criteria), systemic lupus erythematosus, mixed connective tissue disorder, polymyositis, dermatomyositis, eosinophilic fasciitis, primary Sjögren's syndrome, and eosinophilic myalgia syndrome, as determined by the investigator
- Treatment with non-investigational or investigational cell-depleting therapies, including but not limited to alemtuzumab, anti-CD4, anti-CD5, anti-CD3, anti-CD19, and anti-CD20 within 18 months of baseline; or if treatment prior to 18 months from baseline, evidence of peripheral depletion of targeted lymphocyte subset at screening
- Previous treatment with chlorambucil, bone marrow transplantation, or total lymphoid irradiation
- Previous treatment with anti-IL6 therapy (including and not limited to TCZ)
- Previous treatment with thalidomide, antithymocyte globulin, plasmapheresis, or extracorporeal photopheresis
- Treatment with anakinra within 1 week prior to baseline
- Treatment with etanercept within 2 weeks prior to baseline

# Appendix 1
## PROTOCOL SYNOPSIS (cont.)

- Treatment with oral, intramuscular, or intravenous corticosteroids (> 10 mg/day of prednisone or equivalent) within 2 weeks prior to baseline
- Treatment with methotrexate, hydroxychloroquine, cyclosporine A, azathioprine, mycophenolate mofetil, rapamycin, colchicine, or D-penicillamine, within 4 weeks prior to baseline
- Immunization with a live or live attenuated vaccine within 4 weeks prior to baseline
- Treatment with any investigational agent within 5 elimination half-lives of the investigational drug prior to baseline
- Chronic treatment with any of the following within 5 elimination half-lives of the drug prior to baseline:
  - Pirfenidone
  - Nintedanib
  - Endothelin-receptor antagonists, terguride
  - Tyrosine-kinase inhibitors (e.g., imatinib, nilotinib, dasatinib)
  - Janus-kinase inhibitors
- Treatment with IV prostacyclin within 1 week prior to baseline
- Treatment with ultraviolet phototherapy within 6 weeks prior to baseline
- Treatment with infliximab, certolizumab, golimumab, abatacept, or adalimumab within 8 weeks prior to baseline
- Treatment with cyclophosphamide within 6 months prior to baseline
- History of severe allergic or anaphylactic reactions to human, humanized, or murine monoclonal antibodies
- Evidence of moderately severe concurrent nervous system, renal, endocrine, or gastrointestinal (GI) disease not related to SSc, as determined by the investigator
- Pulmonary disease with FVC ≤ 55% of predicted (best of three acceptable and repeatable measurements as described in the site's Pulmonary Function Testing Manual)

  OR

  $DL_{CO}$ ≤ 45% of predicted (corrected for hemoglobin, and the average of the 2 highest acceptable and repeatable measurements as described in the Pulmonary Function Testing Manual)
- Class II or higher pulmonary arterial hypertension (PAH), as defined by the World Health Organization
- Evidence of other moderately severe pulmonary disease (e.g., asthma, emphysema), as determined by the investigator
- Cardiovascular disease with significant arrhythmia, congestive heart failure (New York Heart Association Class II–IV), unstable angina, uncontrolled hypertension, cor pulmonale, or symptomatic pericardial effusion
- History of myocardial infarction in the last 6 months prior to screening
- Current liver disease, as determined by the investigator
- History of diverticulitis or chronic ulcerative lower GI disease, such as Crohn disease, ulcerative colitis, or other symptomatic lower GI conditions that might predispose a patient to perforations
- Known active current or significant history of recurrent bacterial, viral, fungal, mycobacterial, or other infections, including but not limited to atypical mycobacterial disease, hepatitis B or C, herpes zoster, infected digital ulcers, and osteomyelitis

- Any major episode of infection requiring hospitalization or treatment with IV antibiotics within 4 weeks prior to screening or oral antibiotics within 2 weeks prior to screening
- Significant history of recurrent tuberculosis (TB), active TB requiring treatment within the previous 3 years, or untreated latent TB

  Patients should be screened for latent TB, and, if positive, will be eligible for the study after treatment per local standard practices.
- History of or currently active primary or secondary immunodeficiency
- Evidence of malignant disease, or malignancies diagnosed within the previous 5 years (with the exception of local basal or squamous cell carcinoma of the skin or carcinoma in situ of the cervix uteri that has been excised and cured)
- History of alcohol, drug, or chemical abuse within 1 year prior to screening
- Neuropathies or other conditions that might interfere with pain evaluation, as determined by the investigator
- At screening:

  Body weight $>$ 150 kg

  Glomerular filtration rate $<$ 45 mL/min

  Alanine transaminase (ALT) or Aspartate aminotransferase (AST) $>$ 1.5 $\times$ the upper limit of normal (ULN)

  Total bilirubin $>$ ULN

  Platelet count $<$ 100 $\times$ $10^9$/L (100,000/$\mu$L)

  Hemoglobin $<$ 85 g/L (8.5 g/dL; 5.3 mmol/L)

  White blood cell (WBC) count $<$ 3.0 $\times$ $10^9$/L (3000/$\mu$L)

  Absolute neutrophil count (ANC) $<$ 2.0 $\times$ $10^9$/L (2000/$\mu$L)

  Absolute lymphocyte count $<$ 0.5 $\times$ $10^9$/L (500/$\mu$L)

  Positive hepatitis B surface antigen or hepatitis C antibody

**Length of Study**

The length of the study, from screening of the first subject to the end of the study, is expected to be approximately 4 years.

**End of Study**

The end of the study will occur when the last participating patient completes the last scheduled visit of the follow-up period. This is expected to occur 2 years after the last patient is enrolled.

