# Cover Page for Statistical Analysis Plan

| | |
|---|---|
| Sponsor name: | Novo Nordisk A/S |
| NCT number | NCT03926117 |
| Sponsor trial ID: | NN6018-4779 |
| Official title of study: | A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE) |
| Document date | 21 July 2020 |

## APPENDIX 16.1.9: DOCUMENTATION OF STATISTICAL METHODS



*Redacted statistical analysis plan*
*Includes redaction of personal identifiable information only.*

# Patient-Reported Outcomes Statistical Analysis Plan

A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE)

**PROTOCOL NUMBER:** COR-001-02

**Study Sponsor**:

Corvidia Therapeutics Inc.
35 Gatehouse Drive
Waltham MA 02451, USA

**Author**: [redacted]

**Version Number and Date**: V2.0, 21JUL2020

# Table of Contents

**LIST OF FIGURES**



**LIST OF TABLES**

# PRO STATISTICAL ANALYSIS PLAN SIGNATURE PAGE

**Statistical Analysis Plan V2.0 (Dated 21JUL2020) for Protocol COR-001-02.**

| | Name | Signature | Date |
|---|---|---|---|
| **Author:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |

Upon review of this document, the undersigned approves this version of the PRO Statistical Analysis Plan, including the table and figure shells, authorizing that the content is acceptable for the reporting of this study.

| | Name | Signature | Date |
|---|---|---|---|
| **Approved By:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |
| | | | |
| **Approved By:** | [redacted] | [redacted] | |
| **Position:** | [redacted] | | |
| **Company:** | Corvidia | | |

## MODIFICATION HISTORY

| Unique Identifier for this Version | Date of the Document Version | Significant Changes from Previous Authorized Version |
|---|---|---|
| 1.0 | 22APR2020 | Not Applicable – First Version |
| 2.0 | 21JUL2020 | Removal of the probability density function plots.<br>Clarification added regarding the following:<br>• time to definitive deterioration<br>• derivation for the change from baseline in response categories used in the responder analysis |

## LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| ANCOVA | Analysis of Covariance |
| BP | Bodily Pain |
| CARES | Corvidia ePRO |
| CDF | Cumulative Distribution Function |
| CI | Confidence Interval |
| CKD | Chronic Kidney Disease |
| EDC | Electronic Data Capture |
| EOT | End of Treatment |
| ePRO | Electronic Patient-Reported Outcomes |
| GH | General Health Perceptions |
| HRQoL | Health-Related Quality of Life |
| hs-CRP | High-Sensitivity C-Reactive Protein |
| ITT | Intention-To-Treat |
| LS | Least Squares |
| MAR | Missing at Random |
| MCS | Mental Component Summary |
| MH | Mental Health |
| MMRM | Mixed Model Repeated Measures |
| MNAR | Missing Not at Random |
| PAP | Psychometric Analysis Plan |
| PCS | Physical Component Summary |
| PDF | Probability Density Function |
| PF | Physical Functioning |
| PGIC | Patient Global Impression of Change |
| PGIS | Patient Global Impression of Severity |
| PRO | Patient-Report Outcomes |
| PROMIS Sex FS | PROMIS Sexual Function and Satisfaction |
| RE | Role Limitations Due to Emotional Problems |
| REML | Restricted Maximum Likelihood |
| RP | Role Limitations Due to Physical Health |
| SAA | Serum Amyloid A |
| SAP | Statistical Analysis Plan |
| SD | Standard Deviation |
| SF | Social Functioning |
| SMD | Standardized Mean Difference |
| TSS | Total Symptom Score |
| TTDD | Time to First Definitive Deterioration |

| Abbreviation | Definition |
|---|---|
| TTFD | Time to First Deterioration |
| VAS | Visual Analogue Scale |
| VT | Vitality |

# 1. INTRODUCTION

This document describes the rules and conventions to be used in the presentation and analysis of patient-reported outcomes (PRO) data collected in the COR-001-02 study. It describes the data to be summarized and analyzed, including specifics of the statistical analyses to be performed. It is intended to ensure the credibility of the study PRO results by pre-specifying the statistical approaches for the analysis of PRO data prior to database lock.

This statistical analysis plan (SAP) is based on protocol version amendment 5 dated 23 December 2019.

This plan may be revised during the study to accommodate protocol amendments and/or to make changes to adapt to unexpected issues in study execution and/or data that affect planned analyses. The final plan, if revised, will document all changes and be issued prior to database lock.

Due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, an immediate discontinuation of dosing was announced on March 18, 2020 to ensure the safety of trial patients. The announcement required the early termination visit to be scheduled as early as possible for ongoing patients followed by their entry into the safety follow-up period.

At the time of discontinuation of the study, approximately 140 patients had completed the 24-week treatment period and more than 120 patients were non-completers of the treatment period.

## 1.1. OBJECTIVES OF THE PRO ANALYSIS

The aim of PRO analysis is to assess the benefits of ziltivekimab compared to placebo on disease-related symptoms, pain, and health related quality of life (HRQoL) in adult subjects with Stage 3-5 Chronic Kidney Disease (CKD).

Key objectives of the PRO analyses are:

- To assess the impact of ziltivekimab vs. placebo on symptoms (collected in Corvidia ePRO (CARES), Patient-Reported Outcomes Measurement Information System (PROMIS®) Fatigue Short Form 13a and selected items from the PROMIS Item Bank) as measured by mean changes from baseline to Week 13 and Week 24.

Supportive objectives are:

- To assess the time to first clinically meaningful deterioration, as well as time to definitive clinically meaningful deterioration, in symptoms as measured by CARES, PROMIS Fatigue Short Form 13a and selected items from the PROMIS Item Bank.
- To assess responder proportions at Week 13 and Week 24 in symptoms as measured by CARES, PROMIS Fatigue Short Form 13a and selected items from the PROMIS Item Bank.
- To examine the range of the responder definitions by looking into the Cumulative Distribution Function (CDF) plots of ziltivekimab vs placebo.

- To assess the impact of ziltivekimab vs. placebo on functioning and HRQoL as measured by mean changes from baseline at Week 13 and Week 24 on physical and mental component scores as collected in SF-36, as well as on overall HRQoL, as measured by the EQ-5D-5L.

# 2. COR-001-02 STUDY OVERVIEW

## 2.1. STUDY OBJECTIVES

### 2.1.1. PRIMARY OBJECTIVE

The primary objective of the RESCUE study is:

- To evaluate the effects of ziltivekimab compared to placebo on markers of inflammation such as high-sensitivity C-reactive protein (hs-CRP).

### 2.1.2. SECONDARY OBJECTIVES

The secondary objective is:

- To evaluate the effects of ziltivekimab compared to placebo on two markers of inflammation and cardiovascular risk: serum amyloid A (SAA) and fibrinogen.

### 2.1.3. EXPLORATORY OBJECTIVES

Selected exploratory objectives related to PRO are:

- To determine the pharmacokinetic, exploratory pharmacodynamics, pharmacogenetics, and effect of ziltivekimab on inflammatory markers.
- To evaluate the effects of three dose levels of ziltivekimab compared to placebo on PRO: PROMIS Fatigue 13a short form, selected items from the PROMIS fatigue item bank, the Optum SF-36 v2® HealthSurvey, CARES, the PROMIS interest in sexual activity item, the patient global impression of change (PGIC), patient global impression of severity (PGIS), and the EQ-5D-5L.
- To evaluate the psychometric properties of the CARES items, PROMIS Fatigue 13a short form, and selected items from the PROMIS fatigue item bank, in CKD patients.

## 2.2. SAMPLE SIZE

The primary efficacy endpoint is percent change from baseline in hs-CRP (average of the hs-CRP value prior to randomization and Day 1) to Week 13 between each active group and placebo.

Based on the observed treatment difference in percent change from baseline in hs-CRP of

60.74% between combined COR-001-01 active groups and placebo and the associated pooled SD of 16.893% in hs-CRP at Week 4 from the final analysis of study COR-001-SC1, a sample size of 54 per group yields more than 99% power with 2-sided alpha=0.05.

Taking into consideration the dropout rate of 10% by the end of the study, a sample size of 60 per group is planned for this study. Accordingly, approximately 240 patients will be randomized 1:1:1:1 (60 per each treatment group) into the trial. Patient randomization will be stratified by baseline hemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5).

## 2.3. STUDY DESIGN

The RESCUE study is a randomized, double-blind, placebo-controlled trial designed to evaluate the efficacy, safety, and pharmacokinetics of ziltivekimab at three dose levels (7.5 mg, 15 mg or 30 mg) compared to placebo in patients with stage 3-5 CKD, not on dialysis, who have evidence of inflammation with high cardiovascular risk.

The primary, secondary, and exploratory endpoints will be analysed at 13 weeks of dosing and then followed for additional exploratory efficacy analyses through Week 24. Selected efficacy endpoints and safety assessments will be evaluated in the Follow-up Period Week 25 though Week 32 (Figure 1).

Patients will undergo a Screening Period of up to 14-days during which inclusion and exclusion criteria will be evaluated. Patients who meet all inclusion criteria and no exclusion criteria will be randomized to one of three ziltivekimab dose levels (7.5 mg, 15mg, or 30 mg) or placebo for a 24-week Treatment Period. Patient randomization will be stratified by baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4, or 5).

The patient will be randomized on Day 1 and the first dose of study drug should be administered after all assessments are conducted. Doses of study drug will be administered every 28-days for a total of 6 treatments (Weeks 1, 5, 9, 13, 17, and 21).

The test product dose regimens to be examined in this study are:

- Dose #1: Ziltivekimab, 7.5 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21, as a subcutaneous injection.
- Dose #2: Ziltivekimab, 15 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21, as a subcutaneous injection.
- Dose #3: Ziltivekimab, 30 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21, as a subcutaneous injection.

The reference dose regimens to be examined in this study are:

- Matched placebo injections administered subcutaneously 6 times (every 28 days) at the same frequency as the active treatment in a given dose cohort, on Weeks 1, 5, 9, 13, 17, and 21.

**Figure 1: Study Flow Diagram**

Ziltivekimab Dose Level 1, 2 or 3 (N=180); N=60 in each level

Placebo (N=60)

| | Screening | | Treatment Period | | | | | | | | | | | Safety F/Up | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11/ET | 12 | 13/ET |
| Visit Day Window | -14 | -7 | 1 | 8 | 22 | 29-35 | 43-49 | 57-63 | 85-91 | 113-119 | 141-147 | 155-161 | 162-168 | 190-196 | 218-224 |
| Visit Window | | ±2 | 1 | ±2 | ±2 | | | | | | | | | | |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |

## 2.4. PRO INSTRUMENTS

The following PRO instruments are collected in this study:

- CARES
- PROMIS Fatigue 13a short form
- Selected items from the PROMIS fatigue item bank
- PROMIS interest in sexual activity item
- Optum SF-36 v2® Health Survey
- EQ-5D-5L
- PGIS and PGIC.

### 2.4.1. CARES CORVIDIA

The CARES Corvidia ePRO - henceforth CARES - is a new PRO instrument under development. It consists of 19 symptom items, with a 24-hour recall, asking patients to report their worst level of that symptom in the past 24 hours on a numeric rating scale from 0 (no symptom) to 10 (symptom as bad as I can imagine).

The symptom items cover physiological (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea), psychological (Depression, Concentration, Forgetfulness), Pain (Bone / Joints, Nerves, Muscle Cramps) and sensorial (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold) symptoms.

The instrument can be found in Appendix A. CARES.

### 2.4.2. PROMIS Fatigue 13a Short Form and selected items from PROMIS Fatigue Bank, PROMIS Interest in Sexual Activity Item

The PROMIS Fatigue 13a short form contains 13 items that assess symptoms and impacts of fatigue. The recall period is 7 days. Each item is rated on a five-point Likert scale ranging from 0 = "not at all" to 4 = "very much". Higher scores in these items indicate worse fatigue, except from items 7 ("I have energy") and 8 ("I am able to do my usual activities") where higher score indicates less fatigue.

In this study, three additional fatigue items were administered for further testing to account for exhaustion, being physically drained, and impact on alertness, concepts that resonated with patients during patient interviews but are not covered in the PROMIS Fatigue 13a short form. Such combination is from now on referred to as PROMIS Fatigue.

A number of scales and subscales can be derived from the PROMIS Fatigue set:

- The original Fatigue 13a scale
- The fatigue scale using the whole PROMIS Fatigue item set (13 + 3 de novo 16-item short form)
- Two subscales within the PROMIS Fatigue 13a scale: 1) symptoms of fatigue; and 2) impacts of fatigue
- Two subscales within the PROMIS Fatigue item set (13 + 3 de novo 16-item short form): 1) symptoms of fatigue; and 2) impacts of fatigue.

The PROMIS Sexual Function and Satisfaction (PROMIS Sex FS) measure is a customizable, self-reported set of measures that include 79 items covering 11 domains: interest in sexual activity, lubrication, vaginal discomfort, erectile function, global satisfaction with sex life, orgasm, anal discomfort, therapeutic aids, sexual activities, interfering factors, and screener questions. The PROMIS Sex FS uses a 30-day recall period. In this study only one item will be used: interest in sexual activity. This item is rated on a five-point Likert scale ranging from 0 = "not at all" to 4 = "very much". Higher scores in this item indicate more interest in sexual activity.

These three instruments will be administered together as shown in Appendix B. PROMIS Fatigue 13a Short Form, additional fatigue items, sexual activity.

### 2.4.3. Short Form 36-item Health Survey, Version 2 (SF-36v2)

SF-36v2 is a self-report survey of functional health and well-being with 4 weeks recall period (QualityMetric, 2011). Responses to 35 of the 36 items are used to compute an 8-domain profile of functional health and well-being scores.

The remaining item, referred to as the 'Health Transition' item, asks patients to rate how their current state of health compared to their state of health 1 year ago, and is not used to calculate domain scores.

The 8-domain profile consists of the following scales: Physical Functioning (PF), Role Limitations due to Physical Health (RP), Bodily Pain (BP), General Health Perceptions (GH), Vitality (VT), Social Functioning (SF), Role Limitations due to Emotional Problems (RE), and

Mental Health (MH).

Psychometrically-based physical and mental health component summary scores (PCS and MCS, respectively) are computed from subscale scores to give a broader metric of physical and mental health status. The instrument can be found in Appendix C. SF-36.

### 2.4.4. EQ-5D-5L

The EuroQol EQ-5D-5L has been designed as an international, standardized, generic instrument for describing and valuing health-related quality of life, available in over 100 language versions.

The EQ-5D-5L self-report questionnaire consists of 5 items, one per domain, with 5-point scale and one visual analogue scale (VAS) to rate health state from worst (0) to best (100).

The EQ-5D-5L includes domains for each generic health status measure: Mobility, Self-care, Usual Activities, Pain/discomfort and Anxiety/depression. For each question, there are 5 levels of response, corresponding to increasing levels of impairment (no problems, slight, moderate, severe, and extreme problems, or unable to perform activity) and coded 1 to 5.

A higher index indicates better quality of life. The instrument can be found in Appendix D. EQ-5D-5L.

### 2.4.5. Patient Global Impression of Severity (PGIS) and Change (PGIC)

The PGIS is a self-reported measure of patient-perceived overall symptom severity. The PGIS is a 4-point recall scale (from no symptoms to severe symptoms). Patients are asked to rate their overall severity of symptoms in the 7 days prior to a clinic visit.

The PGIC instrument captures the patient's overall evaluation of response to treatment. The patient is asked to report the degree to which they have changed since entering the treatment period using a 7-point scale (Very Much Better to No Changes to Very Much Worse).

The PGIC and PGIS are the most commonly used anchor-based methods of assessing clinically important change and severity in which the external judgment of meaningful change is made by the patient. The instruments can be found in Appendix E. PGIC and PGIS.

### 2.4.6. Assessment Schedule for PRO Instruments

Patient-reported outcomes will be assessed through a personal device supplied by Corvidia:

- At home daily on patient's personal device: CARES
- At the clinic prior to dosing on patient's personal device: PROMIS Fatigue 13a Short Form and selected items from PROMIS Fatigue item bank, PROMIS Interest in Sexual Activity item, SF-36v2, EQ-5D-5L, PGIC, PGIS.

CARES data will be collected at home on patient's personal device or one provided by Corvidia; the questionnaire will be answered daily for 7 days prior to treatment administration

(Screening day -7 to day -1 prior to dose 1 (Visit 1), dose 2 (Visit 4), dose 4 (Visit 7), and for the last 2 weeks after dose 6 (Visit 9), i.e. during weeks 23 and 24.

The SF-36v2 and PROMIS Fatigue items will be collected on patients' device at the clinic prior to treatment administration at Visit 1, Visit 4, Visit 7, Visit 10 and 11, or the early termination visit (whichever occurs first).

The PROMIS interest in sexual activity item and EQ-5D-5L will be collected on patients' device at the clinic prior to treatment administration at Visit 1, Visit 4, Visit 7 and Visit 11, or the early termination visit (whichever occurs first).

The PGIC will be colected on patients' device at the clinic prior to to treatment administration at Visit 4, Visit 7, and Visit 11, or the early termination visit (whichever occurs first).

The PGIS will be collected at home on patients' devices on the last day of CARES assessment period, i.e. day prior to dose 1 (Visit 1), dose 2 (Visit 4), dose 4 (Visit 7), and at Visit 11, or the early termination visit (whichever occurs first).

The schedule of assessment for all PRO instruments collected in different periods of the study are shown in Table 1.

**Table 1 PRO Instruments Assessment Schedule**

| Study Period | Screening | | Treatment | | | | | | | | | | | | Safety F/Up[a] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | N/A | 7 | 8 | 9 | 10 | 11/ET[b] | 12 | 13/ET |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 12 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |
| DOSING | | | X | | | X | | X | | X | X | X | | | | |
| CARES | | X | | | X | | | | X | | | | X | X | | |
| PROMIS Fatigue | | | X | | | X | | | | X | | | X | X | | |
| PROMIS Sex Item | | | X | | | X | | | | X | | | | X | | |
| SF-36v2 | | | X | | | X | | | | X | | | X | X | | |
| EQ-5D-5L | | | X | | | X | | | | X | | | | X | | |
| PGIS | | X | | | X | | | | X | | | | | X | | |
| PGIC | | | | | | X | | | | X | | | | X | | |

[a]Follow-Up.
[b]Early Termination.

## 3. ANALYSIS SETS

All PRO analyses described in this SAP will be performed on the intention-to-treat (ITT) analysis population which includes all randomized patients.

# 4. ANALYSIS VARIABLES

## 4.1. GENERAL VARIABLES AND DERIVATIONS

### 4.1.1. STUDY DAY

Study day will be defined as in the clinical SAP dated 31MAR2020. Specifically, the study day is calculated in reference to the date of the first dose of the study drug. Study day 1 corresponds to the date the patient received the first dose of study drug. The day immediately before Day 1 will be Day -1. For assessments conducted on or after the date of the first dose of study drug, treatment day will be calculated as (assessment date - date of first dose of study drug) + 1.

### 4.1.2. BASELINE

The assessment completed on day 1 will be considered the baseline measurement, if taken prior to first dose of study drug.

If more than one assessment were taken, the closest to enrollment day (Day 1) will be used.

For diary scores (CARES), baseline assessment corresponds to the 7 days prior to first dose of study drug.

The post-baseline value is defined as a measurement taken after initial study drug administration.

### 4.1.3. DERIVED TIMEPOINTS

An "End of Treatment" (EOT) assessment will be derived as described in the clinical SAP.

For the purpose of the PRO analyses defined herein, a Week 24 assessment will be used in the analyses. The Week 24 assessment is defined for each instrument in the correspoding sections below.

### 4.1.4. VISIT WINDOW

Visit windows will be used as provided by the sponsor. According to the clinical SAP dated 31MAR2020, scheduled visits will be assigned to analysis visits as recorded in the EDC system. If a scheduled visit is not available, unscheduled and early termination visits will be assigned to analysis visits using analysis visit windows based on the actual date the assessment took place. The low analysis day of the analysis window will be calculated as the midpoint between the scheduled assessment and previously scheduled assessment for that parameter. The high analysis day of the analysis window will be calculated as the midpoint between the scheduled assessment and the next scheduled assessment for that parameter. Where multiple measurements for a particular parameter appear within an analysis window, the scheduled visit will be used. If no scheduled visit appears in the

analysis window, the result closest to the target day will be used. If equidistant and both are unscheduled and/or early termination visits, the later result will be used for the summary measure.

**Table 2: Visit windows**

| Analysis Visit | Target Day | Low Analysis Day | High Analysis Day |
|---|---|---|---|
| Baseline | 1 | | 1 |
| Week 2 | 8 | 2 | 14 |
| Week 4 | 22 | 15 | 28 |
| Week 5 | 35 | 29 | 42 |
| Week 7 | 49 | 43 | 55 |
| Week 9 | 63 | 56 | 70 |
| Week 13 | 85 | 71 | 99 |
| Week 17 | 114 | 100 | 128 |
| Week 21 | 141 | 129 | 153 |
| Week 23 | 157 | 154 | 161 |
| Week 24 | 165 | 162 | 168 |
| End of Treatment | 161 | 154 | 168 |
| Week 28 | 190 | 169 | 211 |
| Week 32 | 224 | 211 | 236 |

### 4.1.5. Other Derivations

No other derivations besides PRO endpoints will be described in this analysis plan.

To calculate time interval duration, a month is 30.4375 days and a year 365.25 days.

## 4.2. PRO Variables

### 4.2.1. CARES Corvidia ePRO

For each item, weekly scores will be derived as the average of the 7 consecutive daily scores as follows:

- Baseline = average of Day -7 to Day -1
- Week 5 (Visit 4) = average of Day 22 to Day 28
- Week 13 (Visit 7) = average of Day 78 to Day 84
- Week 24 (Visit 11) = average of Day 155 to Day 168

For Baseline, Week 5 and Week 13, if more than 3 daily scores out of the 7 days (>50%) within the weekly period are missing, then the score is set to missing.

For Week 24, if more than 7 daily scores out of the 14 days (>50%) within the bi-weekly period are missing, then the score is set to missing.

Preliminary scoring of the CARES includes the calculation of a Total Symptom Score (TSS). Alternative means of scoring the CARES will be explored in a separate psychometric analysis plan (PAP) and if any derived, these will be included in this SAP for efficacy analysis. Additional scales and subscales may be derived based on the results of the psychometric analysis plan (PAP). A higher score for the TSS or other scales/subscales represent more symptoms.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment. At each post-baseline assessment, the change in score from baseline will be calculated for each scale/item.

Domain scores may result from the psychometric analyses described in a separate PAP. In this PAP, clinically meaningful thresholds for these potential domains will also be derived. Change from baseline in the TSS score, as well as the domain scores that may result from the psychometric analysis, will be categorized at each visit as improvement/stable/deterioration using the clinically meaningful threshold value for change scores derived in the psychometric analyses as follows:

- Improvement: a change from baseline ≤ -threshold points;
- Deterioration: a change from baseline ≥ threshold points
- No change: a change from baseline between (–threshold to threshold).

Sensitivity analyses may be performed if more than one threshold values are suggested by the psychometric analyses.

In addition, patients who experience at least one deterioration during the study will be further classified as follows:

- With definitive deterioration: if the deterioration of at least one-threshold point as compared to the baseline score is also observed at all time points threreafter (e.g., after the first deterioration is observed) or if the patient dropped out after deterioration, resulting in missing data.
- With transient deterioration: otherwise.

### 4.2.2. PROMIS Fatigue 13a Short Form and selected items from PROMIS Fatigue Bank, PROMIS Interest in Sexual Activity Items

The PROMIS Fatigue data is collected as a 7-day recall report from patients at Visit 1, 4, 7, 10, and 11.

The scoring manual for PROMIS Fatigue 13a short form can be found online (Patient-Reported Outcomes Measurement Information System Fatigue scoring manual) (accessed on 16th March 2020).

Each question from the PROMIS Fatigue 13a short form has five response options (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much). To use the scoring tables from the PROMIS fatigue scoring manual, a summed score is calculated. The lowest possible summed score is 13 and the highest possible summed score is 65. All questions must be answered in order to produce a valid score using the scoring tables. Once calculated, the total summed raw score is translated to a T-score. The T-score re-scales the

raw score into a standardized score with a mean of 50 and a standard deviation (SD) of 10. The T-score and standard error are then used to calculate the 95% confidence interval (CI) around the observed score. The 95% CI is calculated as T-Score ±1.96*SE. The T-score will not be calculated if a response is missing for one or more questions from the PROMIS Fatigue 13a short form.

In addition, preliminary scoring of the PROMIS Fatigue 13a includes the calculation of two subdomain scores:

- Symptom score (5 symptom items):
  - I feel fatigued
  - I feel weak all over
  - I feel listless ("washed out")
  - I feel tired
  - I have energy
- Impact score (8 impact items):
  - I have trouble starting things because I'm tired
  - I have trouble finishing things because I'm tired
  - I am able to do my usual activities
  - I need to sleep during the day
  - I am too tired to eat
  - I need help doing usual activities
  - I am frustrated by being too tired to do the things I want to do
  - I have to limit my social activity because I am tired.

The subdomains will be calculated as the sum of the items in the corresponding subdomain. The subdomain score will be calculated only if all items are answered.

Alternative means of scoring, by including also the additional PROMIS Fatigue items from the Item Bank, are described in the PAP and may be derived as follows:

- A total fatigue score summing the items for the PROMIS Fatigue short form 13a and the additional three items from the PROMIS fatigue item bank (de novo 16-item short form)
- Symptom score (8 symptom items):
  - I feel fatigued
  - I feel weak all over
  - I feel listless ("washed out")
  - I feel tired
  - I have energy
  - FATEXP36: How exhausted were you on average?

- FATEXP43: How physically drained were you on average?
- FATEXP52: To what degree did your fatigue make you feel less alert?

If psychometric analyses support the creation of the latter two domains, clinically meaningful thresholds will also be derived for them, as well as the three standard fatigue scores from the 13-item set.

Given *PROMIS* Fatigue items and *PROMIS* Interest in Sexual Activity item are collected at Week 23 and at Week 24, an assessment will be derived for the hypothesized scores as the mean value of the two assessments at Week 23 and Week 24, which will be labeled as Week 24. In case only one assessment is available, this only assessment will be used as Week 24 assessment.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment. At each post-baseline assessment, the change in scores from baseline will be calculated for each scale/item.

Different domain scores may result from the psychometric analyses described in a separate PAP. In this PAP, clinically meaningful thresholds for these potential domains, as well as the standard one for PROMIS fatigue -13a short form will also be derived. Change from baseline in the domain scores that may result from the psychometric analysis will be categorized at each visit as improvement/stable/deterioration using the clinically meaningful threshold value for change scores derived in the psychometric analyses as follows:

- Improvement: a change from baseline ≤ -threshold points;
- Deterioration: a change from baseline ≥ threshold points
- No change: a change from baseline between (–threshold to threshold).

Sensitivity analyses may be performed if more than one threshold values are suggested by the psychometric analyses.

In addition, patients who experience at least one deterioration during the study will be further classified as follows:

- With definitive deterioration: if the deterioration of at least one-threshold point as compared to the baseline score is also observed at all time points threreafter (e.g., after the first deterioration is observed) or if the patient dropped out after deterioration, resulting in missing data.
- With transient deterioration: otherwise.

### 4.2.3. Short Form 36-item Health Survey, Version 2 (SF-36v2)

The scoring for SF-36 v2 instrument will be done according to the procedures described by the developer and license owner in the instruction manual. Scores for the scales (HT, GH, PF, RP, BP, MH, SF, RE, VT) as well as the summary scores (PCS, MCS) will be derived.

Given SF-36 is collected at Week 23 and at Week 24, an assessment will be derived as the mean value of the two assessments at Week 23 and Week 24, which will be labeled as Week 24. In case only one assessment is available, this only assessment will be used as Week 24 assessment.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment. At each post-baseline assessment, the change in score from baseline will be calculated for each scale.

### 4.2.4. EQ-5D-5L

A unique EQ-5D-5L health state is defined by combining 1 level from each of the 5 dimensions: this defines a profile that is primarily reported as a 5-digit number, for instance 11221. A total of 3125 possible health states are defined in this way. For example, state 11111 indicates no problems on any of the 5 dimensions, while state 12345 indicates no problems with mobility, slight problems with washing or dressing, moderate problems with doing usual activities, severe pain or discomfort, and extreme anxiety or depression.

The instrument was specifically designed to provide an overall single number, called a weighted index, for each of the health states resulting from the combination of item responses (Dolan, 1997). The weighted index constitutes a measure of utility, an economics term used to describe consumer preferences or in the present case patient preferences for different HRQoL states. The weighted index can be only derived from patients who have provided a complete 5-response profile. A higher index indicates better QoL.

(Devlin, 2018) published the value set for England (available on the EuroQoL website (http://www.euroqol.org/about-EQ-5D/valuation-of-EQ-5D/EQ-5D-5l-value-sets.html). However, in August 2018, the UK's National Institute of Health and Care Excellence (NICE) has published a position statement (NICE position statement, available at https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisal-guidance/eq5d5l_nice_position_statement.pdf) advising companies, academic groups, and others preparing evidence submissions to NICE not to use the England validation set to derive utility values for their evidence submissions. In their position statement, NICE recommends using the mapping function developed by (van Hout & al., 2012) to be used (presented in Appendix D. EQ-5D-5L). This is the algorithm to be used in this study.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment for the EQ-5D-5L VAS and utility index scores as well as for the individual domain items.

### 4.2.5. Patient Global Impression of Severity (PGIS) and Change (PGIC)

The PGIS and PGIC are single-item questionnaires. Change from baseline defined as post-baseline value minus baseline value will be calculated for the PGIS assessment. At each post-baseline assessment, the change from baseline will be calculated.

Post-baseline PGIS raw scores will be classified according to the following item response categories:

- Worsening 3 points compared to baseline
- Worsening 2 points compared to baseline
- Worsening 1 point compared to baseline

- Stable
- Improved 1 point compared to baseline
- Improved 2 points compared to baseline
- Improved 3 points compared to baseline.

## 4.3. TIME TO PRO DETERIORATION

Time to clinically meaningful symptom worsening or HRQoL deterioration (PRO deterioration) will be analysed separately for each scale of the PRO instruments collected in this study, as appropriate and as described in Section 5. For convenience, a generic term "time to clinically meaningful deterioration" will be used both for symptom worsening and HRQoL deterioration, with an understanding of a specific meaning depending on the scale or subscale analysed. Two definitions will be used and described in the following sections:

- Time to first deterioration (TTFD)
- Time to definitive deterioration (TTDD)

### 4.3.1. TIME TO FIRST CLINICALLY MEANINGFUL DETERIORATION (TTFD)

TTFD will be defined as the duration of time from the date of randomization to the date of the first deterioration in PRO scores of at least one threshold-point (see section 4.2 for the definition of deterioration) as compared to the baseline score.

For those patients who experienced a first clinically meaningful deterioration, TTFD will be computed as follows and then converted to months:

TTFD = Date of assessment when first clinically meaningful deterioration of at least one threshold unit was observed – Date of randomization + 1

Patients who did not experience clinically meaningful deterioration will be censored at the date of the last available PRO assessment (i.e., date of the last non-missing value). Patients with no baseline assessment or patients with no post-baseline assessments will be censored at the date of randomization.

### 4.3.2. TIME TO DEFINITIVE CLINICALLY MEANINGFUL DETERIORATION (TTDD)

TTDD will be defined as the duration of time from the date of randomization to the date of the first deterioration in PRO scores of at least one threshold-point (see section 4.2 for the definition of deterioration) as compared to the baseline score if the deterioration is also observed at all time points thereafter (e.g., after the first deterioration is observed) or if the patient dropped out after deterioration, resulting in missing data.

For those patients who experienced a definitive meaningful deterioration, TTCD will be computed as follows and then converted to months:

TTCD = Date of assessment when first definitive clinically meaningful deterioration was observed – Date of randomization + 1

Patients who did not experience definitive clinically meaningful deterioration will be censored at the date of the last available PRO assessment (i.e., date of the last non-missing value). Patients with no baseline assessment or patients with no post-baseline assessments will be censored at the date of randomization.

# 5. STATISTICAL METHODOLOGY

## 5.1. GENERAL CONSIDERATIONS

Continuous data will be described by the number of observations (N), the number of missing observations (Nmiss), mean, SD, median, minimum (min), and maximum (max). Categorical data will be described by the number (n) and percentage (%) of patients in each category. Missing and invalid observations will be tabulated as separate categories. The calculation of proportions will not include the missing/invalid category.

Statistical comparisons will be made using two-sided tests at the $\alpha = 0.05$ significance level unless stated otherwise. Due to the exploratory nature of the analyses, adjustments for multiple comparisons will not be made.

Unless otherwise specified, all summaries will be presented by treatment group.

## 5.2. PATIENT DISPOSITION

Patient disposition is included in the clinical SAP and will not be repeated herein.

## 5.3. PRO COMPLETION

Instrument completion rate at each timepoint will be reported for each instrument on the ITT population. The following will be provided:

- The number of patients expected to complete a PRO assessment at each timepoint
- Unadjusted completion rate at each timepoint will be calculated as the number of patients meeting at least the minimum requirements for scoring of the instrument divided by the number of patients in the ITT population.
- Adjusted completion rate at each timepoint will be calculated as the number of patients meeting at least the minimum requirements for scoring of the instrument divided by the number of patients who are expected to have PRO assessments.

A patient is expected to complete the PRO instrument as long as the patient is still on treatment and has not discontinued the study.

The minimum requirements for scoring the instruments is as follows:

- CARES: at least 50% of items are answered (e.g. 10 out of the 19 items)
- PROMIS: all 16 items are answered
- SF-36: at least one domain can be calculated
- EQ-5D: at least one item is answered
- PGIC and PGIS: item answered

The completion rates by treatment group at each analysis visit will also be provided graphically by means of a line graph.

## 5.4. Demographics and Baseline Characteristics

Demographics and baseline characteristics are included in the clinical SAP and will not be repeated herein.

## 5.5. Descriptive Analyses

### 5.5.1. Item Level

For all items from the following instruments :CARES, PROMIS Fatigue Short Form 13a and selected items from the PROMIS Item Bank and PROMIS interest in sexual activity item, EQ-5D-5L, PGIS and PGIC, the following will be provided:

- A table with the distribution of response by treatment group for each analysis visit

For the PGIS, a table with the distribution of change in response categories (as defined in section 4.2.5) by treatment group for each analysis visit.

In addition, the following graphical representations will be provided:

- A stacked column chart of the distribution of response by treatment group at each analysis visit
- For the PGIS, a stacked column chart of the distribution of change in response categories (as defined in section 4.2.5) by treatment group.

### 5.5.2. Domain and Overall Score Level

The following domain and overall scores will be analysed:

- CARES: TSS and all other potential subscales resulting from the psychometric analysis
- PROMIS Fatigue: total, symptoms and impact domains from the 13-item set, as well as other potential subscales resulting from the psychometric analysis
- SF-36:
    - Domains: PF, RP, BP, GH, VT, SF, RE and MH
    - Summary scores: PCS and MCS
- EQ-5D-5L: VAS and utility index

The following analyses will be presented on the ITT population:

- A table with descriptive statistics for the PRO scores and a line graph with mean values and corresponding 95% CI will be presented by treatment group and each PRO assessment

- A table with descriptive statistics for the change from baseline in PRO scores will be presented by treatment group and each post-baseline PRO assessment
- A table showing the proportion of patients achieving ≥0%, 10%, 20%, 30%, 40%, 50% and 60% percentage change from baseline during the study will be presented by treatment group. Bar plots depicting the same proportions by treatment group will be created separately for each time point during the study
- A CDF plot showing a continuous plot of the absolute change from baseline during the study for the PRO scores on the X-axis and the cumulative percent of patients experiencing that change on the Y-axis will be presented by treatment group and each PRO assessment

## 5.6. Longitudinal Analysis of Change from Baseline

Change from baseline to Week 13 and Week 24 in PRO scores will be analyzed using a restricted maximum likelihood (REML)-based repeated measures approach (MMRM – Mixed Model Repeated Measures) (Brown & Prescott, 2006) as described in the clinical SAP.

