SLS III Statistical Analysis Plan

TRIAL FULL TITLE	Scleroderma Lung Study III
SAP VERSION	1.0
SAP VERSION DATE	24May2022
TRIAL STATISTICIAN	
Protocol Version (SAP associated with)	Protocol Version 2.5 17NOV2021
TRIAL PRINCIPAL INVESTIGATOR	Michael D. Roth, MD Dinesh Khanna, MD, MSc
SAP AUTHOR(s)	Yihsuan Tu, PhD

1 SAP Signatures

Statistician (Author)

Name: Yihsuan Tu, PhD

Signature:

Date: 24 May 2022

Statistician Reviewer (As applicable)

Name: Cathie Spino, ScD

Signature: Cathie Spino

Date: 24 May 2022

Principal Investigator

Name: Michael D. Roth, MD

Signature: Michael D. Roth MD

Date: 24 May 2022

Co-Principal Investigator

Name: Dinesh Khanna, MD, MSc

Signature: Dinesh Khanna

Date: 24 May 2022

2 Table of Contents

3 Abbreviations and Definitions

AE	Adverse Event
ACR	American College of Rheumatology
ADL	Activities of daily living
ALT	Alanine Aminotransferase
ANC	Absolute Neutrophil Count
ANCOVA	Analysis of Covariance
AST	Aspartate Aminotransferase
ATS	American Thoracic Society
AUC	Area under the curve
BID	Twice daily
BDI	Baseline Mahler Modified Dyspnea Index
BAL	Bronchoalveolar lavage
BUN	Blood Urea Nitrogen
CDC	Center for Disease Control
CFR	Code of Federal Regulations
CLIA	Clinical Laboratory Improvement Amendments
Cmax	Maximum concentration
CMP	Clinical monitoring plan
CMV	Cytolomegalovirus
COVID-19	Coronavirus Disease 2019
Cr	Creatinine
CRF	Case Report Form
CRISS	Combined Response index in Systemic Sclerosis
CTCAE	Common Terminology Criteria for Adverse Events V4.03-2010
CYC	Cyclophosphamide
CYP	Cytochrome P450
DCC	Data Coordinating Center
DLCO	Single-Breath Diffusing Capacity of the Lung for Carbon Monoxide
DLCO-Hb-%	DLCO, adjusted for age, height, gender and hemoglobin
DMARD	Disease-modifying antirheumatic drug
DSMB	Data Safety Monitoring Board
ECG	Electrocardiogram
eCRF	Electronic Case Report Forms
ERS	European Respiratory Society
EULAR	European League Against Rheumatism
FEV1	Forced Expiratory Volume in the first second
FDA	Food and Drug Administration
FVC	Forced vital capacity
FVC-%	Forced vital capacity as a percentage of the age-, height-, gender- and race-adjusted predicted value
GERD	Gastroesophageal reflux disease
GCP	Good Clinical Practice
GGO	Ground glass opacification
GLP	Good Laboratory Practices

GMP	Good Manufacturing Practices
H&P	History and physical
HBV	Hepatitis B virus
HCV	Hepatitis C virus
Hgb	Hemoglobin
HIPAA	Health Insurance Portability and Accountability Act
HRCT	High resolution computerized tomography
HRCT-TLC	HRCT-measured total lung capacity at maximum inspiration
HRQoL	Health-related quality of life
IB	Investigator’s Brochure
IFN-γ	Interferon-gamma
ILD	Interstitial lung disease
IND	Investigational New Drug Application
IPF	Idiopathic pulmonary fibrosis
IRB	Investigational Review Board
JC	Polyomavirus JC
LCQ	Leicester Cough Questionnaire
LFTs	Liver function test
MCP-1	Macrophage chemotactic protein-1
MCTM	Markov Chain Transition Matrix
mITT	Modified intention to treat
MPA	Mycophenolic acid
MPAG	phenolic glucuronide metabolite of MPA
MMF	Mycophenolate mofetil; same as CellCept
MMP	Matrix metalloproteinase
MOP	Manual of Procedures
mRSS	Modified Rodnan Skin Score
NIH	National Institutes of Health
NSIP	Non-specific interstitial pneumonia
OHRP	Office for Human Research Protections
PAH	Pulmonary arterial hypertension
PDGF	Platelet derived growth factor
PFD	Pirfenidone; same as Esbriet
PFT	Pulmonary function test
PI	Principal Investigator
Plac	Placebo
PML	Progressive multifocal leukoencephalopathy associated with JC virus
PP	Per protocol
PPI	Proton pump inhibitors
PRCA	Pure Red Cell Aplasia
PRO	Patient Reported Outcome
PROMIS-29	Patient-reported outcomes measurement information system 29-item health profile
PVAN	Polyomavirus associated nephropathy
QA	Quality Assurance
QC	Quality Control
QGG	Quantitative ground glass

QHC	Quantitative honeycomb change
QIA	Quantitative image analysis
QILD-LM	Quantitative interstitial lung disease score in the lobe of maximal involvement
QILD-WL	Quantitative interstitial lung disease score in the whole lung
QLF-LM	Quantitative lung fibrosis score in the lobe of maximal involvement
QLF-WL	Quantitative lung fibrosis score in the whole lung
SABER	Statistical Analysis of Biomedical and Educational Research Unit
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SGRQ	St. George's Respiratory Questionnaire
SHAQ	Scleroderma Health Assessment Questionnaire
SLS	Scleroderma Lung Study
SMC	Safety Monitoring Committee
SOC	System Organ Class
SOP	Standard Operating Procedure
SSc	Scleroderma (same as Systemic Sclerosis)
SSc-lc	Scleroderma with limited cutaneous features
SSc-dc	Scleroderma with diffuse cutaneous features
SSc-ILD	Scleroderma-related interstitial lung disease
TDI	Transitional Mahler Modified Dyspnea Index
TEAE	Treatment-emergent adverse event
TGF-β1	Transforming growth factor beta-1
TID	Three times daily
UCLA SCTC GIT 2.0	University of California, Los Angeles, Scleroderma Clinical Trials Consortium Gastrointestinal Scale
UIP	Usual interstitial pneumonia
ULN	Upper limit of normal
UP	Unanticipated Problem
US	United States
WBC	White blood cell

4 Introduction

4.1 Preface

This Statistical Analysis Plan (SAP) describes statistical methods and analyses for the SLS III (Scleroderma Lung Study III) trial. This document should be read in tandem with the SLS III Study Protocol version 2.5, dated November 17, 2021.

Scleroderma Lung Study III proposes to investigate a new combination therapy for patients with SSc-ILD that will combine the established effects of immunosuppression, as mediated by MMF, with the anti-fibrotic effects of PFD. The primary hypothesis is that the rapid onset and anti-fibrotic activity of PFD, which have been observed in the treatment of IPF, will complement the delayed anti-inflammatory and immunosuppressive effects of MMF, to produce a significantly more rapid and/or greater improvement in lung function over time than occurs in patients receiving control therapy with MMF (and Plac) alone. A secondary objective is to demonstrate that combination therapy with PFD and MMF is well tolerated, in comparison to MMF alone, and not associated with limiting toxicity that impacts on the overall treatment effect.

The SLS III study is designed as a Phase II multi-center, double-blind, parallel group, randomized and placebo-controlled clinical trial addressing the treatment of patients with active and symptomatic SSc-ILD. Participants were randomized in a 1:1 assignment to receive either oral MMF and Plac (acting as a control arm) or a combination of oral MMF and oral PFD (acting as an experimental arm), with both regimens administered for 18 months. The length of therapy, 18 months, was established based on prior studies with immunosuppressive therapy alone, which demonstrated that 18 months represents the time required to achieve a peak treatment response. The primary outcome will include physiologic measures of lung function, HRCT imaging measures of lung inflammatory and fibrotic changes, dyspnea, assessments of skin inflammation and thickening, and patient reported measures of symptoms and quality of life. Tolerability and toxicity of the two treatments will also be assessed.

4.2 Scope of the Analyses

The purpose of this document is to describe primary, secondary and exploratory statistical analyses to be conducted with regard to the SLS III clinical trial. Ancillary analyses are not covered in this document.

5 Study Objectives and Endpoints

5.1 Study Objectives

5.1.1 Primary Objective

To assess the impact of combined PFD and MMF, as compared to treatment with MMF alone (i.e., combined with placebo), on the overall course of lung function over an 18-month course of therapy.

5.1.2 Secondary Objective

To demonstrate that combination therapy with PFD and MMF is well tolerated, in comparison to MMF alone, and not associated with limiting toxicity that impacts on the overall treatment effect.

5.1.3 Exploratory Objectives

To explore the relationships among participant characteristics, study medication, and clinical outcomes, we define the following pre-specified exploratory objectives:

1. To identify baseline features that predict treatment responsiveness, disease progression and the course of lung and skin disease over time.
2. To identify biomarkers that predict disease features, treatment responsiveness, disease progression and the course of lung and skin disease over time.

5.2 Endpoints

5.2.1 Primary Endpoint

The primary endpoint is the change from baseline, measured at 3-month intervals, in the mean forced vital capacity (represented as the percentage of the age-, height-, gender- and race-adjusted predicted value, i.e. FVC-%) over the course of the 18-month double-blind treatment period.

5.2.2 Secondary Endpoints

1. The change from baseline to 18 months, measured at 3-month intervals, over the course of the 18-month treatment period in:
    a. Single-breath diffusing capacity for carbon monoxide (DLCO), calculated as a percent of the age-, height-, gender-, race- and hemoglobin-adjusted predicted value (DLCOHb-%).
    b. Modified Rodnan Skin Score (mRSS).
    c. Forced vital capacity volume (FVC, in ml)
    d. St. George Respiratory Questionnaire (SGRQ)

2. Transition Dyspnea Index at 18 months, measured at 3-month intervals, over the course of the 18-month treatment period.

3. The change from baseline to 18 months, measured at 6-month intervals, over the course of the 18-month treatment period in Health assessment questionnaire modified for scleroderma (SHAQ) includes:
    a. HAQ-DI calculated without aids/devices
    b. Scleroderma-HAQ-DI visual analogue scales (VAS) assessing burden of pain, burden of digital ulcers, Raynaud’s, GI involvement , breathing, and overall disease

4. The change from screening to month 18 in quantitative HRCT measures of SSc-ILD which specifically include:
    a. Quantitative lung fibrosis score in the whole lung (QLF-WL).
    b. Quantitative lung fibrosis score in the lobe of maximal involvement (QLF-LM).
    c. Quantitative interstitial lung disease score in the whole lung (QILD-WL).
    d. Quantitative interstitial lung disease score in the lobe of maximal involvement (QILD-LM).
    e. Total lung capacity at maximum inspiration (HRCT-TLC)

5. The time (in months) required for each treatment arm to achieve a 3.0% (absolute) or greater improvement from baseline in the FVC-% over the 18-month treatment period.

6. The proportion of participants in each treatment arm achieving greater than a 5% (absolute) improvement from baseline in FVC-% over the 18-month treatment period.

7. The proportion of participants in each treatment arm achieving absolute change of FVC-% from baseline to month 18 as:
    a. Improvements by up to 5%, from 5% to <10% and from 10% to <15% or worsening by up to 5%, from 5% to <10% and from 10% to <15%.
    b. Positive responders (improved at least 3% or more), negative responders (worsened at least 3% or more), and stable (> -3% to < 3%)
    c. Responder (> 0) and non-responders (≤ 0).

8. The proportion of participants in each treatment arm achieving absolute change of mRSS from baseline to month 18 as:
    a. Using 4 points increments: worsen (1 to 4, ≥ 5), no change (=0), improved (≤ -13, -12 to -9, -8 to -5, -4 to -1).
    b. Improved (< -5), no change (-5 to 5), and decreased (> 5).

9. The proportion of participants in each treatment arm achieving the TDI focal score at month 18 as :
    a. Improved by 1-3, 4-6 and 7-9 points, no change (0) and worsened by 1-3, 4-6 and 7-9 points.
    b. Improved (>0), no change, deterioration (<0).

10. Tolerability and toxicity of combined MMF and Plac vs MMF and PFD over the course of the 18-month treatment period.
    a. The time from start of treatment to withdrawal or removal from active drug therapy for any reason.
    b. Number of participants with treatment-emergent adverse events related to study medication as assessed by system organ classification using preferred Medical Dictionary for Regulatory Activities (MedDRA) terms.