**Outcome Measures**

**Efficacy Outcome Measures**

The primary efficacy outcome measure for this study is as follows:

- Change in mRSS from baseline to Week 48

The secondary efficacy outcome measures for this study are as follows:

- Proportions of patients with $\geq$20%, $\geq$40%, and $\geq$60% improvement in mRSS at Week 48 compared with baseline
- Change in FVC from baseline to Week 48
- Change in HAQ-DI from baseline to Week 48
- Change in Patient's Global Assessment from baseline to Week 48
- Change in Physician's Global Assessment from baseline to Week 48

- Time to treatment failure, defined as the time from randomization to the time of one of the following events (whichever occurs first) during the 48-week double-blind treatment period:
  - death,
  - decline in percent-predicted FVC $>10\%$ relative to baseline,
  - $>20\%$ increase in mRSS **and** an increase in mRSS of $\geq 5$ points
  - occurrence of a predefined SSc-related complication as adjudicated by the Clinical Adjudication Committee

**Safety Outcome Measures**

The safety outcome measures for this study are as follows:

- Frequency of deaths
- Nature, frequency, and severity of adverse events
- Incidence of specific laboratory abnormalities
- Change from baseline in digital ulcer count

**Immunogenicity Outcome Measures**

The immunogenicity outcome measures for this study are as follows:

- Incidence of anti-TCZ antibodies during the study relative to the prevalence of anti-TCZ antibodies at baseline
- Correlation between anti-TCZ–antibody status and efficacy, safety, or PK outcome measures

**Pharmacodynamic Outcome Measures**

The PD outcome measure for this study is as follows:

- Predose ESR and serum IL-6, soluble IL-6 receptor (sIL-6R), and CRP levels at baseline and at subsequent timepoints after initiation of study drug

**Pharmacokinetic Outcome Measures**

The PK outcome measures for this study are as follows:

- Predose serum TCZ concentration at baseline and at specified timepoints thereafter
- Correlation between PK parameters for TCZ and efficacy, safety, or immunogenicity outcome measures

**Exploratory Outcome Measures**

The exploratory outcome measures for this study are as follows:

- Proportions of patients who achieve a response, as determined by the investigator using CRISS, at Week 48
- Change in the VAS component of the SHAQ from baseline to Week 24 and baseline to Week 48
- Change in WPAI-GH score from baseline to Week 24 and baseline to Week 48
- Change in EQ-5D-3L score from baseline to Week 24 and baseline to Week 48
- Change in total score and subscores of the SGRQ from baseline to Week 48.
- Change in total and domain scores of the SkinPRO questionnaire from baseline to Week 48 (for North America only)
- Change in FACIT-Fatigue score from baseline to Week 48.
- Change in HRCT fibrosis score from baseline (based on HRCT scan performed within 3 months prior to screening) to Week 48
- Change in $DL_{CO}$ from baseline to Week 48
- Proportion of patients with $\geq 15\%$ decline in observed $DL_{CO}$ at Week 48
- Proportion of patients with $\geq 15\%$ decline in percentage of predicted $DL_{CO}$ at Week 48

- Change in FVC from baseline to Week 24
- Proportion of patients with ≥ 10% decline in observed FVC at Week 24 and at Week 48
- Proportion of patients with ≥ 10% decline in percentage of predicted FVC at Week 24 and at Week 48
- Change in mRSS from baseline to Week 24 and Week 96
- Change in observed and percentage of predicted FVC from baseline to Week 96
- Correlation between non-inherited biomarkers (serum levels of CCL18, sVCAM-1, COMP, and autotaxin; plasma levels of CXCL4; and whole blood gene signatures associated with plasmablasts and IFN) and efficacy, safety, PK, or immunogenicity outcome measures

**Investigational Medicinal Products**

**Test Product (Investigational Drug)**

Patients assigned to the TCZ group will receive *a single* subcutaneous (SC) *injection* of 162 mg of TCZ once weekly (QW) for 48 weeks during the double-blind treatment period. All patients will receive SC injections of 162 mg of TCZ QW for 48 weeks during the open-label treatment period.

**Comparator**

Patients assigned to the placebo group will receive SC injections of placebo QW for 48 weeks during the double-blind treatment period.

**Non-Investigational Medicinal Products**

From Week 24, escape therapy will be permitted for patients with worsening of skin thickening, and from Week 16, escape therapy will be permitted for patients with decline in FVC compared with baseline. Patients may receive other concomitant treatments for SSc, including treatments for new and existing organ complications.

**Statistical Methods**

**Primary Analysis**

The estimand of interest for the primary analysis is the difference between treatment arms in the mean change in the mRSS at Week 48 for the intent to treat (ITT) population. The study has been designed to continue to capture efficacy data on patients who discontinue study drug prematurely or receive escape therapies during the double-blind treatment period. These data will be included in the primary analysis.

**Determination of Sample Size**

A sample size of approximately 105 patients in the TCZ group and 105 patients in the placebo group (a total of 210 patients) will give power in the range of > 75% to 80%, (allowing for an estimated patient dropout rate of approximately 15% to 20%) to detect a between-group difference of 3.55 units (common standard deviation of 8.43) in mean change in mRSS from baseline to Week 48 using a two-group t-test, with a 5% two-sided significance level. The minimal detectable difference in mRSS (smallest treatment difference that would give a statistically significant result) under these assumptions, and with a patient dropout rate of 20%, is approximately 2.6 units.

# Appendix 1
## PROTOCOL SYNOPSIS (cont.)

**Interim Analyses**

The Sponsor will define a futility analysis to which the Sponsor will remain blinded. The futility analysis will be conducted by an external statistical group and reviewed by the iDMC. The futility analysis will be based on the treatment difference for change from baseline in mRSS at Week 24; the stopping boundary will be determined by a beta spending function. The study will be stopped for futility if the endpoint meets the futility criterion.

The futility analysis will be conducted when approximately 76 patients have either reached the Week 24 visit or have withdrawn. Since a repeated measures analysis will be used for the futility criterion, partial data from additional patients enrolled at (but not yet completed) Week 24 will also be utilized in the analysis. Thus, although only approximately one-third of the patients will have reached Week 24, the timing of the futility analysis approximates half of the final expected information (I) at Week 24, where I is the inverse of the expected variance of the treatment difference when all patients have reached Week 24.