The MMRM assumes that the missing observations are missing at random (MAR). That is, MMRM assumes that, given the statistical model and given the observed values of the outcome, missingness is independent of the unobserved values. A corollary is that MAR assumes that a subject's missing values can be estimated based on similar subjects who remained in the study. This infers that withdrawals (who may not receive study medication) have similar symptoms to some who continue to be treated. Given the expectation that PRO scores could be lower (poorer quality of life and/or more disease-related symptoms) after discontinuing the study medication, the MAR's assumption of the similarity of withdrawals and those who stay in the study may not be realistic for all subjects. To address the possibility of the data being missing not at random (MNAR) (e.g., non-ignorable missing data), a second analysis may be implemented using a PMM with sequential modelling with multiple imputation using placebo-based imputation (O'Kelly & Ratitch, 2014), depending on the amount of missing data at each visit. Further details on this exploration can be found in Section 5.9.2. The PMM will be used to assess the robustness of the MMRM estimate with regard to missing data when the MAR assumption is replaced by assumptions that are likely to be relatively less favorable to the experimental treatment. Due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, an immediate discontinuation of dosing was announced on March 18, 2020 to ensure the safety of trial patients. Therefore it is expected that a large number of patients will not have data at Week 24.

Analyses included in the clinical SAP will not be repeated here. The analyses will be performed on the ITT population and for each of the following PRO scores:

- CARES: TSS and other potential subscales resulting from the psychometric analysis
- PROMIS Fatigue: total, symptoms and impact domain from the 13-item set, as well as other potential subscales resulting from the psychometric analysis
- SF-36:

- Domains: PF, RP, BP, GH, VT, SF, RE and MH

- EQ-5D-5L: VAS and utility index.

### 5.6.1. MMRM

All PRO assessments will be included for this analysis The analysis will be based on observed data, i.e., data collected at each timepoint without carrying forward previous values. Data from a limited number of PRO assessments may be used in case of substantial dropout (i.e., analysis will be limited to timepoints at which at least 10% of patients have non-missing data in both treatment groups).

The response variable will be the change from baseline to each PRO assessment. The model will include the treatment arm (ziltivekimab vs. placebo) and timepoint (Week 5, Week 13, and week 24) as fixed-effect categorical factors, the baseline PRO score and stratification factors (baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5)), as well as the baseline PRO score x time and treatment x time interactions. Both main effects and the interaction terms will remain in the model, regardless of significance. The model will present least squares (LS) mean estimates for each dose, least squares mean differences of each study drug dose from placebo, standard errors, 95% CIs and p-values (where applicable) for mean changes from baseline to each visit. A plot of the LS means accompanied by the 95% CI will be produced.

In addition, an overall adjusted mean estimate will be derived that will estimate the average change from baseline across all time points, giving each visit equal weight.

The standardized mean difference (SMD) including 95% CI (Hedges' g) will also be provided.

The MMRM analysis will be conducted using PROC MIXED in SAS. The model will assume unstructured covariance among the within-patient repeated measurements. If the algorithm does not converge, a heterogeneous Toeplitz (the TOEPH option in SAS PROC MIXED) will be tried first and then AR(1) as a covariance structure to achieve convergence. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom.

Variables listed as categorical in the list above will be included in the CLASS statement of the procedure. The unique patient identifier will also be included as a class variable. A REPEATED statement over the visits will be included with the unique patient identifier as the SUBJECT variable in the REPEATED statement.

The normality and homoscedasticity of the residuals will be visually checked. Particularly, the scatter plot of the residuals versus the predicted endpoint values, the histograms, and the normal probability plots of the residuals will be reviewed. Transformation of the raw data will be considered if needed.

### 5.6.2. PMM

Change from baseline in PRO domain and overall scores may be further analyzed using a pattern-mixture model using sequential modelling with multiple imputation as described by O'Kelly (O'Kelly & Ratitch, 2014), depending on exploration of the amount of missing data at post-baseline assessments, as described in Section 5.9.2.

The results from this analysis will be used to judge the validity of the MAR assumption. Similar conclusions from MMRM and PMM would suggest that the results are not overly dependent on the assumptions of the primary analysis with regard to the missing data.

The sensitivity analysis will consider dropout reasons while imputing missing values after the discontinuation. Subjects who discontinued due to lack of efficacy or adverse events in the active arms are assumed to have no treatment effect after the discontinuation. These subjects are assumed to copy the profile in the placebo arm, and missing values are imputed based on the distribution estimated from the placebo group under the missing not at random (MNAR), using copy-reference approach. The rest of missing values in the placebo arm and active arms will be imputed using the observed data in their respective group under the MAR assumption. This includes the patients who were early terminated due to to the impact of the Coronavirus Disease 2019 (COVID19) pandemic. The multiple imputation model will include factors such as treatment arm, baseline hemoglobin category, and CKD stage, in addition to the data outcomes at each visit.

The following steps will be performed:

**Step 1** Intermittent (non-monotone) missing data will be imputed using the MCMC option of SAS PROC MI (using seed=5414).

**Step 2** For monotone missing data, the values for each pattern will be imputed via the chained equation method with range restriction, using SAS PROC MI option MONOTONE REG ().

Under step 2, missing values for subjects who discontinued due to lack of efficacy or adverse events will be imputed under MNAR assumption using copy-reference approach, and missing values for subjects who discontinued due to reasons other than lack of efficacy or adverse events will be imputed under the MAR assumption.

These two steps will be carried out sequentially to construct 100 hypothetical complete data sets.

An analysis of covariance (ANCOVA) model (not repeated measures, and with no random effects) with the following covariates will be performed for each imputation using the SAS Mixed procedure:

- HGB = Baseline Hemoglobin category (>=11 or <11g/dL)
- CKD = CKD Stage (3, or 4/5)
- TRT = Treatment group
- BASE = Baseline hemoglobin value

The SAS MIAnalyze procedure will be used to combine the results of these analyses for the imputations. The overall least square means, standard errors, 95% CI, and p-values will be reported.

Reasons for subject discontinuation that are identified as lack of efficacy will be discussed and finalized prior to database lock.

## 5.7. RESPONDER ANALYSIS

The proportion of patients with improvement, who were stable, or who deteriorated (as defined in section 4.2 using both the primary and the potential sensitivity thresholds if these are suggested by the psychometric analysis) will be summarized at each PRO assessment visit by treatment arms. The denominator in this descriptive analysis will be the number of patients with non-missing data at the particular visit. The analysis will be performed on the ITT population and include only subjects who have an assessment at baseline and at least one post-baseline assessment.

The proportion of patients with definitive deterioration and transient deterioration during the course of the study will be summarized by treatment arms. The denominator in this descriptive analysis will be the number of patients with at least one deterioration during the study.

The analysis will be performed on the ITT population and for each of the following PRO scale scores:

- CARES: TSS and other potential subscales resulting from the psychometric analysis
- PROMIS Fatigue: total, symptoms and impact domain from the 13-item set, as well as other potential subscales resulting from the psychometric analysis

## 5.8. TIME TO EVENT ANALYSIS

For all time to event analyses, the time to deterioration will be analysed in months and will be presented by treatment group.

TTFD and TTCD will be defined as explained in Section 4.3. The non-parametric Kaplan-Meier method will be used to estimate the survival curves for TTFD and TTCD for each PRO scale score. The 50th percentile of Kaplan-Meier estimates will be used to estimate the median duration of TTFD and TTCD. A two-sided 95% CI will be provided for these estimates. The treatment difference in survival will be assessed by the stratified log-rank test (baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5)). Separate log-rank tests will be performed for each of the experimental arms vs. placebo. A Kaplan-Meier plot by treatment group will be presented.

Kaplan-Meier analysis will be performed using PROC LIFETEST (SAS procedure).

A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie., the hazard ratio), using baseline hemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5) as stratification factor. The hazard ratio and its 95% CI from the stratified Cox model for each of the experimental arms vs. placebo will be reported.

The analysis will be performed on the ITT population and for each of the following PRO scale scores:

- CARES: TSSE and other potential subscales resulting from the psychometric

analysis

- PROMIS Fatigue: total, symptoms and impact domain from the 13-item set, as well as other potential subscales resulting from the psychometric analysis

## 5.9. HANDLING OF MISSING DATA

### 5.9.1. MISSING ITEMS

In case of missing items, the scores will be calculated as indicated in the scoring manuals for each of the PRO instruments.

### 5.9.2. MISSING FORMS

It is anticipated that the great majority of missing data in this study will have a monotone pattern, meaning that once a patient has missing data at one visit, data will be missing at all subsequent visits. There may be some small amount of intermittent (non-monotone) missing data (when patient skips intermediate visits but return for evaluations at subsequent visits). The number and percentage of patients for each of the missing data patterns (no missing data, monotone missing data, and intermittent missing data) will be presented by treatment group.

Tabular summaries for the percentage of patients by the reason for discontinuation of study treatment, as well as for withdrawal from the study, are provided in the clinical SAP and will not be repeated herein. A plot of the mean PRO score for selective scores (e.g., CARES TSS and other potential subscales resulting from the psychometric analysis, PROMIS Fatigue: total, symptoms and impact scales from the 13-item set, as well as other potential subscales resulting from the psychometric analysis) over time by selected categories of discontinuation (including completers) will be provided. The reasons of discontinuation will be grouped as follows:

- Death
- Adverse event
- Disease relapse, lack of efficacy, and progressive disease
- Other (lost to follow-up, non-compliance with study drug, pregnancy, protocol deviation, recovery, site terminated by sponsor, technical problems, withdrawal by parent/guardian, withdrawal by subject, other)
- COVID19 pandemic.

# 6. ANALYSIS SOFTWARE

All data processing, summarization, and analyses will be performed using SAS Version 9.4 (SAS Institute, North Carolina) or higher.

# 7. REFERENCES

Brown, H., & Prescott, R. (2006). *Applied Mixed Models in Medicine* (2nd ed.). John Wiley and Sons Ltd.

Devlin, e. a. (2018). An EQ-5D-5L value set for England. *Valuing health-related quality of life*, 7-22.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care 35*, 1095-1108.

Erez, G., Selman, L., & Murtagh, F. (2016). Measuring health-related quality of life in patients with conservatively managed stage 5 chronic kidney disease: limitations of the Medical Outcomes Study Short Form 36: SF-36. *Qual Life Res.*, 2799–2809.

Leaf, D. E., & Goldfarb, D. S. (2009). Interpretation and review of health-related quality of life data in CKD patients receiving treatment for anemia. *Kidney International*, 15-24.

O'Kelly, M., & Ratitch, B. (2014). *Clinical Trials with Missing Data: A Guide for Practitioners.* John Wiley & Sons, Ltd.

*Patient-Reported Outcomes Measurement Information System Fatigue scoring manual.* (n.d.). Retrieved from Health Measures: http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Fatigue_Scoring_Manual.pdf

Pergola, P. E., Pecoits-Filho, R., & Winkelmayer, W. C. (2019). Economic Burden and Health-Related Quality of Life Associated with Current Treatments for Anaemia in Patients with CKD not on Dialysis: A Systematic Review. *PharmacoEconomics*, 463-478.

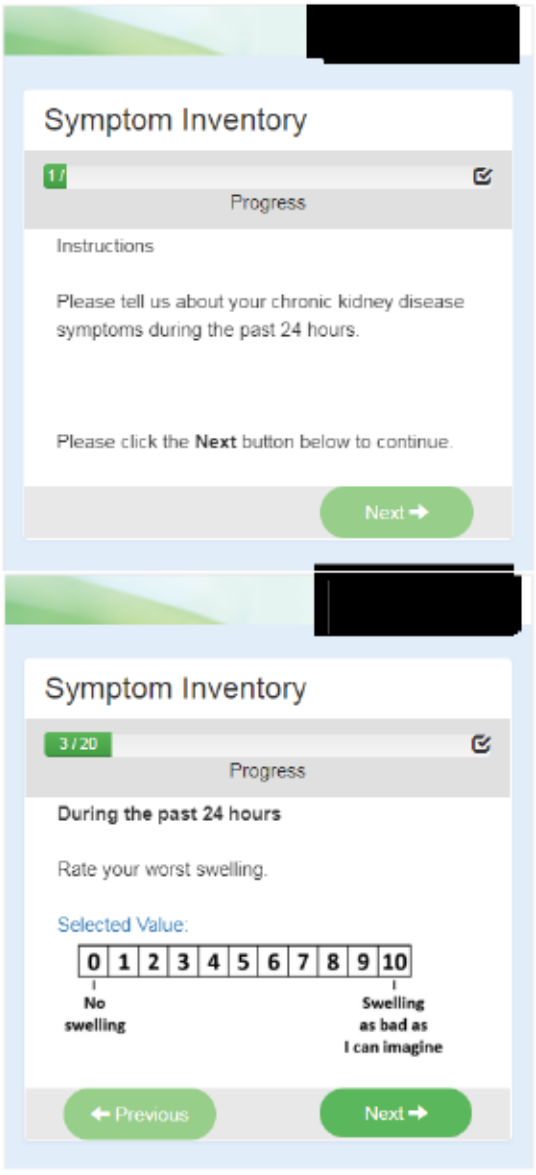
QualityMetric. (2011). User's manual for the SF-36v2 Health Survey (3 ed.). Lincoln, RI.

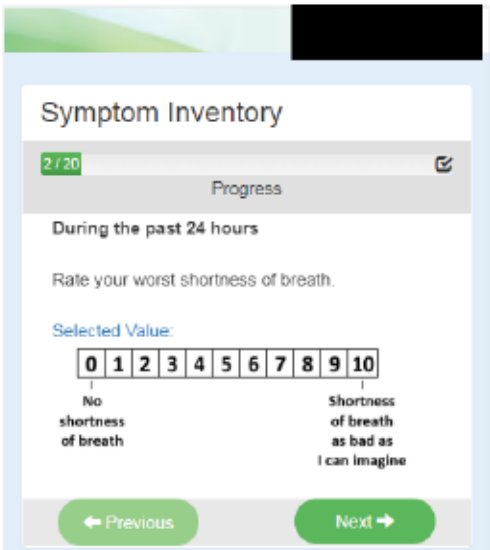
Samsa, G., Edelman, D., & Rothman, M. L. (1999). Determining Clinically Important Differences in Health Status Measures. A General Approach with Illustration to the Health Utilities Index Mark II. *Pharmacoeconomics*, 141-155.

van Hout, B., & al., e. (2012). Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. *Value in Health, 25*, 708-715.

# 8. APPENDICES

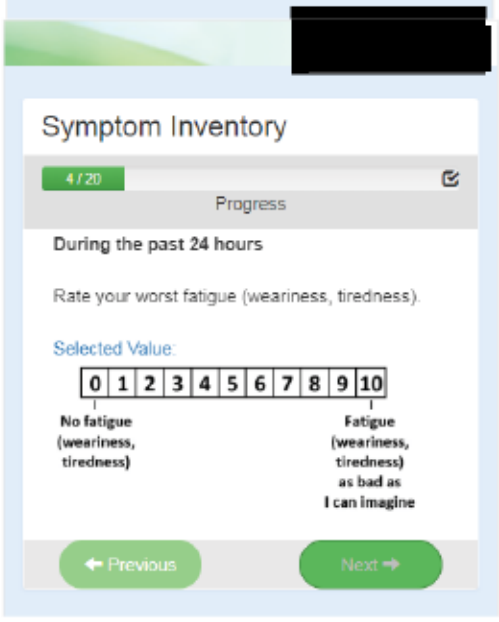
## 8.1. APPENDIX A. CARES

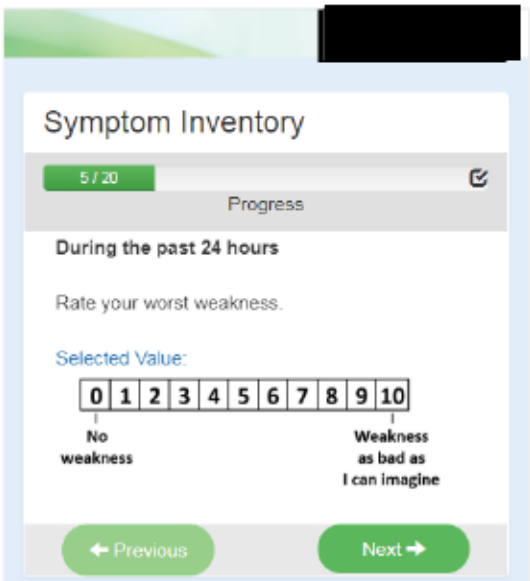


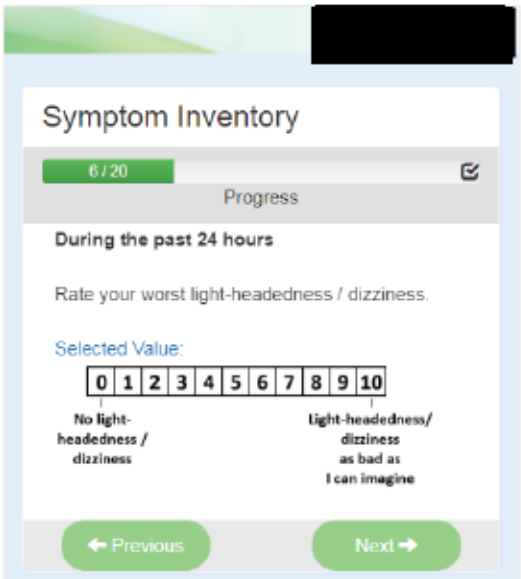



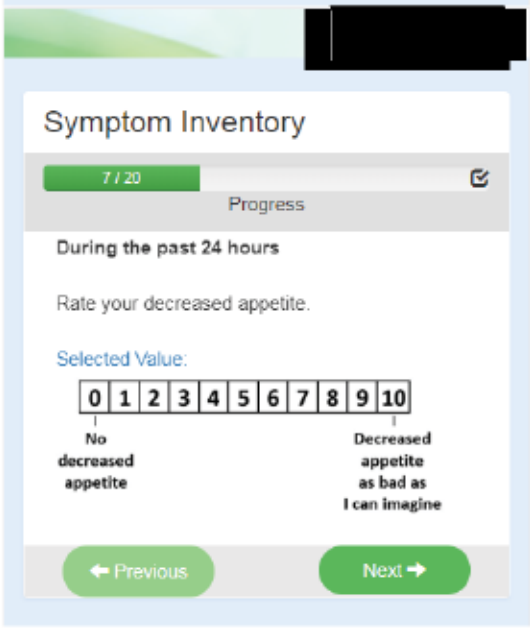


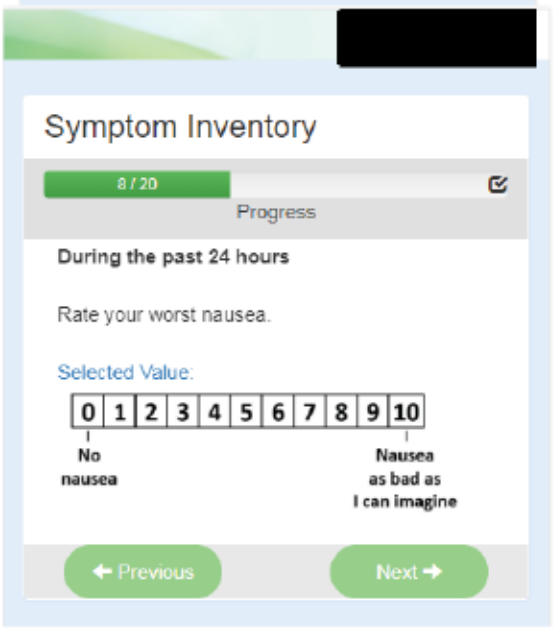


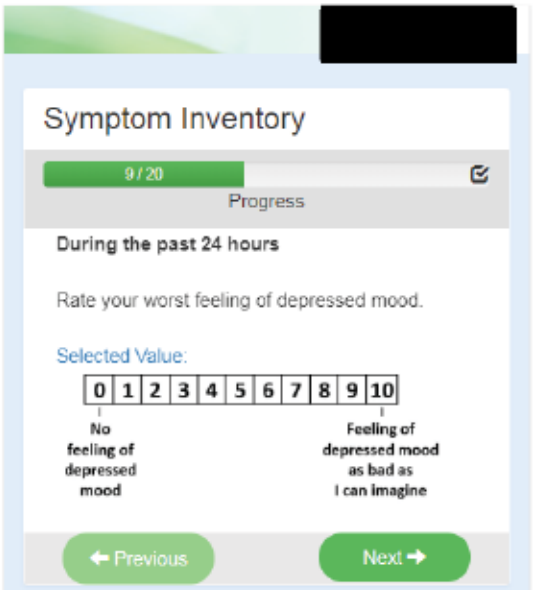


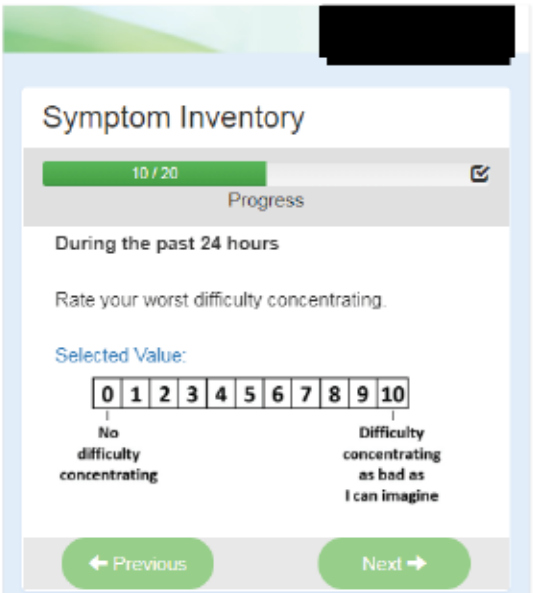


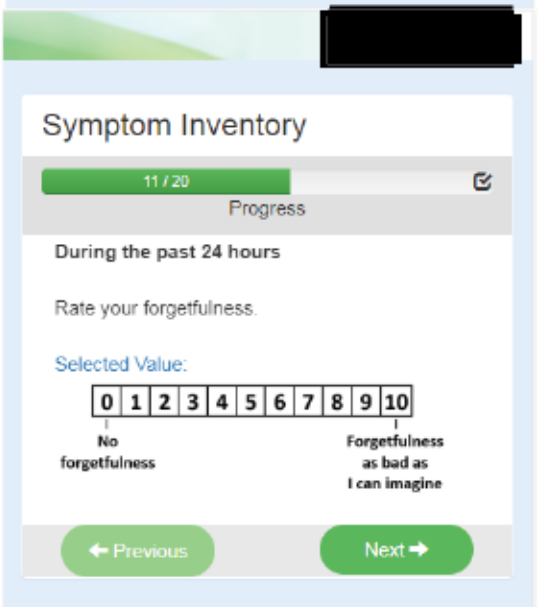



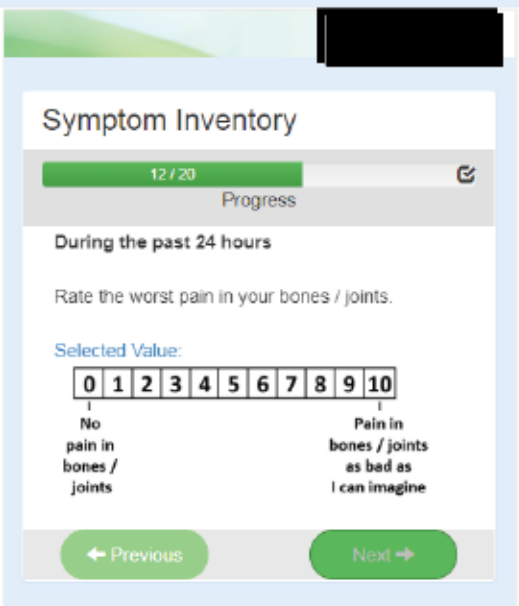



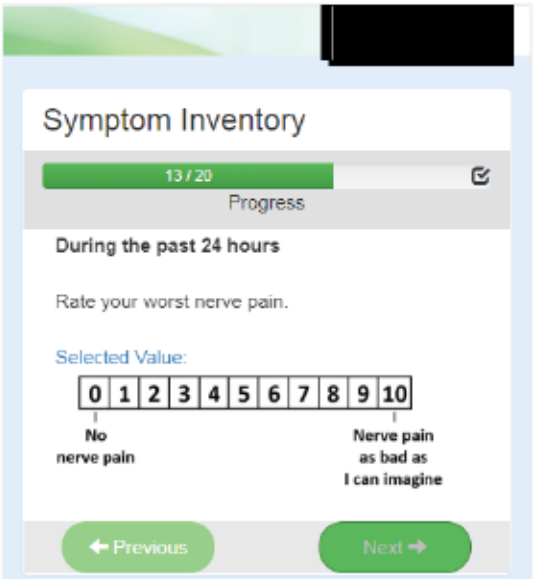



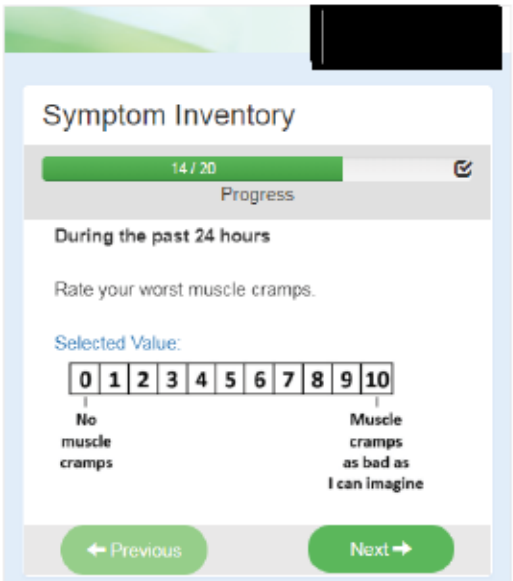


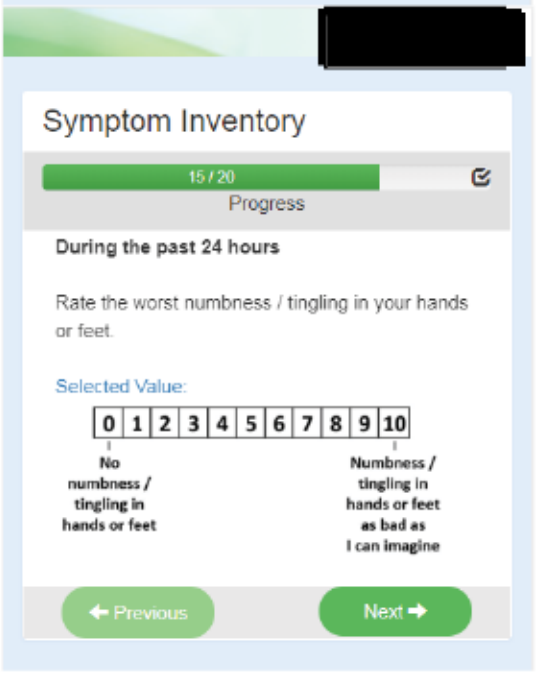



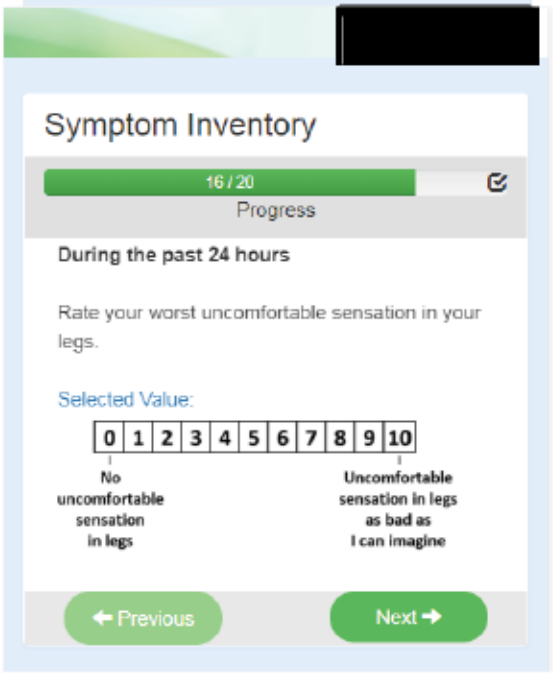



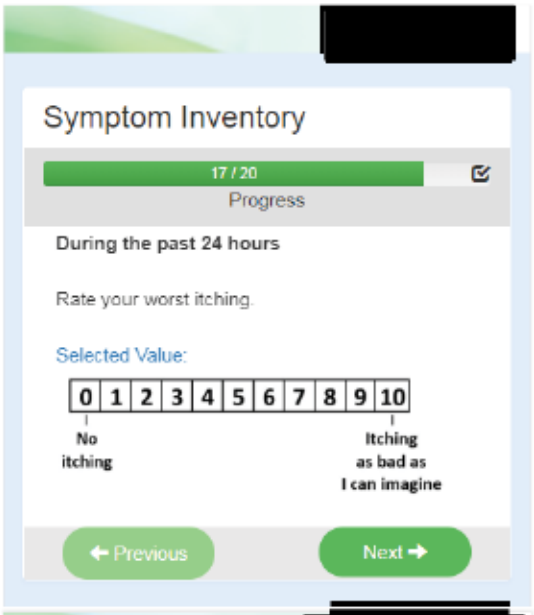



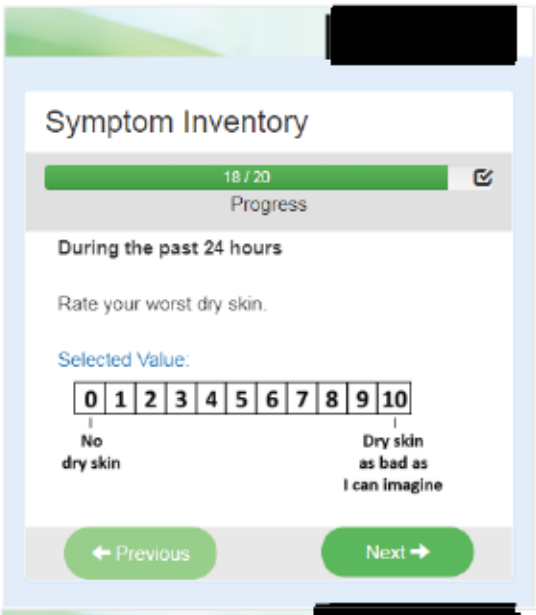


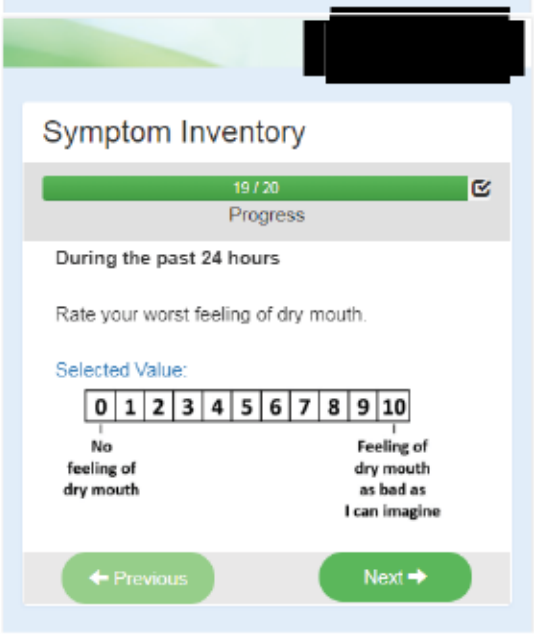



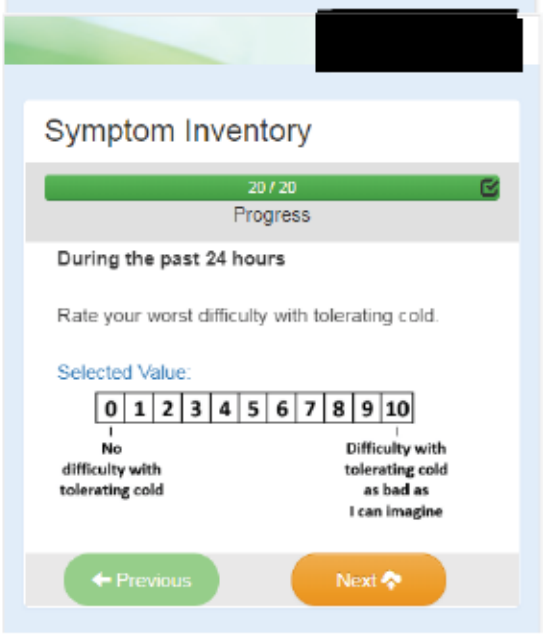





## 8.2. APPENDIX B. PROMIS FATIGUE 13A SHORT FORM, ADDITIONAL FATIGUE ITEMS, SEXUAL ACTIVITY ITEM

### PROMIS Items

PROMIS Short Form v1.0 – Fatigue 13a

| During the past 7 days..... | Not at all | A little bit | Somewhat | Quite a bit | Very much |
|---|---|---|---|---|---|
| I feel fatigued | 1 | 2 | 3 | 4 | 5 |
| I feel weak all over | 1 | 2 | 3 | 4 | 5 |
| I feel listless ("washed out") | 1 | 2 | 3 | 4 | 5 |
| I feel tired | 1 | 2 | 3 | 4 | 5 |
| I have trouble starting things because I am tired | 1 | 2 | 3 | 4 | 5 |
| I have trouble finishing things because I am tired | 1 | 2 | 3 | 4 | 5 |
| I have energy | 1 | 2 | 3 | 4 | 5 |
| I am able to do my usual activities | 1 | 2 | 3 | 4 | 5 |
| I need to sleep during the day | 1 | 2 | 3 | 4 | 5 |
| I am too tired to eat | 1 | 2 | 3 | 4 | 5 |
| I need help doing my usual activities | 1 | 2 | 3 | 4 | 5 |
| I am frustrated by being too tired to do the things I want to do | 1 | 2 | 3 | 4 | 5 |
| I have to limit my social activity because I am tired | 1 | 2 | 3 | 4 | 5 |

PROMIS items (Item identifier in bold)

| | | | | | |
|---|---|---|---|---|---|
| **FATEXP36:** How exhausted were you on average? | 1 | 2 | 3 | 4 | 5 |
| **FATEXP43:** How physically drained were you on average? | 1 | 2 | 3 | 4 | 5 |
| **FATIMP52:** To what degree did your fatigue make you feel less alert? | 1 | 2 | 3 | 4 | 5 |
| **During the past 30 days.....** | **Not at all** | **A little bit** | **Somewhat** | **Quite a bit** | **Very much** |
| **SFINT101:** How interested have you been in sexual activity? | 1 | 2 | 3 | 4 | 5 |

## 8.3. APPENDIX C. SF-36

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| | Your Health and Well-Being | | | | | | |
| | This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. Thank you for completing this survey!<br><br>For each of the following questions, please select the one response that best describes your answer. | | | | | | |
| SF36v2_GH1 | In general, would you say your health is: | Excellent | Very good | Good | Fair | Poor | |
| SF36v2_HT | Compared to one year ago, how would you rate your health in general now? | Much better now than one year ago | Somewhat better now than one year ago | About the same as one year ago | Somewhat worse now than one year ago | Much worse now than one year ago | |
| | The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? | | | | | | |
| SF36v2_PF01 | Does your health now limit you in vigorous activities, such as running, lifting heavy objects, participating in strenuous sports? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF02 | Does your health now limit you in moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF03 | Does your health now limit you in lifting or carrying groceries? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF04 | Does your health now limit you in climbing several flights of stairs? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF05 | Does your health now limit you in climbing one flight of stairs? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF06 | Does your health now limit you in bending, kneeling, or stooping? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF07 | Does your health now limit you in walking more than a mile? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF08 | Does your health now limit you in walking several hundred yards? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF09 | Does your health now limit you in walking one hundred yards? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF10 | Does your health now limit you in bathing or dressing yourself? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| | During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health? | | | | | | |