5.2.3 Other Endpoints

1. The change from baseline to month 18, measured at 6-month intervals, in PROMIS-29 version 2.0 with the following domains: Physical Function, Anxiety, Depression, Fatigue, Sleep Disturbance, Pain Interference, Impact on Social Roles, and a single item on pain intensity.
2. The proportion of participants in each treatment from screening to month 18 with a 2% increment in QILD-WL. It is also categorized as better (<-2%), stable (≥-2%, ≤2%) and worsen (>2%).
3. The proportion of participants in each treatment arm with a 3% or greater decline from baseline in FVC-% over the 18-month treatment period.
4. The proportion of participants in each treatment arm with a 5% or greater decline from baseline in FVC-% over the 18-month treatment period.
5. The change from baseline to month 6, 12, and 18 in patient global assessment of “overall health” in the past week.
6. Patient transition questions of global assessment at month 6, 12, and 18 comparing their i) overall

health and ii) lung involvement to that at the baseline visit.

7. The change from baseline to month 6, 12, and 18 in supplemental questions in Leicester Cough Questionnaire (LCQ) regarding the cough severity, cough frequency and phlegm production.

6 Study Methods

6.1 General Study Design and Plan

This study is designed as a phase II multi-center, double-blind, parallel group, randomized and placebo-controlled clinical trial addressing the treatment of patients with active and symptomatic SSc-ILD. Eligibility for the study was assessed during a 3-month screening period. Eligible participants were randomized in a 1:1 assignment to receive either oral MMF and Plac (acting as a control arm) or a combination of oral MMF and oral PFD (acting as an experimental arm), with both regimens administered for 18 months.

MMF was administered orally twice daily with a 4-step titration at monthly intervals to a target dose of 1500 mg twice daily as tolerated. PFD/PBO was administered orally three times daily with a 3-step titration at 2 week intervals to a target dose of 801 mg three times daily as tolerated. Patients could receive ongoing treatment with a maximum tolerated dose for either drug that was less than the target dose according to rules defined by the protocol. MMF is approved for use as an immunosuppressant in solid organ transplantation (cardiac, hepatic and renal) at doses up to 1.5 g twice daily and for the treatment of Lupus nephritis, an autoimmune manifestation of systemic lupus erythematosus, in the dose range of 1 to 3 g daily. PFD is currently approved for the treatment of IPF at a recommended dose of 801 mg (3 capsules of 267 mg each) 3 times daily. They were administered in this study at doses that fall within these FDA-defined dosing ranges but as experimental therapy for SSc-ILD.

A participant was considered to have completed the study if he or she completed all phases of the study including the last visit, or the last scheduled procedure, regardless of the dose of either study drug. Participants were followed in the study even if study drug was prematurely discontinued unless they had withdrawn their consent to do so. End of study was constituted by completion of the entire potential 18 months in all randomized participants.

6.2 Inclusion-Exclusion Criteria and General Study Population

6.2.1 Inclusion Criteria

In order to be eligible for randomization as a study participant, an individual must have met all of the following inclusion criteria. Patients were provided consent prior to the screening and screening was included as part of the consent process.

<u>Screening criteria that met prior to moving forward to HRCT imaging</u>

1. Age >18 years
2. Scleroderma as determined by the 2013 ACR/EULAR classification criteria.
3. Grade ≥2 on the Magnitude of Task component of the Mahler Modified Dyspnea Index (Becomes short of breath with moderate or average tasks such as walking up a gradual hill, climbing less than three flights of stairs, or carrying a light load on the level.

4. FVC-% of ≤85% at screening. (The original cutpoint was <80%; the cutpoint was changed in protocol amendment version 1.2 [December 1, 2021] to ≤85%.)
5. Onset of the first non-Raynaud manifestation of SSc within the prior 84 months.

Screening HRCT imaging

6. Presence of any ground glass opacification (any GGO) on thoracic HRCT

Final screening criteria fulfilled at Baseline Visit, but prior to randomization

7. Repeat FVC-% at the baseline visit within 10% of the FVC-% value measured at screening. If these criteria are not met, a repeat FVC-% may be obtained within 7 days and the subject may qualify for randomization if the repeat FVC-% agrees within 10% of the FVC-% obtained at screening.

6.2.2 Exclusion Criteria

An individual who met any of the following criteria was excluded from participation in this study. The majority of exclusion criteria were assessed at screening, unless explicitly stated.

1. Disease features supporting the primary diagnosis of another connective tissue disease such as rheumatoid arthritis, systemic lupus erythematosus or mixed connective tissue disease (Features consistent with a secondary Sjogren syndrome or scleroderma-associated myopathy were allowed).
2. FVC-% <45% at either screening or baseline.
3. FEV1/FVC ratio <0.65 at either screening or baseline.
4. DLCOHb-% of <30% at screening or <25% at baseline.
5. Diagnosis of clinically significant resting pulmonary hypertension or mild pulmonary hypertension requiring treatment with more than one oral medication as ascertained prior to study evaluation or as part of a standard of care clinical assessment performed outside of the study protocol.
6. Evidence of uncontrolled congestive heart failure, unstable ischemic heart disease, history of complicated pulmonary embolism impacting on heart or lung function, or unstable cardiac arrhythmia requiring chronic anticoagulation.
7. Clinically significant abnormalities on HRCT not attributable to SSc.
8. Hematologic abnormality at screening including:
    a. Leukopenia (white blood cells [WBC] <4.0x103/µl),
    b. Thrombocytopenia (platelet count <120.0x103/µl),
    c. Clinically significant anemia [Hemoglobin (Hgb) <10.0 g/dl].

    Participants with an identified and correctable etiology may have been eligible if repeat testing within the maximal 90-day screening period met all criteria.
9. A diagnosis of chronic liver disease or abnormal baseline liver function test (LFTs) or total bilirubin that were >2.0 x upper normal limit.
10. Serum creatinine >2.0mg/dl.
11. History of recurrent aspiration, uncontrolled heartburn, or gastroesophageal reflux disease with a reflux scale score of >1.00 as determined by a UCLA Scleroderma Clinical Trial Consortium Gastrointestinal Scale (UCLA SCTC GIT), Version 2.0.

    Participants with uncontrolled heartburn or GERD that was amenable to medical management may have been eligible if repeat testing within the maximal 90-day screening period met this criteria.
12. Known achalasia, esophageal stricture or esophageal dysfunction sufficient to limit the ability to swallow medication.
13. Pregnancy (documented by serum pregnancy test) and/or breast feeding.
14. If of child bearing potential (a female participant < 55 years of age who had not been postmenopausal for ≥ 5 years or who had not had a bilataeral salpingectomy, hysterectomy and/or oophorectomy),

failure to employ two reliable means of contraception which may include surgical sterilization, barrier methods, spermicidals, intrauterine devices, and/or hormonal contraception, unless the participant chooses abstinence (to avoid heterosexual intercourse completely). If a subject chose abstinence, then a second reliable means of contraception was not needed.

15. Prior use of potential disease modifying antirheumatic drugs (DMARDs) according to the following exposure rules:
    a. Use of oral cyclophosphamide (CYC), MMF, azathioprine or other oral or short half-life DMARDs for more than 6 months in the past year, as determined at the time of the initial screening visit.
    b. Treatment with more than three intravenous doses of CYC, more than one course of Rituximab or other intravenous or injectable DMARDs in the past year.
    c. More distant history of treatment with a DMARD was allowed as long as the patient had a new diagnosis/new episode of active SSc-ILD since stopping that treatment and mets the criteria noted in 15a or 15b.
16. Use of CYC, MMF, azathioprine, Rituximab or other DMARD in the 30 days prior to the baseline visit unless the patient is on MMF and the responsible physician indicated that continued use was in the best clinical interest of the patient.
17. Active infection (lung, ulcers or elsewhere) whose management would be compromised by immunosuppression.
18. Other serious concomitant medical illness (e.g., active malignancy within the past 5 years other than surgically-removed local skin cancer such as a basal cell carcinoma), chronic debilitating illness (other than SSc), unreliability or drug abuse that might compromise the patient's participation in the trial.
19. Current use, or use within the 30 days prior to their baseline visit, of prednisone (or equivalent) in doses >10 mg/day.
20. Smoking of cigars, pipes, or cigarettes during the past 6 months.
21. Use of contraindicated medications, including medications with putative disease-modifying properties that do not meet the exposure limits described in Exclusion Criteria #15 and #16, moderate or strong inhibitors of cytochrome P450 (CYP) isozyme 1A2 (CYP1A2) (note ciprofloxacin allowed up to a dose of 500 mg twice daily), and moderate inducers of CYP1A2 (such as tobacco smoke or phenytoin).

6.3 Randomization and Blinding

This study used randomization and masking as two of the cardinal principles of clinical trials to minimize bias.

Randomization. Participants were randomized after all screening assessments were completed and the investigator verified that eligibility criteria were met. Eligible participants were randomized to MMF+PFD or MMF+Plac in a 1:1 manner, stratified by clinical site (pooled into 4 groups) and prior MMF exposure (naïve, >0 to ≤3 months, and >3 months to ≤6 months). Randomization to the different strata were to be capped if necessary to maintain at least 50% of randomized subjects within the treatment naïve strata (no prior therapy with MMF or DMARD) in order to maintain adequate power. The DCC prepared the randomization schedule, using computer-generated block randomization with the block size(s) known only by the DCC. A secure web-based application was built for use by the coordinators to enter participant information (e.g., participant ID, stratification factor) and to obtain the randomization number. The information was printed and sent and/or emailed directly to the site pharmacists. Participants who withdrew from the study prior to completion of the treatment period were not replaced.

Blinding. This is a double-blind study. The study staff (except for select staff at the DCC and Research Pharmacists) and the participant were blinded to the treatment assignment.

6.4 Study Assessments

The Schedule of Activities details the study procedures:

	Screen		**Randomized Double-blind Phase**																				**Exit visit***
Visit #	**S1**	**S2**	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**	**V11**	**V12**	**V13**	**V14**	**V15**	**V16**	**V17**	**V18**	**V19**	**V20**	**V21**	**V22**
Month on Study (Month = 28 days)	Sc-1	Sc-2	0	0.5 (14 days) ±7d	1 (28 days) ±7d	1.5 (42 days) ±7d	2 (56 days) ±10d	3 (84 days) ±10d	4 (112 days) ±10d	5 (140 days) ±10d	6 (168 days) ±10d	7 (196 days) ±10d	8 (224 days) ±10d	9 (252 days) ±10d	10 (280 days) ±10d	11 (308 days) ±10d	12 (336 days) ±14d	13 (364 days) ±10d	14 (392 days) ±10d	15 (420 days) ±14d	16 (448 days) ±10d	17 (476 days) ±10d	18 (504 days) ±14d	Month 19 (532 days) ±10d
Phone contact				X		X						X	X		X	X		X	X		X	X		X
On-Site Visit	X	X	X		X		X	X	X	X	X			X			X			X			X	
Complete H&P	X																X							
F/u SSc-H&P			X		X		X	X	X	X	X			X						X			X	
Vital signs	X		X		X		X	X	X	X	X			X			X			X			X	
mRSS			X					X			X			X			X			X			X	
Study Consent	X																							
Adverse events				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Concomitant Medications	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Dispense meds			X		X		X	X	X	X	X			X			X			X				
Drug reconciliation			X		X		X	X	X	X	X			X			X			X			X	
Mahler BDI/TDI			X					X			X			X			X			X			X	
SGRQ			X					X			X			X			X			X			X	
SHAQ, PROMIS-29, UCLA SCTC GIT, CRISS, LCQ, Global Assessments	UCLA SCTC GIT		X (except CRISS)								X						X						X	
LABS:																								
- CBC, diff, plat	X				X		X	X	X	X	X			X			X			X			X	
- Metabolic/liver	X				X		X	X	X	X	X			X			X			X			X	
- Serum Preg$^+$	X																							
- Urine Preg$^+$			X		X		X	X	X	X	X			X			X			X			X	
HRCT		X																					X	
Spiro/DLCO	X		X‡					X			X			X			X			X			X	
Biorepository			X														X						X	

‡Screen and Baseline FVC value must be within an absolute difference in percent-predicted of 10% - may repeat within 7 days if not and proceed to enroll if criteria met

$^+$For women of childbearing potential, initial serum pregnancy testing will be carried out with subsequent urine testing at each visit using test kits provided by the study.

*Exit visit will also be carried out within 30 days (±10d) of early termination/withdrawal from the protocol. All subjects who terminate/withdraw early will be encouraged to return for the outcome assessments as detailed above for the 12 month (V15) and 18 month (V21) visits.

**Refer to Protocol Section 17.0 (EMERGENCY DISASTER/ PANDEMIC MANAGEMENT PLAN) for allowed adjustments in the event of a disaster/pandemic that disrupts patient or institutional access.