Full statistical details of the futility analysis, along with the rationale, and timing will be documented in the iDMC charter. The iDMC charter will be made available to the relevant health authorities

# Appendix 2
# Schedule of Assessments: Screening, Baseline, and Double-Blind Treatment Period

| Assessment or Procedure | Screen. (up to 40 d) | Double-Blind Treatment Period (± 7 d, except Day 1) | | | | | | | Unsch. Visit | Treat. Discon. [c] | Follow-Up [a] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Day 1, BL | Wk 4 [b] | Wk 8 [b] | Wk 16 [b] | Wk 24 [b] | Wk 36 [b] | Wk 48 [b] | | | Wk 4 | Wk 8 [a] |
| Informed consent | x [d] | | | | | | | | | | | |
| Demographics | x | | | | | | | | | | | |
| Medical history [e] | x | | | | | | | | | | | |
| Review of inclusion and exclusion criteria | x | x | | | | | | | | | | |
| Electronic device training (PROs and study drug compliance) | | x | | | | | | | | | | |
| PRO assessments [f] | | x | | x | x | x | x | x | | | | |
| Review study drug compliance | | | x | x | x | x | x | x | x | x | | |
| Urinalysis [g, h] | x | | x | x | x | x | x | x | x | x | | |
| Pregnancy test [g, i] | x | x | x | x | x | x | x | x | | x | | |
| HBsAg and HCV serology | x | | | | | | | | | | | |
| Tuberculosis screening [j] | x | | | | | | x | | | x | | |
| Serum sample for IL-6 for stratification purposes | x | | | | | | | | | | | |
| Hematology [g, k] | x | x | x | x | x | x | x | x | x | x | | |
| Chemistry panel (serum or plasma) and creatinine clearance [g, l] | x | x | x | x | x | x | x | x | x | x | | |
| Liver profile [g, m] | x | x | x | x | x | x | x | x | x | x | | |
| Lipid panel [g,*n*] | | x | | x | | x | | x | x | x | | |
| ANA sample | | x | | | | | | | | | | |

## Appendix 2
## Schedule of Assessments: Screening, Baseline, and Double-Blind Treatment Period (cont.)

| Assessment or Procedure | Screen. (up to 40 d) | Double-Blind Treatment Period (± 7 d, except Day 1) | | | | | | | Unsch. Visit | Treat. Discon. [c] | Follow-Up [a] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Day 1, BL | Wk 4 [b] | Wk 8 [b] | Wk 16 [b] | Wk 24 [b] | Wk 36 [b] | Wk 48 [b] | | | Wk 4 | Wk 8 [a] |
| SSc-specific auto-antibody panel [g, o] | | x | | | | | | x | x | x | | |
| Serum anti-TCZ antibody sample [g, p] | | x | | x | x | x | x | x | x | x | | x |
| Serum sample for PK analysis [g, p, q] | | x | x | x | x | x | x | x | x | x | | x |
| IL-6 sample [g, q] | | x | x | x | x | x | x | x | | x | | |
| sIL-6R sample [g,p,q] | | x | x | x | x | x | x | x | x | x | | x |
| High-sensitivity CRP [g] | x | x | x | | | x | | x | x | x | | |
| ESR [g] | x | x | x | | | x | | x | x | x | | |
| Serum sample for candidate biomarkers [g] | | x | | | | x | | x | | | | |
| Plasma (EDTA) sample for candidate biomarkers [g] | | x | | | | x | | x | | | | |
| Whole blood sample for RNA extraction [g] | | x | | | | x | | x | | | | |
| Skin biopsies (RCR sample, optional) [r] | | x | | | | | | x | | | | |
| Whole blood RCR sample for DNA exaction (optional) | | x | | | | | | | | | | |
| mRSS | x | x | | x | x | x | x | x | x | x | | |
| Forced vital capacity | x | x | | x | x | x | x | x | x | x | | |
| $DL_{CO}$ | x | x | | | | x | | x | x | x | | |
| Physician's Global Assessment [s] | | x | | x | x | x | x | x | x | x | | |
| High-resolution CT scan [t] | | x | | | | | | x | | | | |
| Physical examination [u] | x | x | | | | | | | x | x | | |

# Appendix 2
# Schedule of Assessments: Screening, Baseline, and Double-Blind Treatment Period (cont.)

| Assessment or Procedure | Screen. (up to 40 d) | Double-Blind Treatment Period (± 7 d, except Day 1) | | | | | | | Unsch. Visit | Treat. Discon. [c] | Follow-Up [a] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Day 1, BL | Wk 4 [b] | Wk 8 [b] | Wk 16 [b] | Wk 24 [b] | Wk 36 [b] | Wk 48 [b] | | | Wk 4 | Wk 8 [a] |
| Height | x [v] | | | | | | | | | | | |
| Body weight [g] | x | x | | | | | | x | *x* | x | | |
| Vital signs [g, w] | x | x | x | x | x | x | x | x | *x* | x | | |
| Digital ulcer count [g] | | x | | x | x | x | x | x | x | x | | |
| ECG | x | | | | | | | | x | | | |
| Echocardiogram | x | | | | | | | | x | x | | |
| Adverse events [g, x, y] | x | x | x | x | x | x | x | x | x | x | x | x |
| Concomitant medications [g, z] | x | x | x | x | x | x | x | x | x | x | x | x |
| Study drug distribution and administration [g, aa] | | x [bb, cc] | x | x | x | x | x | x [bb] | | | | |

ANA = anti-nuclear antibody; BL = baseline; CBC = complete blood count; CRP = C-reactive protein; CT = computed tomography; d = day; Discon. = discontinuation; $DL_{CO}$ = diffusing capacity of the lung for carbon monoxide; eCRF = electronic Case Report Form; ESR = erythrocyte sedimentation rate; EQ-5D-3L = EuroQol 5-Dimension Questionnaire (three levels of severity); FVC = forced vital capacity; HAQ-DI = Health Assessment Questionnaire Disability Index; HBsAg = hepatitis B surface antigen; HCV = hepatitis C virus; HRCT = high-resolution computed tomography; IL = interleukin; mRSS = modified Rodnan Skin Score; PK = pharmacokinetic; PRO = patient-reported outcome; RCR = Roche Clinical Repository; Screen. = screening; SGRQ = Saint George's Respiratory Questionnaire; SHAQ = Scleroderma Health Assessment Questionnaire; SkinPRO = Scleroderma Skin Patient-Reported Outcome; SSc = systemic sclerosis; TCZ = tocilizumab; Treat. = treatment; Unsch. = unscheduled; Wk = week; WPAI-GH = Work Productivity and Activity Impairment—General Health.