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| SF36v2_RP1 | During the past 4 weeks, how much of the time have you cut down on the amount of time you spent on work or other activities as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP2 | During the past 4 weeks, how much of the time have you accomplished less than you would like as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP3 | During the past 4 weeks, how much of the time were you limited in the kind of work or other activities as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP4 | During the past 4 weeks, how much of the time have you had difficulty performing the work or other activities as a result of your physical health (for example, it took extra effort)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| | During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? | | | | | | |
| SF36v2_RE1 | During the past 4 weeks, how much of the time have you cut down on the amount of time you spent on work or other activities as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RE2 | During the past 4 weeks, how much of the time have you accomplished less than you would like as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RE3 | During the past 4 weeks, how much of the time have you done work or other activities less carefully than usual as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_SF1 | During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups? | Not at all | Slightly | Moderately | Quite a bit | Extremely | |
| SF36v2_BP1 | How much bodily pain have you had during the past 4 weeks? | None | Very mild | Mild | Moderate | Severe | Very severe |
| SF36v2_BP2 | During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)? | Not at all | A little bit | Moderately | Quite a bit | Extremely | |
| | These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. | | | | | | |

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| SF36v2_VT1 | How much of the time during the past 4 weeks did you feel full of life? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH1 | How much of the time during the past 4 weeks have you been very nervous? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH2 | How much of the time during the past 4 weeks have you felt so down in the dumps that nothing could cheer you up? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH3 | How much of the time during the past 4 weeks have you felt calm and peaceful? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT2 | How much of the time during the past 4 weeks did you have a lot of energy? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH4 | How much of the time during the past 4 weeks have you felt downhearted and depressed? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT3 | How much of the time during the past 4 weeks did you feel worn out? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH5 | How much of the time during the past 4 weeks have you been happy? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT4 | How much of the time during the past 4 weeks did you feel tired? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_SF2 | During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| | How TRUE or FALSE is each of the following statements for you? | | | | | | |
| SF36v2_GH2 | I seem to get sick a little easier than other people. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH3 | I am as healthy as anybody I know. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH4 | I expect my health to get worse. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH5 | My health is excellent. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| | SF-36v2® Health Survey © 1992, 2000, 2009 Medical Outcomes Trust and QualityMetric Incorporated. All rights reserved. SF-36® is a registered trademark of Medical Outcomes Trust. (SF-36v2® Health Survey Standard, United States (English)) | | | | | | |

## 8.4. APPENDIX D. EQ-5D-5L

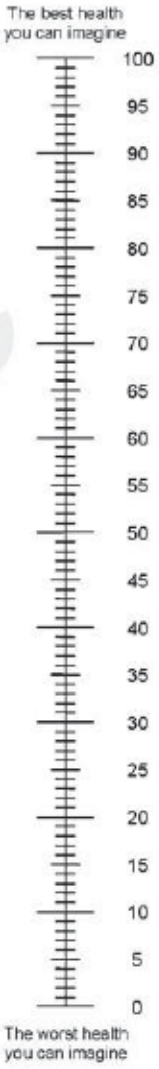
| | |
|---|---|
| EQ-5D-5L | |
| EQ-5D-5L PDA version | |
| English (USA)<br>**Health Questionnaire**<br>**English version for the USA** | Country (Language)<br>Health Questionnaire<br>Version (Target Language)<br>Version (English) |
| On the following screens please tap the statement that best describes your health TODAY. | Instruction |
| **Your mobility TODAY** | **Mobility** |
| I have no problems walking | MB1 |
| I have slight problems walking | MB2 |
| I have moderate problems walking | MB3 |
| I have severe problems walking | MB4 |
| I am unable to walk | MB5 |
| **Your self-care TODAY** | **Self-care** |
| I have no problems washing or dressing myself | SC1 |
| I have slight problems washing or dressing myself | SC2 |
| I have moderate problems washing or dressing myself | SC3 |
| I have severe problems washing or dressing myself | SC4 |
| I am unable to wash or dress myself | SC5 |
| **Your usual activities TODAY** *(e.g. work, study, housework, family or leisure activities)* | **Usual Activities** |
| I have no problems doing my usual activites | UA1 |
| I have slight problems doing my usual activities | UA2 |
| I have moderate problems doing my usual activities | UA3 |
| I have severe problems doing my usual activities | UA4 |
| I am unable to do my usual activities | UA5 |
| **Your pain / discomfort TODAY** | **Pain / Discomfort** |
| I have no pain or discomfort | PD1 |
| I have slight pain or discomfort | PD2 |
| I have moderate pain or discomfort | PD3 |
| I have severe pain or discomfort | PD4 |
| I have extreme pain or discomfort | PD5 |
| **Your anxiety / depression TODAY** | **Anxiety / Depression** |
| I am not anxious or depressed | AD1 |
| I am slightly anxious or depressed | AD2 |
| I am moderately anxious or depressed | AD3 |
| I am severely anxious or depressed | AD4 |
| I am extremely anxious or depressed | AD5 |
| We would like to know how good or bad your health is TODAY. | Vas Line 1 |
| On the next screen you will see a scale numbered 0 to 100. | Vas Line 2 |
| 100 means the best health you can imagine. | Vas Line 3 |
| 0 means the worst health you can imagine. | Vas Line 4 |
| Please tap on the scale to indicate how your health is TODAY. | Vas Line 5 |
| The best health you can imagine | Top Scale |
| The worst health you can imagine | Bottom Scale |
| YOUR HEALTH TODAY | Box Health |



***Disclaimer:*** *This is a preview of the EQ-5D instrument. It demonstrates the text, questions and response options included in this version. This preview does not represent the final product and should not be used as an official EQ-5D instrument.*

- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

YOUR HEALTH TODAY = ☐






3

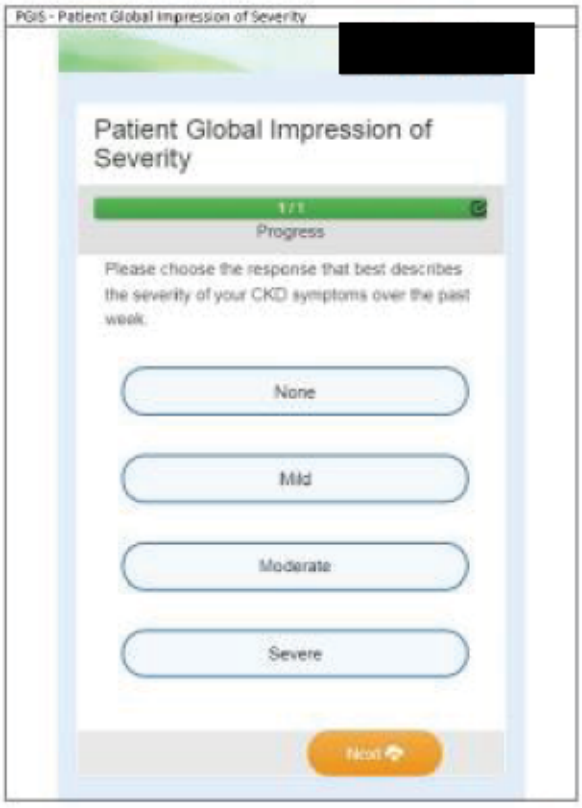


SAS code to calculate the utility index (based on the algorithm published by van Hout 2012) is presented in Table 3.

**Table 3: SAS code for deriving utility index from EQ-5D-5L for UK**

SAS syntax crosswalk values EQ-5D-5L Unite

## 8.5. APPENDIX E. PGIC AND PGIS




## Document Approval Signature(s)

This document has been signed electronically in compliance with [redacted] and [redacted] Companies' electronic signature policies and procedures. This page is a manifestation of the electronic signatures, which may be confirmed within the [redacted] and [redacted] Companies' source system in accordance with [redacted] and [redacted] Companies' records retention policies. Note: the Title field in the signature block has been intentionally left blank.

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

| Statistical Analysis Plan Addendum | |
|---|---|
| Drug Name | Ziltivekimab |
| Protocol Number | COR-001-02 |
| Edition Number | 1.0 |
| Date | 29 April 2020 |

# A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE)

**Investigational Product:** Ziltivekimab
**Protocol Number:** COR-001-02
**Original Protocol Version 1.0:** 05 November 2018
**Protocol Amendment 1:** 05 December 2018
**Protocol Amendment 2:** 24 January 2019
**Protocol Amendment 3:** 16 April 2019
**Protocol Amendment 4:** 25 June 2019
**Protocol Amendment 5:** 23 December 2019

**SAP Version:** 1.0
**SAP Date:** 31 March 2020

**SAP Addendum Version:** 1.0
**SAP Addendum Date:** 29 April 2020

## SAP ADDENDUM SIGNATURE PAGE

# A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE)

We, the understood, have understood and approved the SAP Addendum

| Signature | Date |
|---|---|
| [redacted] | [redacted] |
| [redacted] | [redacted] |
| [redacted] | [redacted] |
| [redacted] | [redacted] |

Corvidia Therapeutics, Inc.

# TABLE OF CONTENTS

# 1. INTRODUCTION

The statistical analysis plan (SAP), version 1.0, for study with protocol number COR-001-02 was finalized on March 31, 2020.

Due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, an immediate discontinuation of dosing was announced on March 18, 2020 to ensure the safety of trial patients. The announcement required the early termination visit to be scheduled as early as possible for ongoing patients followed by their entry into the safety follow-up period. As a result, changes in the planned analysis were documented in Section 5 ("Changes from Protocol-specified Statistical Analyses") of the SAP version 1.0.

The interim database lock for this study is scheduled on 08 May, 2020. Further updates to the planned analysis are documented in this addendum.

Summary of changes:

- Updated definition to clarify baseline for exploratory efficacy endpoints and safety endpoints.
- Inclusion of High-Density Lipoprotein (HDL) as an exploratory efficacy endpoint.
- End of Treatment (EOT) definition for CARES Corvidia ePRO and other Patient Reported Outcome (PRO) endpoints.

# 2. DATA ANALYSIS CHANGES

## 2.1 Definition of Baseline for Exploratory Efficacy Endpoints and Safety Endpoints

The SAP version 1.0 defines the baseline value for the primary efficacy endpoint hs-CRP, secondary efficacy endpoint Serum Amyloid A, and exploratory efficacy endpoints Hemoglobin, Serum Albumin, and NT-pro-BNP as the average of lab value at Day 1 and prior assessment(s). For the CARES Corvidia ePRO endpoint, baseline is defined as the average of 7 consecutive measurements (Day -7 to Day -1) prior to first dose initiation.

Similarly, baseline for all exploratory efficacy endpoints will be defined as follows:

- For endpoints where lab assessments were collected at multiple visits prior to the initiation of study drug, baseline will be calculated as average of all values prior to the first dose.
- For endpoints where lab assessments were collected at Day 1 only, baseline will be defined as the Day 1 value.

For all safety endpoints, baseline will be defined as the Day 1 value. If missing, the last evaluation prior to the initiation of the study drug will be considered the baseline.

## 2.2 Lipid Parameters for Secondary Efficacy Analysis

For lipid parameters, the SAP version 1.0 indicates difference in percent change in LDL-C, Triglycerides, ApoB, ApoA1, ApoB/ApoA1, and lipid profile by NMR spectroscopy from Baseline to Week 13 and to the End of Treatment (Weeks 23 through 24) between each active group and Placebo.

This SAP addendum further clarifies that HDL-C will also be included in this list of endpoints. Descriptive summary by visit as well as statistical analysis using mixed models and analysis of covariance for HDL-C will also be performed along with the specified lipids.

## 2.3 End of Treatment: CARES Corvidia ePRO and other PRO outcomes

According to section 2.3.4.4 of the SAP version 1.0, a bi-weekly score (the End of Treatment) for the CARES Corvidia ePRO will be derived as the average of Day 155 to Day 168. If $\geq 8$ daily scores out of the 14 days (>50%) within the bi-weekly period are missing, then the score is set to missing.

However, for patients who discontinued due to COVID19 after Week 13 and did not complete the end of treatment visit or did not enter their data on their personal device, the end of treatment will be assigned using the average of values from Visit 7 (Day 78 to Day 84). If $\geq 4$ daily scores out of the 7 days (>50%) within the weekly period are missing, then the score is set to missing.

For PRO outcomes PROMIS Fatigue 13a Short Form and Optum SF-36 v2®Health Survey where assessments are entered at Visit 1 (Baseline), Visit 4 (Week 5), Visit 7 (Week 13), Visit 10 (Week 23), and Visit 11 (Week 24), the end of treatment visit will be defined as the average of assessments at Week 23 and Week 24.

For PRO outcomes PROMIS Sexual Function and Satisfaction, EurQol-5D-5L, and Patient Global Impression of Change where assessments are entered at Visit 1, Visit 4 (Week 5), Visit 7 (Week 13), and Visit 11 (Week 24), the end of treatment visit will be defined as the assessment at Week 24.

For PRO outcome Patient Global Impression of Severity where assessments are entered at the end of Week -1, end of Week 4, end of Week 12, and at Week 24, the end of treatment visit will be defined as the assessment at Week 24.

For all PRO outcomes described above, for patients who discontinued due to COVID19 after Week 13, the end of treatment visit will be defined as the last assessment after Week 13. If there are no visits after Week 13, then Week 13 will be considered the End of Treatment visit.

# 3. DATA ANLAYSIS CLARIFICATION

The changes in analysis described in document were applied only to the lab parameters and PRO endpoints described as exploratory efficacy endpoints. This change will not impact the results for the primary endpoint as well as secondary endpoints.

# Psychometric Analysis Plan

## A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE)

Protocol COR-001-02

**Study Sponsor:**

Corvidia Therapeutics Inc.
35 Gatehouse Drive
Waltham MA 02451, USA

**Author:**

[redacted] Patient-Centered Endpoints

**Version Number and Date:** V2.0, 24APR2020





Approved on 28-Apr-2020 (Document ID: [redacted])

# Contents

# PSYCHOMETRIC ANALYSIS PLAN SIGNATURE PAGE

**Psychometric Analysis Plan V2.0 for Protocol COR-001-02.**

| | Name | Signature | Date |
|---|---|---|---|
| **Author:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |

Upon review of this document, the undersigned approves this version of the Psychometric Analysis Plan, authorizing that the content is acceptable for the reporting of this study.

| | Name | Signature | Date |
|---|---|---|---|
| **Approved By:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |
| | | | |
| **Approved By:** | [redacted] | [redacted] | [redacted] |
| **Position:** | [redacted] | | |
| **Company:** | Corvidia | | |
| | | | |

## MODIFICATION HISTORY

| Unique Identifier for this Version | Date of the Document Version | Author | Significant Changes from Previous Authorized Version |
|---|---|---|---|
| 1.0 | 31MAR2020 | [redacted] | Not Applicable – First Version |
| 2.0 | 24APR2020 | [redacted] | Revisions needed to account for the immediate discontinuation of dosing due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic |

## LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| ASCVD | Atherosclerotic Cardiovascular Disease |
| CARES | Core assessment of renal disease symptoms |
| CI | Confidence interval |
| CKD | Chronic kidney disease |
| ECDF | Empirical Cumulative Density Function |
| EFA | Exploratory factor analysis |
| EPDF | Empirical Probability Density Function |
| EQ-5D-5L | 5-Level EuroQol in 5 dimensions |
| FU | Follow-up |
| HRQoL | Health-related quality of life |
| hs-CRP | high-sensitivity C-reactive protein |
| ICC | Intraclass correlation coefficient |
| MCS | Mental Component Score |
| ML | Maximum likelihood |
| NDD-CKD | Non-dialysis-dependent chronic kidney disease |
| PAP | Psychometric analysis plan |
| PCS | Physical Component Score |
| PGIC | Patient Global Impression of Change |
| PGIS | Patient Global Impression of Severity |
| PRO | Patient-reported outcome |
| PROMIS | Patient-Reported Outcomes Measurement Information System |
| SAA | Serum amyloid A |
| SD | Standard deviation |
| SEM | Standard Error of Measurement |
| SF-36 v2 | Short Form 36-item health survey version |
| TSS | Total symptom score |
| ULS | Unweighted least squares |
| Ziltivekimab | Monoclonal antibody to IL-6 |

# 1. INTRODUCTION

This document describes the rules and conventions to be used in the presentation and analysis of patient-reported outcomes (PROs) data for the RESCUE study, Protocol Ziltivekimab-COR-001-02. It describes the data to be summarized and analyzed, including specifics of the psychometric analyses to be performed. The psychometric analysis plan (PAP) will be finalized prior to database lock. Any deviations from the PAP after database lock will be documented in the final psychometric analysis report.

This psychometric analysis plan (PAP) is based on protocol version 5.0 dated 23 December 2019 and Amendments 1 (dated 05 December 2018), 2 (dated 24 January 2019), 3 (dated 16 April 2019), 4 (dated 25 June 2019), and 5 (23 December 2019).

Due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, an immediate discontinuation of dosing was announced on March 18, 2020 to ensure the safety of trial patients. The announcement required the early termination visit to be scheduled as early as possible for ongoing patients followed by their entry into the safety follow-up period.

At the time of discontinuation of the study, approximately 140 patients had completed the 24-week treatment period and more than 120 patients were non-completers of the treatment period.

# 2. RESCUE STUDY OVERVIEW

## 2.1. STUDY OBJECTIVES

### 2.1.1. PRIMARY OBJECTIVE

The primary objective of the RESCUE study is:

- To evaluate the effects of Ziltivekimab compared to placebo on markers of inflammation such as high-sensitivity C-reactive protein (hs-CRP).

### 2.1.2. SECONDARY OBJECTIVES

The secondary objective is:

- To evaluate the effects of Ziltivekimab compared to placebo on two markers of inflammation and cardiovascular risk: serum amyloid A (SAA) and fibrinogen.

### 2.1.3. EXPLORATORY OBJECTIVES

Other objectives are:

- To determine the pharmacokinetic, exploratory pharmacodynamics, pharmacogenetics and effect of Ziltivekimab on inflammatory markers.
- To evaluate the effects of three dose levels of Ziltivekimab compared to placebo on patient reported outcomes (PRO): Patient-Reported Outcomes Measurement Information System (PROMIS®) Fatigue 13a short form, selected items from the PROMIS fatigue item bank, the Optum SF-36 v2® HealthSurvey, a Corvidia PRO, the PROMIS interest in sexual activity item, the patient global impression of change (PGIC), patient global impression of severity (PGIS), and the EQ-5D-5L.
- To evaluate the psychometric properties of the Corvidia electronic PRO (ePRO) items (*CARES*), *PROMIS Fatigue* 13a short form, and selected items from the PROMIS fatigue item bank, in Chronic Kidney Disease (CKD) patients.

## 2.2. SAMPLE SIZE

The primary efficacy endpoint is percent change from baseline in hs-CRP (average of the hs-CRP value prior to randomization and Day 1) to Week 13 between each active group and placebo.

Based on the observed treatment difference in percent change from baseline in hs-CRP of 60.74% between combined COR-001-01 active groups and placebo and the associated pooled SD of 16.893% in hs-CRP at Week 4 from the final analysis of study COR-001-SC1, a sample size of 54 per group yields more than 99% power with 2-sided alpha=0.05.

Taking into consideration the dropout rate of 10% by the end of the study, a sample size of 60 per group is planned for this study. Accordingly, approximately 240 patients will be randomized 1:1:1:1 (60 per each treatment group) into the trial. Patient randomization will be stratified by baseline hemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5).

## 2.3. STUDY DESIGN

The RESCUE study is a randomized, double-blind, placebo-controlled trial designed to evaluate the efficacy, safety, and pharmacokinetics of Ziltivekimab at three dose levels (7.5 mg, 15 mg or 30 mg) compared to placebo in patients with stage 3-5 CKD, not on dialysis, who have evidence of inflammation with high cardiovascular risk.

The primary, secondary, and exploratory endpoints will be analysed at 13 weeks of dosing and then followed for additional exploratory efficacy analyses through Week 24. Selected efficacy

endpoints and safety assessments will be evaluated in the Follow-up Period Week 25 though Week 32 (Figure 1).

Patients will undergo a Screening Period of up to 14-days during which inclusion and exclusion criteria will be evaluated. Patients who meet all inclusion criteria and no exclusion criteria will be randomized to one of three Ziltivekimab dose levels (7.5 mg, 15mg, or 30 mg) or placebo for a 24-week Treatment Period. Patient randomization will be stratified by baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4, or 5).

The patient will be randomized on Day 1 and the first dose of study drug should be administered after all assessments are conducted. Doses of study drug will be administered every 28-days for a total of 6 treatments (Weeks 1, 5, 9, 13, 17, and 21).

**Figure 1: RESCUE Study Flow**

Ziltivekimab Dose Level 1, 2 or 3 (N=180); N=60 in each level

Placebo (N=60)

| | Screening | | Treatment Period | | | | | | | | | | | Safety F/Up | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11/ET | 12 | 13/ET |
| Visit Day Window | -14 | -7 | 1 | 8 | 22 | 29-35 | 43-49 | 57-63 | 85-91 | 113-119 | 141-147 | 155-161 | 162-168 | 190-196 | 218-224 |
| Visit Window | | ±2 | 1 | ±2 | ±2 | | | | | | | | | | |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |

## 2.4. PATIENT-REPORTED OUTCOMES

### 2.4.1. PRO INSTRUMENTS

The following patient-reported outcomes (PRO) are collected in this study:

- Corvidia ePRO (*CARES*)
- Patient-Reported Outcomes Measurement Information System (PROMIS®) Fatigue 13a

short form

- Selected items from the PROMIS fatigue item bank
- PROMIS interest in sexual activity item
- Optum SF-36 v2® Health Survey
- EQ-5D-5L
- Patient global impression of symptoms (PGIS) and Change (PGIC).

#### 2.4.1.1. *CARES* Corvidia

The *CARES* Corvidia ePRO - henceforth *CARES* - is a new PRO instrument under development (Appendix A). It consists of 19 symptom items, with a 24-hour recall, asking patients to report their worst level of that symptom in the past 24 hours on a numeric rating scale from 0 (no symptom) to 10 (symptom as bad as I can imagine).

The symptom items cover physiological (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea), psychological (Depression, Concentration, Forgetfulness), Pain (Bone / Joints, Nerves, Muscle Cramps) and sensorial (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold) symptoms.

The instrument can be found in Appendix A.

#### 2.4.1.2. Short Form *PROMIS Fatigue* 13a and selected items from *PROMIS Fatigue* Bank, *PROMIS* Interest in Sexual Activity Items

The *PROMIS Fatigue* 13a short form contains 13 items that assess symptoms and impacts of fatigue. The recall period is 7 days. Each item is rated on a five-point Likert scale ranging from 0 = "not at all" to 4 = "very much". Higher scores in these items indicate worst fatigue, except for items 7 ("I have energy") and 8 ("I am able to do my usual activities") where higher score indicates less fatigue.

In this study, three additional fatigue items were administered for further testing to account for exhaustion, being physically drained, and impact on alertness, concepts that resonated with patients during patient interviews but are not covered in the *PROMIS Fatigue* 13a short form. Such combination is from now on referred to as *PROMIS Fatigue*.

A number of scales and subscales can be derived from the *PROMIS Fatigue* set:

- The original *PROMIS Fatigue* 13a scale;
- The fatigue scale using the whole *PROMIS Fatigue* item set (*de novo* 16-item short form);
- Two subscales within the *PROMIS Fatigue* 13a scale, e.g.: 1) symptoms of fatigue and

2) impacts of fatigue;

- Two subscales within the *PROMIS Fatigue* set, e.g.: 1) symptoms of fatigue and 2) impacts of fatigue.

The PROMIS Sexual Function and Satisfaction (PROMIS Sex FS) is a customizable, self-reported set of measures that include 79 items covering 11 domains: interest in sexual activity, lubrication, vaginal discomfort, erectile function, global satisfaction with sex life, orgasm, anal discomfort, therapeutic aids, sexual activities, interfering factors, and screener questions. In this study only one item will be used: interest in sexual activity. This item is rated on a five-point Likert scale ranging from 0 = "not at all" to 4 = "very much". Higher scores in this item indicate more interest in sexual activity.

These three instruments will be administered together as shown in Appendix B.

#### 2.4.1.3. Short Form 36-item Health Survey, Version 2 (SF-36v2)

SF-36v2 is a self-report survey of functional health and well-being with 4 weeks recall period [QualityMetric 2011]. Responses to 35 of the 36 items are used to compute an 8-domain profile of functional health and well-being scores.

The remaining item, referred to as the 'Health Transition' item, asks patients to rate how their current state of health compared to their state of health 1 year ago, and is not used to calculate domain scores.

The 8-domain profile consists of the following scales: Physical Functioning (PF), Role Limitations due to Physical Health (RP), Bodily Pain (BP), General Health Perceptions (GH), Vitality (VT), Social Functioning (SF), Role Limitations due to Emotional Problems (RE), and Mental Health (MH).

Psychometrically-based physical and mental health component summary scores (PCS and MCS, respectively) are computed from subscale scores to give a broader metric of physical and mental health status. The instrument can be found in Appendix C.

#### 2.4.1.4. 5-Level EuroQol in 5 dimensions (EQ-5D-5L)

The EuroQol EQ-5D-5L has been designed as an international, standardized, generic instrument for describing and valuing health-related quality of life, available in over 100 language versions.

The EQ-5D-5L self-report questionnaire consists of 5 items, one per domain, with a 5-point scale and one visual analogue scale (EQ VAS) to rate health state from worst (0) to best (100).

The EQ-5D-5L includes domains for each generic health status measure: Mobility, Self-care, Usual Activities, Pain/discomfort and Anxiety/depression. For each question, there are 5 levels of response, corresponding to increasing levels of impairment (no problems, slight, moderate, severe, and extreme problems or unable to perform activity) and coded 1 to 5.

A higher index indicates better quality of life. The instrument can be found in Appendix D.

#### 2.4.1.5. Patient Global Impression of Symptoms (PGIS) and Change (PGIC)

The Patient Global Impression of Severity (PGIS) is a self-reported measure of patient-perceived overall symptom severity. The PGIS is a 4-point recall scale (from no symptoms to severe symptoms). Patients are asked to rate their overall severity of symptoms in the 7 days prior a clinic visit.

The Patient Global Impression of Change (PGIC) instrument captures the patient's overall evaluation of response to treatment. The patient is asked to report the degree to which they have changed since entering the treatment period using a 7-point scale (Very Much Better to No Changes to Very Much Worse).

The PGIC and PGIS are the most commonly used anchor-based methods of assessing clinically important change and severity in which the external judgment of meaningful change is made by the patient. The instruments can be found in Appendix E.

### 2.4.2. ASSESSMENT SCHEDULE FOR PRO INSTRUMENTS

Patient-reported outcomes will be assessed through a personal device supplied by Corvidia:

- At home daily on subject personal device: *CARES*;
- At the clinic prior to dosing on subject personal device: *PROMIS Fatigue*, *PROMIS* Interest in Sexual Activity item, SF-36v2, EQ-5D-5L, PGIC, PGIS .

*CARES* data will be collected at home on subject's personal device or one provided by Corvidia; the questionnaire will be answered daily for 7 days prior to dose (Screening Day -7 to Day -1 prior to Dose 1 (Visit 1), Dose 2 (Visit 4), Dose 4 (Visit 7), and the last 2 weeks after Dose 6 (Visit 9).

The SF-36v2 and PROMIS Fatigue items are to be collected on subjects' devices at the clinic prior to each dosing, e.g., Visit 1, Visit 4, Visit 7, and at Visit 10 and Visit 11, except for PROMIS Sex Item (Only Visit 11) or the early termination visit (whichever occurs first).

The EQ-5D-5L will be collected on patients' device at the clinic prior to treatment administration at Visit 1, Visit 4, Visit 7 and Visit 11.

The PGIC will be colected on patients' device at the clinic prior to to treatment administration at Visit 4, Visit 7, and Visit 11.The PGIS will be collected at home on subjects' devices on the last day of *CARES* assessment period.

The schedule of assessment for all PRO instruments collected in different periods of the study are shown in Table 1.

**Table 1: Schedule of assessment for PRO instruments collected in the RESCUE study-Randomization, treatment period, and follow-up**

| Study Period | Screening | | Treatment | | | | | | | | | | | | Safety F/Up[b] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | N/A | 7 | 8 | 9 | 10 | 11/ET[c] | 12 | 13/ET |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 12 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |
| DOSING | | | X | | | X | | X | | X | X | X | | | | |
| *CARES* | | X | | | X | | | | X | | | | X | X | | |
| *PROMIS Fatigue* | | | X | | | X | | | | X | | | X | X | | |
| PROMIS Sex Item | | | X | | | X | | | | X | | | | X | | |
| SF-36v2 | | | X | | | X | | | | X | | | X | X | | |
| EQ-5D-5L | | | X | | | X | | | | X | | | | X | | |
| PGIS | | X | | | X | | | | X | | | | | X | | |
| PGIC | | | | | | X | | | | X | | | | X | | |

[b]Follow-Up.
[c]Early Termination.

## 3. OBJECTIVES OF THE PSYCHOMETRIC ANALYSIS

The aim of the psychometric analysis is to assess the psychometric properties of the *CARES* and of *PROMIS Fatigue scales*.

Specific objectives are as follows:

- To evaluate item response distributional characteristics;
- To investigate the underlying structure of these PROs and to develop scale and/or subscale scores;
- To evaluate the reliability (internal consistency reliability and test-retest reliability) of *CARES* and of *PROMIS Fatigue scale* and/or subscale scores;
- To evaluate the construct validity of *CARES* and of *PROMIS Fatigue scale* and/or subscale scores;
- To evaluate that changes in the patient's status are reflected in changes in the *CARES*/*PROMIS Fatigue* scale and/or subscale scores (sensitivity to change);
- To estimate a threshold for clinically relevant change withing patients for the *CARES* and of *PROMIS Fatigue* scale and/or subscale scores;
- To derive a definition for a symptomatic patient.

# 4. ANALYSIS TIMING

The psychometric analysis will be performed at the same time as the primary analysis of the study, e.g., once all subjects complete Visit 7 (Week 13). [redacted] will perform the analysis using blinded data and data will be collapsed across treatment groups for the purpose of the psychometric analysis.

After the last subject completes the last visit, the final database will be cleaned and locked and this data will be used to repeat the derivation of the clinically meaningful threshold (see section 7.3.8). However, due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, this analysis will be performed only on the subset of patients with week 24 data. There are no plans to repeat further psychometric analyses on the final database.

# 5. ANALYSIS POPULATION

All PRO analyses described in this PAP will be performed on the ITT analysis population which includes all randomized patients.

# 6. ANALYSIS VARIABLES

## 6.1. PRO VARIABLES

The scoring for each PRO instrument will be done according to the procedures described by each PRO developer and license owner in the respective instruction manuals.

### 6.1.1. VARIABLES GENERATED FROM *CARES*

For each item, weekly scores will be derived as the average of the 7 consecutive daily scores as follows:

Baseline = average of Day -7 to Day -1

Week 5 (Visit 4) = average of Day 22 to Day 28

Week 13 (Visit 7) = average of Day 78 to Day 84

Week 24 (Visit 11) = average of Day 155 to Day 168

For Baseline, Week 5 and Week 13, if more than 3 daily scores out of the 7 days (>50%) within the weekly period are missing, then the score is set to missing.

For Week 24, if more than 7 daily scores out of the 14 days (>50%) within the bi-weekly period

are missing, then the score is set to missing.

Preliminary scoring of the *CARES* includes the calculation of a Total Symptom Score (TSS). Alternative means of scoring the *CARES* that are derived from the factor analyses and conceptual considerations may also be defined if supported by the data and if the research team considers these to be plausible and clinically useful alternatives. The following conceptual domains from CARES will be considered: physiological (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea), psychological (Depression, Concentration, Forgetfulness), pain (Bone / Joints, Nerves, Muscle Cramps), and sensorial (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold) symptoms.

In the preliminary scoring, the daily TSS will be calculated as the average of the 19 daily item scores. A daily TSS score will be calculated if at least 10 out of the 19 items (> 50%) are non-missing. The TSS weekly score will be calculated in the same fashion as the individual items. Higher TSS indicate more symptoms/poorer outcomes.

### 6.1.2. VARIABLES GENERATED FROM *PROMIS FATIGUE* AND *PROMIS* SEXUAL FUNCTION AND SATISFACTION

The *PROMIS Fatigue* data is collected as a 7-day recall report from patients at Visit 1, 4, 7, 10, and 11. The scoring manual for *PROMIS Fatigue* 13a short form can be found at:http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Fatigue_Scoring_Manual.pdf.

Each question from the *PROMIS Fatigue* 13a short form has five response options (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much). To use the scoring tables from the PROMIS fatigue scoring manual, a summed score is calculated. The lowest possible summed score is 13 and the highest possible summed score is 65. All questions must be answered in order to produce a valid score using the scoring tables. Once calculated, the total summed raw score is translated to a T-score. The T-score re-scales the raw score into a standardized score with a mean of 50 and a standard deviation (SD) of 10. The T-score and standard error are then used to calculate the 95% confidence interval around the observed score (as described in the scoring manual available at: http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Fatigue_Scoring_Manual.pdf). The 95% confidence interval is calculated as T-Score ±1.96*SE. The T-score will not be calculated if a response is missing for one or more question from the *PROMIS Fatigue* 13a short form.

In addition, preliminary scoring of the *PROMIS Fatigue* 13a includes the calculation of two subdomain scores:

- Symptom score (5 symptom items):
    - I feel fatigued

  - I feel weak all over
  - I feel listless ("washed out")
  - I feel tired
  - I have energy
- Impact score (8 impact items):
  - I have trouble starting things because I'm tired
  - I have trouble finishing things because I'm tired
  - I am able to do my usual activities
  - I need to sleep during the day
  - I am too tired to eat
  - I need help doing usual activities
  - I am frustrated by being too tired to do the things I want to do
  - I have to limit my social activity because I am tired

The subdomains will be calculated as the sum of the items in the corresponding subdomain. The subdomain score will be calculated only if all items are answered.

Alternative means of scoring, by including also the additional *PROMIS Fatigue* items from the Item Bank, will be derived as follows:

- A total fatigue score using the items for the PROMIS Fatigues 13a and the additional three items from the Item Bank (de novo 16 item short form)
- Symptom score (8 symptom items):
  - I feel fatigued
  - I feel weak all over
  - I feel listless ("washed out")
  - I feel tired
  - I have energy
  - FATEXP36: How exhausted you are on average?
  - FATEXP43: How physically drained were you on average?
  - FATEXP52: To what degree did your fatigue make you feel less alert?

Alternative means of scoring can be derived from the factor analyses using the full set of *PROMIS Fatigue* items (inclusive of items from the Item Bank) to identify plausible and clinically

useful alternatives.

The *PROMIS* Interest in Sexual Activity item will not be included in any of the domain/subdomain scores described above, and will be analysed separately.

Given *PROMIS* Fatigue is collected at Week 23 and at Week 24, an assessment will be derived for the hypothesized scores as the mean value of the two assessments at Week 23 and Week 24, which will be labeled as Week 24. In case only one assessment is available, this will be used as Week 24 assessment.

### 6.1.3. VARIABLES GENERATED FROM SF-36v2

The scoring for SF-36 v2 instrument will be done according to the procedures described by the developer and license owner in the instruction manual. Scores for the scales (HT, GH, PF, RP, BP, MH, SF, RE, VT) as well as the summary scores (PCS, MCS) will be derived. For the psychometric analysis only the baseline scores will be used.

### 6.1.4. VARIABLES GENERATED FROM EQ-5D-5L

No derivation is required for the EQ-5D-5L. The scores for the 5 domains (Mobility, Self-care, Usual Activities, Pain/discomfort and Anxiety/depression) and EQ-5D VAS will be as reported. Missing data for individual items will not be imputed. For the psychometric analysis only the baseline scores will be used.

### 6.1.5. VARIABLES GENERATED FROM PGIS AND PGIC

The PGIS and PGIC are single-item questionnaires. Therefore, they do not require any manipulation to derive scores.

To be noted that PGIS is collected on the last day of *CARES* assessment period. To be in line with the visit label used for CARES, we will label the PGIS assessment from Week 12 as Week 13 (Visit 7). For the psychometric analysis, the baseline scores (PGIS only), Week 13 and Week 24 will be used.