6.4.1 Primary Efficacy Assessment

The primary endpoint is change from baseline in the mean forced vital capacity, measured as the percentage of the age-, height-, gender- and race-adjusted predicted value (FVC-%) over the course of the 18-month treatment period, as reported quarterly (i.e., months 3, 6, 9, 12, 15 and 18.) The change from baseline at each time point is calculated as FVC-% at x months minus FVC-% at baseline. The calculation of percent predicted FVC is based on equations from Hankinson[4] (see table below).

Result name	Formula
Age	(Date of PFT – Date of birth)/365.25 1. For screening, rescreening, baseline, and repeat baseline visits, round down to the nearest integer. This is used for eligibility determination. 2. For all efficacy analyses, use the fractional age (i.e., do not round) for all values including baseline.
FVC Predicted (Reference)	Where H = height in centimeters and Age = Age defines as above Caucasian Males≥20: $FVC(L) = 0.00018642*H^2 + 0.00064*Age – 0.000269*Age^2 – 0.1933$ Males<20: $FVC(L) = 0.00018642*H^2 – 0.20415*Age + 0.010133*Age^2 – 0.2584$ Females≥18: $FVC(L) = 0.00014815*H^2 + 0.01870*Age – 0.000382*Age^2 – 0.3560$ African-American Males≥20: $FVC(L) = 0.00016643*H^2 – 0.01821*Age – 0.1517$ Males<20: $FVC(L) = 0.00016643*H^2 – 0.15497*Age + 0.007701*Age^2 – 0.4971$ Females≥18: $FVC(L) = 0.00013606*H^2 + 0.00536*Age – 0.000265*Age^2 – 0.3039$ NOTE 1: Those subjects who indicate that they are both African American and Mexican-American or Hispanic will use the African American reference equations; those subjects who indicate that they are both Caucasian and Mexican-American or Hispanic will use the Caucasian reference equations. NOTE 2: Those subjects who indicate that they are Asian will use 0.88x the Caucasian values. NOTE 3: If Race is “Unknown or Not Reported”, then use Caucasian values.
FVC Percent Predicted	(FVC Observed / FVC Predicted) * 100

6.4.2 Secondary Efficacy Assessments

6.4.2.1 FVC

FVC is measured in milliliters at baseline and months 3, 6, 9, 12, 15 and 18.

6.4.2.2 DLCOHb-%

Single-breath diffusing capacity for carbon monoxide (DLCO), calculated as a percent of the age-, height-, gender-, and hemoglobin-adjusted predicted value (DLCOHb-%). It is assessed at baseline, and months 3, 6, 9, 12, 15, and 18. The calculation of predicted DLCO is based on equations from Neas and Schwartz[6] and the calculation of hemoglobin-adjusted predicted DLCO is based on equations from Cotes[1] (see table below).

Result name	*Formula*
Age	(Date of PFT – Date of birth)/365.25 1. For screening, rescreening, baseline, and repeat baseline visits, round down to the nearest integer. This is used for eligibility determination. 2. For all efficacy analyses, use the fractional age (i.e., do not round) for all values including baseline.
DLCO Predicted (Reference)	Where H = height in centimeters and Age = Age defined as above. Caucasians Males: DLCO (mL/min/mmHg) = 0.291*H – 0.208*Age – 7.86 Females: DLCO (mL/min/mmHg) = 0.255*H – 0.097*Age – 12.40 African-Americans Males: DLCO (mL/min/mmHg) = 0.323*H – 0.198*Age – 16.46 Females: DLCO (mL/min/mmHg) = 0.136*H – 0.108*Age + 4.77 NOTE 1: For minorities other than African-Americans, the prediction equations for Caucasians will be used. NOTE 2: If “Unknown or Not Reported”, then use Caucasian values.
DLCO, Adjustment for hemoglobin	Men: DLCO_Hgb_adjusted_reference = DLCOreference x (1.7 x Hgb)/ (10.22+ Hgb) Women: DLCO_Hgb_adjusted_reference = DLCOreference x (1.7 x Hgb)/ (9.38+ Hgb) NOTE 1: The hemoglobin will be reported from the study blood draw closest to the DLCO measurement. Ideally this will be within 4 weeks.
DLCO Percent Predicted (Adjusted for Hgb)	(DLCO Observed / DLCO_Hgb_adjusted reference) * 100

6.4.2.3 Modified Rodnan Skin Score (mRSS)

The modified Rodnan skin score (mRSS) calculated by summation of skin thickness in 17 different body sites using a 0–3 scale, where 0 = normal, 1 = mild thickness, 2 = moderate thickness and 3 = severe thickness. Total skin score can range from 0 (no thickening) to 51 (severe thickening in all 17 areas). Each body site needs to be scored otherwise total score should not be calculated.

6.4.2.4 Baseline and Transition Dyspnea Index (BDI and TDI, respectively)

The baseline (BDI) and transition (TDI) dyspnea indices assess breathlessness in domains related to functional impairment, magnitude of task and magnitude of effort. The BDI score ranges from 0 (very severe impairment) to 4 (no impairment) for each domain and are summed to determine the BDI total score (0–12). The TDI score ranges from -3 (major deterioration) to +3 (major improvement) for each domain. The sum of all domains yields the TDI total score (-9 to +9). For both BDI and TDI, each domain needs to be answered otherwise total score should not be calculated.

6.4.2.5 St. George Respiratory Questionnaire (SGRQ)

SGRQ includes three component scores: Symptoms, Activity, and Impacts and are summed to calculate the total score. Each questionnaire response has a unique empirically derived ‘weight’. Symptoms component is the summed weights for the positive responses to questions 1-8 and will tolerate a maximum of 2 missed items. Activity component is the summed weights for the positive responses to questions 11 and 15 and will tolerate a maximum of 4 missed items. Impacts component is the summed weight for the positive responses to questions 9-10, 12-14 and 16-17 and will tolerate a maximum of 6 missed items. The weight for the missed item of symptoms, activity or impacts is subtracted from the total possible weight for the corresponding component and from the total weight. The total score is calculated by summing all positive responses in the questionnaire and expressing the result as a percentage of the total weight for the questionnaire. Each component and total score range from 0 (no impairment) to 100 (maximum impairment).

6.4.2.6 Scleroderma Health assessment questionnaire (SHAQ)

The SHAQ consists of the HAQ-disability index (HAQ-DI, 8 domains and an overall score) and 6 visual analogue scales assessing the burden of pain, burden of digital ulcers, Raynaud's, GI involvement, breathing, and overall disease. There are 41 total questions in HAQ-DI: 20 are 4-point Likert-scale questions, 13 questions assessing use of aids or devices, and 8 questions assessing help received from another. Twenty specific activities are assessed on a 4-point Likert-scale where 0 = without difficulty, 1 = with some difficulty, 2 = with much difficulty and 3 = unable to do. There are 3 steps to scoring the HAQ-DI:

(1) The 20 activities are grouped into 8 functional categories with each category given a single score to the maximum value of their component activities. The score is adjusted for use of aids/help by increasing the category score from 0 or 1 to a 2 if aids/help are used for that category. If the category score is already a 2 or 3, no adjustment is made;

(2) Sum the category scores; and

(3) Divide the final sum by the number of categories answered to obtain the final HAQ-DI score rounded to the nearest value evenly divisible by 0.125.

A complementary scoring method ignores scores for aids and devices when computing the category scores and represents residual disability after compensatory efforts; this measure (without adjustment) will be used in our analyses. HAQ-DI requires a minimum of six categories answered; if less, the score is not calculated. The final HAQ-DI score ranges from 0 to 3. Higher HAQ-DI score indicates more disability.

Category	Variable name
Dressing and Grooming	SHAQ1DRS
	SHAQ2SHM
Arising	SHAQ3SND
	SHAQ4BED
Eating	SHAQ5MEA
	SHAQ6CUP
	SHAQ7MLK
Walking	SHAQ8WLK
	SHAQ9CLB
Hygiene	SHAQ10WAS
	SHAQ11TUB
	SHAQ12TOI
Reach	SHAQ13REA
	SHAQ14BND
Gripping and opening things	SHAQ15DOR
	SHAQ16JAR
	SHAQ17FAU

Errands and chores	SHAQ18SHP
	SHAQ19CAR
	SHAQ20VAC

The VAS scale line from the far left indicates "no pain" to the far right indicate "the most intense pain imaginable". The measurement is between 0 to 10 cm.

6.4.2.7 HRCT measures

Volumetric thoracic HRCT was done at screening and 18 months. We measured the extent of lung involvement by fibrosis (reticulations), ground glass opacity, and honeycombing on thoracic HRCT with a validated, computer-aided scoring method.[12] High resolution CT scores are provided for quantitative extent of lung fibrosis on high resolution for the whole lung and lobe of maximum involvement and quantitative extent of total interstitial lung disease for whole lung and lobe of maximal involvement.

6.4.3 Other Assessments

6.4.3.1 Patient-reported outcomes measurement information system 29-item (PROMIS-29 version 2.0)

The PROMIS-29 assesses each of the seven domains with four questions with an additional pain intensity numeric rating scale (NRS) that is question 29 of the score. The seven domains include Physical Function, Anxiety, Depression, Fatigue, Sleep Disturbance, Pain Interference, and Impact on Social Roles. Each question is scored from 1-5 (with the exception of the NRS which is from 0 = no pain to 10=worst pain imaginable). The sum of each domain results in the raw score, which lies between 4 and 20. There is no total score, but each axis forms its own score. PROMIS assessments use an Item Response Theory (IRT) based score called "Expected A Posteriori" or EAP scores, which are then transformed onto a final T-score metric. This means that the scores are mapped so that the values follow a normal distribution with a population mean T-score of 50 and a standard deviation of 10. It is done by uploading the data to website: HealthMeasures Scoring Service (assessmentcenter.net). Lower scores on Physical Function and Impact on Social Roles indicate better quality of life, while higher scores on other domains indicate better quality of life.

6.4.3.2 Physician and Patient global assessment for overall disease

The Physician Global Assessment employs a 0-10 Likert scale to assess the patient's "overall health" in the past 1 week. The single-item question is anchored from 0 (excellent health) to 10 (extremely poor). Using a similar approach, the Patient Global Assessment employs a 0-10 Likert scale for the patient to assess their overall health in the past 1 week. The single-item question is anchored from 0 (excellent health) to 10 (extremely poor).

6.4.3.3 Physician and Patient global assessment for overall disease follow up

In addition to 6.4.3.2, physician assesses changes in the patient's overall health and lung involvement by transition questions: Compared to the baseline visit, how would you rate your patient's i) overall health and ii) overall lung involvement. The scale ranges from 1=much better to 5=much worse. Using a similar approach, patient assesses changes in their overall health and lung involvement by transition questions: Compared to your baseline visit, how would you rate your i) overall health and ii) overall lung involvement. The scale ranges from 1=much better to 5=much worse.

6.4.3.4 Leicester Cough Questionnaire (LCQ)

Each question has a score from 1 (severe impact of cough) to 7 (no impact of cough). The total score is the average of three domains: physical, psychological, and social. Total score ranges from 3 to 21. Physical domain will tolerate a maximum of 4 missed items; psychological domain will tolerate maximum of 3 missed items; and social domain will tolerate a maximum of 2 missed items. Higher LCQ score indicates less impact of cough. Three supplement questions after the LCQ questionnaire assess cough severity (0=No cough to 3=Severe), cough frequency (1=Infrequent to 3=Persistent), and phlegm production (1=None to 3=Frequent) in past two weeks.

6.4.3.5 Dosing and Adherence

The proportion of participants reaching and sustaining the maximum dose, and the time to reach the maximum dose and the time remaining at the maximum dose in each treatment group. The time to reach the maximum dose is defined as from date of first dose to the date of 12 pills/day for MMF (1500 mg) and 9 pills/day for PFD (801 mg) and placebo. If a participant did not achieve the maximum dose, they were not included in the calculation of the time to reach maximum dose.

Adherence per protocol is defined as the number of pills taken divided by the number prescribed per the protocol for the duration of time the participant received study medication. Adherence prescribed by PI is defined as the number of pills taken divided by the number prescribed per site PI (e.g., temporarily holding study medication administration for tolerability) for the duration of time the participant received study medication.