# Appendix 2
# Schedule of Assessments: Screening, Baseline, and Double-Blind Treatment Period (cont.)

Note: Assessments and procedures should be performed in the sequence that is most practical for the site, as long as PROs are performed first and study drug administration is performed last.

[a] All patients will undergo follow-up assessments within 4 weeks of study drug discontinuation and at 8 weeks after study drug discontinuation. For patients who discontinue study drug prematurely, a follow-up visit can be combined with the next scheduled visit (as outlined in Appendix 3 provided that the timing of the scheduled visit coincides with the specified timing for the follow-up visit. The follow-up visit at 4 weeks may be conducted by telephone.

[b] For patients at participating sites who have provided written informed consent to participate in home nursing services, specified assessments at Weeks 4, 8, 16, 24, 36, and 48, as well as Week 8 of the Follow-up Period may be performed by a trained home nursing professional or (at sites with established teams) by appropriately qualified site personnel at the patient's home or another suitable location.

[c] Patients who discontinue study drug prematurely should undergo assessments as outlined in Appendix 3, with the timing of those visits being relative to baseline. Assessments at the early treatment discontinuation visit should be performed *as soon as possible after discontinuing* study drug.

[d] Informed consent must be documented before any study-specific screening procedure is performed.

[e] Medical history includes clinically significant diseases (including SSc complications) reproductive status, smoking history, and use of alcohol and drugs of abuse.

[f] PRO questionnaires are to be completed prior to all other assessments during the study visit, with the exception of ECGs. Patients will use an electronic PRO device to capture PRO data. The appropriate PRO assessments will be programmed to appear at specific visits. The HAQ-DI, will be completed at baseline and at Weeks 8, 16, 24, 36, and 48. The Patient's Global Assessment, SHAQ, SGRQ, FACIT-Fatigue, and SkinPRO questionnaire will be completed at baseline and at Weeks 8, 16, 24, and 48. The SkinPRO questionnaire will only be administered in North America. The WPAI-GH and EQ-5D-3L will be completed at baseline and at Weeks 24 and 48.

[g] For patients at participating sites who have provided written informed consent to participate in home nursing services, this assessment or procedure may be performed by a trained home nursing professional or (at sites with established teams) by appropriately qualified site personnel at the patient's home or another suitable location.

[h] Urinalysis includes dipstick (pH, specific gravity, glucose, protein, ketones, blood) and microscopic examination (sediment, RBCs, WBCs, casts, crystals, epithelial cells, and bacteria).

[i] All women who are not postmenopausal (≥ 12 months of non-therapy-induced amenorrhea) or surgically sterile will have a serum pregnancy test at screening. Urine pregnancy tests will be performed at specified subsequent visits. If a urine pregnancy test is positive, it must be confirmed by a serum pregnancy test.

[j] Tuberculosis screening *must* be performed *at screening and at Week 36. The screening method (e.g. PPD or QuantiFERON® test) is at the discretion of the investigator.*

## Appendix 2
## Schedule of Assessments: Screening, Baseline, and Double-Blind Treatment Period (cont.)

[k] Hematology includes hemoglobin, hematocrit, RBC count (with morphology), WBC count, platelet count, differential count (neutrophils, eosinophils, basophils, monocytes, lymphocytes, other cells).

[l] Chemistry panel includes total protein, serum albumin, creatine phosphokinase, sodium, potassium, calcium, BUN or urea, serum creatinine, C3, and C4. Creatinine clearance will be calculated by a central laboratory.

[m] Liver profile consists of AST, ALT, alkaline phosphatase, and total bilirubin (direct and indirect bilirubin will be performed if total bilirubin greater than the upper limit of normal).

[n] Overnight fasting (> 8 hours) is required. An additional fasting lipid panel should be obtained 8 weeks after initiation of lipid-lowering therapy.

[o] SSc-specific autoantibody panel includes anti-topoisomerase, anti-RNA polymerase, anti-PM/Scl, anti-histone, anti-U1 snRP, and anti-centromere antibodies.

[p] Additional samples for PK analysis and analysis of anti-TCZ antibodies and sIL-6R will be collected prior to resuming study drug for patients who have missed at least three consecutive doses and at the time of anaphylaxis or a serious hypersensitivity reaction.

[q] Samples for PK analysis and analysis of IL-6, sIL-6R, and candidate biomarkers will be obtained at a single blood draw and aliquoted according to the procedures in the Sample Handling and Logistics Manual.

[r] Two 3-mm punch biopsies are to be obtained from clinically involved skin, preferably at the forearm (optional).

[s] Physician's Global Assessment is to be completed by the investigator on the basis of examination and overall assessment of the patient.

[t] As accepted by the local regulations. Good-quality (as determined by the site radiologist and/or investigator), standard-of-care HRCT scans obtained within 3 months prior to screening and in accordance with study image acquisition guidelines can be used for baseline.

[u] A physical examination will be performed but will not be recorded on the eCRF, if normal; any abnormality will be reported either on the Medical History eCRF (for screening examination) or Adverse Event eCRF (for examinations after the screening).

[v] Height is required at screening only and will be recorded on the Vital Signs eCRF.

[w] Vital signs will include measurements of respiratory rate, pulse rate, and systolic and diastolic blood pressures while the patient is in a seated position, and temperature. Additional measurements may be performed in the event of an adverse event, at the discretion of the investigator. Temperature readings (as part of vital signs) will be measured but will not be recorded on the eCRF, if normal; any abnormal body temperature will be reported on the Adverse Event eCRF.

## Appendix 2
## Schedule of Assessments: Screening, Baseline, and Double-Blind Treatment Period (cont.)

[x] After informed consent has been obtained but prior to initiation of study drug, only serious adverse events caused by a protocol-mandated intervention should be reported. After initiation of study drug, all adverse events will be reported until 8 weeks after the last dose of study drug or until completion of the last scheduled visit, whichever occurs later. However, for patients who discontinue study drug prematurely but continue scheduled visits, all adverse events will be reported until completion of the last scheduled visit or 8 weeks after the last dose of study drug, whichever occurs later. After this period, the investigator should report any serious adverse events that are believed to be related to prior study drug treatment (see Section 5.6). The investigator should follow each adverse event until the event has resolved to baseline grade or better, the event is assessed as stable by the investigator, the patient is lost to follow-up, or the patient withdraws consent. Every effort should be made to follow all serious adverse events considered to be related to study drug or trial-related procedures until a final outcome can be reported.