## 6.2. CLINICAL OUTCOME VARIABLES

Demographic and clinical data will be used to characterize the sample, as follows:

- Age (years)
- Sex
- Race

- Baseline Hemoglobin (>=11 or <11 g/dL)
- CKD Stage (Stage 3, Stage 4, or Stage 5)

The clinical variable that will be used in the psychometric evaluation is the total plasma level for the hs-CRP Protein.

## 6.3. General Variables and Definitions

### 6.3.1. Baseline

The baseline for *CARES* will be defined as the weekly-average of the daily measurements over the 7-day period immediately preceding randomization day (i.e., study days -7 through -1), e.g., the second week of the screening period.

For all other PRO instruments, except PGIC, the baseline is defined as the measurements carried out at clinic prior to first dose of the study drug, scheduled for the 1st day of the Treatment Period. There will be no baseline assessment for PGIC.

### 6.3.2. Derived Timepoints

No other timepoints are derived.

# 7. Statistical Methodology

## 7.1. General Considerations

Summary statistics for continuous data will include the number of observations (N), mean, standard deviation (SD), median, first and third quartiles, minimum (min), and maximum (max). Categorical data will be described by the number (n) and percentage (%) of patients in each category. Missing and invalid observations will be tabulated as separate categories. The calculation of proportions will not include the missing/invalid category, unless specified otherwise.

Statistical comparisons will be made using two-sided tests at the $\alpha = 0.05$ significance level unless stated otherwise. Due to the exploratory nature of these analyses, no adjustments for multiplicity will be performed. For point estimates, 95% confidence intervals will be used.

The psychometric analysis will be performed on the pooled treatment arms.

## 7.2. STUDY SUBJECTS

### 7.2.1. PRO COMPLETION

The completion rate for *CARES* will be defined as the total number of actual completed diary entries divided by the expected number of diary entries in a given time period (total number of days X number of patients expected to complete the instrument). Specifically, completion rate will be evaluated for the baseline period (Day -7 to Day -1), in the weeks prior to each clinic visit for administration of dosage (Day 22 to 28 [Week 5] and 78 to 84 [Week 13]), and Day 155 to 168 [Week 24].

For the *PROMIS* Fatigue, *PROMIS* Sexual Interest, PGIC and PGIS, completion rates will be calculated as the total number of actual completed clinic visit entries divided by the expected number of patients taking part in the scheduled clinic visit.

### 7.2.2. DEMOGRAPHICS AND BASELINE CHARACTERISTICS

The demographic and patient baseline characteristics are analyzed in the clinical SAP and will not be repeated herein.

## 7.3. MEASUREMENT PROPERTIES ANALYSES

The analyses described below follow best practice guidelines for the psychometric analysis of PRO measures according to the FDA guidance [FDA 2009, FDA 2019].

### 7.3.1. DESCRIPTIVE ANALYSES

Descriptive statistics for *CARES, PROMIS* Fatigue, PROMIS sexual function and satisfaction, PGIC, and PGIS will be assessed by evaluating the following at baseline, Visit 4, and Visit 7:

- Counts and percentages for each response option per item, including a stacked column chart of the distribution of responses
- Summary statistics for each item of *CARES, PROMIS* Fatigue, PROMIS sexual function and satisfaction
- Summary statistics for each scale/domain and total score of *CARES* and *PROMIS* Fatigue

Item descriptive analyses will be examined for *CARES* and *PROMIS* Fatigue for item removal, combined with qualitative evaluation. Specifically, items with floor or ceiling effect (e.g., tems

that are disproportionately rated at either extreme of their rating scales by patients at baseline and/or after treatment) will be considered as candidates for item reduction.

### 7.3.2. ITEM-TO-ITEM CORRELATIONS FOR CARES AND PROMIS FATIGUE ITEMS

The purpose of item-to-item correlation analyses is to evaluate inter-relationships among items to determine the extent and degree of correspondence and discordance between items, and whether those patterns are consistent with hypothesized expectations based on the item content.

Spearman rank correlation coefficient will be calculated using baseline data to examine the inter-relationships among items for *CARES* (Items 1 to 19) and among *PROMIS Fatigue* Items (Fatigue 13a Items 1 to 13, Fatigue Bank Items FATEXP36, FATEXP43 and FATEXP52).

Correlations greater than 0.4 may provide support for combining items into a multi-item scale. Items with coefficients greater than 0.9 and/or less than 0.1 may be considered for item removal.

#### 7.3.2.1. CARES

The *CARES* instrument is organized thematically in different sets of symptoms. Items 1 to 7 describe a set of physiological symptoms (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea). Items 8 to 10 cover psychological issues (Depression, Concentration, Forgetfulness). Item 11 to 13 describe pain (Nerve, Bones/Joints, Muscle Cramps).

The remainder of the items cover a variety of sensorial symptoms that can be loosely described as a sensation of uncomfortableness, mostly associated with the skin and terminal nervous system (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold). Correlations may vary accordingly in a non-predictable manner.

#### 7.3.2.2. PROMIS Fatigue

The combined set of fatigue items is quite homogeneous with regards to the topics investigated. Specifically, the fatigue Bank additional items may be reasonably perceived as 'general' feelings of fatigue/exhaustion and therefore be strongly correlated with the short form 13a Items. It is expected that these items will show strong correlations (above 0.50) across the items.

### 7.3.3. EXPLORATORY FACTOR ANALYSIS

An exploratory factor analysis (EFA) will be carried out for the items of *CARES* and *PROMIS Fatigue* to test the measurement model. EFA is a model of the measurement of a latent variable. This latent variable cannot be directly measured with a single variable. Instead, it is seen through the relationships it *causes* in a set of variables. The relationships between the

factor(s) and each variable are weighted, and factor analysis calculates the optimal weights.

The Kaiser-Meyer-Olkin measure of sampling adequacy will be used to evaluate the strength of the linear association among the items in the correlation matrix. The KMO measure can range between 0 and 1. The standard threshold value for sample adequacy is above 0.6, with 0.8 and above being ideal [Pett et al. 2003].

Both the maximum likelihood (ML) and the unweighted least squares (ULS) extraction method will be used in this study. ML is the most frequently used factor extraction methods. However, it does require the assumption of multivariate normal distribution of the variables. If the assumption of multivariate normality is severely violated (e.g., skewness < -1 or > 1), Fabrigar et al. 1999 and Nunnally and Bernstein 1994 recommend using ULS.

The symptoms investigated in the two PROs under investigation are expected to be associated which each other. We expect the underlying domains to be correlated. Accordingly, the EFA will be executed using a PROMAX rotation. The number of factors will be determined using established methods (Eigenvalues assessments, scree-plots).

The PROC FACTOR procedure in SAS will be used to perform the EFA.

A total summary scale or set of scales will be proposed in light of the factor loadings and clinical and conceptual considerations. The score(s) developed as psychometrically sound will be used to evaluate efficacy.

Data distributions and correlations with other PROs will then be carried out for the proposed *CARES* and *PROMIS Fatigue* scales and/or subscales as described in sections 7.3.5 and 7.3.6.

### 7.3.4. ITEM-TOTAL CORRELATIONS FOR *CARES* AND PROMIS FATIGUE

Correlations between items and scales derived using the preliminary algorithm and possible alternative structures selected from the factor analysis will be assessed using baseline data. The multitrait-multimethod approach will be applied to assess the association of items with their hypothesized scales (item removed) and the association of items with other scales. Spearman rank correlations will be estimated for this analysis. Items are expected to have a correlation coefficient >0.4 with their hypothesized scales (convergent validity) and to have a higher correlation with their hypothesized scale than with any other scale (divergent validity).

Item-total correlations will be considered adequate if the Spearman rank correlation coefficient values are at least 0.40. Any low item-total correlations will be flagged, and item removal will be considered, as it may indicate inadequate scale validity. Any large correlation coefficient (e.g., ≥ 0.90) might suggest redundancy, marking the item as a candidate for elimination or modification.

### 7.3.5. RELIABILITY

#### 7.3.5.1. Test-Retest Reliability

Test-retest reliability (i.e., the stability of an instrument over time) will be assessed using intra-class correlation coefficients (ICCs) between two time points ("test" and "retest"), employing form 2.1 as described by Shrout and Fleiss 1979. For *CARES,* two sets of data will be examined: for the 7 days prior to treatment on Day 1 for all patients and Visit 4 period for patients considered to be in a stable condition, as described below. For *PROMIS Fatigue* only the second set of data will be used, e.g., using patients considered to be in a stable condition.

**Systematically selected days**

The last 7 days before Day 1 will be considered to define the test and retest time points for *CARES* collected via eDiary. Test and retest periods will be defined using two consecutive days from the baseline period (Day -6 to Day 1). The following pairs of consecutive days will be evaluated: Days -6 and -5; -5 and -4; -4 and -3; -3 and -2; -2 and -1; -1 and 1.. This analysis will be performed for *CARES* only.

**Stable condition defined using PGIC**

Only those subjects reporting "no change" from baseline on the PGIC at Visit 4 will be included. Period 1 (i.e., test) will defined as baseline and Period 2 (i.e., retest) will be defined as Visit 4.

**Stable condition defined using PGIS**

Only those patients reporting no change on the PGIS, i.e., those that report the same severity level from baseline to Visit 4, will be included for test-retest investigation. Period 1 (i.e., test) will be defined as baseline and period 2 (i.e., retest) will be defined as Visit 4.

Test-retest reliability will be tested for each multi-item scale/domain as defined by the review of EFA results and total scores. ICC values of 0.40–0.75 will be considered to represent fair to good reliability and values >0.75 represent excellent reliability.

#### 7.3.5.2. Internal Consistency Reliability

Internal consistency reliability of each multi-item scale/domain score as identified by EFA results will be estimated using Cronbach's Alpha. Cronbach's Alpha, which ranges from 0 to 1, where "1" equals perfect reliability, is based on the average inter-item correlation and the number of items. Minimum values equal to or greater than 0.70 have been recommended for group level comparisons. Internal consistency will be tested for each domain of the *CARES* and *PROMIS Fatigue*, as well as for their total scores.

Internal consistency will also be assessed with the Cronbach Alpha statistics with Item Removal. Alpha will be estimated for each item and compared with the same statistic for the full set of

items. Items that do exhibit a strong drop in the Alpha value compared to the full set value should be selected for retention. Conversely items with alpha values for which there is small or no change from the full set value should be considered for removal.

### 7.3.6. CONSTRUCT VALIDITY

Validity refers to the evidence that the *CARES* and *PROMIS Fatigue* measure what they are designed to measure in the population for which it is intended. Construct validity is an experimental demonstration that the *CARES* and *PROMIS Fatigue* measures the core constructs of the symptoms of NDD-CKD patients at advanced stage. Three forms of construct validity will be examined: convergent, divergent, and known group validity.

#### 7.3.6.1. Convergent and Divergent Validity

The Spearman Rank correlation coefficient will be used to evaluate convergent and divergent validity. Convergent validity refers to how well constructs that theoretically should be related to each other are observed to be related. Divergent validity can be viewed as the counterpart to convergent validity; it assumes that constructs theoretically unrelated to each other will be observed to be unrelated and will have low correlations.

Correlations between the *CARES* scores (items, TSS and any other scale/domain score as identified by EFA results) and *PROMIS Fatigue* scores (all scores identified in section 6.1.2) with the SF 36V2, EQ-5D-5L, and PGIS will be examined at Baseline. It is expected that there would be moderate to high correlations between domain and total scores of similar content across the assessments, and low correlations for non-overlapping domains.

*PROMIS Fatigue* items 7 and 8 are keyed in the opposite direction of the remaining items and of those in the other instruments. Accordingly, those items should, regardless of the magnitude, show a negative correlation with other items and instruments.

**Moderate Correlations**

These are expected for *CARES* Tiredness (items 1, 2, 3, 4, 5) with the following domains:

- SF-36 v2: VT, GH, PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Mobility, Self-care, Usual activities, Pain, Anxiety/Depression
- PGIS

and for *CARES* Items Gastrointestinal (6, 7) with the following domains:

- SF-36 v2: VT, GH, PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Pain, Anxiety/Depression
- PGIS

and for *PROMIS Fatigue* Items 1 to 4 with:

- SF-36 v2: PF, RP, PCS, MH, RE, VT, MCS
- EQ-5D-5L: Mobility, Self-care, Usual activities, Anxiety/Depression

**Low to Moderate Correlations**

These are expected for *CARES* Items Mental Health (8, 9, 10) and Pain (11 12, 13) with the following domains:

- SF-36 v2: VT, GH, PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Mobility, Self-care,Usual activities, Pain, Anxiety/Depression
- PGIS

and for *PROMIS Fatigue* Items 5 to 16 with:

- SF-36 v2: PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Self-care, Usual activities, Anxiety/Depression

**Low Correlations**

These are expected for *CARES* Items Discomfort (14 to 19) with the following domains:

- EQ-5D-5L: Mobility, Self-care,Usual activities,

and for *PROMIS Fatigue* scores with:

- SF-36 v2: VT, GH, BP
- EQ-5D-5L: Pain
- PGIS

#### 7.3.6.2. Known-Groups Validity

The purpose of the known-groups validity is to assess the degree to which the measure can distinguish among groups of subjects hypothesized to be different in concepts of interest. In the case of the RESCUE trial there are several concepts of interests.

The known-group analysis will accordingly refer to the following three grouping criteria:

- Hemoglobin: High (≥11 g/dL) Vs. Low ( <11 g/dL)
- CKD Stages: 3 vs. 4 or 5
- PGIS: groups will be defined at baseline by PGIS (None, Mild, Moderate, Severe)

For the PGIS classification, adjacent groups may be combined if the sample size of a particular category is <10 patients.

Known-group validity for the Hemoglobin and CKD Stage groups will be tested with a group mean t-test with alpha = 0.05 level for the *CARES* and *PROMIS Fatigue* total and subscale scores.

PGIS groups will serve as the independent variable in an analysis of variance (ANOVA) with alpha = 0.05 level, and the *CARES* and *PROMIS Fatigue* total and subscale scores will serve as dependent variables. Kruskal-Wallis test, a non-parametric alternative to ANOVA, will be conducted if group data for any group criteria are not reasonably balanced.

Box-plots of the *CARES* and *PROMIS Fatigue* total and subscale scores for each group will be also presented to highlight the degree of separation and heterogeneity between the scores' distributions across groups.

### 7.3.7. SENSITIVITY TO CHANGE

Sensitivity to change is the ability of an instrument to measure change in a state regardless of whether it is relevant or meaningful to the decision maker. Sensitivity to change will be examined using one-way analysis of covariance (ANCOVA). Subjects with baseline and Visit 7 data will be included in the analysis.

The dependent variable will be the change from baseline to Visit 7 in the *CARES* and *PROMIS Fatigue* total and subscale scores. The responder group will be formed using the PGIC and PGIS scores. The model will include the "responder" factor as a fixed factor and the analysis will be controlled for the baseline PRO scores. Separate models will be considered for each CARES and *PROMIS Fatigue* total score and EFA-based scales.

Using the PGIC, the following three groups will be created using the PGIC score at Visit 7:

- 'Improved' group will include those participants who answered "Very much better", "Moderately better", and "A little better".
- 'No change' group will include those participants who answered 'No change'.
- 'Worsened' group will include those participants who answered "Very much worse", "Moderately Worse", and "A little worse" .

Using the PGIS, the following three groups will be created using the PGIS score at Visit 7:

- Improved is defined as patients with an improvement on their PGIS of at least one level from baseline to Visit 7
- Worsening is defined as patients with a worsening of at least one level from baseline to Visit 7
- Unchanged is defined as patients with same PGIS score at Visit 7 and at baseline

The analysis will be performed on groups with ≥10 patients and adjacent groups may be

combined if there are <10 patients. Separate models will be considered for each "responder" group definition.

The model will present least squares (LS) mean estimates, standard errors, 95% confidence intervals (CIs) and p-values. The analysis of variance will be conducted using the PROC GLM procedure in SAS.

### 7.3.8. MEANINGFUL CHANGE

Individual-level thresholds of meaningful change will be explored. To identify patients who experienced a significant improvement in their symptoms over the course of treatment, a responder definition will be determined to characterize a meaningful change in the scores of the *CARES* and *PROMIS Fatigue* measures. Anchor-based and distribution-based methods, as well as graphical displays, will be used to support the interpretation of possible treatment benefits as reflected in the *CARES* and *PROMIS Fatigue* scores. These approaches will help characterize the amount of change important to patients on these dimensions (as recommended in the FDA PRO guidance [FDA 2009; FDA 2019]).

Thresholds will be estimated for the *CARES* and *PROMIS Fatigue* total score and scales and/or subscales obtained from the EFA.

#### 7.3.8.1. Distribution-Based Approach

The following two distribution-based approaches will be applied to the baseline *CARES* and *PROMIS Fatigue* total scores and EFA-based scales.

**One-half standard deviation (SD):**

SD of Baseline PRO scores will be computed and divided by 2

**1 standard error of measurement (SEM):**

$$SEM = SD_{baseline} * \sqrt{1 - r_{xx}}$$

where $SD_{baseline}$ = SD at Baseline and $r_{xx}$ = the reliability (internal consistency) of PRO scale score at Baseline

#### 7.3.8.2. Anchor-Based Approach

In the anchor-based approach, the PGIS and PGIC will be used. Specifically, subjects will be classified in groups to provide a clearer difference between subjects who have and have not experienced meaningful change according to the anchors. Meaningful change thresholds will be derived using each level of the PGIS and PGIC. If the sample size within a response category is too small, grouping of the response categories will also be considered as indicated in Table 2.

**Table 2 Categories for change in PGIS and PGIC**

| Instrument | Category | Definition |
|---|---|---|
| **PGIS at Visit 7** | **Un-collapsed Categories** | |
| | Improved 3 categories | CFB =-3 |
| | Improved 2 categories | CFB =-2 |
| | Improved 1 category | CFB =-1 |
| | No Change | CFB =0 |
| | Worsened 1 category | CFB =+1 |
| | Worsened 2 categories | CFB =+2 |
| | Worsened 3 categories | CFB =+3 |
| | **Collapsed Categories** | |
| | Improvement | CFB ≤ -1 point |
| | No Change | CFB=0 |
| | Worsening | CFB ≥+1 point |
| **PGIC at Visit 7** | **Un-collapsed Categories** | |
| | Very much improved | Rating "very much better" |
| | Much improved | Rating "moderately better" |
| | Minimally improved | Rating "a little better" |
| | No change | Rating "no change" |
| | Minimally worse | Rating "a little worse" |
| | Much worse | Rating "moderately worse" |
| | Very much worse | Rating "very much worse" |
| | **Collapsed Categories, 5 categories** | |
| | Very much improved | Rating "very much better" |
| | Improvement | Rating of "a little better" and "moderately better" |
| | No change | Rating of "no change" |
| | Worsening | Rating of "a little worse" or "moderately worse" |
| | Very much worse | Rating "very much worse" |
| | **Collapsed Categories, 3 categories** | |
| | Improvement | Rating of "a little better", "moderately better" or "very much better" |
| | No change | Rating of "no change" |
| | Worsening | Rating of "a little worse" or "moderately worse" or "very much worse" |

CFB=Change from baseline at Visit 7.

The change in the *CARES* total score and EFA based scales from baseline (average of Day -7 to Day -1) to the Visit 7 (average of daily measurements from day 78 to day 84) will be used in the anchor-based analyses.

The same procedure will be applied to the *PROMIS Fatigue* total score and EFA-based scales using the matching 7-day recall measurement at baseline and Visit 7.

The adequacy of the anchors proposed above is explored with the Spearman correlations between the change from baseline to Visit 7 in *CARES* and *PROMIS Fatigue* total score and EFA-based scales. If the correlation coefficient >0.30, it will be emphasized in the interpretation of the results.

The following descriptive analyses will be conducted:

- Descriptive statistics for changes in the *CARES* total score and EFA-based scales from baseline (average of Day -7 to Day -1) to the Visit 7 (average of day 78 to 84). Descriptive statistics will be conducted for all categories as described in Table 2 Categories for change in PGIS and PGIC
- . Descriptive statistics for changes in the *PROMIS Fatigue* total score and EFA-based scales from baseline to the Visit 7. Descriptive statistics will be conducted for all categories as described in Table 2 Categories for change in PGIS and PGIC
- Empirical cumulative distribution function plots (eCDF) and smooth Probability Density Functions (PDF) will be presented for each CARES and PROMIS Fatigue total score and EFA-based scales.These curves will be generated for all categories as described in Table 2.

Furthermore, two descriptive tables will be generated to display change in the PRO scores from baseline at Visit 7 by baseline PGIS. This will be conducted separately for patients who achieved a 1-category and 2-category PGIS decrease at Visit 7 from baseline. Change scores will be categorized by percentiles: 10th, 25th, median, 75th, and 90th. Three baseline PGIS categories (i.e., mild, moderate, severe) will be used for the 1-category PGIS decrease analysis, while two categories for baseline PGIS (i.e., moderate, severe) will be used for the 2-category PGIS decrease analysis.

The various estimates from the different streams of evidence (distribution methods and anchor methods,) will be tabulated and will be examined for convergence in an effort to triangulate onto a single threshold value that represents meaningful within-patient worsening/improvement. However, if this is not possible, a range of thresholds will be considered.

The results of the anchor-based analyses (original, un-collapsed PGIS categories) will be considered primary when making decisions regarding the clinically meaningful worsening and clinically meaningful improvement thresholds and will be supplemented with eCDF and PDF curves. The primary anchor to establish the meaningfulness of deterioration or improvement in patient scores over time is PGIS. Specifically, our *a priori* definition of clinically meaningful deterioration threshold will be change from baseline of +1 points on the PGIS; and the *a priori* definition of clinically meaningful improvement threshold will be change from baseline -1 points on the PGIS. However, the results from other analyses will also be considered.

The eCDF and ePDF plots will be visually examined to explore the data across a range of thresholds to ensure there is separation in the curves between the anchor groups and to examine the proportion of subjects experiencing up to that change at various thresholds. The eCDF displays a continuous plot of the change from baseline on the horizontal axis and the cumulative percent of patients experiencing up to that change on the vertical axis. The eCDFs and PDFs are mathematically related, displaying a distribution of PRO change score in different but complementary ways. eCDF plots optimally display the cumulative probabilities of change across the entire distribution and provide insight into the percentiles (e.g., median) of the PRO changes. PDF plots optimally display the moments of a distribution, such as the mean and variability.

### 7.3.9. CLASSIFICATION BY PATIENT SYMPTOMATOLOGY

One of the objectives of the interim blinded PRO data analysis of RESCUE data is to characterize patient symptomology, such that two classes of patients can be defined: low symptom burden (or non-symptomatic) and high symptom burden (or symptomatic).

The investigation will rely on quantitative approaches and will be also reviewed on the basis of qualitative considerations. This analysis will be exploratory in nature and iterative. Two methods will be considered:

- Cluster analysis (see section 7.3.9.1)
- Anchor analysis (see section 7.3.9.2)

The two symptom classes resulting from the above methods will be investigated for differential response patterns, by means of profile analysis (see section 7.3.9.3).

#### 7.3.9.1. Cluster analysis

Cluster analysis will be performed to identify subgroups of patients who differ meaningfully on symptoms at baseline. The items and/or scales to be used for cluster analysis will be determined following an examination of the results of the descriptive analyses, item-to-item correlations, and EFA. The analysis will be performed separately for *CARES* and *PROMIS* Fatigue.

Two Cluster Analysis methods may be used: hierarchical clustering and k-means (or k-medoids) clustering. The initial cluster centres may be based on prototypical (clinically meaningful) "symptomatic", "non-symptomatic" etc. cluster centres. The algorithm from the k-means cluster analysis establishes an initial set of cluster means then assigns each case (i.e., patient) to the closest cluster mean. Alternative clustering algorithms may be used as appropriate given the features of the response. A two-cluster solution will be specified to aid in interpretation.

#### 7.3.9.2. Anchor analysis

An external anchor, specifically PGIS will be used to inform the definition of a sufficiently

symptomatic patient. Specifically, the median baseline score for those who reported "mild", "moderate" or "severe" severity on the PGIS at screening will used as a *threshold* for significant symptom burden. Patients will be further classified as "non-symptomatic" if their baseline score is < *threshold* and as "symptomatic" if their baseline score is ≥ *threshold*.

The analysis will be performed separately for each scale/subscale of *CARES* and *PROMIS* Fatigue.

#### 7.3.9.3. Profile analysis

The cluster analysis and the anchor analysis will identify patient profiles (symptom classes) based on shared symptoms at baseline. Differential response patterns will be assessed. Specifically, the following will be presented:

- A table with descriptive statistics for the PRO scores and a line graph with mean values and corresponding 95% confidence intervals (CI) by symptom classes and each visit (e.g., at baseline, Visit 4 and Visit 7).
- A table with descriptive statistics for the change from baseline in PRO scores by symptom classes and each visit (e.g., at baseline, Visit 4 and Visit 7).
- A cumulative distribution plot showing a continuous plot of the absolute change from baseline during the study for the PRO scores on the X-axis and the cumulative percent of patients experiencing that change on the Y-axis will be presented by symptom classes at Visit 7.

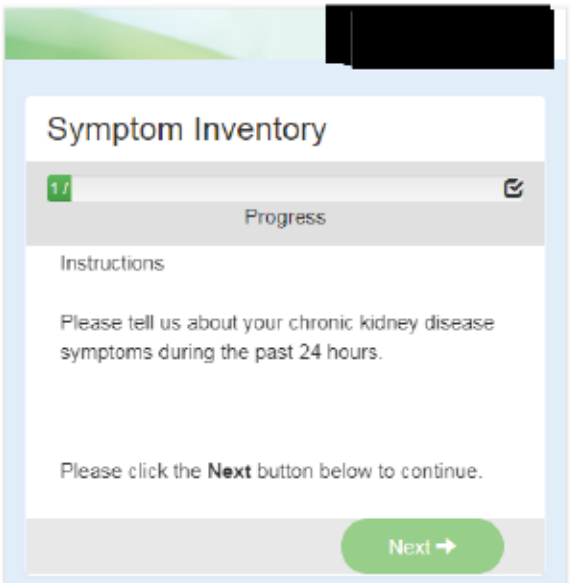
In addition, a mixed model for repeated measures (MMRM) will be used to evaluate change from baseline to Visit 7. The model will include variables for baseline hemoglobin (>= 11 or < 11 g/dL), CKD Stage (3, 4 or 5), symptom class group, visit and treatment group-by-visit interaction as categorical fixed effects, baseline value and baseline-by-visit interaction will be included as covariates. The least squares means for each symptom class group, the least squares mean differences between symptom class groups along with the associated 95% confidence intervals (CIs) and p-values will be presented.

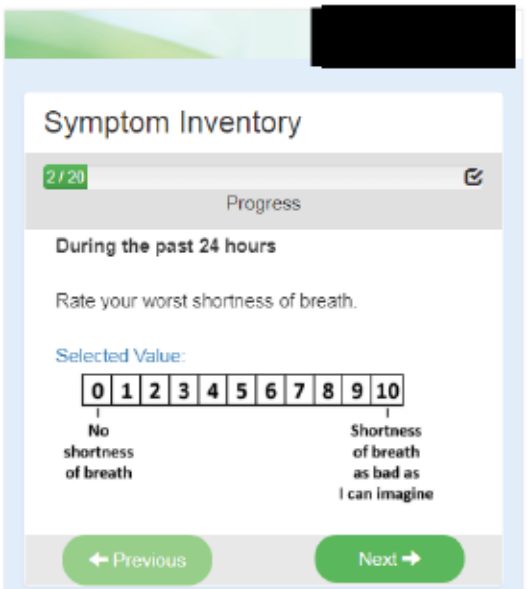
The analyses will be conducted using PROC MIXED in SAS. An unstructured covariance matrix will be used to model the within-subject correlation. The Kenward-Roger approximation will be used to adjust the denominator degrees of freedom. The analysis will be performed based on all observed post-baseline scores without any imputation of missing data. In the case when the MMRM fails to converge using an unstructured covariance matrix in any stage, a less stringent covariance matrix (e.g., autoregressive 1) will be used.

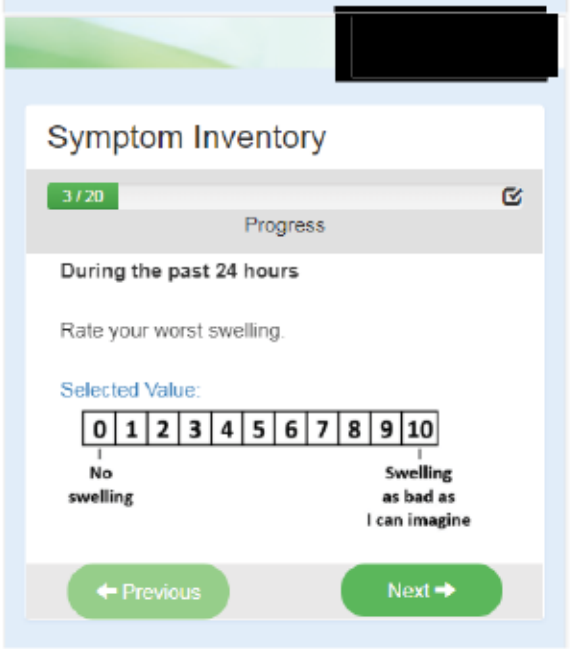
# 8. REFERENCES

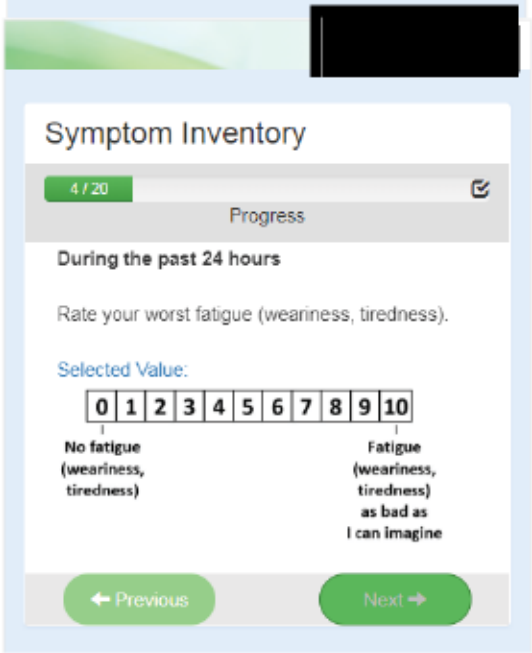

| | |
|---|---|
| Cohen 1992 | Cohen J. A power primer. Psychol Bull. 1992 Jul;112(1):155-9. PubMed PMID:19565683 |
| Fabrigar et al. 1999 | Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological methods, 4(3), 272. |
| Fayers et al. 2007 | Fayers PM, Machin D. Quality of life. The assessment, analysis and interpretation of patient-reported outcomes. Second edition. Wiley 2007. |
| FDA 2009 | US Department of Health and Human Services (USDHHS). Guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. December.2009. |
| FDA 2019 | Patient-focused drug development guidance public workshop. October 15-16, 2018. https://www.fda.gov/downloads/Drugs/NewsEvents/UCM620711.pdf. Accessed 28 March 2019. |
| Liang 2000 | Liang, Matthew. (2000). Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments. Medical care. 38. II84-90. |
| Nunnally and Bernstein 1994 | Nunnally, J. C., and Bernstein, I. H. (1994). Validity. Psychometric theory, 3, 99-132. |
| QualityMetric 2011 | User's manual for the SF-36v2 Health Survey (3 ed.). Lincoln, RI. Strand, V., Schett, G., Hu, C., & Randall, M. (2013). Patient-reported Health-related Quality of Life with Apremilast for Psoriatic Arthritits: A Phase II, Randomized, Controlled Study. (Vol. 40). The Journal of Rheumatology. doi:10.3899/jrheum.121200 |
| Pett et al. 2003 | Pett M, Lackey N, Sullivan J, editors. Making Sense of Factor Analysis. Thousand Oaks, CA: Sage; 2003 |
| Shrout and Fleiss 1979 | Shrout PE and Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, Vol 86(2), Mar 1979, 420-428. |

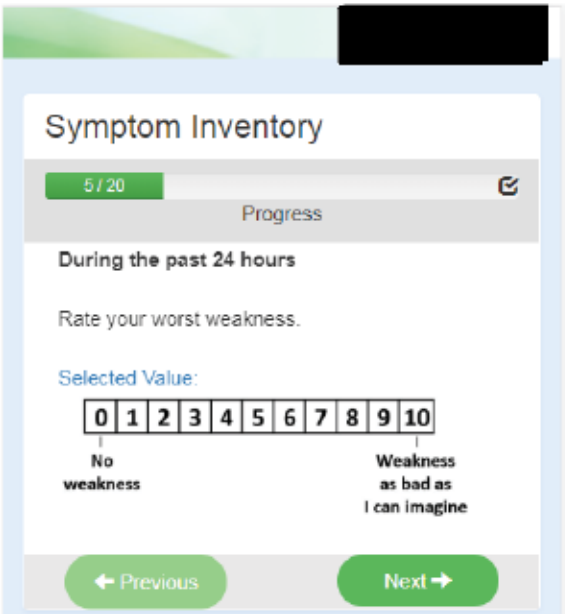
# APPENDIX A: CORVIDIA EPRO (CARES)