7 Sample Size

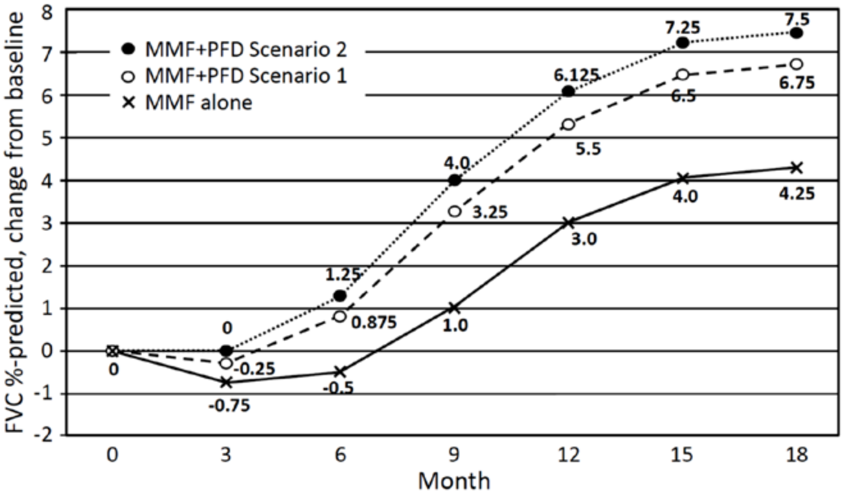
An initial sample size target of 150 participants (up to 190 consented and screened subjects to achieve 75 randomized per treatment group) was identified based on recent clinical trial experience and logistical considerations including the number of clinical centers with appropriate leadership and infrastructure to be considered as state of the art for treating SSc-ILD. Statistical analysis, using a clinical trials simulation approach, was then employed to estimate the power to detect various treatment effects on the primary endpoint with a two-sided Type I error of 5% and an 18-month attrition estimate of 24%.

Data from the SLS I and II clinical trials[7, 9-10] were used to estimate the baseline characteristics of the proposed study population and their response over time to treatment with MMF, including the impact of prior treatment with MMF (occurring prior to study randomization), on the predicted course of lung function changes over time. Note that SLS III will employ essentially the same patient recruiting centers and investigators as those that participated in the prior studies (with some new additions) and that the enrollment criteria have been relatively conserved over time to assure reproducibility. Furthermore, both the control arm and experimental arm of the proposed study will be treated with a course of cytotoxic therapy that is essentially identical to that reported from the SLS I and II studies. For the impact of PFD, safety and tolerability data from the LOTUSS trial[5] and outcome data from studies in which PFD was used to treat IPF[2,3] were considered informative.

Two outcome scenarios were considered based on the two potential mechanisms of action by which the addition of PFD might improve the response to treatment with MMF alone. Both scenarios assume that the average response pattern to MMF alone (the control arm) would approximate the average response pattern to cytotoxic therapy observed from SLS II. In that study, the overall improvement from baseline in the FVC-% was approximately 3.0% at the 12 month time point and 4.25% at 18 months. Two different response scenarios were then proposed to represent the range of possible responses in patients treated with combined MMF+PFD (the experimental arm).

- ***In Scenario 1***, it is assumed that treatment with PFD is associated with an early slowing of the decline in lung function normally observed during the first 3-6 months of cytotoxic therapy and that there is only a small additive effect of PFD to the later improvement in lung function. The collective result is a 2.5% greater response to treatment at 12 months in the experimental versus control arms and this difference between the groups remains the same for the duration of the 18-month study.
- ***In Scenario 2***, it is assumed that treatment with PFD is associated with both an early slowing of the decline in lung function and a definite additive effect on the overall improvement in FVC-% that continues throughout the entire treatment period. The collective result is a 3.125% greater response to treatment at 12 months in the experimental versus control arms which continues to slowly increase to a final 3.25% difference between the two groups at 18 months.

Figure 7.1. Hypothesized Response Trajectories over the 18-Month Double-Blind Treatment Period by Treatment Group with two scenarios considered, all participants naïve to therapy.



The power to detect a response based on each Scenario was then calculated by carrying out 1,000 clinical trials simulations for each scenario in SAS 9.4, using a linear mixed model with participant-month in the study (3, 6, 9, 12, 15 and 18) as the unit of analysis and the change from baseline in FVC-% as the outcome, with terms for treatment group, baseline FVC-%, month and the interaction of treatment group with month as a fixed covariate and participant as a random effect to account for the correlation of outcomes over time within a participant. (This is a simplification of the model used for the primary analysis of the primary endpoint described above.) We used a compound symmetry variance-covariance structure. The average power is calculated as the number of simulations where the F-test used to test the hypothesis that the mean change from baseline during the double-blind treatment period differs between the two treatment groups is rejected ($p \leq 0.05$) divided by the number of simulations. 95% Cis for power are calculated based on the exact binomial proportion.

Table 7.1. Power to detect differences in the response pattern from baseline to 18 months for the primary outcome (change in FVC-%) when comparing the course of change in lung function over time for the MMF+PFD arm and the MMF+Plac arm according to the two proposed outcome scenarios. Results assume a 5% two-sided type I error and are provided for three different sample sizes (N = 75/arm; 70/arm and 65/arm)

Assumptions:

- Drop-out rates, variances and data inter-correlations presented below are derived from SLS I and II data and assumed to be identical for the two defined scenarios
- A 5% two-sided type I error is assumed
- The difference in the outcome measure between treatment arms (relative difference in the change from baseline in absolute FVC %-predicted) for each scenario was estimated based on the hypothetical response to PFD proposed for each scenario as detailed in **Figure 7.5.4**.

Scenario 1 Difference between treatment arms (FVC-%)	**Scenario 2** Difference between treatment arms (FVC-%)	**Dropout (Cumulative %)**	**SD Estimate**	**Assumed Correlations (AR1):**
Mo 3: 0.5 Mo 6: 1.375 Mo 9: 2.25 Mo 12: 2.5 Mo 15: 2.5 Mo 18: 2.5	Mo 3: 0.75 Mo 6: 1.75 Mo 9: 3.0 Mo 12: 3.125 Mo 15: 3.25 Mo 18: 3.25	Mo 3: 6.5% Mo 6: 13% Mo 9: 17% Mo 12: 20% Mo 15: 23% Mo 18: 24%	Mo 3: 4.1 Mo 6: 5.1 Mo 9: 6.2 Mo 12: 6.9 Mo 15: 6.2 Mo 18: 6.2	3-mo intervals: 0.866 6-mo intervals: 0.834 9-mo intervals: 0.802 12-mo intervals: 0.790 15-mo intervals: 0.780

Power Calculations Scenario 1:

Evaluation of Statistical Power by sample size, adjusting for drop outs over time, based on 1000 clinical trial outcome simulations:

N per arm at baseline	**N per arm at 18 mo**	**Estimated Statistical Power**	**95% Lower Bound of Statistical Power**	**95% Upper Bound of Statistical Power**
75	57	82.3%	79.8%	84.6%
70	53	78.6%	75.9%	81.1%
65	49	76.3%	73.5%	78.9%

Power Calculations Scenario 2:

Evaluation of Statistical Power by sample size, adjusting for drop outs over time, based on 1000 clinical trial outcome simulations:

N per arm at baseline	**N per arm at 18 mo**	**Estimated Statistical Power**	**95% Lower Bound of Statistical Power**	**95% Upper Bound of Statistical Power**
75	57	95.6%	94.1%	96.8%
70	53	92.6%	90.8%	94.2%
65	49	91.8%	89.9%	93.4%

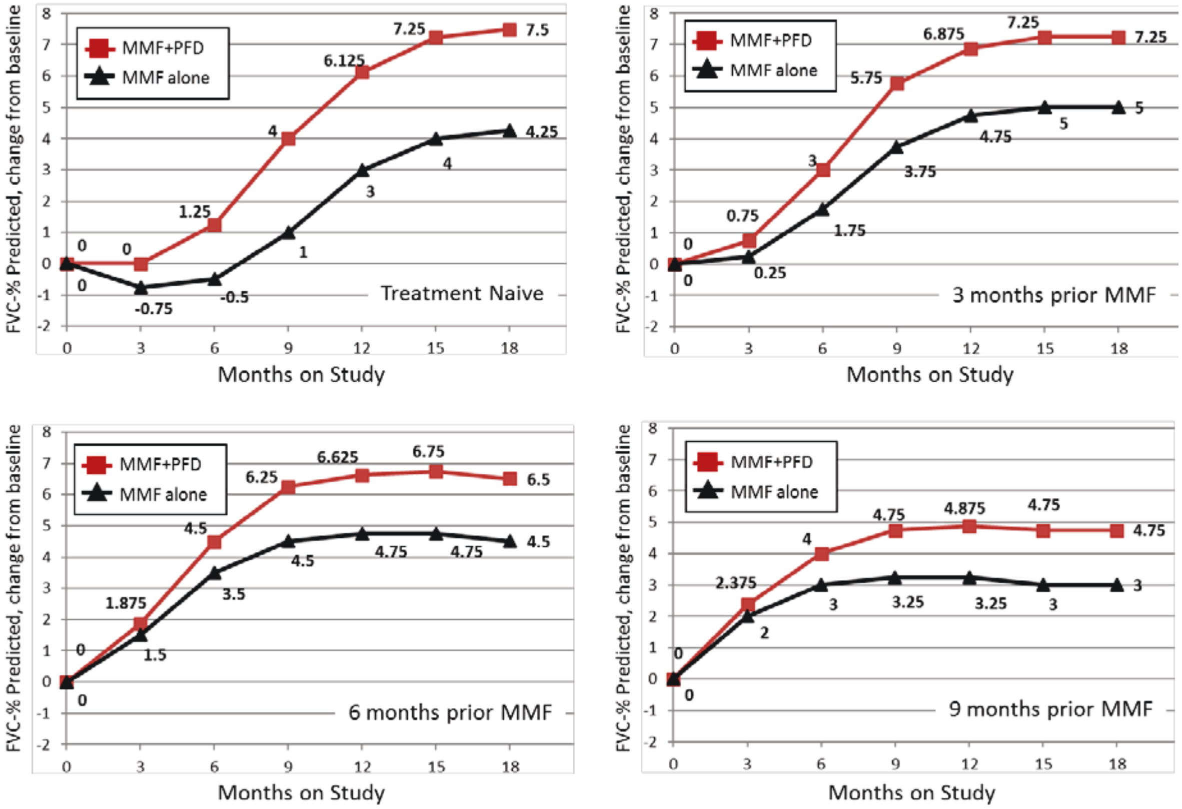
Note that with the given m-ITT design, the linear mixed effects model will actually include data from all subjects who have at least a baseline and one follow-up measure of FVC-%. In this setting, even with an overall 24% attrition by the end of 18 months, it is estimated that 83% of the maximal possible study outcome data points (assuming no attrition) would be included in the analysis.

According to Scenario 1, which is strictly based on the documented capacity for PFD to slow the rate of decline of lung function in patients with IPF who are not receiving any other treatment, a sample size of 75 patients per treatment arm is required to obtain a minimum power of 80% (i.e., 80% is within the 95% CI for this prediction). If PFD has novel effects when combined with MMF so that it both reduces the rate of lung decline and works in concert with MMF to improve overall lung function (i.e., Scenario 2), then the predicted outcome will be much more reliably detected and a smaller sample size would be sufficient.

Having established adequate estimates of study power in a population of treatment naïve patients, the clinical trial Scenarios were then adjusted to assess the impact of a mixed starting population of patients in which some of the participants are treatment naïve and others have been on prior treatment with MMF (or DMARD with similar activity) for a period of 3, 6 or 9 months. Additional assumptions associated with these modified Scenarios include:

1) The mean change in FVC-% over a given 3-month interval and the corresponding impact on the standard deviation of the FVC-% measurement will always reflect the total time that a patient has been on therapy, including therapy that was administered before randomization and therapy that was administered following randomization.
2) The treatment effect is measured as the change from baseline, in which baseline is determined at the time of randomization (baseline visit). The baseline FVC-% measurement therefore establishes the zero reference value for evaluating change over time in response to the time that a patient is on the protocol. In this setting, patients who have been on 3 months or more of MMF will not be expected to experience the initial decline in lung function that occurs when enrolling treatment naïve patients.
3) As one of the hypothesized benefits of starting PFD and MMF at the same time in treatment naïve patients is the prevention of an early decline in lung function, there is a relative treatment penalty in the predicted PFD+MMF arm when patients have been on prior therapy with MMF. This results in less and less separation between the course of the two treatment arms as the length of time on prior MMF increases. This impact must be accounted for when modeling treatment outcomes and predicting study power.

Figure 7.2. Hypothesized Response Trajectories over the 18-Month Double-Blind Treatment Period by Treatment Group for Scenario 2 based on length of time on MMF therapy prior to randomization.