[y] For all serious infectious adverse events, CBC, differentials, and platelets should be determined during the disease episode. Every effort should be made to collect appropriate specimens for serology, polymerase chain reaction, or culture to identify the infectious organism. The results of all laboratory assessments performed locally, except for CRP, should be reported on the eCRF.

[z] Concomitant therapy includes any medication (e.g., prescription drugs, over-the-counter drugs, herbal or homeopathic remedies, nutritional supplements) used by a patient from 4 weeks prior to screening until 8 weeks after the last dose of study drug or until completion of the last scheduled visit, whichever occurs later. In addition, all medications taken for SSc since diagnosis will be recorded at screening.

[aa] If for any reason the weekly schedule cannot be kept (e.g., SC injections have to be administered to patients during a site visit), injections may be given a minimum of 5 days and a maximum of 11 days apart.

[bb] The first SC injection in both study periods (double-blind and open-label) will be administered to patients at the site under close supervision. The Week 48 injection is the first dose for the open-label treatment period.

[cc] Patients (and patient caregivers, if applicable) will be trained on how to perform SC injections at the Day 1 visit. For patients and caregivers at applicable sites who require additional training, study drug may be administered (or guidance provided) at Weeks 1 and 2 by a home nursing professional or by appropriately qualified site personnel.

# Appendix 3
# Schedule of Assessments: Open-Label Treatment Period

| Assessment or Procedure | Open-Label Treatment Period (± 7 d) | | | | | Unsch. Visit | Treat. Discon. [b] | Follow-Up [c] | |
|---|---|---|---|---|---|---|---|---|---|
| | Wk 52 [a] | Wk 60 [a] | Wk 72 [a] | Wk 84 [a] | Wk 96 [a] | | | Wk 4 | Wk 8 [a] |
| PRO assessments [d] | | | x | | x | | | | |
| Review study drug compliance | x | x | x | x | x | x | x | | |
| Urinalysis [e,f] | | x | x | x | x | x | x | | |
| Pregnancy test [e,g] | x | x | x | x | x | | x | | |
| Tuberculosis screening [h] | | | | | | | x | | |
| Hematology [e,i] | x | x | x | x | x | x | x | | |
| Chemistry panel (serum or plasma) and creatinine clearance [e,j] | x | x | | x | | x | x | | |
| Liver profile [e,k] | x | x | x | x | x | x | x | | |
| Lipid panel [e,l] | | x | x | | x | x | x | | |
| SSc-specific auto-antibody panel [e,m] | | | | | x | x | x | | |
| Serum anti-TCZ antibody sample [e,n] | | | | | x | x | x | | x |
| Serum sample for PK analysis [e,n,o] | | | | | x | x | x | | x |
| IL-6 sample [e,o] | | | | | x | | $x$ | | |
| sIL-6R sample [e,n,o] | | | | | x | x | x | | x |
| High-sensitivity CRP [e] | | | x | | x | x | x | | |
| ESR [e] | | | x | | x | x | x | | |
| mRSS | | | x | | x | x | x | | |
| Forced vital capacity | | | x | | x | x | x | | |
| $DL_{CO}$ | | | x | | x | x | x | | |

# Appendix 3
# Schedule of Assessments: Open-Label Treatment Period (cont.)

| Assessment or Procedure | Open-Label Treatment Period (± 7 d) | | | | | Unsch. Visit | Treat. Discon. [b] | Follow-Up [c] | |
|---|---|---|---|---|---|---|---|---|---|
| | Wk 52 [a] | Wk 60 [a] | Wk 72 [a] | Wk 84 [a] | Wk 96 [a] | | | Wk 4 | Wk 8 [a] |
| Physician's Global Assessment [p] | | | x | | x | x | x | | |
| Physical examination [q] | | | | | | x | x | | |
| Body weight [e] | | | | | x | *x* | x | | |
| Vital signs [e,r] | x | x | x | x | x | x | x | | |
| Digital ulcer count [e] | | | x | | x | x | x | | |
| Echocardiogram | | | | | | x | x | | |
| Adverse events [e,s,t] | x | x | x | x | x | x | x | x | x |
| Concomitant medications [e,u] | x | x | x | x | x | x | x | x | x |
| Study drug distribution and administration [e,v] | *x*[w] | x | x | x | | | | | |

CBC = complete blood count; CRP = C-reactive protein; d = day; Discon. = discontinuation; $DL_{CO}$ = diffusing capacity of the lung for carbon monoxide; eCRF = electronic Case Report Form; ESR = erythrocyte sedimentation rate; EQ-5D-3L = EuroQol 5-Dimension Questionnaire (three levels of severity); FVC = forced vital capacity; HAQ-DI = Health Assessment Questionnaire Disability Index; HBsAg = hepatitis B surface antigen; HCV = hepatitis C virus; HRCT = high-resolution computed tomography; IL = interleukin; mRSS = modified Rodnan Skin Score; PK = pharmacokinetic; PRO = patient-reported outcome; RCR = Roche Clinical Repository; SHAQ = Scleroderma Health Assessment Questionnaire; SGRQ = Saint George's Respiratory Questionnaire; SkinPRO = Scleroderma Skin Patient-Reported Outcome; SSc = systemic sclerosis; TCZ = tocilizumab; Treat. = treatment; Unsch. = unscheduled; Wk = week; WPAI-GH = Work Productivity and Activity Impairment—General Health.

Note: Assessments and procedures should be performed in the sequence that is most practical for the site, as long as PROs are performed first and study drug administration is performed last.

[a] For patients at participating sites who have provided written informed consent to participate in home nursing services, specified assessments at Weeks 52, 60, 72, 84, and 96, as well as Week 8 of the Follow-up Period may be performed by a trained home nursing professional or (at sites with established teams) by appropriately qualified site personnel at the patient's home or another suitable location.

[b] Discontinued patients should undergo *a treatment discontinuation visit as soon as possible after discontinuing study drug.*

# Appendix 3
# Schedule of Assessments: Open-Label Treatment Period (cont.)

[c] All patients will undergo follow-up assessments within 4 weeks of study drug discontinuation and at 8 weeks after study drug discontinuation, *except for patients who reach Week 96 but who transition onto locally provided TCZ prior to the follow up visit.* The follow-up visit at 4 weeks may be conducted by telephone.