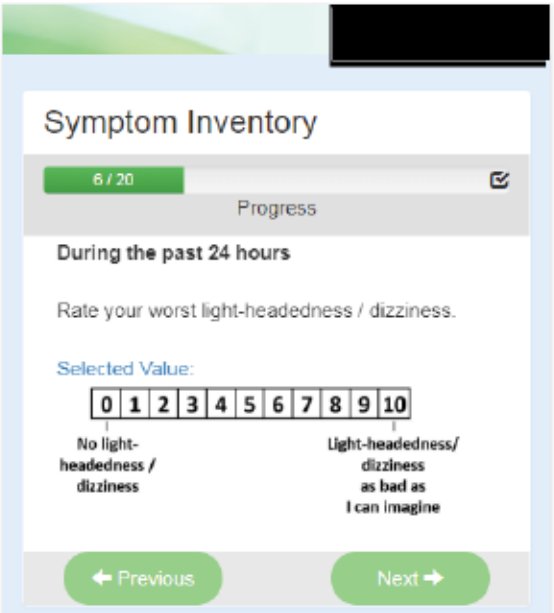


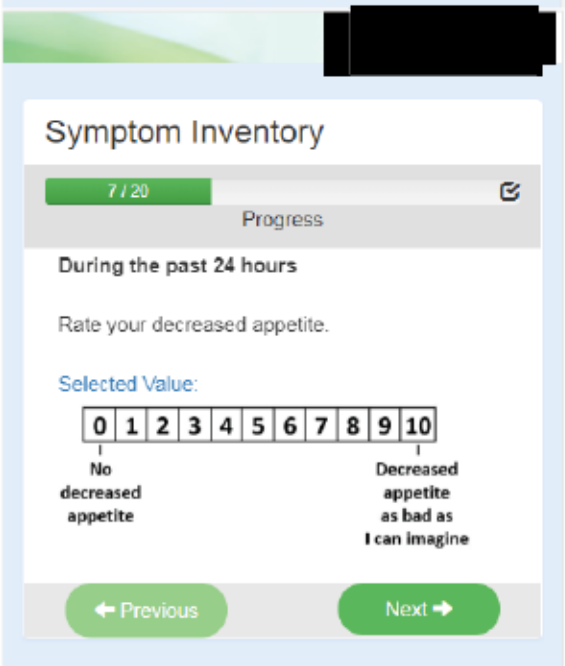



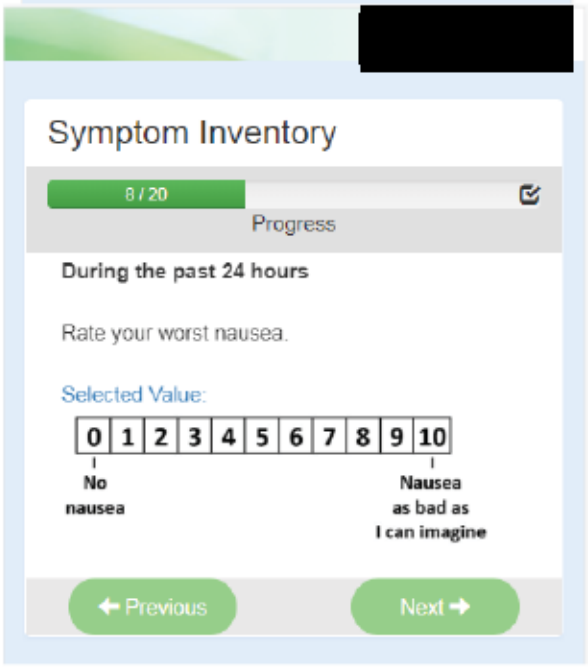



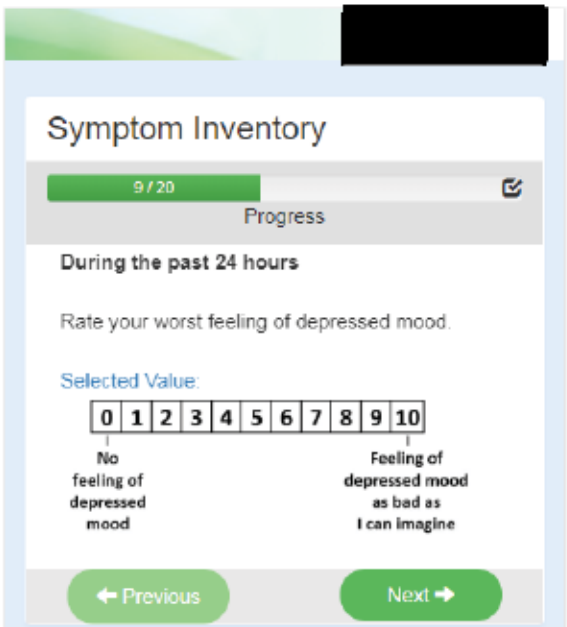



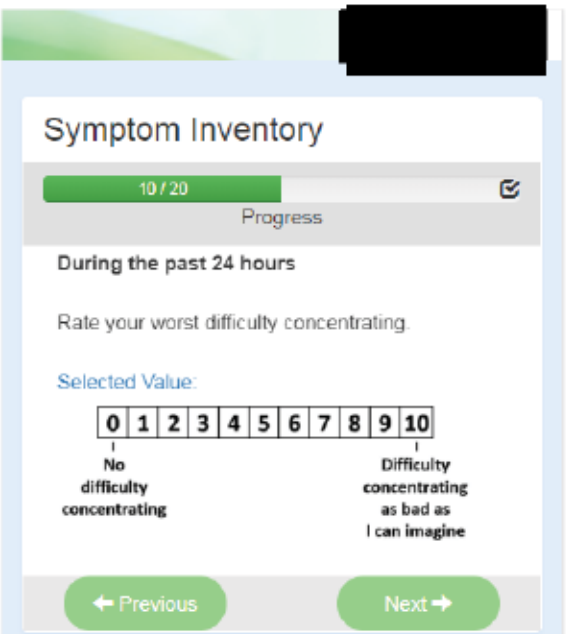


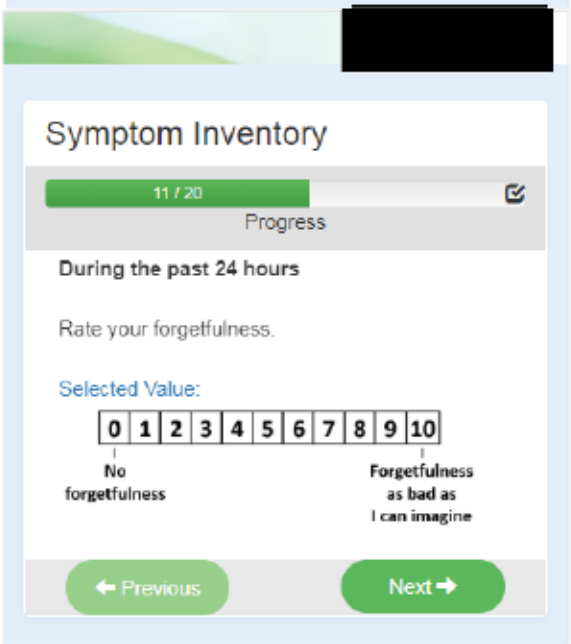



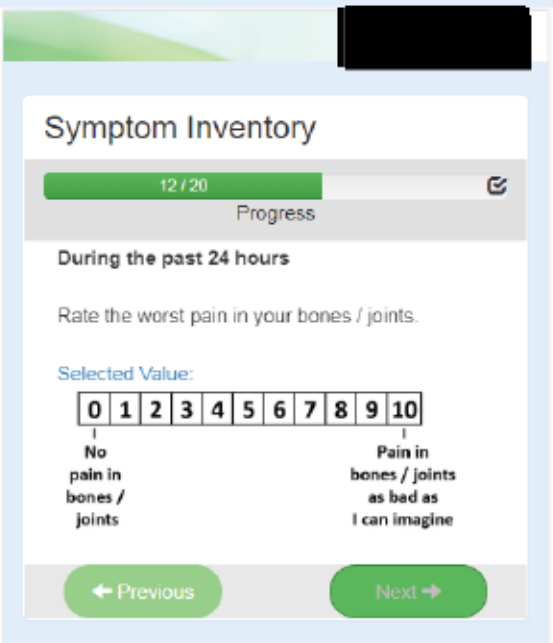



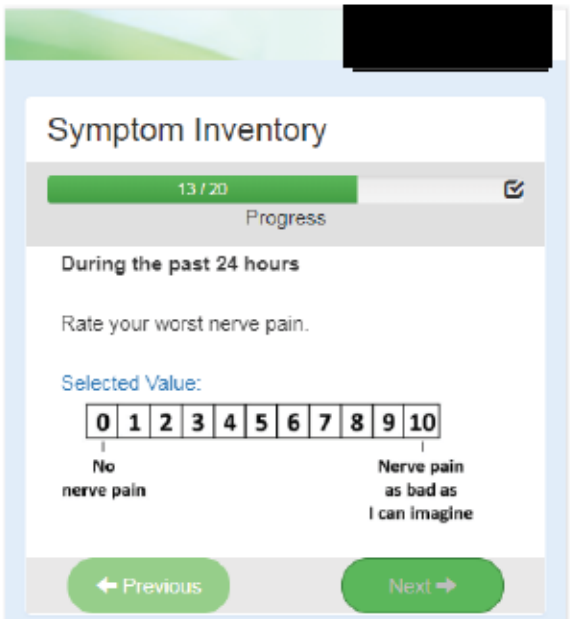



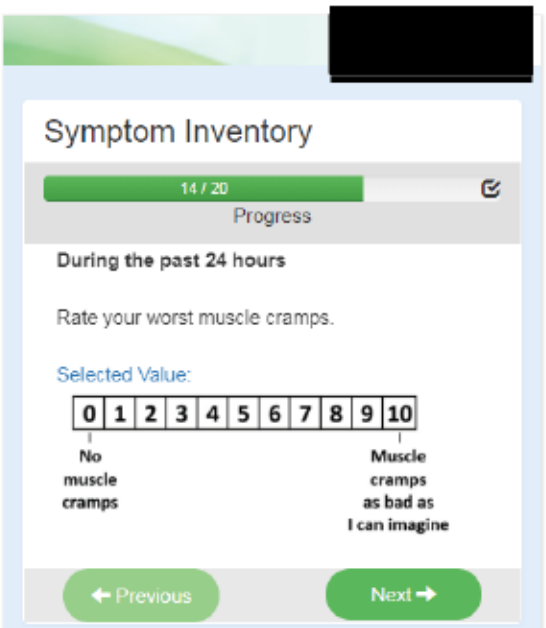


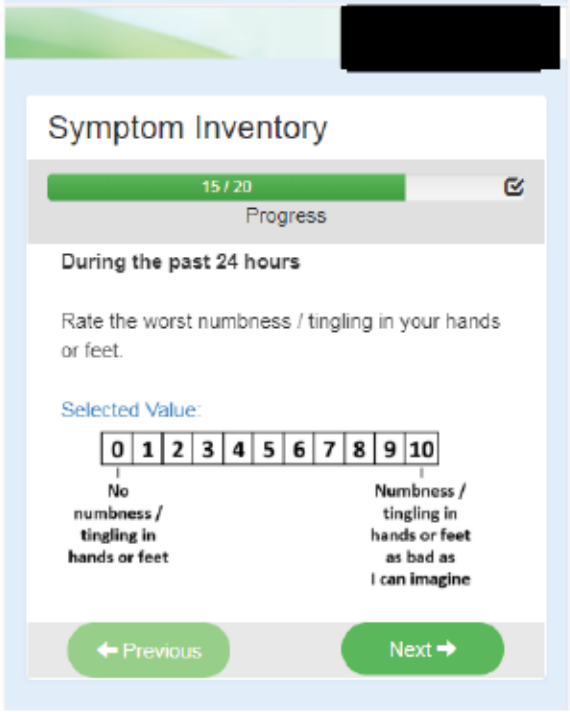



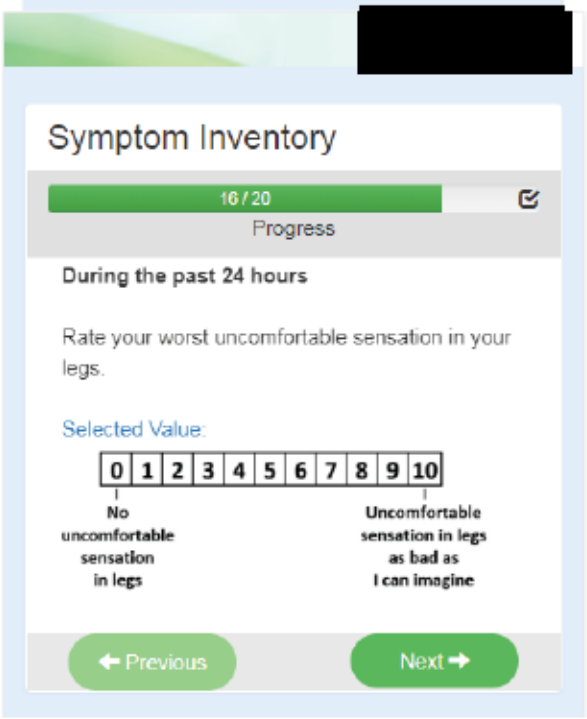



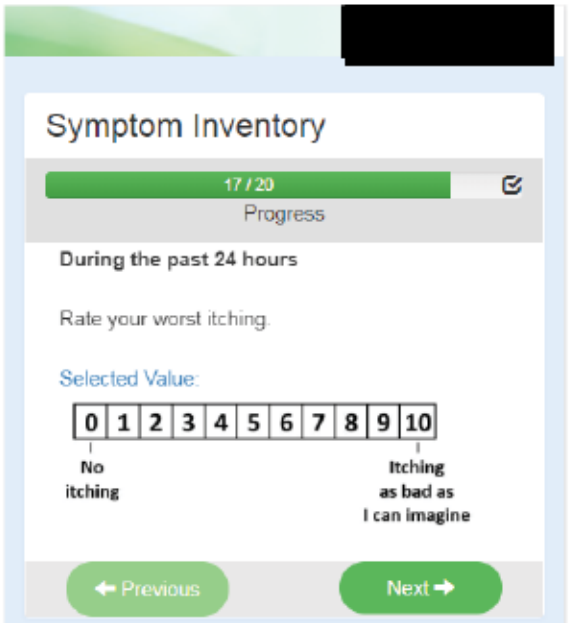



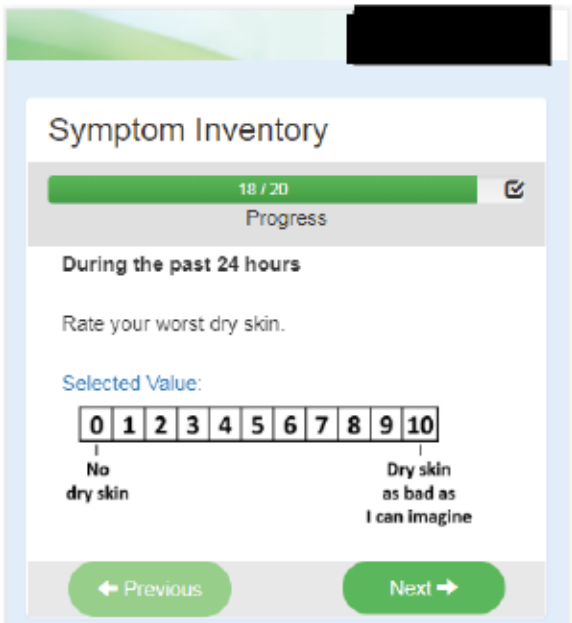


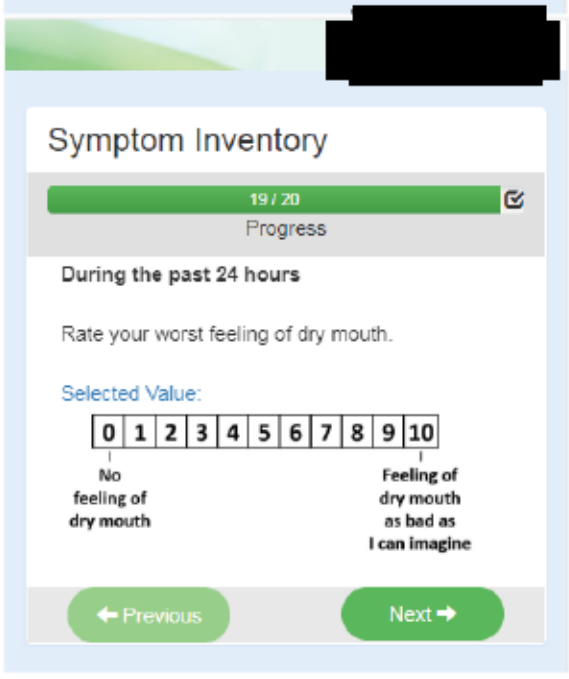



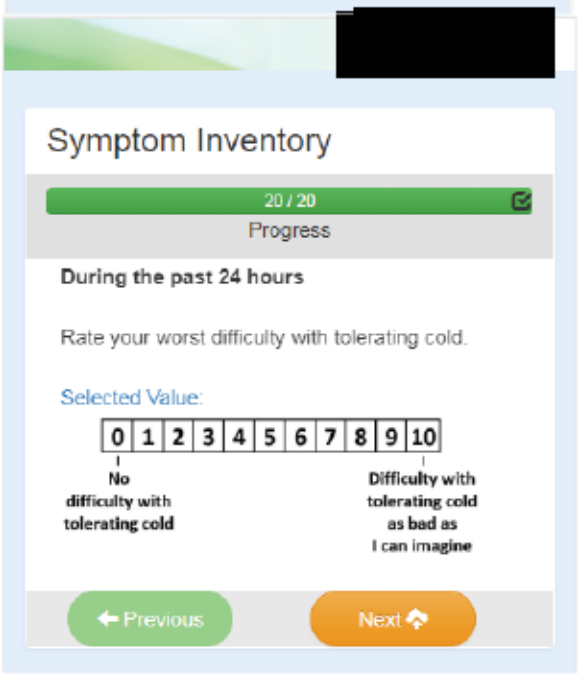





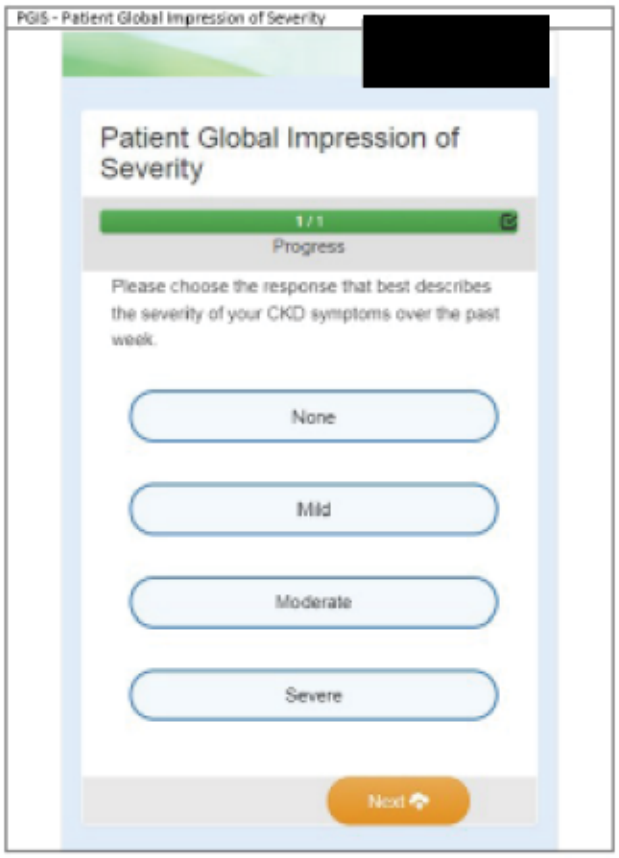
# APPENDIX C: SHORT FORM 36-ITEM HEALTH SURVEY V2 (SF-36V2)

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| | Your Health and Well-Being | | | | | | |
| | This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. Thank you for completing this survey!<br><br>For each of the following questions, please select the one response that best describes your answer. | | | | | | |
| SF36v2_GH1 | In general, would you say your health is: | Excellent | Very good | Good | Fair | Poor | |
| SF36v2_HT | Compared to one year ago, how would you rate your health in general now? | Much better now than one year ago | Somewhat better now than one year ago | About the same as one year ago | Somewhat worse now than one year ago | Much worse now than one year ago | |
| | The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? | | | | | | |
| SF36v2_PF01 | Does your health now limit you in vigorous activities, such as running, lifting heavy objects, participating in strenuous sports? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF02 | Does your health now limit you in moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF03 | Does your health now limit you in lifting or carrying groceries? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF04 | Does your health now limit you in climbing several flights of stairs? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF05 | Does your health now limit you in climbing one flight of stairs? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF06 | Does your health now limit you in bending, kneeling, or stooping? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF07 | Does your health now limit you in walking more than a mile? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF08 | Does your health now limit you in walking several hundred yards? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF09 | Does your health now limit you in walking one hundred yards? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF10 | Does your health now limit you in bathing or dressing yourself? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| | During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health? | | | | | | |

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| SF36v2_RP1 | During the past 4 weeks, how much of the time have you cut down on the amount of time you spent on work or other activities as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP2 | During the past 4 weeks, how much of the time have you accomplished less than you would like as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP3 | During the past 4 weeks, how much of the time were you limited in the kind of work or other activities as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP4 | During the past 4 weeks, how much of the time have you had difficulty performing the work or other activities as a result of your physical health (for example, it took extra effort)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| | During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? | | | | | | |
| SF36v2_RE1 | During the past 4 weeks, how much of the time have you cut down on the amount of time you spent on work or other activities as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RE2 | During the past 4 weeks, how much of the time have you accomplished less than you would like as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RE3 | During the past 4 weeks, how much of the time have you done work or other activities less carefully than usual as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_SF1 | During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups? | Not at all | Slightly | Moderately | Quite a bit | Extremely | |
| SF36v2_BP1 | How much bodily pain have you had during the past 4 weeks? | None | Very mild | Mild | Moderate | Severe | Very severe |
| SF36v2_BP2 | During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)? | Not at all | A little bit | Moderately | Quite a bit | Extremely | |
| | These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. | | | | | | |

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| SF36v2_VT1 | How much of the time during the past 4 weeks did you feel full of life? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH1 | How much of the time during the past 4 weeks have you been very nervous? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH2 | How much of the time during the past 4 weeks have you felt so down in the dumps that nothing could cheer you up? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH3 | How much of the time during the past 4 weeks have you felt calm and peaceful? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT2 | How much of the time during the past 4 weeks did you have a lot of energy? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH4 | How much of the time during the past 4 weeks have you felt downhearted and depressed? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT3 | How much of the time during the past 4 weeks did you feel worn out? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH5 | How much of the time during the past 4 weeks have you been happy? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT4 | How much of the time during the past 4 weeks did you feel tired? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_SF2 | During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| | How TRUE or FALSE is each of the following statements for you? | | | | | | |
| SF36v2_GH2 | I seem to get sick a little easier than other people. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH3 | I am as healthy as anybody I know. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH4 | I expect my health to get worse. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH5 | My health is excellent. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| | SF-36v2® Health Survey © 1992, 2000, 2009 Medical Outcomes Trust and QualityMetric Incorporated. All rights reserved. SF-36® is a registered trademark of Medical Outcomes Trust. (SF-36v2® Health Survey Standard, United States (English)) | | | | | | |

# APPENDIX D: PGIC AND PGIS




# APPENDIX D: EQ-5D-5L

| EQ-5D-5L | |
|---|---|
| EQ-5D-5L PDA version | |
| English (USA) | Country (Language) |
| **Health Questionnaire** | Health Questionnaire |
| **English version for the USA** | Version (Target Language)<br>Version (English) |
| On the following screens please tap the statement that best describes your health TODAY. | Instruction |
| **Your mobility TODAY** | **Mobility** |
| I have no problems walking | MB1 |
| I have slight problems walking | MB2 |
| I have moderate problems walking | MB3 |
| I have severe problems walking | MB4 |
| I am unable to walk | MB5 |
| **Your self-care TODAY** | **Self-care** |
| I have no problems washing or dressing myself | SC1 |
| I have slight problems washing or dressing myself | SC2 |
| I have moderate problems washing or dressing myself | SC3 |
| I have severe problems washing or dressing myself | SC4 |
| I am unable to wash or dress myself | SC5 |
| **Your usual activities TODAY** *(e.g. work, study, housework, family or leisure activities)* | **Usual Activities** |
| I have no problems doing my usual activities | UA1 |
| I have slight problems doing my usual activities | UA2 |
| I have moderate problems doing my usual activities | UA3 |
| I have severe problems doing my usual activities | UA4 |
| I am unable to do my usual activities | UA5 |
| **Your pain / discomfort TODAY** | **Pain / Discomfort** |
| I have no pain or discomfort | PD1 |
| I have slight pain or discomfort | PD2 |
| I have moderate pain or discomfort | PD3 |
| I have severe pain or discomfort | PD4 |
| I have extreme pain or discomfort | PD5 |
| **Your anxiety / depression TODAY** | **Anxiety / Depression** |
| I am not anxious or depressed | AD1 |
| I am slightly anxious or depressed | AD2 |
| I am moderately anxious or depressed | AD3 |
| I am severely anxious or depressed | AD4 |
| I am extremely anxious or depressed | AD5 |
| We would like to know how good or bad your health is TODAY. | Vas Line 1 |
| On the next screen you will see a scale numbered 0 to 100. | Vas Line 2 |
| 100 means the best health you can imagine. | Vas Line 3 |
| 0 means the worst health you can imagine. | Vas Line 4 |
| Please tap on the scale to indicate how your health is TODAY. | Vas Line 5 |
| The best health you can imagine | Top Scale |
| The worst health you can imagine | Bottom Scale |
| YOUR HEALTH TODAY | Box Health |



***Disclaimer:** This is a preview of the EQ-5D instrument. It demonstrates the text, questions and response options included in this version. This preview does not represent the final product and should not be used as an official EQ-5D instrument.*

# APPENDIX D: PROMIS FATIGUE 13A, FATIGUE, SEX INTEREST

## PROMIS Items

PROMIS Short Form v1.0 – Fatigue 13a

| During the past 7 days..... | Not at all | A little bit | Somewhat | Quite a bit | Very much |
|---|---|---|---|---|---|
| I feel fatigued | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I feel weak all over | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I feel listless ("washed out") | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I feel tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I have trouble starting things because I am tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I have trouble finishing things because I am tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I have energy | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I am able to do my usual activities | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I need to sleep during the day | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I am too tired to eat | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I need help doing my usual activities | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I am frustrated by being too tired to do the things I want to do | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| I have to limit my social activity because I am tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |

PROMIS items (Item identifier in bold)

| | | | | | |
|---|---|---|---|---|---|
| **FATEXP36:** How exhausted were you on average? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| **FATEXP43:** How physically drained were you on average? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| **FATIMP52:** To what degree did your fatigue make you feel less alert? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| **During the past 30 days.....** | **Not at all** | **A little bit** | **Somewhat** | **Quite a bit** | **Very much** |
| **SFINT101:** How interested have you been in sexual activity? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |

## Document Approval Signature(s)

This document has been signed electronically in compliance with [redacted] and [redacted] Companies' electronic signature policies and procedures. This page is a manifestation of the electronic signatures, which may be confirmed within the [redacted] and [redacted] Companies' source system in accordance with [redacted] and [redacted] Companies' records retention policies. Note: the Title field in the signature block has been intentionally left blank.

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

# Patient-Reported Outcomes Statistical Analysis Plan

A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE)

**PROTOCOL NUMBER:** COR-001-02

**Study Sponsor**:

Corvidia Therapeutics Inc.
35 Gatehouse Drive
Waltham MA 02451, USA

**Author**: [redacted]

**Version Number and Date**: V1.0, 22APR2020

# Table of Contents

**LIST OF FIGURES**



**LIST OF TABLES**

# PRO STATISTICAL ANALYSIS PLAN SIGNATURE PAGE

**Statistical Analysis Plan V1.0 (Dated 18MAR2020) for Protocol COR-001-02.**

| | Name | Signature | Date |
|---|---|---|---|
| **Author:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |

Upon review of this document, the undersigned approves this version of the PRO Statistical Analysis Plan, including the table and figure shells, authorizing that the content is acceptable for the reporting of this study.

| | Name | Signature | Date |
|---|---|---|---|
| **Approved By:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |
| | | | |
| **Approved By:** | [redacted] | [redacted] | [redacted] |
| **Position:** | [redacted] | | |
| **Company:** | Corvidia | | |

## MODIFICATION HISTORY

| Unique Identifier for this Version | Date of the Document Version | Significant Changes from Previous Authorized Version |
|---|---|---|
| 1.0 | 22APR2020 | Not Applicable – First Version |

## LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| ANCOVA | Analysis of Covariance |
| BP | Bodily Pain |
| CARES | Corvidia ePRO |
| CDF | Cumulative Distribution Function |
| CI | Confidence Interval |
| CKD | Chronic Kidney Disease |
| EDC | Electronic Data Capture |
| EOT | End of Treatment |
| ePRO | Electronic Patient-Reported Outcomes |
| GH | General Health Perceptions |
| HRQoL | Health-Related Quality of Life |
| hs-CRP | High-Sensitivity C-Reactive Protein |
| ITT | Intention-To-Treat |
| LS | Least Squares |
| MAR | Missing at Random |
| MCS | Mental Component Summary |
| MH | Mental Health |
| MMRM | Mixed Model Repeated Measures |
| MNAR | Missing Not at Random |
| PAP | Psychometric Analysis Plan |
| PCS | Physical Component Summary |
| PDF | Probability Density Function |
| PF | Physical Functioning |
| PGIC | Patient Global Impression of Change |
| PGIS | Patient Global Impression of Severity |
| PRO | Patient-Report Outcomes |
| PROMIS Sex FS | PROMIS Sexual Function and Satisfaction |
| RE | Role Limitations Due to Emotional Problems |
| REML | Restricted Maximum Likelihood |
| RP | Role Limitations Due to Physical Health |
| SAA | Serum Amyloid A |
| SAP | Statistical Analysis Plan |
| SD | Standard Deviation |
| SF | Social Functioning |
| SMD | Standardized Mean Difference |
| TSS | Total Symptom Score |
| TTCD | Time to First Confirmed Deterioration |

| Abbreviation | Definition |
|---|---|
| TTFD | Time to First Deterioration |
| VAS | Visual Analogue Scale |
| VT | Vitality |

# 1. INTRODUCTION

This document describes the rules and conventions to be used in the presentation and analysis of patient-reported outcomes (PRO) data collected in the COR-001-02 study. It describes the data to be summarized and analyzed, including specifics of the statistical analyses to be performed. It is intended to ensure the credibility of the study PRO results by pre-specifying the statistical approaches for the analysis of PRO data prior to database lock.

This statistical analysis plan (SAP) is based on protocol version amendment 5 dated 23 December 2019.

This plan may be revised during the study to accommodate protocol amendments and/or to make changes to adapt to unexpected issues in study execution and/or data that affect planned analyses. The final plan, if revised, will document all changes and be issued prior to database lock.

Due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, an immediate discontinuation of dosing was announced on March 18, 2020 to ensure the safety of trial patients. The announcement required the early termination visit to be scheduled as early as possible for ongoing patients followed by their entry into the safety follow-up period.

At the time of discontinuation of the study, approximately 140 patients had completed the 24-week treatment period and more than 120 patients were non-completers of the treatment period.

## 1.1. OBJECTIVES OF THE PRO ANALYSIS

The aim of PRO analysis is to assess the benefits of ziltivekimab compared to placebo on disease-related symptoms, pain, and health related quality of life (HRQoL) in adult subjects with Stage 3-5 Chronic Kidney Disease (CKD).

Key objectives of the PRO analyses are:

- To assess the impact of ziltivekimab vs. placebo on symptoms (collected in Corvidia ePRO (CARES), Patient-Reported Outcomes Measurement Information System (PROMIS®) Fatigue Short Form 13a and selected items from the PROMIS Item Bank) as measured by mean changes from baseline to Week 13 and Week 24.

Supportive objectives are:

- To assess the time to first clinically meaningful deterioration, as well as time to first confirmed clinically meaningful deterioration, in symptoms as measured by CARES, PROMIS Fatigue Short Form 13a and selected items from the PROMIS Item Bank.
- To assess responder proportions at Week 13 and Week 24 in symptoms as measured by CARES, PROMIS Fatigue Short Form 13a and selected items from the PROMIS Item Bank.
- To examine the range of the responder definitions by looking into the Cumulative Distribution Function (CDF) and Probability Density Function (PDF) plots of ziltivekimab vs placebo.

- To assess the impact of ziltivekimab vs. placebo on functioning and HRQoL as measured by mean changes from baseline at Week 13 and Week 24 on physical and mental component scores as collected in SF-36, as well as on overall HRQoL, as measured by the EQ-5D-5L.

# 2. COR-001-02 STUDY OVERVIEW

## 2.1. STUDY OBJECTIVES

### 2.1.1. PRIMARY OBJECTIVE

The primary objective of the RESCUE study is:

- To evaluate the effects of ziltivekimab compared to placebo on markers of inflammation such as high-sensitivity C-reactive protein (hs-CRP).

### 2.1.2. SECONDARY OBJECTIVES

The secondary objective is:

- To evaluate the effects of ziltivekimab compared to placebo on two markers of inflammation and cardiovascular risk: serum amyloid A (SAA) and fibrinogen.

### 2.1.3. EXPLORATORY OBJECTIVES

Selected exploratory objectives related to PRO are:

- To determine the pharmacokinetic, exploratory pharmacodynamics, pharmacogenetics, and effect of ziltivekimab on inflammatory markers.
- To evaluate the effects of three dose levels of ziltivekimab compared to placebo on PRO: PROMIS Fatigue 13a short form, selected items from the PROMIS fatigue item bank, the Optum SF-36 v2® HealthSurvey, CARES, the PROMIS interest in sexual activity item, the patient global impression of change (PGIC), patient global impression of severity (PGIS), and the EQ-5D-5L.
- To evaluate the psychometric properties of the CARES items, PROMIS Fatigue 13a short form, and selected items from the PROMIS fatigue item bank, in CKD patients.

## 2.2. SAMPLE SIZE

The primary efficacy endpoint is percent change from baseline in hs-CRP (average of the hs-CRP value prior to randomization and Day 1) to Week 13 between each active group and placebo.

Based on the observed treatment difference in percent change from baseline in hs-CRP of

60.74% between combined COR-001-01 active groups and placebo and the associated pooled SD of 16.893% in hs-CRP at Week 4 from the final analysis of study COR-001-SC1, a sample size of 54 per group yields more than 99% power with 2-sided alpha=0.05.

Taking into consideration the dropout rate of 10% by the end of the study, a sample size of 60 per group is planned for this study. Accordingly, approximately 240 patients will be randomized 1:1:1:1 (60 per each treatment group) into the trial. Patient randomization will be stratified by baseline hemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5).

## 2.3. STUDY DESIGN

The RESCUE study is a randomized, double-blind, placebo-controlled trial designed to evaluate the efficacy, safety, and pharmacokinetics of ziltivekimab at three dose levels (7.5 mg, 15 mg or 30 mg) compared to placebo in patients with stage 3-5 CKD, not on dialysis, who have evidence of inflammation with high cardiovascular risk.

The primary, secondary, and exploratory endpoints will be analysed at 13 weeks of dosing and then followed for additional exploratory efficacy analyses through Week 24. Selected efficacy endpoints and safety assessments will be evaluated in the Follow-up Period Week 25 though Week 32 (Figure 1).

Patients will undergo a Screening Period of up to 14-days during which inclusion and exclusion criteria will be evaluated. Patients who meet all inclusion criteria and no exclusion criteria will be randomized to one of three ziltivekimab dose levels (7.5 mg, 15mg, or 30 mg) or placebo for a 24-week Treatment Period. Patient randomization will be stratified by baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4, or 5).

The patient will be randomized on Day 1 and the first dose of study drug should be administered after all assessments are conducted. Doses of study drug will be administered every 28-days for a total of 6 treatments (Weeks 1, 5, 9, 13, 17, and 21).

The test product dose regimens to be examined in this study are:

- Dose #1: Ziltivekimab, 7.5 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21, as a subcutaneous injection.
- Dose #2: Ziltivekimab, 15 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21, as a subcutaneous injection.
- Dose #3: Ziltivekimab, 30 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21, as a subcutaneous injection.

The reference dose regimens to be examined in this study are:

- Matched placebo injections administered subcutaneously 6 times (every 28 days) at the same frequency as the active treatment in a given dose cohort, on Weeks 1, 5, 9, 13, 17, and 21.

**Figure 1: Study Flow Diagram**

Ziltivekimab Dose Level 1, 2 or 3 (N=180); N=60 in each level

Placebo (N=60)

| | Screening | | Treatment Period | | | | | | | | | | | Safety F/Up | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11/ET | 12 | 13/ET |
| Visit Day Window | -14 | -7 | 1 | 8 | 22 | 29-35 | 43-49 | 57-63 | 85-91 | 113-119 | 141-147 | 155-161 | 162-168 | 190-196 | 218-224 |
| Visit Window | | ±2 | 1 | ±2 | ±2 | | | | | | | | | | |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |

## 2.4. PRO INSTRUMENTS

The following PRO instruments are collected in this study:

- CARES
- PROMIS Fatigue 13a short form
- Selected items from the PROMIS fatigue item bank
- PROMIS interest in sexual activity item
- Optum SF-36 v2® Health Survey
- EQ-5D-5L
- PGIS and PGIC.

### 2.4.1. CARES CORVIDIA

The CARES Corvidia ePRO - henceforth CARES - is a new PRO instrument under development. It consists of 19 symptom items, with a 24-hour recall, asking patients to report their worst level of that symptom in the past 24 hours on a numeric rating scale from 0 (no symptom) to 10 (symptom as bad as I can imagine).

The symptom items cover physiological (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea), psychological (Depression, Concentration, Forgetfulness), Pain (Bone / Joints, Nerves, Muscle Cramps) and sensorial (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold) symptoms.

The instrument can be found in Appendix A. CARES.

### 2.4.2. PROMIS FATIGUE 13A SHORT FORM AND SELECTED ITEMS FROM PROMIS FATIGUE BANK, PROMIS INTEREST IN SEXUAL ACTIVITY ITEM

The PROMIS Fatigue 13a short form contains 13 items that assess symptoms and impacts of fatigue. The recall period is 7 days. Each item is rated on a five-point Likert scale ranging from 0 = "not at all" to 4 = "very much". Higher scores in these items indicate worse fatigue, except from items 7 ("I have energy") and 8 ("I am able to do my usual activities") where higher score indicates less fatigue.