The power to detect a response according to Scenario 2, when taking into account the length of prior treatment with MMF (naïve, 3 mo, 6 mo and 9 mo strata) and the percentage of subjects within each of these strata, was then calculated by carrying out 1,000 clinical trials simulations for each situation in SAS 9.4 as already described. We varied the proportion of patients from 100% naïve down to 50% naïve, with the distribution across the prior treatment with MMF strata varied as described below. Estimates of the differences between treatment arms and the standard deviation for the measurements at a given time were adjusted according to the scenario curves shown in Figure 7.2.

Table 7.2. Power to detect differences in the response pattern from baseline to 18 months for the primary outcome (change in FVC-%) when comparing the course of change in lung function over time for the MMF+PFD arm and the MMF+Plac arm according to Scenario 2 with different mixtures of treatment naïve patients and those who were on prior MMF therapy for 3, 6 or 9 months as indicated. Results assume a 5% two-sided type I error and are provided for three different sample sizes (N = 75/arm; 70/arm and 65/arm)

a. **Scenario 2: N = 75 per arm**

% per stratum (length of prior MMF)				**Estimated statistical power based on 1000 clinical trial outcome simulations**		
Naive	**3 mo**	**6 mo**	**9 mo**	**Power**	**95% Lower Bound**	**95% Upper Bound**
1.00	0.00			**95.6%**	94.1%	96.8%

0.80	0.20			**93.8%**	92.1%	95.2%
0.75	0.25			**92.2%**	90.4%	93.8%
0.70	0.30			**90.8%**	88.8%	92.5%
0.60	0.40			**90.5%**	88.5%	92.2%
0.50	0.50			**86.1%**	83.8%	88.2%
0.80	0.10	0.10		**91.4%**	89.5%	93.1%
0.70	0.20	0.10		**89.5%**	87.4%	91.3%
0.70	0.10	0.20		**90.2%**	88.2%	92.0%
0.60	0.30	0.10		**89.0%**	86.9%	90.9%
0.60	0.20	0.20		**88.2%**	86.0%	90.1%
0.60	0.10	0.30		**85.2%**	82.8%	87.3%
0.50	0.40	0.10		**84.7%**	82.3%	86.9%
0.50	0.30	0.20		**82.7%**	80.2%	85.0%
0.50	0.20	0.30		**82.7%**	80.2%	85.0%
0.50	0.10	0.40		**79.7%**	77.1%	82.2%
0.70	0.10	0.10	0.10	**86.4%**	84.1%	88.5%
0.60	0.20	0.10	0.10	**86.2%**	83.9%	88.3%
0.60	0.10	0.20	0.10	**83.7%**	81.3%	85.9%
0.50	0.30	0.10	0.10	**79.9%**	77.3%	82.3%
0.50	0.20	0.20	0.10	**79.3%**	76.7%	81.8%

b. **Scenario 2: N = 70 per arm**

% per stratum (length of prior MMF)				**Estimated statistical power based on 1000 clinical trial outcome simulations**		
Naive	**3 mo**	**6 mo**	**9 mo**	**Power**	**95% Lower Bound**	**95% Upper Bound**
1.00	0.00			**91.8%**	89.9%	93.4%
0.80	0.20			**91.3%**	89.4%	93.0%
0.70	0.30			**88.1%**	85.9%	90.0%
0.60	0.40			**86.0%**	83.7%	88.1%
0.50	0.50			**84.6%**	82.2%	86.8%
0.80	0.10	0.10		**89.4%**	87.3%	91.2%
0.70	0.20	0.10		**88.7%**	86.6%	90.6%
0.70	0.10	0.20		**85.6%**	83.3%	87.7%
0.60	0.30	0.10		**85.7%**	83.4%	87.8%
0.60	0.20	0.20		**82.9%**	80.4%	85.2%
0.60	0.10	0.30		**81.5%**	79.0%	83.9%
0.50	0.40	0.10		**83.0%**	80.5%	85.3%
0.50	0.30	0.20		**81.5%**	79.0%	83.9%

0.50	0.20	0.30		**79.9%**	77.3%	82.3%
0.50	0.10	0.40		**79.9%**	77.3%	82.3%

c. **Scenario 2: N = 65 per arm**

% per stratum (length of prior MMF)				**Estimated statistical power based on 1000 clinical trial outcome simulations**		
Naive	**3 mo**	**6 mo**	**9 mo**	**Power**	**95% Lower Bound**	**95% Upper Bound**
1.00	0.00			**91.8%**	89.9%	93.4%
0.80	0.20			**88.2%**	86.0%	90.1%
0.70	0.30			**85.5%**	83.2%	87.6%
0.60	0.40			**81.4%**	78.9%	83.8%
0.50	0.50			**79.7%**	77.1%	82.2%
0.80	0.10	0.10		**87.1%**	84.9%	89.1%
0.70	0.20	0.10		**85.5%**	83.2%	87.6%
0.70	0.10	0.20		**83.2%**	80.7%	85.5%
0.60	0.30	0.10		**78.4%**	75.7%	80.9%
0.60	0.20	0.20		**77.5%**	74.8%	80.0%

According to these predictions for Clinical Trial Scenario 2, representing the primary outcome to be investigated, the study is adequately powered (80% power or greater) to detect a difference between the two treatment arms when the sample size is 150 and the patient population contains at least 50% treatment naïve patients. If the remaining 50% of enrolled patients are limited to no more than 6 months of prior therapy with MMF, then adequate power is maintained across the entire range of anticipated distributions between subjects with up to 3 mo of prior therapy and those with between 3 to 6 mo of prior therapy.

As a result, the target study population will consist of at least 50% of patients who are treatment naïve and up to 50% of patients who recently started on therapy within 6 months of entering the study (as detailed in Table 7.2 above). With this mixture of patients, the power remains adequate even if only 70 patients are enrolled in each arm (total randomization 140 patients), providing optimal flexibility for the enrollment phase of the study.

When a similar statistical approach is applied to Clinical Trial Scenario 1, in which only a small additive effect of PFD is modeled in addition the underlying treatment response to MMF, the power to detect a difference between the two treatment arms is adequate (80% power or greater) when the sample size is 150 and the patient population contains at least 75% treatment naïve patients. When the percentage of treatment naïve patients is reduced to 50%, the predicted power is at best 75% (assumes % per stratum of 50% naïve, 40% 3 mo, 10% 6 mo).

8 General Analysis Considerations

8.1 Timing of Analyses

The final analyses of the clinical trial will be performed after

1) All enrolled participants either (a) completed 18 months of follow-up from enrollment or (b) discontinued early from the study.
2) Any chart reviews or investigations needed to confirm dates of elements of the primary composite endpoint were identified and completed.
3) All study data were entered, locked and unblinded as per SABER SOPs.

This SAP document was finalized and approved prior to the double-blind database lock and unblinding.

8.2 Analysis Populations

8.2.1 Modified Intention to Treat Population

The main population for efficacy is the modified intention-to-treat population (m-ITT), defined as all participants randomized, receiving at least one dose of study medication, and having at least one post-baseline efficacy assessment. Subjects were analyzed by assigned treatment.

8.2.2 Per Protocol Population

The Per Protocol (PP) analysis set consists of all participants in the m-ITT population who completed 18 months of study medication (i.e., both MMF and PFD/Plac) and do not have a major protocol violation, defined as violation of entry criteria or taking prohibited medications. The study window for month 18 visit is ± 6 weeks. Membership in the PP analysis population was determined before study unblinding.

8.2.3 Completers Population

The Completers analysis set consists of all subjects in the m-ITT population who had an 18-month efficacy outcome (regardless of completion of study medication). It can be different depending on the outcome that is analyzed.

8.2.4 Safety Population

The Safety Population is defined as all participants who were randomized and received at least one dose of the study medication. The Safety Population was used for all safety analyses. Subjects were analyzed by assigned treatment.

8.3 Covariates and Subgroups

Prior MMF therapy (stratification factor: naïve, >0 to $\leq$3 months, and >3 months to 6 months) and FVC-% at baseline were included as covariates in the analysis of the primary efficacy endpoint. Several other covariates are incorporated in specific analyses that are denoted in the analysis specifications described below (e.g., secondary analysis of primary endpoint including baseline mRSS as a covariate). No subgroup analyses were planned.

8.4 Missing Data

Our primary analytic approach assumes a missing-at-random mechanism[8]. Sensitivity analyses are performed to assess how subjects who withdrew might affect conclusions of the analysis for the primary endpoint.

8.5 Interim Analyses and Data Monitoring

No formal interim analyses were planned nor carried out for this study. The study was overseen by a Data and Safety Monitoring Board (DSMB) that reviewed the pooled and by-treatment subject disposition, study conduct and safety data approximately every 6 months.

8.6 Multiple Testing

Statistical testing is conducted at the 0.05 significance level using two-tailed tests; two-sided p-values are reported. No adjustments for multiplicity were planned since there is one primary analysis of the primary endpoint at the end of the study. Secondary analyses of the primary endpoint and analyses of the secondary endpoints serve to assess the robustness of the results and consistency of treatment effect on clinical outcomes important in this disease.

9 Summary of Study Conduct Data

Descriptive summary statistics are tabulated for baseline patient demographics and clinical characteristics, separately by treatment group and overall. Intervention groups are characterized as “MMF+PFD” and “MMF+Plac”. For pooled summaries, “Overall” is used as the column heading. All tables are annotated with the total population size relevant to that table, including any missing observations.

Continuous variables are summarized using descriptive statistics including n, mean, median, standard deviation, range (e.g., minimum and maximum) and using boxplots to examine the distributions of baseline characteristics by treatment group. Qualitative variables are summarized using counts and percentages. Unless otherwise specified, statistical analyses are performed using SAS Version 9 or higher.

9.1 Subject Disposition

Participant disposition is summarized descriptively. The number and percentage of participants randomized, completed, and withdrawing, along with reasons for withdrawal, are tabulated and summarized in a CONSORT diagram overall, and by treatment group. The number of participants in each analysis population are reported. Other disposition and study conduct information, including major protocol violations are summarized. Duration of the study follow-up is summarized overall and by treatment group.

9.2 Protocol Deviation

A protocol deviation is any noncompliance with the clinical trial protocol, GCP, or MOP requirements. The noncompliance may be either on the part of the participant, the investigator, or the study site staff. Protocol deviations were classified as major or not major by the study team; major protocol deviations are incorporated into the definition of the Per Protocol Analysis Set. Protocol deviations are summarized overall and by treatment group. Major protocol deviations are provided in a listing by treatment group.

9.3 Demographic and Baseline Variables

Demographic and baseline variables for participants include: age in years at consent, dichotomized age (18-35 years, >35-55 years, >55-75 years, >75 years), sex (Male/Female), ethnicity (Hispanic or Latino/not Hispanic or Latino), race (White/Black or African American/Asian/American Indian or Alaskan native/native Hawaiian or other Pacific islander), % predicted FVC, FVC (ml), % predicted DLCO (corrected for hemoglobin), time from first non-Raynaud manifestation of SSc in months at screening, modified Rodnan Skin Score (mRSS), Scleroderma classification, Mahler Baseline Dyspnea Index (BDI), St. George's Respiratory Questionnaire (SGRQ) – total and each domain score, Scleroderma Scleroderma Health Assessment Questionnaire (SHAQ), including HAD-DI without Aids/Devices and SHAQ-DI VAS domains, Leicester Cough Questionnaire (LCQ) – total and three supplemental questions, PROMIS-29 domains, UCLA SCTC GIT, physician global assessment, patient global assessment, height, weight, HRCT at screening – quantitative lung fibrosis score in the whole lung (QLF-LM), quantitative lung fibrosis score in the lobe of maximal involvement (QLF-WL), quantitative interstitial lung disease score in the whole lung (QILD-LM), quantitative interstitial lung disease score in the lobe of maximal involvement (QILD-WL), and total lung capacity at maximum inspiration (HRCT-TLC), and laboratory data at screening – sodium, potassium, chloride, CO2, BUN, creatinine, glucose, protein, albumin, bilirubin, alkaline phosphate, AST, ALT, cholesterol, WBC, hemoglobin, hematocrit, platelets, neutrophil percent, absolute neutrophil count, lymphocyte percent, absolute lymphocyte count.

9.4 Treatment Compliance

Compliance with study medication (MMF and PFD/Plac) is assessed and summarized, including

- the proportion of participants who adhered to study treatment, overall and at each study visit,
- the median duration of adherence to study treatment (defined in section 11.1),
- the proportion of participants reaching maximum dose,
- the proportion of participants who sustain their maximum dose,
- the time to reach the maximum dose, and
- the time remaining at the maximum dose.