[d] PRO questionnaires are to be completed prior to all other assessments during the study visit. Patients will use an electronic PRO device to capture PRO data. The appropriate PRO assessments will be programmed to appear at specific visits. The HAQ-DI, Patient's Global Assessment, SHAQ, SGRQ, FACIT-Fatigue, and SkinPRO questionnaire will be completed at Weeks 72 and 96. The SkinPRO questionnaire will only be administered in North America. The WPAI-GH and EQ-5D-3L will be completed at Week 96.

[e] For patients at participating sites who have provided written informed consent to participate in home nursing services, this assessment or procedure may be performed by a trained home nursing professional or (at sites with established teams) by appropriately qualified site personnel at the patient's home or another suitable location.

[f] Urinalysis includes dipstick (pH, specific gravity, glucose, protein, ketones, blood) and microscopic examination (sediment, RBCs, WBCs, casts, crystals, epithelial cells, and bacteria).

[g] Urine pregnancy tests will be conducted for all women who are not postmenopausal (≥ 12 months of non-therapy-induced amenorrhea) or surgically sterile. If a urine pregnancy test is positive, it must be confirmed by a serum pregnancy test.

[h] Tuberculosis screening *must* be performed *at screening and at Week 36. The screening method (e.g. PPD or QuantiFERON® test) is at the discretion of the investigator.*

[i] Hematology includes hemoglobin, hematocrit, RBC count (with morphology), WBC count, platelet count, differential count (neutrophils, eosinophils, basophils, monocytes, lymphocytes, other cells).

[j] Chemistry panel includes total protein, serum albumin, creatine phosphokinase, sodium, potassium, calcium, BUN or urea, serum creatinine, C3, and C4. Assessment of creatinine clearance is to be performed every 12 weeks and will be calculated by the central laboratory. Creatinine clearance will be calculated by a central laboratory.

[k] Liver profile consists of AST, ALT, alkaline phosphatase, and total bilirubin (direct and indirect bilirubin will be performed if total bilirubin greater than the upper limit of normal).

[l] Overnight fasting (> 8 hours) is required. An additional fasting lipid panel should be obtained 8 weeks after initiation of lipid-lowering therapy.

[m] SSc-specific autoantibody panel includes anti-topoisomerase, anti-RNA polymerase, anti-PM/Scl, anti-histone, anti-U1 snRP, and anti-centromere antibodies.

[n] Additional samples for PK analysis and analysis of anti-TCZ antibodies and sIL-6R will be collected prior to resuming study drug for patients who have missed at least three consecutive doses and at the time of anaphylaxis or a serious hypersensitivity reaction.

[o] Samples for PK analysis and analysis of IL-6 and sIL-6R will be obtained at a single blood draw and aliquoted according to the procedures in the Sample Handling and Logistics Manual.

# Appendix 3
# Schedule of Assessments: Open-Label Treatment Period (cont.)

[p] Physician's Global Assessment is to be completed by the investigator on the basis of examination and overall assessment of the patient.

[q] A physical examination will be performed but will not be recorded on the eCRF, if normal; any abnormality will be reported on the Adverse Event eCRF.

[r] Vital signs will include measurements of respiratory rate, pulse rate, and systolic and diastolic blood pressures while the patient is in a seated position, and temperature. Additional measurements may be performed in the event of an adverse event, at the discretion of the investigator. Temperature readings (as part of vital signs) will be measured but will not be recorded on the eCRF, if normal; any abnormal body temperature will be reported on the Adverse Event eCRF.

[s] After initiation of study drug, all adverse events will be reported until 8 weeks after the last dose of study drug or until completion of the last scheduled visit, whichever occurs later. However, for patients who discontinue study drug prematurely but continue scheduled visits, all adverse events will be reported until completion of the last scheduled visit or 8 weeks after the last dose of study drug, whichever occurs later. After this period, the investigator should report any serious adverse events that are believed to be related to prior study drug treatment (see Section 5.6). The investigator should follow each adverse event until the event has resolved to baseline grade or better, the event is assessed as stable by the investigator, the patient is lost to follow-up, or the patient withdraws consent. Every effort should be made to follow all serious adverse events considered to be related to study drug or trial-related procedures until a final outcome can be reported.

[t] For all serious infectious adverse events, CBC, differentials and platelets should be determined during the disease episode. Every effort should be made to collect appropriate specimens for serology, polymerase chain reaction, or culture to identify the infectious organism. The results of all laboratory assessments performed locally, except for CRP, should be reported on the eCRF.

[u] Concomitant therapy includes any medication (e.g., prescription drugs, over-the-counter drugs, herbal or homeopathic remedies, nutritional supplements) used by a patient from 4 weeks prior to screening until 8 weeks after the last dose of study drug or until completion of the last scheduled visit, whichever occurs later.

[v] If for any reason the weekly schedule cannot be kept (e.g., SC injections have to be administered to patients during a site visit), injections may be given a minimum of 5 days and a maximum of 11 days apart.

[w] Study drug administration only; no study drug distribution at this visit.

# Appendix 4
# Schedule of Assessments: Patients Who Have Discontinued Study Drug Prematurely

| Assessment or Procedure | Timing of Visit Relative to Baseline (Day 1) | | | | |
|---|---|---|---|---|---|
| | Wk 8 [a] (±7d) | Wk 16 [a] (±7d) | Wk 24 [a] (±7d) | Wk 36 [a] (±7d) | Wk 48 [a, e] (±7d) |
| mRSS | x | x | x | x | x |
| Forced vital capacity | x | x | x | x | x |
| HAQ-DI | x | x | x | x | x |
| Adverse events [a,b,c] | x | x | x | x | x |
| Concomitant medications [a,d] | x | x | x | x | x |

## Appendix 4
## Schedule of Assessments: Patients Who Have Discontinued Study Drug Prematurely (cont.)

CBC = complete blood count; CRP = C-reactive protein; d; eCRF = electronic Case Report Form; HAQ-DI = Health Assessment Questionnaire Disability Index; mrSS = modified Rodnan Skin Score; Wk = week.