In this study, three additional fatigue items were administered for further testing to account for exhaustion, being physically drained, and impact on alertness, concepts that resonated with patients during patient interviews but are not covered in the PROMIS Fatigue 13a short form. Such combination is from now on referred to as PROMIS Fatigue.

A number of scales and subscales can be derived from the PROMIS Fatigue set:

- The original Fatigue 13a scale
- The fatigue scale using the whole PROMIS Fatigue item set (13 + 3 de novo 16-item short form)
- Two subscales within the PROMIS Fatigue 13a scale: 1) symptoms of fatigue; and 2) impacts of fatigue
- Two subscales within the PROMIS Fatigue item set (13 + 3 de novo 16-item short form): 1) symptoms of fatigue; and 2) impacts of fatigue.

The PROMIS Sexual Function and Satisfaction (PROMIS Sex FS) measure is a customizable, self-reported set of measures that include 79 items covering 11 domains: interest in sexual activity, lubrication, vaginal discomfort, erectile function, global satisfaction with sex life, orgasm, anal discomfort, therapeutic aids, sexual activities, interfering factors, and screener questions. The PROMIS Sex FS uses a 30-day recall period. In this study only one item will be used: interest in sexual activity. This item is rated on a five-point Likert scale ranging from 0 = "not at all" to 4 = "very much". Higher scores in this item indicate more interest in sexual activity.

These three instruments will be administered together as shown in Appendix B. PROMIS Fatigue 13a Short Form, additional fatigue items, sexual activity.

### 2.4.3. SHORT FORM 36-ITEM HEALTH SURVEY, VERSION 2 (SF-36V2)

SF-36v2 is a self-report survey of functional health and well-being with 4 weeks recall period (QualityMetric, 2011). Responses to 35 of the 36 items are used to compute an 8-domain profile of functional health and well-being scores.

The remaining item, referred to as the 'Health Transition' item, asks patients to rate how their current state of health compared to their state of health 1 year ago, and is not used to calculate domain scores.

The 8-domain profile consists of the following scales: Physical Functioning (PF), Role Limitations due to Physical Health (RP), Bodily Pain (BP), General Health Perceptions (GH), Vitality (VT), Social Functioning (SF), Role Limitations due to Emotional Problems (RE), and

Mental Health (MH).

Psychometrically-based physical and mental health component summary scores (PCS and MCS, respectively) are computed from subscale scores to give a broader metric of physical and mental health status. The instrument can be found in Appendix C. SF-36.

### 2.4.4. EQ-5D-5L

The EuroQol EQ-5D-5L has been designed as an international, standardized, generic instrument for describing and valuing health-related quality of life, available in over 100 language versions.

The EQ-5D-5L self-report questionnaire consists of 5 items, one per domain, with 5-point scale and one visual analogue scale (VAS) to rate health state from worst (0) to best (100).

The EQ-5D-5L includes domains for each generic health status measure: Mobility, Self-care, Usual Activities, Pain/discomfort and Anxiety/depression. For each question, there are 5 levels of response, corresponding to increasing levels of impairment (no problems, slight, moderate, severe, and extreme problems, or unable to perform activity) and coded 1 to 5.

A higher index indicates better quality of life. The instrument can be found in Appendix D. EQ-5D-5L.

### 2.4.5. Patient Global Impression of Severity (PGIS) and Change (PGIC)

The PGIS is a self-reported measure of patient-perceived overall symptom severity. The PGIS is a 4-point recall scale (from no symptoms to severe symptoms). Patients are asked to rate their overall severity of symptoms in the 7 days prior to a clinic visit.

The PGIC instrument captures the patient's overall evaluation of response to treatment. The patient is asked to report the degree to which they have changed since entering the treatment period using a 7-point scale (Very Much Better to No Changes to Very Much Worse).

The PGIC and PGIS are the most commonly used anchor-based methods of assessing clinically important change and severity in which the external judgment of meaningful change is made by the patient. The instruments can be found in Appendix E. PGIC and PGIS.

### 2.4.6. Assessment Schedule for PRO Instruments

Patient-reported outcomes will be assessed through a personal device supplied by Corvidia:

- At home daily on patient's personal device: CARES
- At the clinic prior to dosing on patient's personal device: PROMIS Fatigue 13a Short Form and selected items from PROMIS Fatigue item bank, PROMIS Interest in Sexual Activity item, SF-36v2, EQ-5D-5L, PGIC, PGIS.

CARES data will be collected at home on patient's personal device or one provided by Corvidia; the questionnaire will be answered daily for 7 days prior to treatment administration

(Screening day -7 to day -1 prior to dose 1 (Visit 1), dose 2 (Visit 4), dose 4 (Visit 7), and for the last 2 weeks after dose 6 (Visit 9), i.e. during weeks 23 and 24.

The SF-36v2 and PROMIS Fatigue items will be collected on patients' device at the clinic prior to treatment administration at Visit 1, Visit 4, Visit 7, Visit 10 and 11, or the early termination visit (whichever occurs first).

The PROMIS interest in sexual activity item and EQ-5D-5L will be collected on patients' device at the clinic prior to treatment administration at Visit 1, Visit 4, Visit 7 and Visit 11, or the early termination visit (whichever occurs first).

The PGIC will be colected on patients' device at the clinic prior to to treatment administration at Visit 4, Visit 7, and Visit 11, or the early termination visit (whichever occurs first).

The PGIS will be collected at home on patients' devices on the last day of CARES assessment period, i.e. day prior to dose 1 (Visit 1), dose 2 (Visit 4), dose 4 (Visit 7), and at Visit 11, or the early termination visit (whichever occurs first).

The schedule of assessment for all PRO instruments collected in different periods of the study are shown in Table 1.

**Table 1 PRO Instruments Assessment Schedule**

| Study Period | Screening | | Treatment | | | | | | | | | | | | Safety F/Up[a] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | N/A | 7 | 8 | 9 | 10 | 11/ET[b] | 12 | 13/ET |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 12 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |
| DOSING | | | X | | | X | | X | | X | X | X | | | | |
| CARES | | X | | | X | | | | X | | | | X | X | | |
| PROMIS Fatigue | | | X | | | X | | | | X | | | X | X | | |
| PROMIS Sex Item | | | X | | | X | | | | X | | | | X | | |
| SF-36v2 | | | X | | | X | | | | X | | | X | X | | |
| EQ-5D-5L | | | X | | | X | | | | X | | | | X | | |
| PGIS | | X | | | X | | | | X | | | | | X | | |
| PGIC | | | | | | X | | | | X | | | | X | | |

[a]Follow-Up.
[b]Early Termination.

## 3. ANALYSIS SETS

All PRO analyses described in this SAP will be performed on the intention-to-treat (ITT) analysis population which includes all randomized patients.

# 4. ANALYSIS VARIABLES

## 4.1. GENERAL VARIABLES AND DERIVATIONS

### 4.1.1. STUDY DAY

Study day will be defined as in the clinical SAP dated 31MAR2020. Specifically, the study day is calculated in reference to the date of the first dose of the study drug. Study day 1 corresponds to the date the patient received the first dose of study drug. The day immediately before Day 1 will be Day -1. For assessments conducted on or after the date of the first dose of study drug, treatment day will be calculated as (assessment date - date of first dose of study drug) + 1.

### 4.1.2. BASELINE

The assessment completed on day 1 will be considered the baseline measurement, if taken prior to first dose of study drug.

If more than one assessment were taken, the closest to enrollment day (Day 1) will be used.

For diary scores (CARES), baseline assessment corresponds to the 7 days prior to first dose of study drug.

The post-baseline value is defined as a measurement taken after initial study drug administration.

### 4.1.3. DERIVED TIMEPOINTS

An "End of Treatment" (EOT) assessment will be derived as described in the clinical SAP.

For the purpose of the PRO analyses defined herein, a Week 24 assessment will be used in the analyses. The Week 24 assessment is defined for each instrument in the correspoding sections below.

### 4.1.4. VISIT WINDOW

Visit windows will be used as provided by the sponsor. According to the clinical SAP dated 31MAR2020, scheduled visits will be assigned to analysis visits as recorded in the EDC system. If a scheduled visit is not available, unscheduled and early termination visits will be assigned to analysis visits using analysis visit windows based on the actual date the assessment took place. The low analysis day of the analysis window will be calculated as the midpoint between the scheduled assessment and previously scheduled assessment for that parameter. The high analysis day of the analysis window will be calculated as the midpoint between the scheduled assessment and the next scheduled assessment for that parameter. Where multiple measurements for a particular parameter appear within an analysis window, the scheduled visit will be used. If no scheduled visit appears in the

analysis window, the result closest to the target day will be used. If equidistant and both are unscheduled and/or early termination visits, the later result will be used for the summary measure.

**Table 2: Visit windows**

| Analysis Visit | Target Day | Low Analysis Day | High Analysis Day |
|---|---|---|---|
| Baseline | 1 | | 1 |
| Week 2 | 8 | 2 | 14 |
| Week 4 | 22 | 15 | 28 |
| Week 5 | 35 | 29 | 42 |
| Week 7 | 49 | 43 | 55 |
| Week 9 | 63 | 56 | 70 |
| Week 13 | 85 | 71 | 99 |
| Week 17 | 114 | 100 | 128 |
| Week 21 | 141 | 129 | 153 |
| Week 23 | 157 | 154 | 161 |
| Week 24 | 165 | 162 | 168 |
| End of Treatment | 161 | 154 | 168 |
| Week 28 | 190 | 169 | 211 |
| Week 32 | 224 | 211 | 236 |

### 4.1.5. Other Derivations

No other derivations besides PRO endpoints will be described in this analysis plan.

To calculate time interval duration, a month is 30.4375 days and a year 365.25 days.

## 4.2. PRO Variables

### 4.2.1. CARES Corvidia ePRO

For each item, weekly scores will be derived as the average of the 7 consecutive daily scores as follows:

- Baseline = average of Day -7 to Day -1
- Week 5 (Visit 4) = average of Day 22 to Day 28
- Week 13 (Visit 7) = average of Day 78 to Day 84
- Week 24 (Visit 11) = average of Day 155 to Day 168

For Baseline, Week 5 and Week 13, if more than 3 daily scores out of the 7 days (>50%) within the weekly period are missing, then the score is set to missing.

For Week 24, if more than 7 daily scores out of the 14 days (>50%) within the bi-weekly period are missing, then the score is set to missing.

Preliminary scoring of the CARES includes the calculation of a Total Symptom Score (TSS). Alternative means of scoring the CARES will be explored in a separate psychometric analysis plan (PAP) and if any derived, these will be included in this SAP for efficacy analysis. Additional scales and subscales may be derived based on the results of the psychometric analysis plan (PAP).

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment. At each post-baseline assessment, the change in score from baseline will be calculated for each scale/item.

Domain scores may result from the psychometric analyses described in a separate PAP. In this PAP, clinically meaningful thresholds for these potential domains will also be derived. Change from baseline in the TSS score, as well as the domain scores that may result from the psychometric analysis, will be categorized at each visit as improvement/stable/deterioration using the clinically meaningful threshold value for change scores derived in the psychometric analyses. Sensitivity analyses may be performed if more than one threshold values are suggested by the psychometric analyses.

In addition, patients who experience at least one deterioration during the study will be further classified as follows:

- With definitive deterioration: if the deterioration of at least one-threshold point as compared to the baseline score is also observed at all time points threreafter (e.g., after the first deterioration is observed) or if the patient dropped out after deterioration, resulting in missing data.
- With transient deterioration: otherwise.

### 4.2.2. PROMIS FATIGUE 13A SHORT FORM AND SELECTED ITEMS FROM PROMIS FATIGUE BANK, PROMIS INTEREST IN SEXUAL ACTIVITY ITEMS

The PROMIS Fatigue data is collected as a 7-day recall report from patients at Visit 1, 4, 7, 10, and 11.

The scoring manual for PROMIS Fatigue 13a short form can be found online (Patient-Reported Outcomes Measurement Information System Fatigue scoring manual) (accessed on 16th March 2020).

Each question from the PROMIS Fatigue 13a short form has five response options (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much). To use the scoring tables from the PROMIS fatigue scoring manual, a summed score is calculated. The lowest possible summed score is 13 and the highest possible summed score is 65. All questions must be answered in order to produce a valid score using the scoring tables. Once calculated, the total summed raw score is translated to a T-score. The T-score re-scales the raw score into a standardized score with a mean of 50 and a standard deviation (SD) of 10. The T-score and standard error are then used to calculate the 95% confidence interval (CI) around the observed score. The 95% CI is calculated as T-Score ±1.96*SE. The T-score will not be calculated if a response is missing for one or more questions from the PROMIS Fatigue 13a short form.

In addition, preliminary scoring of the PROMIS Fatigue 13a includes the calculation of two

subdomain scores:

- Symptom score (5 symptom items):
  - I feel fatigued
  - I feel weak all over
  - I feel listless ("washed out")
  - I feel tired
  - I have energy
- Impact score (8 impact items):
  - I have trouble starting things because I'm tired
  - I have trouble finishing things because I'm tired
  - I am able to do my usual activities
  - I need to sleep during the day
  - I am too tired to eat
  - I need help doing usual activities
  - I am frustrated by being too tired to do the things I want to do
  - I have to limit my social activity because I am tired.

The subdomains will be calculated as the sum of the items in the corresponding subdomain. The subdomain score will be calculated only if all items are answered.

Alternative means of scoring, by including also the additional PROMIS Fatigue items from the Item Bank, are described in the PAP and may be derived as follows:

- A total fatigue score summing the items for the PROMIS Fatigue short form 13a and the additional three items from the PROMIS fatigue item bank (de novo 16-item short form)
- Symptom score (8 symptom items):
  - I feel fatigued
  - I feel weak all over
  - I feel listless ("washed out")
  - I feel tired
  - I have energy
  - FATEXP36: How exhausted were you on average?
  - FATEXP43: How physically drained were you on average?
  - FATEXP52: To what degree did your fatigue make you feel less alert?

If psychometric analyses support the creation of the latter two domains, clinically meaningful thresholds will also be derived for them, as well as the three standard fatigue scores from the 13-item set.

Given *PROMIS* Fatigue items and *PROMIS* Interest in Sexual Activity item are collected at Week 23 and at Week 24, an assessment will be derived for the hypothesized scores as the mean value of the two assessments at Week 23 and Week 24, which will be labeled as Week 24. In case only one assessment is available, this only assessment will be used as Week 24 assessment.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment. At each post-baseline assessment, the change in scores from baseline will be calculated for each scale/item.

Different domain scores may result from the psychometric analyses described in a separate PAP. In this PAP, clinically meaningful thresholds for these potential domains, as well as the standard one for PROMIS fatigue -13a short form will also be derived. Change from baseline in the domain scores that may result from the psychometric analysis will be categorized at each visit as improvement/stable/deterioration using the clinically meaningful threshold value for change scores derived in the psychometric analyses. Sensitivity analyses may be performed if more than one threshold values are suggested by the psychometric analyses.

In addition, patients who experience at least one deterioration during the study will be further classified as follows:

- With definitive deterioration: if the deterioration of at least one-threshold point as compared to the baseline score is also observed at all time points threreafter (e.g., after the first deterioration is observed) or if the patient dropped out after deterioration, resulting in missing data.
- With transient deterioration: otherwise.

### 4.2.3. SHORT FORM 36-ITEM HEALTH SURVEY, VERSION 2 (SF-36V2)

The scoring for SF-36 v2 instrument will be done according to the procedures described by the developer and license owner in the instruction manual. Scores for the scales (HT, GH, PF, RP, BP, MH, SF, RE, VT) as well as the summary scores (PCS, MCS) will be derived.

Given SF-36 is collected at Week 23 and at Week 24, an assessment will be derived as the mean value of the two assessments at Week 23 and Week 24, which will be labeled as Week 24. In case only one assessment is available, this only assessment will be used as Week 24 assessment.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment. At each post-baseline assessment, the change in score from baseline will be calculated for each scale.

### 4.2.4. EQ-5D-5L

A unique EQ-5D-5L health state is defined by combining 1 level from each of the 5 dimensions: this defines a profile that is primarily reported as a 5-digit number, for instance 11221. A total of 3125 possible health states are defined in this way. For example, state 11111 indicates no problems on any of the 5 dimensions, while state 12345 indicates no problems with mobility, slight problems with washing or dressing, moderate problems with doing usual activities, severe pain or discomfort, and extreme anxiety or depression.

The instrument was specifically designed to provide an overall single number, called a weighted index, for each of the health states resulting from the combination of item responses (Dolan, 1997). The weighted index constitutes a measure of utility, an economics term used to describe consumer preferences or in the present case patient preferences for different HRQoL states. The weighted index can be only derived from patients who have provided a complete 5-response profile. A higher index indicates better QoL.

(Devlin, 2018) published the value set for England (available on the EuroQoL website (http://www.euroqol.org/about-EQ-5D/valuation-of-EQ-5D/EQ-5D-5l-value-sets.html). However, in August 2018, the UK's National Institute of Health and Care Excellence (NICE) has published a position statement (NICE position statement, available at https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisal-guidance/eq5d5l_nice_position_statement.pdf) advising companies, academic groups, and others preparing evidence submissions to NICE not to use the England validation set to derive utility values for their evidence submissions. In their position statement, NICE recommends using the mapping function developed by (van Hout & al., 2012) to be used (presented in Appendix D. EQ-5D-5L). This is the algorithm to be used in this study.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment for the EQ-5D-5L VAS and utility index scores as well as for the individual domain items.

### 4.2.5. Patient Global Impression of Severity (PGIS) and Change (PGIC)

The PGIS and PGIC are single-item questionnaires. Change from baseline defined as post-baseline value minus baseline value will be calculated for the PGIS assessment. At each post-baseline assessment, the change from baseline will be calculated.

Post-baseline PGIS raw scores will be classified according to the following item response categories:

- Worsening 3 points compared to baseline
- Worsening 2 points compared to baseline
- Worsening 1 point compared to baseline
- Stable
- Improved 1 point compared to baseline
- Improved 2 points compared to baseline
- Improved 3 points compared to baseline.

## 4.3. Time to PRO deterioration

Time to clinically meaningful symptom worsening or HRQoL deterioration (PRO deterioration) will be analysed separately for each scale of the PRO instruments collected in this study, as appropriate and as described in Section 5. For convenience, a generic term "time to clinically meaningful deterioration" will be used both for symptom worsening and

HRQoL deterioration, with an understanding of a specific meaning depending on the scale or subscale analysed. Two definitions will be used and described in the following sections:

- Time to first deterioration (TTFD)
- Time to first confirmed deterioration (TTCD)

### 4.3.1. Time to First Clinically Meaningful Deterioration (TTFD)

TTFD will be defined as the duration of time from the date of randomization to the date of the first deterioration in PRO scores of at least one threshold-point (see section 4.2 for the definition of deterioration) as compared to the baseline score.

For those patients who experienced a first clinically meaningful deterioration, TTFD will be computed as follows and then converted to months:

TTFD = Date of assessment when first clinically meaningful deterioration of at least one threshold unit was observed – Date of randomization + 1

Patients who did not experience clinically meaningful deterioration will be censored at the date of the last available PRO assessment (i.e., date of the last non-missing value). Patients with no baseline assessment or patients with no post-baseline assessments will be censored at the date of randomization.

### 4.3.2. Time to First Confirmed Clinically Meaningful Deterioration (TTCD)

TTCD will be defined as the duration of time from the date of randomization to the date of the first deterioration in PRO scores of at least one threshold-point (see section 4.2 for the definition of deterioration) as compared to the baseline score if the deterioration is also observed at all time points thereafter (e.g., after the first deterioration is observed) or if the patient dropped out after deterioration, resulting in missing data.

For those patients who experienced a definitive meaningful deterioration, TTCD will be computed as follows and then converted to months:

TTCD = Date of assessment when definitive clinically meaningful deterioration was observed – Date of randomization + 1

Patients who did not experience confirmed clinically meaningful deterioration will be censored at the date of the last available PRO assessment (i.e., date of the last non-missing value). Patients with no baseline assessment or patients with no post-baseline assessments will be censored at the date of randomization.

# 5. STATISTICAL METHODOLOGY

## 5.1. GENERAL CONSIDERATIONS

Continuous data will be described by the number of observations (N), the number of missing observations (Nmiss), mean, SD, median, minimum (min), and maximum (max). Categorical data will be described by the number (n) and percentage (%) of patients in each category. Missing and invalid observations will be tabulated as separate categories. The calculation of proportions will not include the missing/invalid category.

Statistical comparisons will be made using two-sided tests at the $\alpha = 0.05$ significance level unless stated otherwise. Due to the exploratory nature of the analyses, adjustments for multiple comparisons will not be made.

Unless otherwise specified, all summaries will be presented by treatment group.

## 5.2. PATIENT DISPOSITION

Patient disposition is included in the clinical SAP and will not be repeated herein.

## 5.3. PRO COMPLETION

Instrument completion rate at each timepoint will be reported for each instrument on the ITT population. The following will be provided:

- The number of patients expected to complete a PRO assessment at each timepoint
- Unadjusted completion rate at each timepoint will be calculated as the number of patients meeting at least the minimum requirements for scoring of the instrument divided by the number of patients in the ITT population.
- Adjusted completion rate at each timepoint will be calculated as the number of patients meeting at least the minimum requirements for scoring of the instrument divided by the number of patients who are expected to have PRO assessments.

The completion rates by treatment group at each analysis visit will also be provided graphically by means of a line graph.

## 5.4. DEMOGRAPHICS AND BASELINE CHARACTERISTICS

Demographics and baseline characteristics are included in the clinical SAP and will not be repeated herein.

## 5.5. DESCRIPTIVE ANALYSES

### 5.5.1. ITEM LEVEL

For all items from the following instruments (CARES, PROMIS Fatigue Short Form 13a and selected items from the PROMIS Item Bank and PROMIS interest in sexual activity item, EQ-5D-5L, PGIS and PGIC), the following will be provided:

- A table with the distribution of response by treatment group for each analysis visit
- A table with the distribution of change in response by treatment group for each analysis visit

In addition, the following graphical representations will be provided:

- A stacked column chart of the distribution of response by treatment group at each analysis visit
- For the PGIS, a stacked column chart of the distribution of change in response categories (as defined in section 4.2.5) by treatment group.

### 5.5.2. DOMAIN AND OVERALL SCORE LEVEL

The following domain and overall scores will be analysed:

- CARES: TSS and all other potential subscales resulting from the psychometric analysis
- PROMIS Fatigue: total, symptoms and impact domains from the 13-item set, as well as other potential subscales resulting from the psychometric analysis
- SF-36:
    - Domains: PF, RP, BP, GH, VT, SF, RE and MH
    - Summary scores: PCS and MCS
- EQ-5D-5L: VAS and utility index

The following analyses will be presented on the ITT population:

- A table with descriptive statistics for the PRO scores and a line graph with mean values and corresponding 95% CI will be presented by treatment group and each PRO assessment
- A table with descriptive statistics for the change from baseline in PRO scores will be presented by treatment group and each post-baseline PRO assessment
- A table showing the proportion of patients achieving ≥0%, 10%, 20%, 30%, 40%, 50% and 60% percentage change from baseline during the study will be presented by treatment group. Bar plots depicting the same proportions by treatment group will be created separately for each time point during the study
- A CDF plot showing a continuous plot of the absolute change from baseline during

the study for the PRO scores on the X-axis and the cumulative percent of patients experiencing that change on the Y-axis will be presented by treatment group and each PRO assessment

- A PDF plot showing a continuous plot of the absolute change from baseline during the study for the PRO scores on the X-axis and the probability density function of that change on the Y-axis will be presented by treatment group and each PRO assessment

## 5.6. LONGITUDINAL ANALYSIS OF CHANGE FROM BASELINE

Change from baseline to Week 13 and Week 24 in PRO scores will be analyzed using a restricted maximum likelihood (REML)-based repeated measures approach (MMRM – Mixed Model Repeated Measures) (Brown & Prescott, 2006) as described in the clinical SAP.

The MMRM assumes that the missing observations are missing at random (MAR). That is, MMRM assumes that, given the statistical model and given the observed values of the outcome, missingness is independent of the unobserved values. A corollary is that MAR assumes that a subject's missing values can be estimated based on similar subjects who remained in the study. This infers that withdrawals (who may not receive study medication) have similar symptoms to some who continue to be treated. Given the expectation that PRO scores could be lower (poorer quality of life and/or more disease-related symptoms) after discontinuing the study medication, the MAR's assumption of the similarity of withdrawals and those who stay in the study may not be realistic for all subjects. To address the possibility of the data being missing not at random (MNAR) (e.g., non-ignorable missing data), a second analysis may be implemented using a PMM with sequential modelling with multiple imputation using placebo-based imputation (O'Kelly & Ratitch, 2014), depending on the amount of missing data at each visit. Further details on this exploration can be found in Section 5.9.2. The PMM will be used to assess the robustness of the MMRM estimate with regard to missing data when the MAR assumption is replaced by assumptions that are likely to be relatively less favorable to the experimental treatment. Due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, an immediate discontinuation of dosing was announced on March 18, 2020 to ensure the safety of trial patients. Therefore it is expected that a large number of patients will not have data at Week 24.

Analyses included in the clinical SAP will not be repeated here. The analyses will be performed on the ITT population and for each of the following PRO scores:

- CARES: TSS and other potential subscales resulting from the psychometric analysis
- PROMIS Fatigue: total, symptoms and impact domain from the 13-item set, as well as other potential subscales resulting from the psychometric analysis
- SF-36:
  - Domains: PF, RP, BP, GH, VT, SF, RE and MH
- EQ-5D-5L: VAS and utility index.

### 5.6.1. MMRM

All PRO assessments will be included for this analysis The analysis will be based on observed data, i.e., data collected at each timepoint without carrying forward previous values. Data from a limited number of PRO assessments may be used in case of substantial dropout (i.e., analysis will be limited to timepoints at which at least 10% of patients have non-missing data in both treatment groups).

The response variable will be the change from baseline to each PRO assessment. The model will include the treatment arm (ziltivekimab vs. placebo) and timepoint (Week 5, Week 13, and week 24) as fixed-effect categorical factors, the baseline PRO score and stratification factors (baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5)), as well as the baseline PRO score x time and treatment x time interactions. Both main effects and the interaction terms will remain in the model, regardless of significance. The model will present least squares (LS) mean estimates for each dose, least squares mean differences of each study drug dose from placebo, standard errors, 95% CIs and p-values (where applicable) for mean changes from baseline to each visit. A plot of the LS means accompanied by the 95% CI will be produced.

In addition, an overall adjusted mean estimate will be derived that will estimate the average change from baseline across all time points, giving each visit equal weight.

The standardized mean difference (SMD) including 95% CI (Hedges' g) will also be provided.

The MMRM analysis will be conducted using PROC MIXED in SAS. The model will assume unstructured covariance among the within-patient repeated measurements. If the algorithm does not converge, a heterogeneous Toeplitz (the TOEPH option in SAS PROC MIXED) will be tried first and then AR(1) as a covariance structure to achieve convergence. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom.

Variables listed as categorical in the list above will be included in the CLASS statement of the procedure. The unique patient identifier will also be included as a class variable. A REPEATED statement over the visits will be included with the unique patient identifier as the SUBJECT variable in the REPEATED statement.

The normality and homoscedasticity of the residuals will be visually checked. Particularly, the scatter plot of the residuals versus the predicted endpoint values, the histograms, and the normal probability plots of the residuals will be reviewed. Transformation of the raw data will be considered if needed.

### 5.6.2. PMM

Change from baseline in PRO domain and overall scores may be further analyzed using a pattern-mixture model using sequential modelling with multiple imputation as described by O'Kelly (O'Kelly & Ratitch, 2014), depending on exploration of the amount of missing data at post-baseline assessments, as described in Section 5.9.2.

The results from this analysis will be used to judge the validity of the MAR assumption. Similar conclusions from MMRM and PMM would suggest that the results are not overly dependent on the assumptions of the primary analysis with regard to the missing data.

The sensitivity analysis will consider dropout reasons while imputing missing values after the discontinuation. Subjects who discontinued due to lack of efficacy or adverse events in the active arms are assumed to have no treatment effect after the discontinuation. These subjects are assumed to copy the profile in the placebo arm, and missing values are imputed based on the distribution estimated from the placebo group under the missing not at random (MNAR), using copy-reference approach. The rest of missing values in the placebo arm and active arms will be imputed using the observed data in their respective group under the MAR assumption. This includes the patients who were early terminated due to to the impact of the Coronavirus Disease 2019 (COVID19) pandemic. The multiple imputation model will include factors such as treatment arm, baseline hemoglobin category, and CKD stage, in addition to the data outcomes at each visit.

The following steps will be performed:

**Step 1** Intermittent (non-monotone) missing data will be imputed using the MCMC option of SAS PROC MI (using seed=5414).

**Step 2** For monotone missing data, the values for each pattern will be imputed via the chained equation method with range restriction, using SAS PROC MI option MONOTONE REG ().

Under step 2, missing values for subjects who discontinued due to lack of efficacy or adverse events will be imputed under MNAR assumption using copy-reference approach, and missing values for subjects who discontinued due to reasons other than lack of efficacy or adverse events will be imputed under the MAR assumption.

These two steps will be carried out sequentially to construct 100 hypothetical complete data sets.

An analysis of covariance (ANCOVA) model (not repeated measures, and with no random effects) with the following covariates will be performed for each imputation using the SAS Mixed procedure:

- HGB = Baseline Hemoglobin category (>=11 or <11g/dL)
- CKD = CKD Stage (3, or 4/5)
- TRT = Treatment group
- BASE = Baseline hemoglobin value

The SAS MIAnalyze procedure will be used to combine the results of these analyses for the imputations. The overall least square means, standard errors, 95% CI, and p-values will be reported.

Reasons for subject discontinuation that are identified as lack of efficacy will be discussed and finalized prior to database lock.

## 5.7. RESPONDER ANALYSIS

The proportion of patients with improvement, who were stable, or who deteriorated (as defined in section 4.2 using both the primary and the potential sensitivity thresholds if these are suggested by the psychometric analysis) will be summarized at each PRO assessment

visit by treatment arms. The denominator in this descriptive analysis will be the number of patients with non-missing data at the particular visit. The analysis will be performed on the ITT population and include only subjects who have an assessment at baseline and at least one post-baseline assessment.

The proportion of patients with confirmed deterioration and transient deterioration during the course of the study will be summarized by treatment arms. The denominator in this descriptive analysis will be the number of patients with at least one deterioration during the study.

The analysis will be performed on the ITT population and for each of the following PRO scale scores:

- CARES: TSS and other potential subscales resulting from the psychometric analysis
- PROMIS Fatigue: total, symptoms and impact domain from the 13-item set, as well as other potential subscales resulting from the psychometric analysis

## 5.8. TIME TO EVENT ANALYSIS

For all time to event analyses, the time to deterioration will be analysed in months and will be presented by treatment group.

TTFD and TTCD will be defined as explained in Section 4.3. The non-parametric Kaplan-Meier method will be used to estimate the survival curves for TTFD and TTCD for each PRO scale score. The 50th percentile of Kaplan-Meier estimates will be used to estimate the median duration of TTFD and TTCD. A two-sided 95% CI will be provided for these estimates. The treatment difference in survival will be assessed by the stratified log-rank test (baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5)). Separate log-rank tests will be performed for each of the experimental arms vs. placebo. A Kaplan-Meier plot by treatment group will be presented.

Kaplan-Meier analysis will be performed using PROC LIFETEST (SAS procedure).

A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie., the hazard ratio), using baseline hemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5) as stratification factor. The hazard ratio and its 95% CI from the stratified Cox model with a single treatment covariate will be reported.

The analysis will be performed on the ITT population and for each of the following PRO scale scores:

- CARES: TSSE and other potential subscales resulting from the psychometric analysis
- PROMIS Fatigue: total, symptoms and impact domain from the 13-item set, as well as other potential subscales resulting from the psychometric analysis

## 5.9. HANDLING OF MISSING DATA

### 5.9.1. MISSING ITEMS

In case of missing items, the scores will be calculated as indicated in the scoring manuals for each of the PRO instruments.

### 5.9.2. MISSING FORMS

It is anticipated that the great majority of missing data in this study will have a monotone pattern, meaning that once a patient has missing data at one visit, data will be missing at all subsequent visits. There may be some small amount of intermittent (non-monotone) missing data (when patient skips intermediate visits but return for evaluations at subsequent visits). The number and percentage of patients for each of the missing data patterns (no missing data, monotone missing data, and intermittent missing data) will be presented by treatment group.

Tabular summaries for the percentage of patients by the reason for discontinuation of study treatment, as well as for withdrawal from the study, are provided in the clinical SAP and will not be repeated herein. A plot of the mean PRO score for selective scores (e.g., CARES TSS and other potential subscales resulting from the psychometric analysis, PROMIS Fatigue: total, symptoms and impact scales from the 13-item set, as well as other potential subscales resulting from the psychometric analysis) over time by selected categories of discontinuation (including completers) will be provided. The reasons of discontinuation will be grouped as follows:

- Death
- Adverse event
- Disease relapse, lack of efficacy, and progressive disease
- Other (lost to follow-up, non-compliance with study drug, pregnancy, protocol deviation, recovery, site terminated by sponsor, technical problems, withdrawal by parent/guardian, withdrawal by subject, other)
- COVID19 pandemic.

# 6. ANALYSIS SOFTWARE

All data processing, summarization, and analyses will be performed using SAS Version 9.4 (SAS Institute, North Carolina) or higher.

# 7. REFERENCES

Brown, H., & Prescott, R. (2006). *Applied Mixed Models in Medicine* (2nd ed.). John Wiley and Sons Ltd.

Devlin, e. a. (2018). An EQ-5D-5L value set for England. *Valuing health-related quality of life*, 7-22.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care 35*, 1095-1108.

Erez, G., Selman, L., & Murtagh, F. (2016). Measuring health-related quality of life in patients with conservatively managed stage 5 chronic kidney disease: limitations of the Medical Outcomes Study Short Form 36: SF-36. *Qual Life Res.*, 2799–2809.

Leaf, D. E., & Goldfarb, D. S. (2009). Interpretation and review of health-related quality of life data in CKD patients receiving treatment for anemia. *Kidney International*, 15-24.

O'Kelly, M., & Ratitch, B. (2014). *Clinical Trials with Missing Data: A Guide for Practitioners.* John Wiley & Sons, Ltd.

*Patient-Reported Outcomes Measurement Information System Fatigue scoring manual.* (n.d.). Retrieved from Health Measures: http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Fatigue_Scoring_Manual.pdf

Pergola, P. E., Pecoits-Filho, R., & Winkelmayer, W. C. (2019). Economic Burden and Health-Related Quality of Life Associated with Current Treatments for Anaemia in Patients with CKD not on Dialysis: A Systematic Review. *PharmacoEconomics*, 463-478.

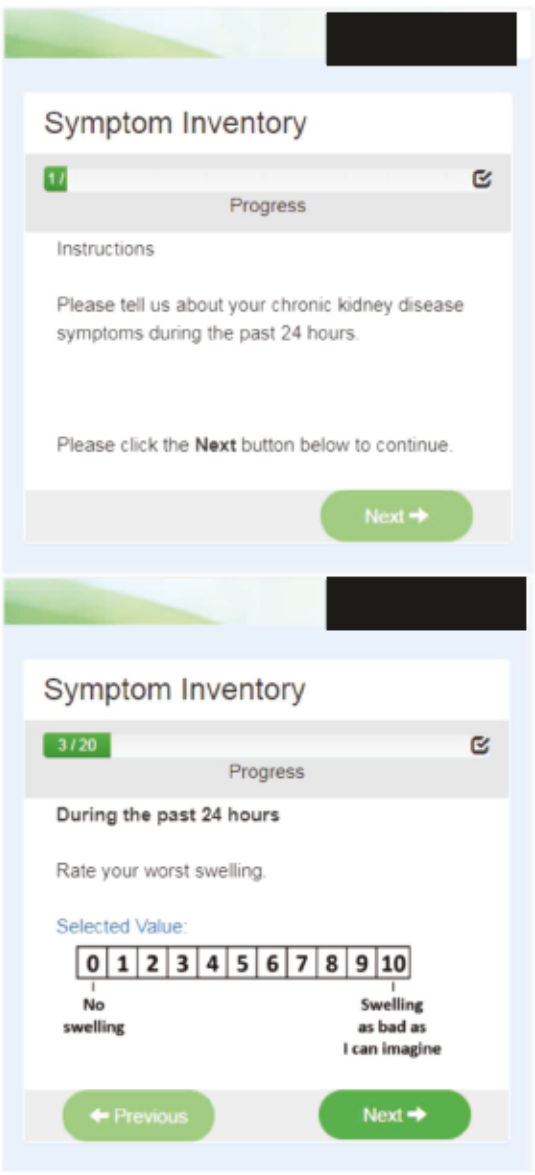
QualityMetric. (2011). User's manual for the SF-36v2 Health Survey (3 ed.). Lincoln, RI.