These are calculated and summarized overall and by treatment group. The maximum dose is 12 pills/day for MMF (1500 mg) and 9 pills/day for PFD (801 mg)/Plac. The proportion of the maximum dose that the participants reached during study period is summarized overall and by treatment group. Participants were expected to take study medication unless it was permanently discontinued due to an AE. The study medication log (Form 027), adverse event form (Form 044), serious adverse event form (Form 045) and final status form (Form 035) are used to derive the compliance measure.

The proportion of participants who permanently discontinued study medication will be summarized with descriptive statistics, overall and by treatment group. A Kaplan-Meier curve will present the time to permanent discontinuation of study medication by treatment group. A listing of these participants with their reason for permanent discontinuation of study medication will be provided.

10 Efficacy Analyses

Continuous variables will be summarized using descriptive statistics including n, mean, median, standard deviation, and range (e.g., minimum and maximum). Qualitative variables will be summarized using counts and percentages. Summaries will be provided by treatment groups and overall.

10.1 Primary Efficacy Analysis

The primary endpoint is change from baseline in the mean forced vital capacity, measured as the percentage of the age-, height-, gender- and race-adjusted predicted value (FVC-%) over the course of the 18-month treatment period, as reported quarterly (i.e., months 3, 6, 9, 12, 15 and 18). The primary efficacy analysis tests the null hypothesis that the differences between the MMF+PFD and MMF+Plac is zero. The primary analysis of the primary efficacy endpoint uses the m-ITT analysis set. The endpoint is analyzed using a linear mixed model with participant-month in the study (3, 6, 9, 12, 15, 18) as the unit of analysis and the change from baseline in FVC-% as the outcome, with terms for baseline FVC-%, treatment group, month, the interaction of month and treatment group, and prior MMF therapy (stratification factor: naïve, >0 to ≤3 months, and >3 months to 6 months), as fixed covariates. Study participant is treated as a random effect to account for the correlation of outcomes over time within a participant.

The model is summarized below.

$$\text{FVC-\%}_{it} = \mu_0 + \mu_1 \text{TRT}_i + \mu_2 \text{ MONTH}_{it} + \mu_3 \text{TRT}_i \times \text{MONTH}_{it} + \mathbf{x}_\mathbf{i}' \, \boldsymbol{\beta} + \epsilon_{it},$$

where FVC-%$_{it}$ is the change of FVC-% from baseline at time t for patient i (that is, FVC-% at time *t* - FVC-% at baseline), $i = 1, \dots, 51$; $t = 1, \dots, 6$ (corresponding to 3, 6, 9, 12, 15 and 18 months). TRT_i is treatment group (equal to 1 for the PFD and MMF and 0 for Plac and MMF) for patient i. MONTH$_{it}$ is the study month at time *t* for patient *i*. Two dummy variables for prior MMF therapy (stratification factor: naïve, >0 to ≤3 months, and >3 months to 6 months), and baseline FVC-% for subject i are included in a vector $\mathbf{x}_\mathrm{i}$ of 3 covariates (x_1, x_2, x_3) and associated fixed effects are stored in vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. We assume that residual errors ϵ_{it} $(\mathrm{t} = 1, \dots, 6)$ for subject i are normally distributed with zero mean and 6x6 general/unstructured variance-covariance matrix.

In the context of the primary analysis of the primary endpoint, we are interested in parameter μ_3, representing the trajectories of the primary endpoint adjusted for stratifying variables and baseline covariates.

The model generates adjusted estimates of change from baseline in the FVC-% for each treatment group and month, and an F-test is used to test the hypothesis that the mean change from baseline during the treatment period differs between the two treatment groups.

The model-based adjusted mean change from baseline in FVC-% will be presented graphically by treatment group by study month, with 95% confidence intervals (CIs) provided at each month.

Analysis	Primary Analysis of the Primary Endpoint: mean FVC-% change from baseline during the treatment period differs between the two treatment groups
Analysis Set	mITT Population
Methods	Linear mixed model for repeated measures with correlated errors
Dependent Variable	Change of FVC-% measured at 3, 6, 9, 12, 15 and 18 months (FVC-% at x months – FVC-% at baseline, where x=3, 6, 9, 12, 15 and 18)
Model	Fixed effects: • Treatment • Study month • Interaction of treatment and study month

	• Stratifying variable: prior MMF therapy with 3 categories • Baseline FVC-%
Results	• Estimates of mean change from baseline in FVC-% during the treatment period (slope) by treatment group and treatment difference • 95% confidence interval for treatment estimates and treatment difference • P-value for treatment difference

10.2 Secondary Efficacy Analyses

10.2.1 Secondary Analyses of Primary Efficacy Endpoint

Several sensitivity analyses are performed to assess how alternative approaches to missing data and model assumptions affect the conclusions of the analysis of the primary outcome:

1. Analysis of the primary efficacy variable as described in section 10.1 is performed using the PP population.

2. Analysis of the primary efficacy variable described in section 10.1 with the m-ITT analysis set except that HRCT fibrosis scores in the lobe of maximal involvement (QLF-LM) at screening is added as a covariate.

3. Analysis of the primary efficacy variable described in section 10.1 with the m-ITT analysis set except that baseline mRSS is added as a covariate.

4. Analysis of covariance (ANCOVA) is used to analyze the change of FVC-% from baseline to month 18 (that is, FVC-% at month 18 - FVC-% at baseline). The model includes the treatment group, baseline FVC-% and prior MMF therapy.

5. In addition to the assessment of the overall trajectory of pulmonary response which integrates the timing of the onset of response and the overall magnitude of effect at 18 months, we evaluate the individual components of this response. This is accomplished by using the linear mixed effects model with same covariates described for the analysis of primary efficacy variable and provide adjusted estimates of change from baseline in FVC-% for each treatment group and treatment difference for each time point (e.g., 3, 6, 9, 12, 15 and 18 months).

6. Descriptive statistics are provided to assess more granular changes in FVC-% by treatment adherence (e.g., premature discontinuation of study treatment prior to 3, 6, 9, 12, 15 and 18 months) in the completers population.

10.2.2 Secondary Efficacy Analyses

The m-ITT analysis set will be used in the analyses of secondary endpoints if not otherwise indicated.

1. The linear mixed model as described for the primary analysis of the primary endpoint is performed to compare the two treatment groups for the change from baseline to 18 months, measured by

3-month intervals, in the following outcomes. The model is adjusted by treatment, study month, interaction of treatment and study month, prior MMF therapy and baseline outcome. For FVC (measured in ml), the model also includes age (measured as a real number and not an integer), height, sex and race (i.e., the same variables used in the percent predicted equations). The model generates adjusted estimates of the change from baseline for each treatment group and month, and an F-test is used to test the hypothesis that the mean change from baseline during the treatment period differs between the two treatment groups. The model-based adjusted mean change from baseline will be presented graphically by treatment group and study month, with 95% confidence intervals (CIs) provided at each month.

   a. Hemoglobin-adjusted DLCOHb-%
   b. mRSS
   c. FVC (in ml)
   d. SGRQ: Total score, Symptoms, Activity, Impacts

2. The linear mixed model as described for the primary analysis of the primary endpoint is performed to compare the two treatment groups for the change from baseline to 18 months, measured at 3-month intervals, in dyspnea. The model is adjusted by treatment, study month, interaction of treatment and study month, prior MMF therapy, and BDI. The model generates adjusted estimates of TDI for each treatment group and month, and an F-test is used to test the hypothesis that the mean change from baseline during the treatment period differs between the two treatment groups. The model-based adjusted mean TDI will be presented graphically by treatment group and study month, with 95% confidence intervals (CIs) provided at each month.

3. The linear mixed model as described for the primary analysis of the primary endpoint is performed to compare the two treatment groups for the change from baseline to 18 months, measured at 6-month interval, in SHAQ. The model is adjusted by treatment, study month, interaction of treatment and study month, prior MMF therapy and baseline SHAQ. The model generates adjusted estimates of the change from baseline in SHAQ for each treatment group and month, and an F-test is used to test the hypothesis that the mean change from baseline during the treatment period differs between the two treatment groups. The model-based adjusted mean change from baseline in SHAQ will be presented graphically by treatment group and study month, with 95% confidence intervals (CIs) provided at each month. SHAQ includes:
   a. HAQ-DI without Aids/Devices: Total score
   b. SHAQ-DI VAS assessing burden of pain, burden of digital ulcers, Raynaud's, GI involvement, breathing, and overall disease

4. For each HRCT outcomes – QLF-LM, QLF-WL, QILD-LM, QILD-WL, and HRCT-TLC, analysis of covariance (ANCOVA) is used to analyze the change of the HRCT outcome from screening to month 18. The model includes the treatment group, HRCT outcome at screening, and prior MMF therapy.

5. For time (in months) required to achieve a 3.0% (absolute) or greater improvement from baseline in the FVC-% over the 18-month treatment period, Kaplan-Meier methods are used to graphically present treatment group differences. The median time to event, with 95% CIs, will be presented by treatment group, and a stratified (by prior MMF therapy) log-rank test p-value are presented. The time (in months) is defined as the time from baseline until the time of the first change of FVC-%

from baseline is greater or equal to 3%. Participants are censored at date of FVC-% collected at month 18 visit or at the last date of FVC-% collected if the participants permanently discontinued study.

6. Logistic regression is performed to evaluate the proportion of participants who achieve greater than a 5% improvement from baseline in the FVC-% over the 18-month treatment period. The model includes treatment group, baseline FVC-%, and prior MMF therapy as covariates. A participant with FVC-% at 18 months minus FVC-% at baseline that is greater than 5% defines a 5% improvement or greater. . The odds ratio with corresponding 95% CI is presented; the p-value for the test of treatment group differences is also presented.

7. Box plots are used to summarize the frequency distributions of changes from baseline to 18 months by treatment group for key secondary endpoints: Hemoglobin-adjusted DLCOHb-%, mRSS (for all patients and for diffuse cutaneous scleroderma [dcSSc] patients), SGRQ (total score, Symptoms, Activity and Impacts subscores), HAQ-DI without aids/devices score, and Scleroderma-HAQ-DI visual analogue scales (VAS) scores. A box plot is also used to summarize the frequency distribution of TDI during the study period.

8. A frequency table and bar chart are used to show the frequency distribution of the change of FVC-% from baseline to month 18 with a 5% increment (e.g., improved by up to 5%, from 5% to <10% and from 10% to <15% or worsened by up to 5%, from 5% to <10%, and from 10% to <15%), overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month FVC-%.

9. A frequency table is used to show the frequency distribution of the change of FVC-% from baseline to month 18 in negative responders (≤ -3%) and stable (> -3% to < 3%) overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month FVC-%.

10. A frequency table is used to show the frequency distribution of the change of FVC-% from baseline to month 18 in responder (>0) and non-responder (≤ 0) overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month FVC-%.

11. A frequency table and bar chart are used to show the frequency distribution of the change of mRSS from baseline to month 18 categorized in 4 points increments: worsen (1 to 4, ≥ 5), no change (=0), and improved (≤ -13, -12 to -9, -8 to -5, -4 to -1) overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month mRSS. A frequency distribution of the change of mRSS is also shown in patients with dcSSc in the completers population. The proportions of participants with improved (< -5), no change (-5 to 5), and decreased (> 5) in the change from baseline in mRSS, overall and in each treatment group, are also presented.

12. A frequency table and bar chart are used to show the frequency distribution of the TDI at month 18 categorized as improved by 1-3, 4-6 and 7-9 points, no change (0), and worsened by 1-3, 4-6 and 7-9 points, overall and in each treatment group in the completers population, and separately for those who completed the entire treatment period and those who prematurely discontinued treatment but returned for the 18-month TDI. The proportions of participants with improved (>0), no change, and deterioration (<0), overall and in each treatment group, are also presented.

11 Safety Analyses

Safety data, including AEs, clinical laboratory tests, vital signs, physical examinations, and concomitant medication usage will be summarized descriptively overall and by treatment group for the Safety Population. AEs include frequency of events and/or incidence of study subjects experiencing events, by study arm for the following:

a) Study-defined treatment failures
b) Study-defined treatment emergent AEs of special Interest
c) All treatment-emergent AEs, total and by organ system classification
d) All treatment-emergent AEs related to study medication, total and by organ system classification
e) All treatment-emergent SAEs, total and by organ system classification
f) All treatment-emergent SAEs related to study medication, total and by organ system classification
g) All treatment-emergent AEs and SAEs that result in discontinuation of study medication
h) Time to treatment failure
i) Time to treatment-emergent AEs and SAEs resulting in premature discontinuation of study medication

For categorical safety outcomes, numbers and percentages will be used. For continuous safety outcomes, number, mean, standard deviation, median, interquartile range, minimum and maximum are used to summarize changes from baseline to each study visit in laboratory tests and vital signs. Percent change from baseline for laboratory values are summarized. The primary organ system associated with each adverse event was coded by the medical monitor. Kaplan-Meier methods are used to summarize graphically the time to treatment failure, and treatment-emergent AEs and SAEs that result in premature discontinuation of study medication.