Note: Patients who discontinue study drug prematurely but continue to attend scheduled study visits should follow a reduced assessment schedule as outlined above, starting with the first scheduled visit following discontinuation of study drug. *These patients should also undergo a treatment discontinuation (TD) visit as soon as possible after discontinuing study drug, and follow-up assessments within 4 weeks of study drug discontinuation and at 8 weeks after study drug discontinuation. A TD and/or follow-up visit may be combined with the next scheduled visit (as outlined above in Appendix 3), provided that the timing of the scheduled visit coincides with the specified timing for the TD or follow-up visit.*

[a] For patients at participating sites who have provided written informed consent to participate in home nursing services, specified assessments at each visit may be performed by a trained home nursing professional or (at sites with established teams) by appropriately qualified site personnel at the patient's home or another suitable location.

[b] After initiation of study drug, all adverse events will be reported until 8 weeks after the last dose of study drug or until completion of the last scheduled visit, whichever occurs later. However, for patients who discontinue study drug prematurely but continue scheduled visits, all adverse events will be reported until completion of the last scheduled visit or 8 weeks after the last dose of study drug, whichever occurs later. After this period, the investigator should report any serious adverse events that are believed to be related to prior study drug treatment (see Section 5.6). The investigator should follow each adverse event until the event has resolved to baseline grade or better, the event is assessed as stable by the investigator, the patient is lost to follow-up, or the patient withdraws consent. Every effort should be made to follow all serious adverse events considered to be related to study drug or trial-related procedures until a final outcome can be reported.

[c] For all serious infectious adverse events, CBC, differentials and platelets should be determined during the disease episode. Every effort should be made to collect appropriate specimens for serology, polymerase chain reaction, or culture to identify the infectious organism. The results of all laboratory assessments performed locally, except for CRP, should be reported on the eCRF.

[d] Concomitant therapy includes any medication (e.g., prescription drugs, over-the-counter drugs, herbal or homeopathic remedies, nutritional supplements) used by a patient from 4 weeks prior to screening until 8 weeks after the last dose of study drug or until completion of the last scheduled visit, whichever occurs later. *Patients who start other medications for systemic sclerosis (e.g., DMARDs) between treatment discontinuation and Week 48 may remain on these medications during the open-label period at the discretion of the investigator.*

[e] *Patients who enter the open-label period at Week 48 must complete the full Week 48 schedule of assessments in Appendix 1. If these patients discontinue after TCZ treatment in the open-label period they should undergo a second TD visit and complete the 4 and 8 week follow-up visits.*

# Appendix 5
# HAQ-DI

The HAQ-DI is a self-completed patient questionnaire specific for RA. It consists of 20 questions referring to eight domains: dressing/grooming, arising, eating, walking, hygiene, reach, grip, and common daily activities. There are four possible responses for each component:

0=without any difficulty

1=with some difficulty

2=with much difficulty

3=unable to do

A domain score is determined from the highest score of the components in that domain (except when aids and devices are taken into account; see below). For example, if there are three components in a domain and the responses were 1, 2 and 0 to the components in the domain, then the score for the domain would be 2.

If a domain consists of only two questions and has one missing response, then the non-missing response will be used as the value for the domain. If a domain consists of three questions and has one missing response, then the domain score will be the higher of the two responses. If a domain consists of three questions and has two missing responses, then the domain will be considered missing.

To calculate the HAQ-DI, the patient must have a domain score for at least six of the eight domains. The HAQ-DI is the sum of the domain scores, divided by the number of domains that have a score.

The HAQ-DI takes into account the patient's use of aids or devices in the scoring for a domain. For each of the eight domains, there is an aids or devices companion variable(s) that is used to record any assistance the patient uses. Where aids or devices are indicated by a patient for a domain or help from another person is required for a domain, if the maximum score for the domain is $<2$, the domain score is increased to 2 to reflect the use of an aid or device, or help. If the maximum value is $\geq 2$, the score is not modified. In the event that a domain score is missing but a corresponding aid or device is listed, then the score for that domain will reflect the use of the aid or device (i.e., it will be scored as 2).

Where "other" is ticked for use of aids or devices, the use of the aid or device will not be assigned to a domain and will therefore not be reflected in the domain scoring.

HAQ-DI scores can range from 0 to 3.

# Appendix 6
# St. George's Respiratory Questionnaire

## OUTLINE OF SCORING ALGORITHM

Three component scores are calculated: Symptoms; Activity; Impacts. One Total score is also calculated.

## PRINCIPLE OF CALCULATION

Each questionnaire response has a unique empirically derived 'weight'. The lowest possible weight is zero and the highest is 100. Each component of the questionnaire is scored separately in three steps:

1. The weights for all items with a positive response are summed.
2. The weights for missed items are deducted from the maximum possible weight for each component. The weights for all missed items are deducted from the maximum possible weight for the Total score.
3. The score is calculated by dividing the summed weights by the adjusted maximum possible weight for that component and expressing the result as a percentage:

Total and scale scores will be calculated as follows:

Total Score=100*((Σ item weights)/(maximum total weight))

Symptom Score=100*((Σ item 1, 2, 3, 4, 5, 6, 7, 8 weights)/(maximum symptom total weight))

Activity Score=100*((Σ item 11, 15 weights)/(maximum activity total weight))

Impact Score=100*((Σ item 9, 10, 12, 13, 14, 16, 17 weights)/(maximum impact total weight)

Sum of maximum possible weights for each component and Total:

- Symptoms: 662.5
- Activity: 1209.1
- Impacts: 2117.8
- Total: 3989.4

(Note: these are the maximum possible weights that could be obtained for the worst possible state of the patient).

It will be noted that the questionnaire requests a single response to questions 1-7, 9-10 and 17. If multiple responses are given to one of these questions then the weights for all positive responses for that question will be averaged.

# Appendix 6
# St. George's Respiratory Questionnaire (cont.)

For information regarding question weights, please refer to the St George's Respiratory Questionnaire Manual, Version 2.3.

Missing individual items for the scoring of the SGRQ will be handled as follows: The symptoms component will be considered missing if more than 2 of the items are missing. The activity component will be considered missing if more than 4 of the items are missing. The impacts component will be considered missing if more than 6 of the items are missing. The total score will be considered missing if any component is missing. For a valid component questionnaire with unanswered questions, the predefined weight and score for a particular missing question will be used to calculate the aggregated score for the component (St George's Respiratory Questionnaire Manual, Version 2.3). Missing data will not be imputed.