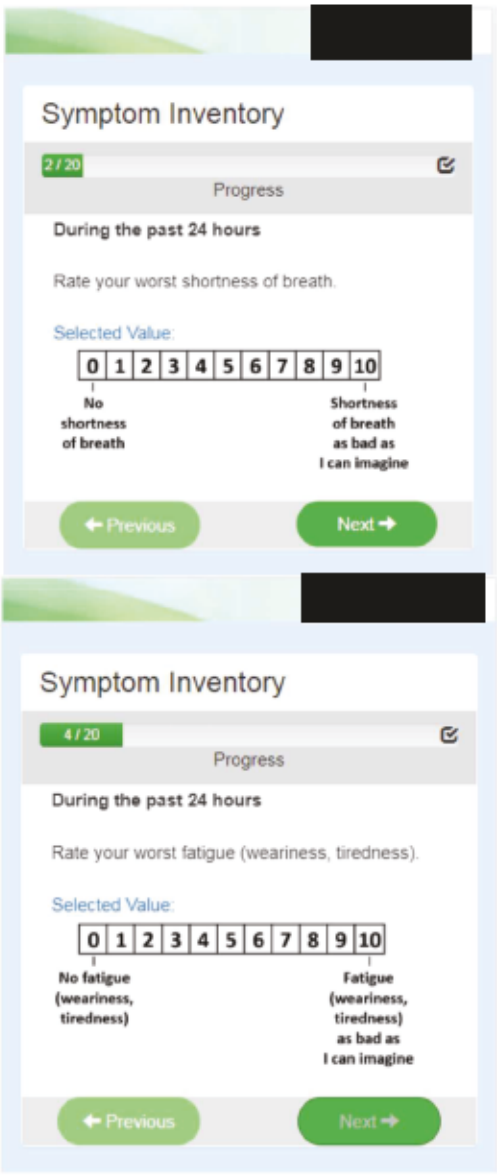
Samsa, G., Edelman, D., & Rothman, M. L. (1999). Determining Clinically Important Differences in Health Status Measures. A General Approach with Illustration to the Health Utilities Index Mark II. *Pharmacoeconomics*, 141-155.

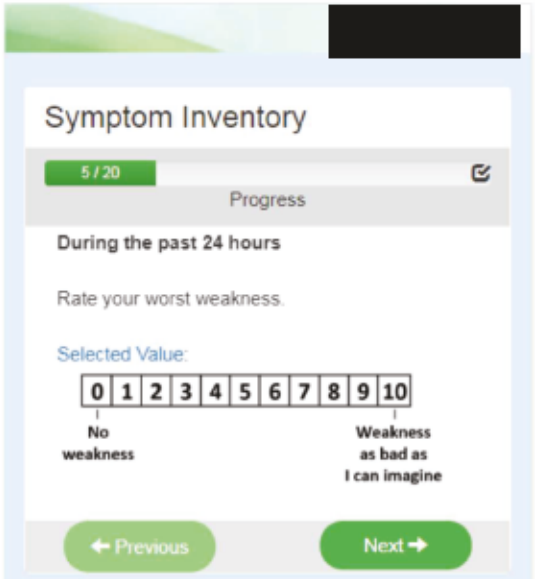
van Hout, B., & al., e. (2012). Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. *Value in Health, 25*, 708-715.

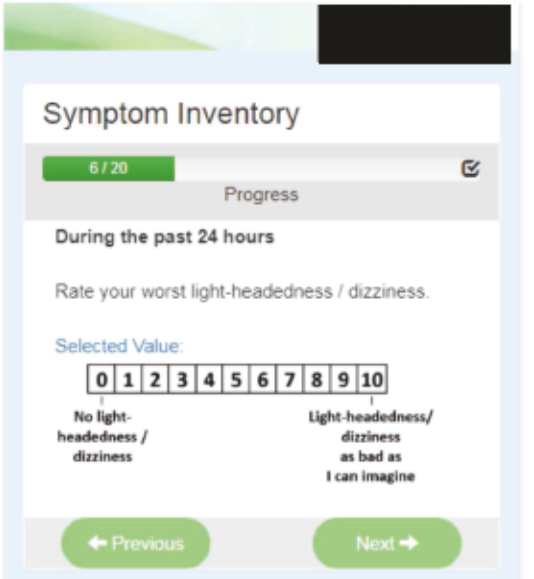
# 8. APPENDICES

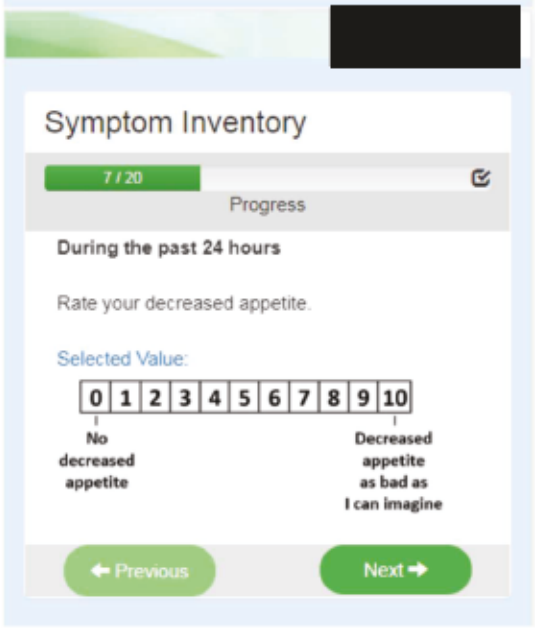
## 8.1. APPENDIX A. CARES

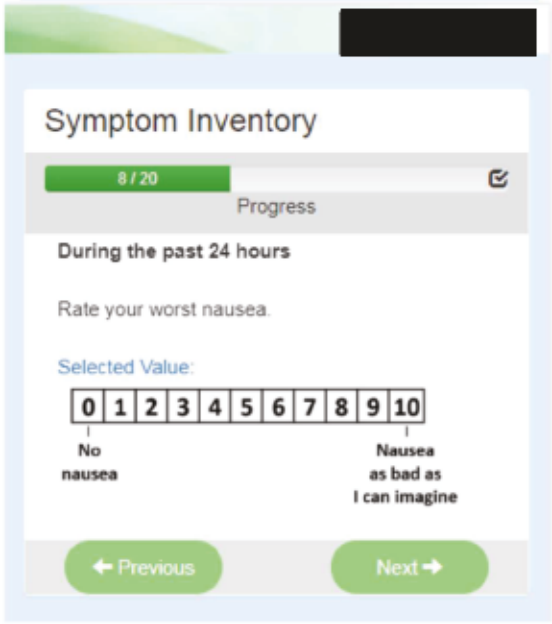


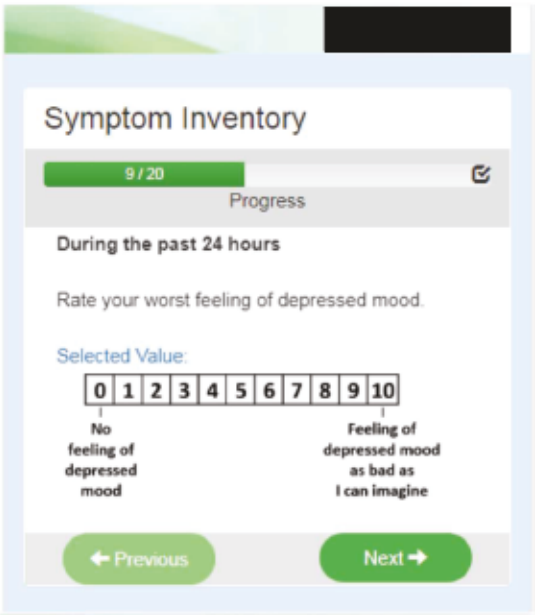



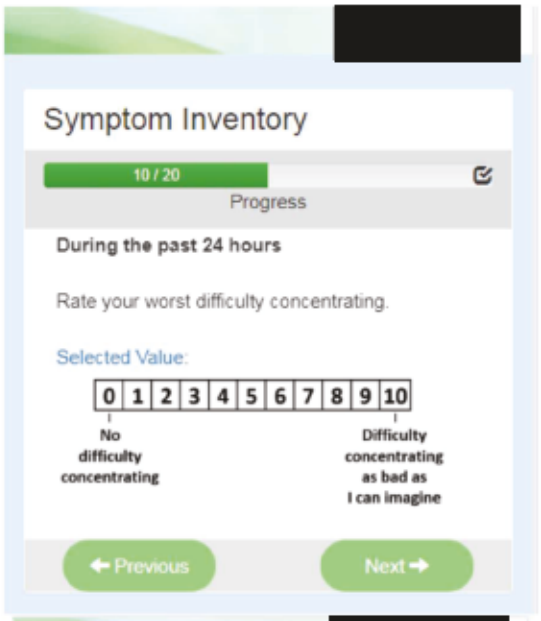


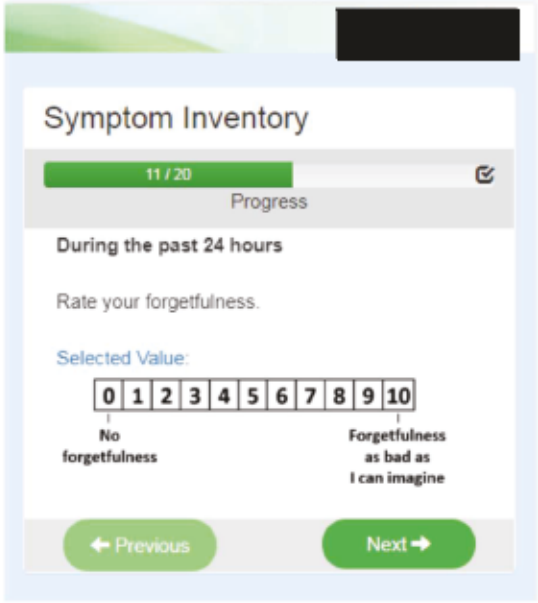



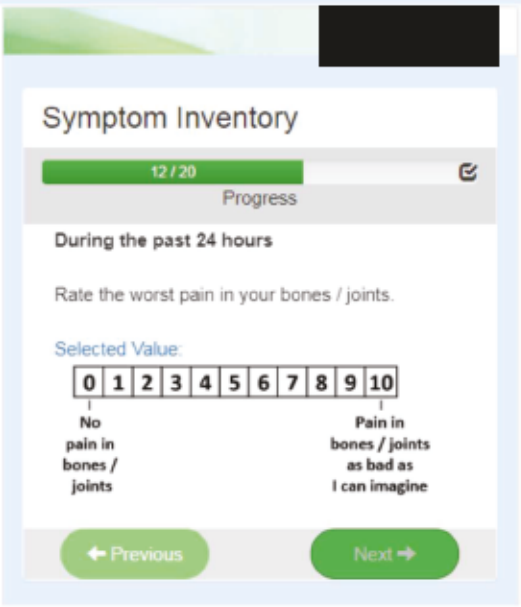



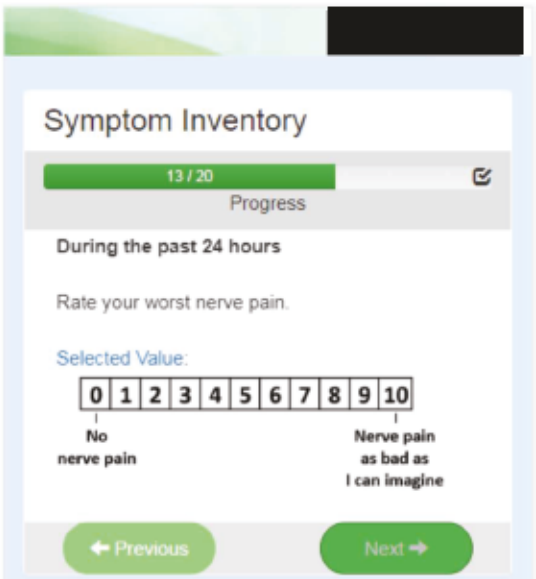


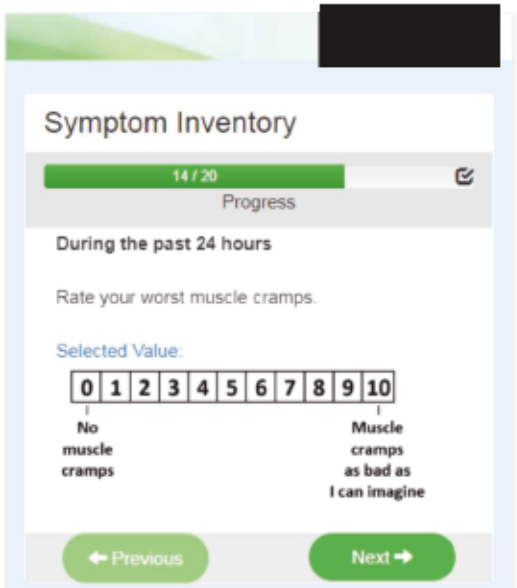



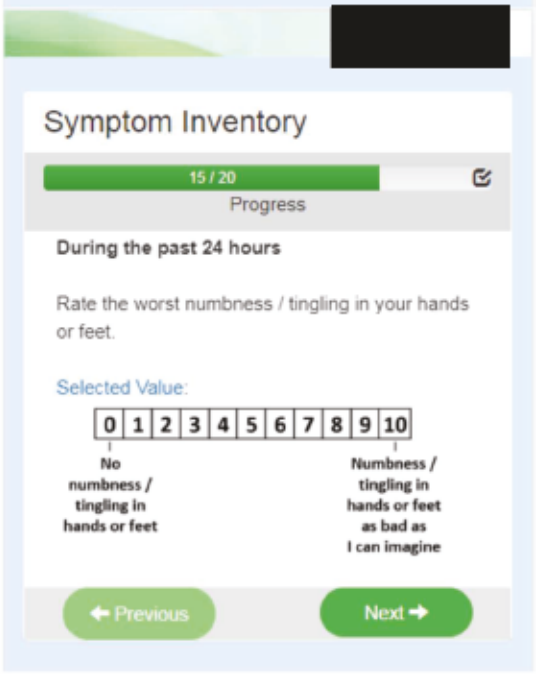
[redacted]



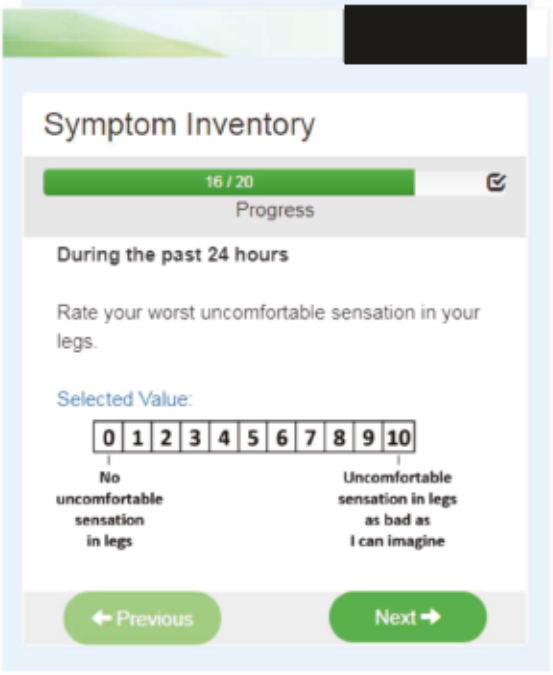

[redacted]

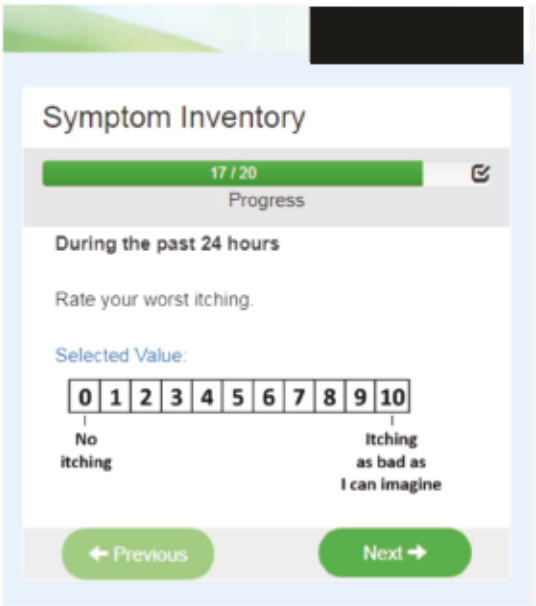


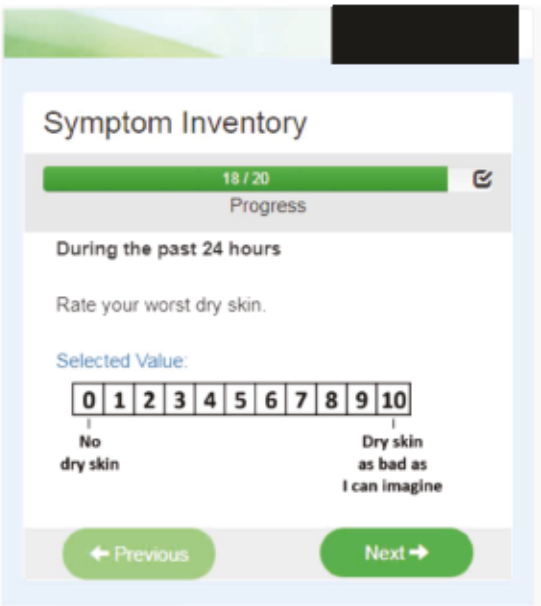

[redacted]

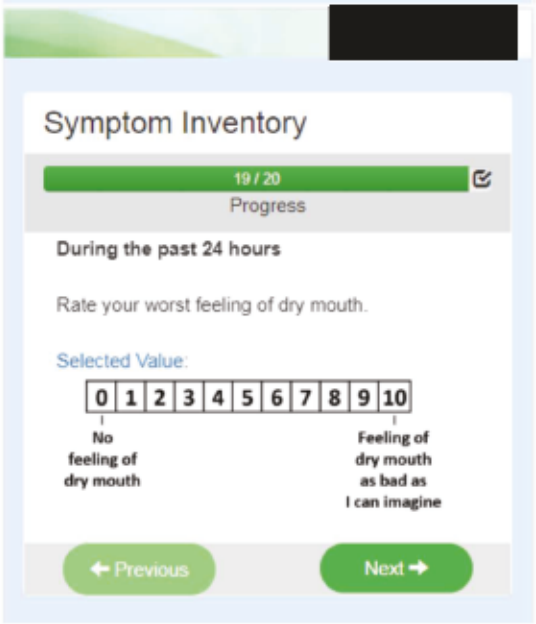


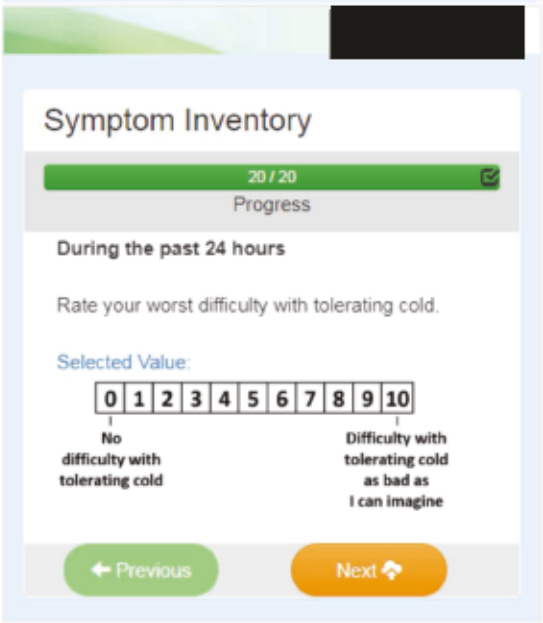

[redacted]





## 8.2. APPENDIX B. PROMIS FATIGUE 13A SHORT FORM, ADDITIONAL FATIGUE ITEMS, SEXUAL ACTIVITY ITEM

### PROMIS Items

PROMIS Short Form v1.0 – Fatigue 13a

| During the past 7 days..... | Not at all | A little bit | Somewhat | Quite a bit | Very much |
|---|---|---|---|---|---|
| I feel fatigued | 1 | 2 | 3 | 4 | 5 |
| I feel weak all over | 1 | 2 | 3 | 4 | 5 |
| I feel listless ("washed out") | 1 | 2 | 3 | 4 | 5 |
| I feel tired | 1 | 2 | 3 | 4 | 5 |
| I have trouble starting things because I am tired | 1 | 2 | 3 | 4 | 5 |
| I have trouble finishing things because I am tired | 1 | 2 | 3 | 4 | 5 |
| I have energy | 1 | 2 | 3 | 4 | 5 |
| I am able to do my usual activities | 1 | 2 | 3 | 4 | 5 |
| I need to sleep during the day | 1 | 2 | 3 | 4 | 5 |
| I am too tired to eat | 1 | 2 | 3 | 4 | 5 |
| I need help doing my usual activities | 1 | 2 | 3 | 4 | 5 |
| I am frustrated by being too tired to do the things I want to do | 1 | 2 | 3 | 4 | 5 |
| I have to limit my social activity because I am tired | 1 | 2 | 3 | 4 | 5 |

PROMIS items (Item identifier in bold)

| | | | | | |
|---|---|---|---|---|---|
| **FATEXP36:** How exhausted were you on average? | 1 | 2 | 3 | 4 | 5 |
| **FATEXP43:** How physically drained were you on average? | 1 | 2 | 3 | 4 | 5 |
| **FATIMP52:** To what degree did your fatigue make you feel less alert? | 1 | 2 | 3 | 4 | 5 |
| **During the past 30 days.....** | **Not at all** | **A little bit** | **Somewhat** | **Quite a bit** | **Very much** |
| **SFINT101:** How interested have you been in sexual activity? | 1 | 2 | 3 | 4 | 5 |

## 8.3. APPENDIX C. SF-36

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| | Your Health and Well-Being | | | | | | |
| | This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. Thank you for completing this survey!<br><br>For each of the following questions, please select the one response that best describes your answer. | | | | | | |
| SF36v2_GH1 | In general, would you say your health is: | Excellent | Very good | Good | Fair | Poor | |
| SF36v2_HT | Compared to one year ago, how would you rate your health in general now? | Much better now than one year ago | Somewhat better now than one year ago | About the same as one year ago | Somewhat worse now than one year ago | Much worse now than one year ago | |
| | The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? | | | | | | |
| SF36v2_PF01 | Does your health now limit you in vigorous activities, such as running, lifting heavy objects, participating in strenuous sports? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF02 | Does your health now limit you in moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF03 | Does your health now limit you in lifting or carrying groceries? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF04 | Does your health now limit you in climbing several flights of stairs? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF05 | Does your health now limit you in climbing one flight of stairs? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF06 | Does your health now limit you in bending, kneeling, or stooping? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF07 | Does your health now limit you in walking more than a mile? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF08 | Does your health now limit you in walking several hundred yards? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF09 | Does your health now limit you in walking one hundred yards? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| SF36v2_PF10 | Does your health now limit you in bathing or dressing yourself? If so, how much? | Yes, limited a lot | Yes, limited a little | No, not limited at all | | | |
| | During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health? | | | | | | |

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| SF36v2_RP1 | During the past 4 weeks, how much of the time have you cut down on the amount of time you spent on work or other activities as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP2 | During the past 4 weeks, how much of the time have you accomplished less than you would like as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP3 | During the past 4 weeks, how much of the time were you limited in the kind of work or other activities as a result of your physical health? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RP4 | During the past 4 weeks, how much of the time have you had difficulty performing the work or other activities as a result of your physical health (for example, it took extra effort)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| | During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? | | | | | | |
| SF36v2_RE1 | During the past 4 weeks, how much of the time have you cut down on the amount of time you spent on work or other activities as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RE2 | During the past 4 weeks, how much of the time have you accomplished less than you would like as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_RE3 | During the past 4 weeks, how much of the time have you done work or other activities less carefully than usual as a result of any emotional problems (such as feeling depressed or anxious)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_SF1 | During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups? | Not at all | Slightly | Moderately | Quite a bit | Extremely | |
| SF36v2_BP1 | How much bodily pain have you had during the past 4 weeks? | None | Very mild | Mild | Moderate | Severe | Very severe |
| SF36v2_BP2 | During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)? | Not at all | A little bit | Moderately | Quite a bit | Extremely | |
| | These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. | | | | | | |

| Item Name | Question Text | Answer Text 1 | Answer Text 2 | Answer Text 3 | Answer Text 4 | Answer Text 5 | Answer Text 6 |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| SF36v2_VT1 | How much of the time during the past 4 weeks did you feel full of life? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH1 | How much of the time during the past 4 weeks have you been very nervous? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH2 | How much of the time during the past 4 weeks have you felt so down in the dumps that nothing could cheer you up? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH3 | How much of the time during the past 4 weeks have you felt calm and peaceful? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT2 | How much of the time during the past 4 weeks did you have a lot of energy? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH4 | How much of the time during the past 4 weeks have you felt downhearted and depressed? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT3 | How much of the time during the past 4 weeks did you feel worn out? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_MH5 | How much of the time during the past 4 weeks have you been happy? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_VT4 | How much of the time during the past 4 weeks did you feel tired? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| SF36v2_SF2 | During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)? | All of the time | Most of the time | Some of the time | A little of the time | None of the time | |
| | How TRUE or FALSE is each of the following statements for you? | | | | | | |
| SF36v2_GH2 | I seem to get sick a little easier than other people. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH3 | I am as healthy as anybody I know. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH4 | I expect my health to get worse. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| SF36v2_GH5 | My health is excellent. | Definitely true | Mostly true | Don't know | Mostly false | Definitely false | |
| | SF-36v2® Health Survey © 1992, 2000, 2009 Medical Outcomes Trust and QualityMetric Incorporated. All rights reserved. SF-36® is a registered trademark of Medical Outcomes Trust. (SF-36v2® Health Survey Standard, United States (English)) | | | | | | |

## 8.4. APPENDIX D. EQ-5D-5L

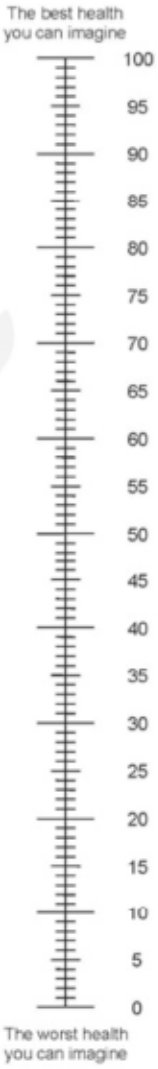
| | |
|---|---|
| EQ-5D-5L | |
| EQ-5D-5L PDA version | |
| English (USA)<br>**Health Questionnaire**<br>**English version for the USA** | Country (Language)<br>Health Questionnaire<br>Version (Target Language)<br>Version (English) |
| On the following screens please tap the statement that best describes your health TODAY. | Instruction |
| **Your mobility TODAY** | **Mobility** |
| I have no problems walking | MB1 |
| I have slight problems walking | MB2 |
| I have moderate problems walking | MB3 |
| I have severe problems walking | MB4 |
| I am unable to walk | MB5 |
| **Your self-care TODAY** | **Self-care** |
| I have no problems washing or dressing myself | SC1 |
| I have slight problems washing or dressing myself | SC2 |
| I have moderate problems washing or dressing myself | SC3 |
| I have severe problems washing or dressing myself | SC4 |
| I am unable to wash or dress myself | SC5 |
| **Your usual activities TODAY** (*e.g. work, study, housework, family or leisure activities*) | **Usual Activities** |
| I have no problems doing my usual activites | UA1 |
| I have slight problems doing my usual activities | UA2 |
| I have moderate problems doing my usual activities | UA3 |
| I have severe problems doing my usual activities | UA4 |
| I am unable to do my usual activities | UA5 |
| **Your pain / discomfort TODAY** | **Pain / Discomfort** |
| I have no pain or discomfort | PD1 |
| I have slight pain or discomfort | PD2 |
| I have moderate pain or discomfort | PD3 |
| I have severe pain or discomfort | PD4 |
| I have extreme pain or discomfort | PD5 |
| **Your anxiety / depression TODAY** | **Anxiety / Depression** |
| I am not anxious or depressed | AD1 |
| I am slightly anxious or depressed | AD2 |
| I am moderately anxious or depressed | AD3 |
| I am severely anxious or depressed | AD4 |
| I am extremely anxious or depressed | AD5 |
| We would like to know how good or bad your health is TODAY. | Vas Line 1 |
| On the next screen you will see a scale numbered 0 to 100. | Vas Line 2 |
| 100 means the best health you can imagine. | Vas Line 3 |
| 0 means the worst health you can imagine. | Vas Line 4 |
| Please tap on the scale to indicate how your health is TODAY. | Vas Line 5 |
| The best health you can imagine | Top Scale |
| The worst health you can imagine | Bottom Scale |
| YOUR HEALTH TODAY | Box Health |



***Disclaimer:*** *This is a preview of the EQ-5D instrument. It demonstrates the text, questions and response options included in this version. This preview does not represent the final product and should not be used as an official EQ-5D instrument.*

- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

YOUR HEALTH TODAY = ☐






3

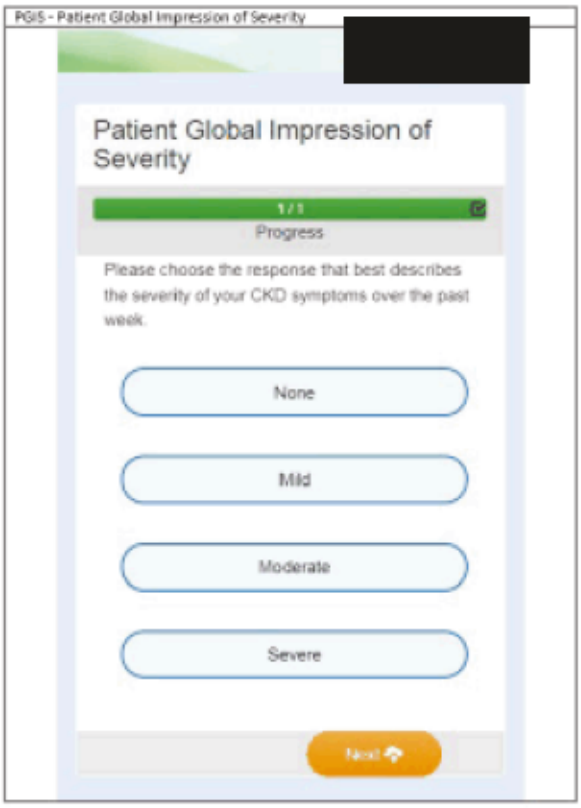


SAS code to calculate the utility index (based on the algorithm published by van Hout 2012) is presented in Table 3.

**Table 3: SAS code for deriving utility index from EQ-5D-5L for UK**

SAS syntax crosswalk values EQ-5D-5L Unite

## 8.5. APPENDIX E. PGIC AND PGIS




## Document Approval Signature(s)

This document has been signed electronically in compliance with [redacted] and [redacted] Companies' electronic signature policies and procedures. This page is a manifestation of the electronic signatures, which may be confirmed within the [redacted] and [redacted] Companies' source system in accordance with [redacted] and [redacted] Companies' records retention policies. Note: the Title field in the signature block has been intentionally left blank.

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

# Psychometric Analysis Plan

## A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE)

Protocol COR-001-02

**Study Sponsor**:

Corvidia Therapeutics Inc.
35 Gatehouse Drive
Waltham MA 02451, USA

**Author**:

[redacted]
[redacted]
[redacted] Patient-Centered Endpoints

**Version Number and Date**: V1.0, 31MAR2020





Approved on 02-Apr-2020 (Document ID: [redacted])

# Contents

# PSYCHOMETRIC ANALYSIS PLAN SIGNATURE PAGE

**Psychometric Analysis Plan V1.0 for Protocol COR-001-02.**

| | Name | Signature | Date |
|---|---|---|---|
| **Author:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |

Upon review of this document, the undersigned approves this version of the Psychometric Analysis Plan, authorizing that the content is acceptable for the reporting of this study.

| | Name | Signature | Date |
|---|---|---|---|
| **Approved By:** | [redacted] | *See appended electronic signature page* | |
| **Position:** | [redacted] | | |
| **Company:** | [redacted] | | |
| | | | |
| **Approved By:** | [redacted] | [redacted] | |
| **Position:** | [redacted] | | |
| **Company:** | Corvidia | | |

## MODIFICATION HISTORY

| Unique Identifier for this Version | Date of the Document Version | Author | Significant Changes from Previous Authorized Version |
|---|---|---|---|
| 1.0 | 09Mar2020 | [redacted] | Not Applicable – First Version |

## LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| ASCVD | Atherosclerotic Cardiovascular Disease |
| CARES | Core assessment of renal disease symptoms |
| CI | Confidence interval |
| CKD | Chronic kidney disease |
| ECDF | Empirical Cumulative Density Function |
| EFA | Exploratory factor analysis |
| EOT | End of treatment |
| EPDF | Empirical Probability Density Function |
| EQ-5D-5L | 5-Level EuroQol in 5 dimensions |
| FU | Follow-up |
| HRQoL | Health-related quality of life |
| hs-CRP | high-sensitivity C-reactive protein |
| ICC | Intraclass correlation coefficient |
| MCS | Mental Component Score |
| ML | Maximum likelihood |
| NDD-CKD | Non-dialysis-dependent chronic kidney disease |
| PAP | Psychometric analysis plan |
| PCS | Physical Component Score |
| PGIC | Patient Global Impression of Change |
| PGIS | Patient Global Impression of Severity |
| PRO | Patient-reported outcome |
| PROMIS | Patient-Reported Outcomes Measurement Information System |
| SAA | Serum amyloid A |
| SD | Standard deviation |
| SEM | Standard Error of Measurement |
| SF-36 v2 | Short Form 36-item health survey version |
| TSS | Total symptom score |
| ULS | Unweighted least squares |
| Ziltivekimab | Monoclonal antibody to IL-6 |

# 1. INTRODUCTION

This document describes the rules and conventions to be used in the presentation and analysis of patient-reported outcomes (PROs) data for the RESCUE study, Protocol Ziltivekimab-COR-001-02. It describes the data to be summarized and analyzed, including specifics of the psychometric analyses to be performed. The psychometric analysis plan (PAP) will be finalized prior to database lock. Any deviations from the PAP after database lock will be documented in the final psychometric analysis report.

This psychometric analysis plan (PAP) is based on protocol version 5.0 dated 23 December 2019 and Amendments 1, 05 December 2018, 2, 24 January 2019, 3 16 April 2019, 4, 25 June 2019, and 5, 23 December 2019.

# 2. RESCUE STUDY OVERVIEW

## 2.1. STUDY OBJECTIVES

### 2.1.1. PRIMARY OBJECTIVE

The primary objective of the RESCUE study is:

- To evaluate the effects of Ziltivekimab compared to placebo on markers of inflammation such as high-sensitivity C-reactive protein (hs-CRP).

### 2.1.2. SECONDARY OBJECTIVES

The secondary objective is:

- To evaluate the effects of Ziltivekimab compared to placebo on two markers of inflammation and cardiovascular risk: serum amyloid A (SAA) and fibrinogen.