11.1 Extent of Exposure

Total participant months of exposure to study medication is defined as the number of days from date of last dose of study medication minus the date of the first dose of study medication plus one. It is converted to months by divided the total number of days by 28. It is summarized descriptively by treatment group. Both full and partial amount of drug received is counted in the assessment of exposure.

11.2 Definition of Adverse Events

Adverse event means any untoward medical occurrence, such as an abnormal laboratory finding, physical sign, symptom, or diagnosis of a disease state, that is temporally associated with the use of an interventional treatment or procedure in a human subject regardless of whether or not it is considered intervention-related [consistent with 21 CFR 312.32 (a)].

11.2.1 Treatment-emergent AEs (TEAEs)

TEAEs are those that start or worsen after the start of study treatment and up to 7 days (for AEs) and 30 days (for serious AEs) after the last dose of study treatment. This AE definition would include the following:

- Any pre-existing condition that increases in severity or changes in nature during or as a consequence of the study treatment administration
- Complications resulting from protocol-mandated procedures
- AEs occurring as a result of product withdrawal, abuse or overdose
- A change in a laboratory variable if considered by the investigator to be clinically significant or if it is caused (or should have caused) the investigator to reduce or discontinue the use of the product or initiate a non-protocol therapy or procedure

11.2.2 AEs of Special Interest

AEs of Special Interest represent a specific subset of all AEs and are defined by the study protocol based on two distinct features: 1) their likely association with the investigational drug(s), and 2) the presence of pre-specified study drug management guidelines that are to be followed when the AE occurs.

AEs meeting the following criteria will be classified and reported as AEs of Special Interest in addition to their inclusion in the overall frequency and incidence of AEs.

- Leukopenia, defined as WBC $\leq 2.5 \times 10^3$/µl of blood
- Neutropenia, defined as an absolute neutrophil count < 1.0×10^3/µl of blood
- Clinically significant anemia, defined as blood hemoglobin < 10.0 gm/dl or a drop in hemoglobin to < 9.0 gm/dl if the baseline hemoglobin was < 11.0 gm/dl
- Liver enzyme elevations that exceed 3× the upper limit of normal (ULN) ALT and/or AST.
- New onset or worsening of existing gastrointestinal symptoms that do not respond to medical management and are of sufficient clinical importance to warrant drug dose adjustment (including nausea, diarrhea, vomiting, and dyspepsia).
- Documentation of gastrointestinal ulcer, gastrointestinal bleeding or abdominal emergency
- Infection requiring hospitalization, intravenous antibiotics, or judged as requiring the withdrawal of immune suppression for effective treatment.
- New or reactivated viral infections including PVAN, JC virus associated PML, CMV infections, reactivation of HBV or HCV.
- Photosensitivity skin reactions that do not respond to sunscreen, avoidance, and as-needed use of over the counter topical creams, and are of sufficient clinical importance to warrant drug dose adjustment.
- Development of a proven malignancy other than basal cell cancer of the skin or cervical carcinoma in situ removed entirely by biopsy.
- Angioedema

11.3 Definition of Serious Adverse Events (SAEs)

An AE or suspected adverse reaction is considered "serious" if, in the view of either the investigator or sponsor, it results in any of the following outcomes:

- Death,

- A life-threatening AE
- Inpatient hospitalization or prolongation of existing hospitalization
- A persistent or significant incapacity or substantial disruption of the ability to conduct normal life functions
- A congenital anomaly/birth defect.

Important medical events that may not result in death, be life-threatening, or require hospitalization may be considered serious when, based upon appropriate medical judgment, they may jeopardize the patient or subject and may require medical or surgical intervention to prevent one of the outcomes listed in this definition. Examples of such medical events include allergic bronchospasm requiring intensive treatment in an emergency room or at home, blood dyscrasias or convulsions that do not result in inpatient hospitalization, or the development of drug dependency or drug abuse. [Consistent with 21 CFR 312.32 (a)]

11.4 Definition of Unanticipated Problems (UP – unexpected AEs & Problems)

An AE or suspected adverse reaction is considered "unanticipated" if it is not listed or characterized in the Package Insert or in the current Investigator Brochure or is not listed at the specificity or severity that has been observed; or, if an investigator brochure is not required or available, is not consistent with the risk information described in the general investigational plan or elsewhere in the current application, as amended. For example, under this definition, hepatic necrosis would be unexpected (by virtue of greater severity) if the investigator brochure referred only to elevated hepatic enzymes or hepatitis. Similarly, cerebral thromboembolism and cerebral vasculitis would be unexpected (by virtue of greater specificity) if the investigator brochure listed only cerebral vascular accidents. "Unexpected," as used in this definition, also refers to AEs or suspected adverse reactions that are mentioned in the investigator brochure as occurring with a class of drugs or as anticipated from the pharmacological properties of the drug, but are not specifically mentioned as occurring with the particular drug under investigation. [Consistent with 21 CFR 312.32 (a)]

11.5 Classification of an Adverse Event

11.5.1 Severity of Event

The intensity of all AEs will be graded using a five-point grading scale in which the following descriptions of severity will apply. Note that for some AEs, Grades 4 and/or 5 may not be applicable. In those cases only 3 different grades (Grade 1-3) are to be considered:

8.2.1.1 **Grade 1:** “Mild”; asymptomatic or mild symptoms; clinical or diagnostic observations only; intervention not indicated.

8.2.1.2 **Grade 2:** “Moderate”; minimal, local or noninvasive intervention indicated; limiting age-appropriate instrumental ADL*.

8.2.1.3 **Grade 3:** “Severe”; medically significant but not immediately life-threatening; hospitalization or prolongation of hospitalization indicated; disabling; limiting self care ADL**.

8.2.1.4 **Grade 4:** “Life-threatening consequences”; urgent intervention indicated.

8.2.1.5 **Grade 5:** “Death related to AE”.

*Instrumental ADL refer to preparing meals, shopping for groceries or clothes, using the telephone, managing money, etc.

**Self care ADL refer to bathing, dressing and undressing, feeding self, using the toilet, taking medications, and not bedridden.

11.5.2 Relationship to Study Intervention

For all collected AEs, the clinician who examines and evaluates the participant will determine the AE's causality based on temporal relationship and his/her clinical judgment. A binary assessment (related/not related) will be made and take into consideration the natural history of the underlying disease, concurrent illness, concomitant therapy, study-related procedures, accidents, and other external factors. While the relationship to the study drug (related/not related) is part of the documentation process, it is not a factor in determining what is or is not reported in the study. All AEs are recorded regardless of relatedness.

Related The AE is known to occur with the study agent, there is a reasonable possibility that the study agent caused the AE, or there is a temporal relationship between the study agent and event. Reasonable possibility means that there is evidence to suggest a causal relationship between the study agent and the AE. An AE can be deemed related even if other factors may have contributed to the event.

Not Related There is not a reasonable possibility that the administration of the study agent caused the event, there is no temporal relationship between the study agent and event onset, or an alternate etiology has been established or appears to provide a plausible explanation (e.g. the participant's clinical condition, underlying disease or concomitant treatments).

11.5.3 Expectedness

For all SAEs, the Clinical Site Principal Investigator, in consultation with a designated Medical Monitor whenever possible, will be responsible for determining whether the SAE is expected or unexpected using the definition from section 11.4 above.

11.5.4 Treatment Failures

Treatment failures are participants who, after >3 months of study, demonstrate

- an absolute fall in FVC-% of ≥ 15% from their baseline determination, or
- an FVC-% ≤ 35%, regardless of the absolute change from baseline.

To meet these definitions, participants must have two FVC-% measurements greater than 15 days apart, both showing an absolute decrement of ≥ 15% from baseline and/or a FVC-% of ≤ 35%. Participants with treatment failures will be withdrawn from active drug treatment (both PFD/Plac and MMF). Participants who fail treatment will be encouraged to return for key outcome determinations at 12 and 18 months.

11.6 Pregnancies

A listing of all pregnancies occurring after the start of study medication is provided.

11.7 Clinical Laboratory Evaluations

Clinical laboratory evaluations are summarized using number, mean, standard deviation, median, interquartile range, minimum and maximum. Absolute and relative changes from baseline to each study visit are similarly summarized. These summaries are reported overall and by treatment group.

Laboratory values collected during the study are:

Type	Name	Units	Values	Abnormal (checked yes/no)	applicable upper or lower limit of normal
Metabolic panel	Sodium	Mmol/L	X	X	X
	Potassium	Mmol/L	X	X	X
	Chloride	Mmol/L	X	X	X
	CO2	Mmol/L	X	X	X
	BUN	Mg/dL	X	X	X
	Creatinine	Mg/dL	X	X	X
	Glucose	Mg/dL	X	X	X
	Protein – total	g/dL	X	X	X
	Albumin	g/dL	X	X	X
	Bilirubin – total	Mg/dL	X	X	X
	Alkaline phosphatase	IU/L	X	X	X
	AST	IU/L	X	X	X
	ALT	IU/L	X	X	X
	Cholesterol	Mg/dL	X	X	X
CBC with differential	WBC	K/uL	X	X	X
	Hemoglobin	g/dL	X	X	X
	Hematocrit	%	X	X	X
	Platelets	K/uL	X	X	X
	Neutrophil percent	%	X	X	X
	Absolute neutrophil count	K/uL	X	X	X
	Lymphocyte percent	%	X	X	X
	Absolute lymphocyte count	K/uL	X	X	X

11.8 Prior and Concurrent Medications

Prior and concurrent medications are summarized descriptively by time point, overall, and by treatment group.

11.9 Other Safety Measures

Vital signs including weight, heart rate, blood pressure, and oxygen saturation are summarized using similar methods to those given in Section 11.7.

12 Other Analyses

Analyses of the following outcomes will be summarized and allow for further interpretation of the study results. The m-ITT analysis set is used in the analyses.

1. The linear mixed model as described for the primary analysis of the primary endpoint is performed to compare two treatment groups on the change of the PROMIS-29 version 2.0 (in Physical Function, Anxiety, Depression, Fatigue, Sleep Disturbance, Pain Interference, Impact on Social Roles, and a single item on pain intensity) from baseline to month 18, measured by 6-month intervals. The model is adjusted by treatment, study month, interaction of treatment and study month, prior MMF therapy and baseline PROMIS-29. The model generates adjusted estimates of change from baseline in PROMIS-29 for each treatment group and month, and an F-test is used to test the hypothesis that the mean change from baseline during the treatment period differs between the two treatment groups. The model-based adjusted mean change from baseline in PROMIS-29 will be presented graphically by treatment group by study month, with 95% confidence intervals (CIs) provided at each month.
2. A frequency table and bar chart are used to show the frequency distribution of the change of QILD-WL from screening to month 18 categorized in 2% increments. The proportions of participants with better (<-2%), stable (≥ -2% to ≤ 2%) and worsening (>2%) QILD-WL in each treatment group are also presented.
3. A frequency table is used to evaluate the proportion of participants with a 3% or greater decline from baseline in FVC-% over the 18-month treatment period between the two treatment groups.
4. A frequency table is used to evaluate the proportion of participants with a 5% or greater decline from baseline in FVC-% over the 18-month treatment period between the two treatment groups.
5. Descriptive statistics are used to show the change from baseline to month 6, 12, and 18 in patient global assessment of “overall health” in the past week, overall and in each treatment group.
6. Descriptive statistics are used to show the patient transition questions of global assessment at month 6, 12, and 18 comparing patient’s (i) overall health and (ii) lung involvement to that at the baseline visit, overall and in each treatment group.
7. Descriptive statistics are used to show the change from baseline to month 6, 12, and 18 in cough severity, cough frequency, and phlegm production, overall and in each treatment group.

13 Reporting Conventions

P-values ≥0.001 will be reported to 3 decimal places; p-values less than 0.001 will be reported as “<0.001”. The mean, standard deviation, and any other statistics other than quantiles, will be reported to one decimal place greater than the original data. Quantiles, such as median, or minimum and maximum will use the same number of decimal places as the original data. Estimated parameters, not on the same scale as raw observations (e.g., regression coefficients) will be reported to 3 significant figures.

14 Summary of Changes to the Protocol and/or SAP

a. Additional Secondary Endpoint

FVC in ml is added to the analysis as one of the secondary endpoints. It will help us to compare SLSIII results with other studies, including SENCIS and TRAIL-1.