# Appendix 7
# FACIT-FATIGUE

The FACIT-Fatigue questionnaire consists of 13 statements designed to measure the degree of fatigue experienced by the patient in the previous 7 days. For each question there are five possible responses: 0 (not at all), 1 (a little bit), 2 (somewhat), 3 (quite a bit), 4 (very much). Statements 1 to 6 and 9 to 13 are worded so that higher scores correspond to greater fatigue, while statements 7 and 8 are worded so that higher scores correspond to less fatigue. All scores except those for statements 7 and 8 will therefore be "reversed" on the 0 to 4 scale (i.e., a response of 0 will receive a score of 4, a response of 1 will receive a score of 3, etc.), so that for all questions higher scores will reflect improvement.

For each questionnaire, if there are less than 7 responses recorded, then the total fatigue score will be considered missing. If there are 7 or more responses recorded, then the total fatigue score for that questionnaire will be calculated as the average of the non-missing scores multiplied by 13. FACIT-Fatigue scores range from 0 to 52.

# Appendix 8
# SkinPRO

The SkinPRO is a patient-reported outcome instrument developed to assess the skin-related quality of life in patients with Ssc. Initially, the SkinPRO contained 22 items, but following further validation work, 4 items from the SkinPRO were dropped (2, 4, 6, 22) resulting in an 18 item final version referred to as the SSPRO (Man A et al. 2017). The SSPRO has 4 subscales: physical effects, physical limitations, emotional effects, and social effects. Each item is scored on a 7-point Likert scale (0–6) with verbal anchors at 0 (Not at all) and 6 (very much).

For the version of the SSPRO in protocol WA29767, items 1, 3, 5, 7–21 contribute to the total score. This has a range of 0 to 108, which is then transformed to a 0–100 scale (the score is divided by 108 and then multiplied by 100). Higher scores indicate greater severity of symptoms/impacts. For the subscales, the items included are as follows: physical effects (Items 1, 3, 5, 7, 8), physical limitations (Items 9-14), emotional effects (items 15–18), and social effects (items 19–21). The subscale scores are also transformed to a 0-100 scale for ease of comparison.

# Appendix 9
# EQ-5D-3L™

The EQ VAS records the respondents self-rated health status on a vertical graduated (0-100) visual analogue scale. Participants draw a line from a box to the point on the thermometer-like scale corresponding to their health state (100=Best health state). Scores for the visual analogue scale reflect the position where participant's line crosses the thermometer-like scale.

The EQ-5D-3L descriptive system comprises 5 dimensions of health:

1. mobility
2. self-care
3. usual activities
4. pain/discomfort
5. anxiety/depression

Each dimension comprises three levels (no problems, some problems, and extreme problems). Participants are asked to indicate their level of health by checking one of the three responses for each domain.

# Appendix 10
# WPAI-GH

WPAI outcomes are expressed as impairment percentages, with higher numbers indicating greater impairment and less productivity, i.e., worse outcomes, as follows:

**Questions**:

1. currently employed
2. hours missed due to health problems
3. hours missed other reasons
4. hours actually worked
5. degree health affected productivity while working
6. degree health affected regular activities

**Scores:**

Multiply scores by 100 to express in percentages

Percent work time missed due to health: $Q2/(Q2+Q4)$

Percent impairment while working due to health: Q5/10

Percent overall work impairment due to health:
$Q2/(Q2+Q4)+[(1-(Q2/(Q2+Q4)))\times(Q5/10)]$

Percent activity impairment due to health: Q6/10

# Appendix 11
# Cochran-Mantel-Haenszel Test

- The weighted difference in proportions is the difference in the response rates in the experimental treatment group compared with the co ntrol treatment group, adjusted for any stratification factors. With two stratification factors, the number of patients in each strata is defined as nijk where i is the level of the first stratification factor and j is the level of the second stratification factor and k is treatment group (experimental or control). The number of events in each strata is denoted by xijk, where i, j and k are as above. The proportion of responders in each strata will be calculated by:

$$p_{ijk} = \frac{x_{ijk}}{n_{ijk}}$$ where i, j and k are as above

- The difference in proportions for each strata will then be calculated as the proportion of patients in each strata in the experimental treatment group (EXP) minus the proportion of patients in each strata in the control treatment group (CON) and denoted $d_{ij} = p_{ijEXP} - p_{ijCON}$, for i and j as above.
- The weights for each strata (i, j) will be calculated as follows:

$$w_{ij} = \frac{n_{ijEXP} * n_{ijCON}}{n_{ijEXP} + n_{ijCON}}$$

- Within each strata, the weighted differences in the proportions in each of the treatment groups will be calculated as follows:

$$wd_{ij} = w_{ij} d_{ij}$$

- and then summed:

$$WD = \sum_i \sum_j wd_{ij}$$

- After calculation of the weighted difference in proportions, the calculation of the 95% confidence interval is as follows;
- Continuity-corrected Proportions

$$p^{\#}_{ijk} = \frac{x_{ijk} + 0.5}{n_{ijk} + 1}$$

- Variances

$$Upvar_{ij} = w_{ij}^2 \left[ p^{\#}_{ijEXP} \frac{\left(1 - p^{\#}_{ijEXP}\right)}{n_{ijEXP}} + p^{\#}_{ijCON} \frac{\left(1 - p^{\#}_{ijCON}\right)}{n_{ijCON}} \right]$$

# Appendix 11
# Cochran-Mantel-Haenszel Test (cont.)

- To calculate the sum of the weights and variances over all strata:

Sum over Strata

$$W = \sum_i \sum_j w_{ij} \quad \text{(sum of weights)}$$

$$Var = \sum_i \sum_j Up \operatorname{var}_{ij} \quad \text{(sum of variances)}$$

Point Estimate and Standard Error

$$d = \frac{WD}{W}; \quad se = \sqrt{\frac{Var}{W^2}}$$

Stratified 95% Confidence Intervals

$$\text{Lower Limit} = d - 1.96se$$

$$\text{Upper Limit} = d + 1.96se$$