PROTOCOL:

N/A

SAP:

Section 5.2.2 Secondary Endpoints

1. The change from baseline to 18 months, measured at 3-month intervals, over the course of the 18-month treatment period in:

   c. Forced vital capacity volume (FVC, in ml).

b. Remove Other Endpoints.

Physician global assessments and UCLA Scleroderma Clinical Trial Consortium GIT 2.0 will not be analyzed due to small sample size.

PROTOCOL:

Section 4.2.3 Other Important Endpoints

4.2.3.1 Physician and Patient Global Assessments

a) Physician global assessment of patient's "overall health" in the past week on a Likert scale; 6 month intervals

b) Physician transition questions comparing patient's i) overall health and ii) lung involvement to that at the baseline visit; 6 month intervals

4.2.3.2 UCLA Scleroderma Clinical Trial Consortium GIT 2.0 (UCLA SCTC GIT 2.0); 6 month intervals

SAP:

Not included.

c. Additional Other Endpoints

Because we had to stop enrollment before the planned sample size was achieved, we changed the classification of "other important endpoints" to "other endpoints"; modified the classification of PROMIS-19 endpoints to "other"; and added or clarified the additional other endpoints: categorized QILD-WL and two FVC-% endpoints based on the meaningful clinical importance difference; remove frequency distribution of DLCOHb-% because it is the most variable of the PFTs, and changed LCQ to the supplemental questions in LCQ. These changes clarify the importance of interpreting endpoints in the context of the reduced sample size.

PROTOCOL:

Section 4.2.2 Secondary Endpoints

4.2.2.4 The change from baseline to 18 months, measured at 3-month or 6-month intervals, in Patient Reported Outcomes (PROs), which provide subjective measures of dyspnea and quality of life based on patient responses to standardized patient questionnaires which include:

c) Patient-reported outcomes measurement information system 29-item health profile (PROMIS-29 version 2.0); 6 month intervals

4.2.2.6 Differences in the frequency distribution of individual patient responses when grouped into defined intervals of improvement or worsening (defined by the change in an outcome measure from baseline to 18 months) for the following outcome measures:

b) Single-breath diffusing capacity for carbon monoxide (DLCO), calculated as a percent of the age, height, gender-, race- and hemoglobin-adjusted predicted value (DLCOHb-%).

Section 4.2.3 Other Important Endpoints

4.2.3.3 Leicester Cough Questionnaire (LCQ); 6 month intervals

SAP:

Section 5.2.3 Other Endpoints

1. The change from baseline to month 6, 12, and 18 in PROMIS-29 version 2.0 with the following domains: physical function, anxiety, depression, fatigue, sleep disturbance, pain interference, ability to participate in social roles and activities, and a single item on pain intensity.
2. The proportion of participants in each treatment from screening to month 18 with a 2% increment in QILD-WL. It is also categorized as better (<-2%), stable (≥-2%, ≤2%) and worsen (>2%).
3. The proportion of participants in each treatment arm with a 3% or greater decline from baseline in FVC-% over the 18-month treatment period.
4. The proportion of participants in each treatment arm with a 5% or greater decline from baseline in FVC-% over the 18-month treatment period.
7. The change from baseline to month 6, 12, and 18 in supplemental questions in Leicester Cough Questionnaire (LCQ) regarding the cough severity, cough frequency and phlegm production.

d. Remove Exploratory Endpoints and Analyses of Exploratory Outcomes.

The pre-specified exploratory endpoints and implicit associated analyses in the protocol will not be conducted owing to the small sample size and ensuing limitations on multivariable analyses.

PROTOCOL:

4.2.4 Exploratory Endpoints

Exploratory endpoints will include the following:

4.2.4.1 Identification of baseline features that predict treatment responsiveness, disease progression and the course of lung and skin disease over time.

4.2.4.2 Identification of biomarkers that predict disease features, treatment responsiveness, disease progression and the course of lung and skin disease over time.

4.2.4.3 Composite outcome measures that distinguish early and late treatment responses

4.2.4.4 Performance of CRISS index at 6, 12 and 18-months

10.4.3.2 Analyses of Exploratory Outcomes.

Comparable methods as described above for primary and secondary outcomes will be used for the exploratory aims described in Section 4.2.3. For example, identification of baseline features that predicted treatment responsiveness, disease progression and course of lung and skin disease over time will employ the appropriate models with the baseline covariate and potentially the interaction of treatment and baseline as covariates. Separate models would be assessed for each outcome and baseline. Biomarker identification would be handled similarly. The identification of composite outcome measures that distinguish early and late treatment responses will be detailed in the SAP. These analyses will be considered exploratory and hypothesis-generating.

SAP:

Not included

e. Add Completers Population

Add Completers analysis set that wasn't in the protocol to provide an analysis population for descriptive statistics of select efficacy endpoints.

PROTOCOL:

N/A

SAP:

8.2.3 Completers Population

The Completers analysis set consists of all subjects in the m-ITT population who had an 18-month efficacy outcome (regardless of completion of study medication). It can be different depending on the outcome that is analyzed.

f. Change Analyses of Secondary Endpoints

Due to small sample size, instead of considering binary secondary endpoints and using logistic regression, we consider clinically meaningful cutpoints and using the frequency table and bar chart to demonstrate the frequency distribution of the secondary endpoints.

PROTOCOL:

Section 10.4.3.1 Analyses of Secondary Endpoints

There are three categories of secondary outcomes:

2. Dichotomous outcomes such as the proportion of subjects who report improvement on the TDI at 18 months during the treatment period.

For dichotomous secondary outcomes, logistic or Poisson regression will be used with treatment group, site, and prior MMF therapy (stratification factor: naïve, >0 to $\leq$3 months and >3 months to 6 months) included as covariates.

SAP:

Section 10.2.2

8. A frequency table and bar chart is used to show the frequency distribution of the change of FVC-% from baseline to month 18 with a 5% increment (e.g., improved by up to 5%, from 5% to <10% and from 10% to <15% or worsened by up to 5%, from 5% to <10%, and from 10% to <15%) overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month FVC-%.
9. A frequency table is used to show the frequency distribution of the change of FVC-% from baseline to month 18 in negative responders (≤ -3%) and stable (> -3% to < 3%) overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month FVC-%.
10. A frequency table is used to show the frequency distribution of the change of FVC-% from baseline to month 18 in responder (>0) and non-responder (≤ 0) overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month FVC-%.
11. A frequency table and bar chart are used to show the frequency distribution of the change of mRSS from baseline to month 18 categorized in 4 points increments: worsen (1 to 4, ≥ 5), no change (=0), and improved (≤ -13, -12 to -9, -8 to -5, -4 to -1) overall and in each treatment group in the completers population, and separately for those who completed the treatment period and those who prematurely discontinued treatment but returned for the 18-month mRSS. A frequency distribution of the change of mRSS is also shown in patients with dcSSc in the completers population. The proportions of participants with improved (< -5), no change (-5 to 5), and decreased (> 5) in the change from baseline in mRSS, overall and in each treatment group, are also presented.
12. A frequency table and bar chart are used to show the frequency distribution of the TDI at month 18 categorized as improved by 1-3, 4-6 and 7-9 points, no change (0), and worsened by 1-3, 4-6 and 7-9 points, overall and in each treatment group in the completers population, and separately for those who completed the entire treatment period and those who prematurely discontinued treatment but returned for the 18-month TDI. The proportions of participants with improved (>0), no change, and deterioration (<0), overall and in each treatment group, are also presented.

g. Remove One of Analysis of the Primary Efficacy Endpoint

Because we had to stop enrollment before the planned sample size was achieved, we don't have sufficient data to perform the analysis to access the change point for response.

PROTOCOL:

10.4.2 Analysis of the Primary Efficacy Endpoint

Using the same model as for the primary endpoint, we will assess the change point for response to evaluate the onset of response for the two treatment groups, and the change from baseline to month 18 to evaluate the overall magnitude of effect at the end of the treatment period.

SAP:

Not included.

h. Elimination of reports regarding death

There were no deaths in the study and so the planned descriptive statistics and Kaplan-Meier plot are not presented.

PROTOCOL:

8.1 Specification of Safety Parameters

g) All treatment-emergent deaths

i) Time to Time to treatment failure, treatment-emergent AEs and SAEs resulting in premature discontinuation of study medication, and death

10.4.4 Safety Analyses

Kaplan-Meier methods will be used to summarize graphically the time to treatment failure, treatment emergent AEs and SAEs that result in discontinuation of treatment, and death.

SAP:

Not included.

i. Biomarkers will not be analyzed

Because we had to stop enrollment before the planned sample size was achieved, we will not assess biomarkers in this primary analysis.

PROTOCOL:

4.2.4 Exploratory Endpoints

Exploratory endpoints will include the following:

4.2.4.2 Identification of biomarkers that predict disease features, treatment responsiveness, disease progression and the course of lung and skin disease over time.

10.4.3.2 Analyses of Exploratory Outcomes.

Biomarker identification would be handled similarly. The identification of composite outcome measures that distinguish early and late treatment responses will be detailed in the SAP. These analyses will be considered exploratory and hypothesis-generating.

SAP:

Not included.

15 References

1. Cotes JE, Dabbs JM, Elwood PC, Hall AM, McDonald A, Saunders MJ. Iron-deficiency anaemia: its effect on transfer factor for the lung (diffusing capacity) and ventilation and cardiac frequency during sub-maximal exercise. Clinical Science. 1972; 42: 325–335.
2. Esbriet Package Insert, NDA-022535, Genentech Inc., Approved Oct 15, 2014.
3. Fan MH, Feghali-Bostwick CA, Silver RM. Update on scleroderma-associated interstitial lung disease. Curr Opin Rheumatol. 2014;26(6):630-6.
4. Hankinson JL OJ, Fedan KB. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med. 1999;159(1):179–87.
5. Khanna D, Albera C, Fischer A, N. Khalidi, G. Raghu, L. Chung, D. Chen, E. Schiopu, 8, E. Gorina, M. Tagliaferri, J. R. Seibold. Safety And Tolerability Of Pirfenidone In Patients With Systemic Sclerosis-Associated Interstitial Lung Disease - The Lotuss Study. Am J Respir Crit Care Med. 2015; 191: A1175
6. Neas LM, and Schwartz J. The determinants of pulmonary diffusing capacity in a national sample of U.S. adults. Am J Respir Crit Care Med. 1996; 153 (2): 656-664.
7. Roth MD, Tseng CH, Clements PJ, Furst DE, Tashkin DP, Goldin JG, Khanna D, Kleerup EC, Li N, Elashoff D, Elashoff RM; Scleroderma Lung Study Research Group. Predicting treatment outcomes and responder subsets in scleroderma-related interstitial lung disease. Arthritis Rheum. 2011; 63(9):2797-808.
8. Rubin D. (1976). Inference and missing data (with discussion). Biometrika 63:581-592.
9. Tashkin DP, Elashoff R, Clements PJ, et al. for the Scleroderma Lung Study (SLS) Research Group. Effects of 1-year treatment with cyclophosphamide on outcomes at 2 years in scleroderma lung disease. Am J Respir Crit Care Med 2007 177(10): 1026.
10. Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, Arriola E, Silver R, Strange C, Bolster M, Seibold JR, Riley DJ, Hsu VM, Varga J, Wigley F, White B, Steen V, Read C, Mayes M, Parsley E, Mubarak K, Connolly MK, Golden J, Olman M, Fessler B, Rothfield N, Metersky M for the Scleroderma Lung Study Research Group. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med 2006; 354:2655-66.
11. Tashkin DP, Roth MD, Clements PJ, Furst DE, Khanna D, MD, Kleerup EC, Goldin J, Arriola E, Volkmann ER, Kafaja S, Silver R, Steen V, Strange C, Wise R, Wigley F, Mayes M, Riley DJ, Hussain S, Assassi S, Hsu VM, Patel B, Phillips K, Martinez F, Golden J, Connolly MK, Varga, MD J, Dematte J, Hinchcliff ME, Fischer A, Swigris J, Meehan R, Theodore A, Simms R, Volkov S, Schraufnagel DE, Scholand MB, Frech T, Molitor JA, Highland K, Read CA, Fritzler MJ, Kim GHJ, Tseng C-H, Elashoff RM, for the Sclerodema Lung Study II Investigators. Mycophenolate mofetil versus oral cyclophosphamide in scleroderma-related interstitial lung disease (SLS II): a randomised controlled, double-blind, parallel group trial. Lancet Respir Med. 2016 Sep;4(9):708-19.
12. Kim HJ, Tashkin DP, Clements PJ, et al. A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. Clin Exp Rheumatol 2010; 5: S26–35.

16 Planned Tables, Listings and Figures

See separate document.