### 2.1.3. EXPLORATORY OBJECTIVES

Other objectives are:

- To determine the pharmacokinetic, exploratory pharmacodynamics, pharmacogenetics and effect of Ziltivekimab on inflammatory markers.
- To evaluate the effects of three dose levels of Ziltivekimab compared to placebo on patient reported outcomes (PRO): Patient-Reported Outcomes Measurement

Information System (PROMIS®) Fatigue 13a short form, selected items from the PROMIS fatigue item bank, the Optum SF-36 v2® HealthSurvey, a Corvidia PRO, the PROMIS interest in sexual activity item, the patient global impression of change (PGIC), patient global impression of severity (PGIS), and the EQ-5D-5L.

- To evaluate the psychometric properties of the Corvidia electronic PRO (ePRO) items (*CARES*), *PROMIS Fatigue* 13a short form, and selected items from the PROMIS fatigue item bank, in Chronic Kidney Disease (CKD) patients.

## 2.2. SAMPLE SIZE

The primary efficacy endpoint is percent change from baseline in hs-CRP (average of the hs-CRP value prior to randomization and Day 1) to Week 13 between each active group and placebo.

Based on the observed treatment difference in percent change from baseline in hs-CRP of 60.74% between combined COR-001-01 active groups and placebo and the associated pooled SD of 16.893% in hs-CRP at Week 4 from the final analysis of study COR-001-SC1, a sample size of 54 per group yields more than 99% power with 2-sided alpha=0.05.

Taking into consideration the dropout rate of 10% by the end of the study, a sample size of 60 per group is planned for this study. Accordingly, approximately 240 patients will be randomized 1:1:1:1 (60 per each treatment group) into the trial. Patient randomization will be stratified by baseline hemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4 or 5).

## 2.3. STUDY DESIGN

The RESCUE study is a randomized, double-blind, placebo-controlled trial designed to evaluate the efficacy, safety, and pharmacokinetics of Ziltivekimab at three dose levels (7.5 mg, 15 mg or 30 mg) compared to placebo in patients with stage 3-5 CKD, not on dialysis, who have evidence of inflammation with high cardiovascular risk.

The primary, secondary, and exploratory endpoints will be analysed at 13 weeks of dosing and then followed for additional exploratory efficacy analyses through Week 24. Selected efficacy endpoints and safety assessments will be evaluated in the Follow-up Period Week 25 though Week 32 (Figure 1).

Patients will undergo a Screening Period of up to 14-days during which inclusion and exclusion criteria will be evaluated. Patients who meet all inclusion criteria and no exclusion criteria will be randomized to one of three Ziltivekimab dose levels (7.5 mg, 15mg, or 30 mg) or placebo for a 24-week Treatment Period. Patient randomization will be stratified by baseline haemoglobin (≥11 or <11 g/dL) and CKD stage (3, 4, or 5).

The patient will be randomized on Day 1 and the first dose of study drug should be administered after all assessments are conducted. Doses of study drug will be administered every 28-days for a total of 6 treatments (Weeks 1, 5, 9, 13, 17, and 21).

**Figure 1: RESCUE Study Flow**

Ziltivekimab Dose Level 1, 2 or 3 (N=180); N=60 in each level

Placebo (N=60)

| | Screening | | Treatment Period | | | | | | | | | | | Safety F/Up | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11/ET | 12 | 13/ET |
| Visit Day Window | -14 | -7 | 1 | 8 | 22 | 29-35 | 43-49 | 57-63 | 85-91 | 113-119 | 141-147 | 155-161 | 162-168 | 190-196 | 218-224 |
| Visit Window | | ±2 | 1 | ±2 | ±2 | | | | | | | | | | |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |

## 2.4. PATIENT-REPORTED OUTCOMES

### 2.4.1. PRO INSTRUMENTS

The following patient-reported outcomes (PRO) are collected in this study:

- Corvidia ePRO (*CARES*)
- Patient-Reported Outcomes Measurement Information System (PROMIS®) Fatigue 13a short form
- Selected items from the PROMIS fatigue item bank
- PROMIS interest in sexual activity item
- Optum SF-36 v2® Health Survey
- EQ-5D-5L

- Patient global impression of symptoms (PGIS) and Change (PGIC).

#### 2.4.1.1. *CARES* Corvidia

The *CARES* Corvidia ePRO - henceforth *CARES* - is a new PRO instrument under development (Appendix A). It consists of 19 symptom items, with a 24-hour recall, asking patients to report their worst level of that symptom in the past 24 hours on a numeric rating scale from 0 (no symptom) to 10 (symptom as bad as I can imagine).

The symptom items cover physiological (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea), psychological (Depression, Concentration, Forgetfulness), Pain (Bone / Joints, Nerves, Muscle Cramps) and sensorial (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold) symptoms.

The instrument can be found in Appendix A.

#### 2.4.1.2. Short Form *PROMIS Fatigue* 13a and selected items from *PROMIS Fatigue* Bank, *PROMIS* Interest in Sexual Activity Items

The *PROMIS Fatigue* 13a short form contains 13 items that assess symptoms and impacts of fatigue. The recall period is 7 days. Each item is rated on a five-point Likert scale ranging from 0 = "not at all" to 4 = "very much". Higher scores in these items indicate worst fatigue, except for items 7 ("I have energy") and 8 ("I am able to do my usual activities") where higher score indicates less fatigue.

In this study, three additional fatigue items were administered for further testing to account for exhaustion, being physically drained, and impact on alertness, concepts that resonated with patients during patient interviews but are not covered in the *PROMIS Fatigue* 13a short form. Such combination is from now on referred to as *PROMIS Fatigue*.

A number of scales and subscales can be derived from the *PROMIS Fatigue* set:

- The original *PROMIS Fatigue* 13a scale;
- The fatigue scale using the whole *PROMIS Fatigue* item set (*de novo* 16-item short form);
- Two subscales within the *PROMIS Fatigue* 13a scale, e.g.: 1) symptoms of fatigue and 2) impacts of fatigue;
- Two subscales within the *PROMIS Fatigue* set, e.g.: 1) symptoms of fatigue and 2) impacts of fatigue.

The PROMIS Sexual Function and Satisfaction (PROMIS Sex FS) is a customizable, self-reported set of measures that include 79 items covering 11 domains: interest in sexual activity, lubrication, vaginal discomfort, erectile function, global satisfaction with sex life, orgasm, anal discomfort, therapeutic aids, sexual activities, interfering factors, and screener questions. In this

study only one item will be used: interest in sexual activity.

These three instruments will be administered together as shown in Appendix B.

#### 2.4.1.3. Short Form 36-item Health Survey, Version 2 (SF-36v2)

SF-36v2 is a self-report survey of functional health and well-being with 4 weeks recall period [QualityMetric 2011]. Responses to 35 of the 36 items are used to compute an 8-domain profile of functional health and well-being scores.

The remaining item, referred to as the 'Health Transition' item, asks patients to rate how their current state of health compared to their state of health 1 year ago, and is not used to calculate domain scores.

The 8-domain profile consists of the following scales: Physical Functioning (PF), Role Limitations due to Physical Health (RP), Bodily Pain (BP), General Health Perceptions (GH), Vitality (VT), Social Functioning (SF), Role Limitations due to Emotional Problems (RE), and Mental Health (MH).

Psychometrically-based physical and mental health component summary scores (PCS and MCS, respectively) are computed from subscale scores to give a broader metric of physical and mental health status. The instrument can be found in Appendix C.

#### 2.4.1.4. 5-Level EuroQol in 5 dimensions (EQ-5D-5L)

The EuroQol EQ-5D-5L has been designed as an international, standardized, generic instrument for describing and valuing health-related quality of life, available in over 100 language versions.

The EQ-5D-5L self-report questionnaire consists of 5 items, one per domain, with a 5-point scale and one visual analogue scale (EQ VAS) to rate health state from worst (0) to best (100).

The EQ-5D-5L includes domains for each generic health status measure: Mobility, Self-care, Usual Activities, Pain/discomfort and Anxiety/depression. For each question, there are 5 levels of response, corresponding to increasing levels of impairment (no problems, slight, moderate, severe, and extreme problems or unable to perform activity) and coded 1 to 5.

A higher index indicates better quality of life. The instrument can be found in Appendix D.

#### 2.4.1.5. Patient Global Impression of Symptoms (PGIS) and Change (PGIC)

The Patient Global Impression of Severity (PGIS) is a self-reported measure of patient-perceived overall symptom severity. The PGIS is a 5-point recall scale (from no symptoms to very severe symptoms). Patients are asked to rate their overall severity of symptoms in the 7 days prior a clinic visit.

The Patient Global Impression of Change (PGIC) instrument captures the patient's overall

evaluation of response to treatment. The patient is asked to report the degree to which they have changed since entering the treatment period using a 7-point scale (Very Much Improved to No Changes to Very Much Worse).

The PGIC and PGIS are the most commonly used anchor-based methods of assessing clinically important change and severity in which the external judgment of meaningful change is made by the patient. The instruments can be found in Appendix E.

### 2.4.2. ASSESSMENT SCHEDULE FOR PRO INSTRUMENTS

Patient-reported outcomes will be assessed through a personal device supplied by Corvidia:

- At home daily on subject personal device: *CARES*;
- At the clinic PRIOR to dosing on subject personal device: *PROMIS Fatigue*, PROMIS Interest in Sexual Activity item, SF-36v2, EQ-5D-5L, PGIC, PGIS .

*CARES* data will be collected at home on subject's personal device or one provided by Corvidia; the questionnaire will be answered daily for 7 days prior to dose (Screening Day -7 to Day -1 prior to Dose 1 (Visit 1), Dose 2 (Visit 4), Dose 4 (Visit 7), and the last 2 weeks after Dose 6 (Visit 9).

The SF-36v2, EQ-5D-5L and PGIC are to be collected on subjects' devices at the clinic prior to each dosing, e.g., Visit 1, Visit 4, Visit 7, and the last 2 weeks after Visit 9, except for PROMIS Sex Item (Only Week 24).

The PGIS will be collected at home on subjects' devices on the last day of *CARES* assessment period.

The schedule of assessment for all PRO instruments collected in different periods of the study are shown in Table 1.

**Table 1: Schedule of assessment for PRO instruments collected in the RESCUE study-Randomization, treatment period, and follow-up**

| Study Period | Screening | | Treatment | | | | | | | | | | | | Safety F/Up[b] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visit Number | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | N/A | 7 | 8 | 9 | 10 | 11/ET[a] | 12 | 13/ET |
| Visit Week | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 12 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |
| DOSING | | | X | | | X | | X | | X | X | X | | | | |
| *CARES* | | X | | | X | | | | X | | | | X | X | | |
| *PROMIS Fatigue* | | | X | | | X | | | | X | | | X | X | | |
| PROMIS Sex Item | | | X | | | X | | | | X | | | | X | | |
| SF-36v2 | | | X | | | X | | | | X | | | X | X | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EQ-5D-5L | | X | | X | | X | | X | | | |
| PGIS | X | | X | | X | | X | | | | |
| PGIC | | | | X | | X | | X | | | |

[b]Follow-Up.
[c]Early Termination.

# 3. OBJECTIVES OF THE PSYCHOMETRIC ANALYSIS

The aim of the psychometric analysis is to assess the psychometric properties of the *CARES* and of *PROMIS Fatigue scales*.

Specific objectives are as follows:

- To evaluate item response distributional characteristics;
- To investigate the underlying structure of these PROs and to develop scale and/or subscale scores;
- To evaluate the reliability (internal consistency reliability and test-retest reliability) of *CARES* and of *PROMIS Fatigue scale* and/or subscale scores;
- To evaluate the construct validity of *CARES* and of *PROMIS Fatigue scale* and/or subscale scores;
- To evaluate that changes in the patient's status are reflected in changes in the *CARES*/*PROMIS Fatigue* scale and/or subscale scores (sensitivity to change);
- To estimate a threshold for clinically relevant change withing patients for the *CARES* and of *PROMIS Fatigue* scale and/or subscale scores;
- To derive a definition for a symptomatic patient.

# 4. ANALYSIS TIMING

The psychometric analysis will be performed at the same time as the primary analysis of the study, e.g., once all subjects complete Visit 7 (Week 13). [redacted] will perform the analysis using blinded data and data will be collapsed across treatment groups for the purpose of the psychometric analysis.

After the last subject completes the last visit, the final database will be cleaned and locked and this data will be used to repeat the derivation of the clinically meaningful threshold (see section 7.3.4). There are no plans to repeat further psychometric analyses on the final database.

# 5. ANALYSIS POPULATION

All PRO analyses described in this PAP will be performed on the ITT analysis population which includes all randomized patients.

# 6. ANALYSIS VARIABLES

## 6.1. PRO VARIABLES

The scoring for each PRO instrument will be done according to the procedures described by each PRO developer and license owner in the respective instruction manuals.

### 6.1.1. VARIABLES GENERATED FROM *CARES*

For each item, weekly scores will be derived as the average of the 7 consecutive daily scores as follows:

Baseline = average of Day -7 to Day -1

Week 5 (Visit 4) = average of Day 22 to Day 28

Week 13 (Visit 7) = average of Day 78 to Day 84

If more than 3 daily scores out of the 7 days (>50%) within the weekly period are missing, then the score is set to missing.

A bi-weekly score (the End of Treatment [EOT]; Week 24) will also be derived as the average of Day 155 to Day 168. If more than 7 or more daily scores out of the 14 days (≥50%) within the bi-weekly period are non-missing, then the score is set to missing.

Preliminary scoring of the *CARES* includes the calculation of a Total Symptom Score (TSS). Alternative means of scoring the *CARES* that are derived from the factor analyses and conceptual considerations may also be defined if supported by the data and if the research team considers these to be plausible and clinically useful alternatives. The following conceptual domains from CARES will be considered: physiological (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea), psychological (Depression, Concentration, Forgetfulness), pain (Bone / Joints, Nerves, Muscle Cramps), and sensorial (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold) symptoms.

In the preliminary scoring, the daily TSS will be calculated as the average of the 19 daily item scores. A daily TSS score will be calculated if at least 10 out of the 19 items (> 50%) are non-missing. The TSS weekly score will be calculated in the same fashion as the individual items. Higher TSS indicate more symptoms/poorer outcomes.

### 6.1.2. VARIABLES GENERATED FROM *PROMIS FATIGUE* AND *PROMIS* SEXUAL FUNCTION AND SATISFACTION

The *PROMIS Fatigue* data is collected as a 7-day recall report from patients at Visit 1, 4, 7, 10, and 11. The scoring manual for *PROMIS Fatigue* 13a short form can be found at:http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Fatigue_Scoring_Manual.pdf.

Each question from the *PROMIS Fatigue* 13a short form has five response options (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much). To use the scoring tables from the PROMIS fatigue scoring manual, a summed score is calculated. The lowest possible summed score is 13 and the highest possible summed score is 65. All questions must be answered in order to produce a valid score using the scoring tables. Once calculated, the total summed raw score is translated to a T-score. The T-score re-scales the raw score into a standardized score with a mean of 50 and a standard deviation (SD) of 10. The T-score and standard error are then used to calculate the 95% confidence interval around the observed score. The 95% confidence interval is calculated as T-Score ±1.96*SE. The T-score will not be calculated if a response is missing for one or more question from the *PROMIS Fatigue* 13a short form.

In addition, preliminary scoring of the *PROMIS Fatigue* 13a includes the calculation of two subdomain scores:

- Symptom score (5 symptom items):
  - I feel fatigued
  - I feel weak all over
  - I feel listless (“washed out”)
  - I feel tired
  - I have energy
- Impact score (8 impact items):
  - I have trouble starting things because I’m tired
  - I have trouble finishing things because I’m tired
  - I am able to do my usual activities
  - I need to sleep during the day
  - I am too tired to eat
  - I need help doing usual activities
  - I am frustrated by being too tired to do the things I want to do

- I have to limit my social activity because I am tired

The subdomains will be calculated as the sum of the items in the corresponding subdomain. The subdomain score will be calculated only if all items are answered.

Alternative means of scoring, by including also the additional *PROMIS Fatigue* items from the Item Bank, will be derived as follows:

- A total fatigue score using the items for the PROMIS Fatigues 13a and the additional three items form the Item Bank (de novo 16 item short form)
- Symptom score (8 symptom items):
  - I feel fatigued
  - I feel weak all over
  - I feel listless ("washed out")
  - I feel tired
  - I have energy
  - FATEXP36: How exhausted you are on average?
  - FATEXP43: How physically drained were you on average?
  - FATEXP52: To what degree did your fatigue make you feel less alert?

Alternative means of scoring can be derived from the factor analyses using the full set of *PROMIS Fatigue* items (inclusive of items from the Item Bank) to identify plausible and clinically useful alternatives.

An EOT assessment (Weeks 23 through 24) will be derived for the hypothesized scores as the mean value of the two assessments at Week 23 and Week 24. In case only one assessment is available, this will be used as the EOT assessment.

### 6.1.3. VARIABLES GENERATED FROM SF-36v2

The scoring for SF-36 v2 instrument will be done according to the procedures described by the developer and license owner in the instruction manual. Scores for the scales (HT, GH, PF, RP, BP, MH, SF, RE, VT) as well as the summary scores (PCS, MCS) will be derived. For the psychometric analysis only the baseline scores will be used.

### 6.1.4. VARIABLES GENERATED FROM EQ-5D-5L

No derivation is required for the EQ-5D-5L. The scores for the 5 domains (Mobility, Self-care, Usual Activities, Pain/discomfort and Anxiety/depression) and EQ-5D VAS will be as reported.

Missing data for individual items will not be imputed. For the psychometric analysis only the baseline scores will be used.

### 6.1.5. VARIABLES GENERATED FROM PGIS AND PGIC

The PGIS and PGIC are single-item questionnaires. Therefore, they do not require any manipulation to derive scores.

To be noted that PGIS is collected on the last day of *CARES* assessment period. To be in line with the visit label used for CARES, we will label the PGIS assessment as Week 13 (Visit 7) as well.

## 6.2. CLINICAL OUTCOME VARIABLES

Demographic and clinical data will be used to characterize the sample, as follows:

- Age (years)
- Sex
- Race
- Baseline Hemoglobin (>=11 or <11 g/dL)
- CKD Stage (Stage 3, Stage 4, or Stage 5)

The clinical variable that will be used in the psychometric evaluation is the total plasma level for the hs-CRP Protein.

## 6.3. GENERAL VARIABLES AND DEFINITIONS

### 6.3.1. BASELINE

The baseline for *CARES* will be defined as the weekly-average of the daily measurements over the 7-day period immediately preceding randomization day (i.e., study days -7 through -1), e.g., the second week of the screening period.

For all other PRO instruments, except PGIC, the baseline is defined as the measurements carried out at clinic prior to first dose of the study drug, scheduled for the 1st day of the Treatment Period. There will be no baseline assessment for PGIC.

### 6.3.2. DERIVED TIMEPOINTS

The EOT assessment for *CARES* will be defined as the bi-weekly average of the daily measurements over the 14-day period leading to the end of the treatment period (i.e., study days 155 through 168).

For the *PROMIS Fatigue* items, the EOT assessment will be calculated by average of the data collected at Visit 10 and Visit 11 (Weeks 23 and 24). In case only one assessment is available, this will be used.

For the PGIC and PGIS instruments, the EOT assessment will be represented by the measurements carried out at clinic visit in the last week of the treatment period (Week 24).

# 7. STATISTICAL METHODOLOGY

## 7.1. GENERAL CONSIDERATIONS

Summary statistics for continuous data will include the number of observations (N), mean, standard deviation (SD), median, first and third quartiles, minimum (min), and maximum (max). Categorical data will be described by the number (n) and percentage (%) of patients in each category. Missing and invalid observations will be tabulated as separate categories. The calculation of proportions will not include the missing/invalid category, unless specified otherwise.

Statistical comparisons will be made using two-sided tests at the $\alpha = 0.05$ significance level unless stated otherwise. Due to the exploratory nature of these analyses, no adjustments for multiplicity will be performed. For point estimates, 95% confidence intervals will be used.

The psychometric analysis will be performed on the pooled treatment arms.

## 7.2. STUDY SUBJECTS

### 7.2.1. PRO COMPLETION

The completion rate for *CARES* will be defined as the total number of actual completed diary entries divided by the expected number of diary entries in a given time period (total number of days X number of patients expected to complete the instrument). Specifically, completion rate will be evaluated for the baseline period (Day -7 to Day -1), in the weeks prior to each clinic visit for administration of dosage (Day 22 to 28 [Week 5] and 78 to 84 [Week 13]).

For the *PROMIS* Fatigue, *PROMIS* Sexual Interest, PGIC and PGIS, completion rates will be calculated as the total number of actual completed clinic visit entries divided by the expected number of patients taking part in the scheduled clinic visit.

### 7.2.2. Demographics and Baseline Characteristics

The demographic and patient baseline characteristics are analyzed in the clinical SAP and will not be repeated herein.

- 

## 7.3. Measurement Properties Analyses

The analyses described below follow best practice guidelines for the psychometric analysis of PRO measures according to the FDA guidance [FDA 2009, FDA 2019].

### 7.3.1. Descriptive Analyses

Descriptive statistics for *CARES, PROMIS* Fatigue, PROMIS sexual function and satisfaction, PGIC, and PGIS will be assessed by evaluating the following at baseline, Visit 4, and Visit 7:

- Counts and percentages for each response option per item, including a stacked column chart of the distribution of responses
- Summary statistics for each item of *CARES, PROMIS* Fatigue, PROMIS sexual function and satisfaction
- Summary statistics for each scale/domain and total score of *CARES* and *PROMIS* Fatigue

Item descriptive analyses will be examined for *CARES* and *PROMIS* Fatigue for item removal, combined with qualitative evaluation. Specifically, items with floor or ceiling effect (e.g., tems that are disproportionately rated at either extreme of their rating scales by patients at baseline and/or after treatment) will be considered as candidates for item reduction.Item-to-Item Correlations for *CARES* and Promis Fatigue Items

The purpose of item-to-item correlation analyses is to evaluate inter-relationships among items to determine the extent and degree of correspondence and discordance between items, and whether those patterns are consistent with hypothesized expectations based on the item content.

Spearman rank correlation coefficient will be calculated using baseline data to examine the inter-relationships among items for *CARES* (Items 1 to 19) and among *PROMIS Fatigue* Items

(Fatigue 13a Items 1 to 13, Fatigue Bank Items FATEXP36, FATEXP43 and FATEXP52).

Correlations greater than 0.4 may provide support for combining items into a multi-item scale. Items with coefficients greater than 0.9 and/or less than 0.1 may be considered for item removal.

#### 7.3.1.1. CARES

The *CARES* instrument is organized thematically in different sets of symptoms. Items 1 to 7 describe a set of physiological symptoms (Short Breath, Swelling, Fatigue, Weakness, Light-headedness, Decreased Appetite, Nausea). Items 8 to 10 cover psychological issues (Depression, Concentration, Forgetfulness). Item 11 to 13 describe pain (Nerve, Bones/Joints, Muscle Cramps).

The remainder of the items cover a variety of sensorial symptoms that can be loosely described as a sensation of uncomfortableness, mostly associated with the skin and terminal nervous system (Hand numbness/tingling, Uncomfortable legs, Itching, Dry skin, Dry mouth, Cold). Correlations may vary accordingly in a non-predictable manner.

#### 7.3.1.2. PROMIS Fatigue

The combined set of fatigue items is quite homogeneous with regards to the topics investigated. Specifically, the fatigue Bank additional items may be reasonably perceived as 'general' feelings of fatigue/exhaustion and therefore be strongly correlated with the short form 13a Items. It is expected that these items will show strong correlations (above 0.50) across the items.

### 7.3.2. EXPLORATORY FACTOR ANALYSIS

An exploratory factor analysis (EFA) will be carried out for the items of *CARES* and *PROMIS Fatigue* to test the measurement model. EFA is a model of the measurement of a latent variable. This latent variable cannot be directly measured with a single variable. Instead, it is seen through the relationships it *causes* in a set of variables. The relationships between the factor(s) and each variable are weighted, and factor analysis calculates the optimal weights.

The Kaiser-Meyer-Olkin measure of sampling adequacy will be used to evaluate the strength of the linear association among the items in the correlation matrix. The KMO measure can range between 0 and 1. The standard threshold value for sample adequacy is above 0.6, with 0.8 and above being ideal [Pett et al. 2003].

Both the maximum likelihood (ML) and the unweighted least squares (ULS) extraction method will be used in this study. ML is the most frequently used factor extraction methods. However, it does require the assumption of multivariate normal distribution of the variables. If the assumption of multivariate normality is severely violated (e.g., skewness < -1 or > 1), Fabrigar et al. 1999 and Nunnally and Bernstein 1994 recommend using ULS.

The symptoms investigated in the two PROs under investigation are expected to be associated

which each other. We expect the underlying domains to be correlated. Accordingly, the EFA will be executed using a PROMAX rotation. The number of factors will be determined using established methods (Eigenvalues assessments, scree-plots).

The PROC FACTOR procedure in SAS will be used to perform the EFA.

A total summary scale or set of scales will be proposed in light of the factor loadings and clinical and conceptual considerations. The score(s) developed as psychometrically sound will be used to evaluate efficacy.

Data distributions and correlations with other PROs will then be carried out for the proposed *CARES* and *PROMIS Fatigue* scales and/or subscales as described in sections 7.3.1 and 7.3.2.

### 7.3.3. ITEM-TOTAL CORRELATIONS FOR *CARES* AND PROMIS FATIGUE

Correlations between items and scales derived using the preliminary algorithm and possible alternative structures selected from the factor analysis will be assessed using baseline data. The multitrait-multimethod approach will be applied to assess the association of items with their hypothesized scales (item removed) and the association of items with other scales. Spearman rank correlations will be estimated for this analysis. Items are expected to have a correlation coefficient >0.4 with their hypothesized scales (convergent validity) and to have a higher correlation with their hypothesized scale than with any other scale (divergent validity).

Item-total correlations will be considered adequate if the Spearman rank correlation coefficient values are at least 0.40. Any low item-total correlations will be flagged, and item removal will be considered, as it may indicate inadequate scale validity. Any large correlation coefficient (e.g., ≥ 0.90) might suggest redundancy, marking the item as a candidate for elimination or modification.

### 7.3.1. RELIABILITY

#### 7.3.1.1. Test-Retest Reliability

Test-retest reliability (i.e., the stability of an instrument over time) will be assessed using intra-class correlation coefficients (ICCs) between two time points ("test" and "retest"), employing form 2.1 as described by Shrout and Fleiss 1979. For *CARES,* two sets of data will be examined: for the 7 days prior to treatment on Day 1 for all patients and Visit 4 period for patients considered to be in a stable condition, as described below. For *PROMIS Fatigue* only the second set of data will be used, e.g., using patients considered to be in a stable condition.

**Systematically selected days**

The last 7 days before Day 1 will be considered to define the test and retest time points for *CARES* collected via eDiary. Test and retest periods will be defined using two consecutive days

from the baseline period (Day -6 to Day 1). The following pairs of consecutive days will be evaluated: Days -6 and -5; -5 and -4; -4 and -3; -3 and -2; -2 and -1; -1 and 1.. This analysis will be performed for *CARES* only.

**Stable condition defined using PGIC**

Only those subjects reporting "no change" from baseline on the PGIC at Visit 4 will be included. Period 1 (i.e., test) will defined as baseline and Period 2 (i.e., retest) will be defined as Visit 4.

**Stable condition defined using PGIS**

Only those patients reporting no change on the PGIS, i.e., those that report the same severity level from baseline to Visit 4, will be included for test-retest investigation. Period 1 (i.e., test) will be defined as baseline and period 2 (i.e., retest) will be defined as Visit 4.

Test-retest reliability will be tested for each multi-item scale/domain as defined by the review of EFA results and total scores. ICC values of 0.40–0.75 will be considered to represent fair to good reliability and values >0.75 represent excellent reliability.

#### 7.3.1.2. Internal Consistency Reliability

Internal consistency reliability of each multi-item scale/domain score as identified by EFA results will be estimated using Cronbach's Alpha. Cronbach's Alpha, which ranges from 0 to 1, where "1" equals perfect reliability, is based on the average inter-item correlation and the number of items. Minimum values equal to or greater than 0.70 have been recommended for group level comparisons.**Error! Bookmark not defined.** Internal consistency will be tested for each d omain of the *CARES* and *PROMIS Fatigue*, as well as for their total scores.

Internal consistency will also be assessed with the Cronbach Alpha statistics with Item Removal. Alpha will be estimated for each item and compared with the same statistic for the full set of items. Items that do exhibit a strong drop in the Alpha value compared to the full set value should be selected for retention. Conversely items with alpha values for which there is small or no change from the full set value should be considered for removal.

### 7.3.2. CONSTRUCT VALIDITY

Validity refers to the evidence that the *CARES* and *PROMIS Fatigue* measure what they are designed to measure in the population for which it is intended. Construct validity is an experimental demonstration that the *CARES* and *PROMIS Fatigue* measures the core constructs of the symptoms of NDD-CKD patients at advanced stage. Three forms of construct validity will be examined: convergent, divergent, and known group validity.

#### 7.3.2.1. Convergent and Divergent Validity

The Spearman Rank correlation coefficient will be used to evaluate convergent and divergent validity. Convergent validity refers to how well constructs that theoretically should be related to each other are observed to be related. Divergent validity can be viewed as the counterpart to convergent validity; it assumes that constructs theoretically unrelated to each other will be observed to be unrelated and will have low correlations.

Correlations between the *CARES* and *PROMIS Fatigue* with the SF 36V2, EQ-5D-5L, and PGIS will be examined at Baseline. It is expected that there would be moderate to high correlations between domain and total scores of similar content across the assessments, and low correlations for non-overlapping domains.

*PROMIS Fatigue* items 7 and 8 are keyed in the opposite direction of the remaining items and of those in the other instruments. Accordingly, those items should, regardless of the magnitude, show a negative correlation with other items and instruments.

#### 7.3.2.2. Moderate Correlations

These are expected for *CARES* Tiredness (1, 2, 3, 4, 5) with the following domains:

- SF-36 v2: HT, GT, PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Mobility, Selfcare, Activity, Pain, MH
- PGIS

and for *CARES* Items Gastrointestinal (6, 7) with the following domains:

- SF-36 v2: HT, GT, PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Pain, MH
- PGIS

and for *PROMIS Fatigue* Items 1 to 4 with:

- SF-36 v2: PF, RP, PCS, MH, RE, VT, MCS
- EQ-5D-5L: Mobility, Selfcare, Activity, Mental Health

#### 7.3.2.3. Low to Moderate Correlations

These are expected for *CARES* Items Mental Health (8, 9, 10) and Pain (11 12, 13) with the following domains:

- SF-36 v2: HT, GT, PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Mobility, Selfcare, Activity, Pain, Mental Health

- PGIS

and for *PROMIS Fatigue* Items 5 to 16 with:

- SF-36 v2: PF, RP, PCS, BP, MH , SF, RE, MCS
- EQ-5D-5L: Selfcare, Activity, Mental Health

#### 7.3.2.4. Low Correlations

These are expected for *CARES* Items Discomfort (14 to 19) with the following domains:

- EQ-5D-5L: Mobility, Selfcare, Activity

and for *PROMIS Fatigue* Items 1 to 16 with:

- SF-36 v2: HT, GH, BP
- EQ-5D-5L: Pain
- PGIS

#### 7.3.2.5. Known-Groups Validity

The purpose of the known-groups validity is to assess the degree to which the measure can distinguish among groups of subjects hypothesized to be different in concepts of interest. In the case of the RESCUE trial there are several concepts of interests.

The known-group analysis will accordingly refer to the following three grouping criteria:

- Hemoglobin: High (≥11 g/dL) Vs. Low ( <11 g/dL)
- CKD Stages: 3 vs. 4 or 5
- PGIS: groups will be defined at baseline by PGIS (None, Mild, Moderate, Severe, Very Severe)

For the PGIS classification, adjacent groups may be combined if the sample size of a particular category is <10 patients.

Known-group validity for the Hemoglobin and CKD Stage groups will be tested with a group mean t-test with alpha = 0.05 level for the *CARES* and *PROMIS Fatigue* total and subscale scores.

PGIS groups will serve as the independent variable in an analysis of variance (ANOVA) with alpha = 0.05 level, and the *CARES* and *PROMIS Fatigue* total and subscale scores will serve as dependent variables. Kruskal-Wallis test, a non-parametric alternative to ANOVA, will be conducted if group data for any group criteria are not reasonably balanced.

Box-plots of the *CARES* and *PROMIS Fatigue* total and subscale scores for each group will be also presented to highlight the degree of separation and heterogeneity between the scores'

distributions across groups.

### 7.3.3. SENSITIVITY TO CHANGE

Sensitivity to change is the ability of an instrument to measure change in a state regardless of whether it is relevant or meaningful to the decision maker. Sensitivity to change will be examined using one-way analysis of covariance (ANCOVA). Subjects with baseline and Visit 7 data will be included in the analysis.

The dependent variable will be the change from baseline to Visit 7 in the *CARES* and *PROMIS Fatigue* total and subscale scores. The responder group will be formed using the PGIC and PGIS scores. The model will include the "responder" factor as a fixed factor and the analysis will be controlled for the baseline PRO scores. Separate models will be considered for each CARES and *PROMIS Fatigue* total score and EFA-based scales.

Using the PGIC, the following three groups will be created using the PGIC score at Visit 7:

- 'Improved' group will include those participants who answered "Very much better", "Moderately better", and "A little better".
- 'No change' group will include those participants who answered '0= No change'.
- 'Worsened' group will include those participants who answered "Very much worse", "Moderately Worse", and "A little worse" .

Using the PGIS, the following three groups will be created using the PGIS score at Visit

- Improved is defined as patients with an improvement on their PGIS of at least one level from baseline to Visit 7
- Worsening is defined as patients with a worsening of at least one level from baseline to Visit 7
- Unchanged is defined as patients with same PGIS score at Visit 7 and at baseline

The analysis will be performed on groups with ≥10 patients and adjacent groups may be combined if there are <10 patients. Separate models will be considered for each "responder" group definition.

The model will present least squares (LS) mean estimates, standard errors, 95% confidence intervals (CIs) and p-values. The analysis of variance will be conducted using the PROC GLM procedure in SAS.

### 7.3.4. MEANINGFUL CHANGE

Individual-level thresholds of meaningful change will be explored. To identify patients who

experienced a significant improvement in their symptoms over the course of treatment, a responder definition will be determined to characterize a meaningful change in the scores of the *CARES* and *PROMIS Fatigue* measures. Anchor-based and distribution-based methods, as well as graphical displays, will be used to support the interpretation of possible treatment benefits as reflected in the *CARES* and *PROMIS Fatigue* scores. These approaches will help characterize the amount of change important to patients on these dimensions (as recommended in the FDA PRO guidance [FDA 2009; FDA 2019]).

Thresholds will be estimated for the *CARES* and *PROMIS Fatigue* total score and scales and/or subscales obtained from the EFA.

#### 7.3.4.1. Distribution-Based Approach

The following two distribution-based approaches will be applied to the baseline *CARES* and *PROMIS Fatigue* total scores and EFA-based scales.

**One-half standard deviation (SD)**:

SD of Baseline PRO scores will be computed and divided by 2

**1 standard error of measurement (SEM)**:

$$SEM = SD_{baseline} * \sqrt{1 - r_{xx}}$$

where $SD_{baseline}$ = SD at Baseline and $r_{xx}$ = the reliability (internal consistency) of PRO scale score at Baseline

#### 7.3.4.2. Anchor-Based Approach

In the anchor-based approach, the PGIS and PGIC will be used. Specifically, subjects will be classified in groups to provide a clearer difference between subjects who have and have not experienced meaningful change according to the anchors. Meaningful change thresholds will be derived using each level of the PGIS and PGIC. If the sample size within a response category is too small, grouping of the response categories will also be considered as indicated in Table .

**Table 2 Categories for change in PGIS and PGIC**

| Instrument | Category | Definition |
|---|---|---|
| **PGIS at Visit 7** | **Un-collapsed Categories** | |
| | Improved 4 categories | CFB =-4 |
| | Improved 3 categories | CFB =-3 |
| | Improved 2 categories | CFB =-2 |
| | Improved 1 categories | CFB =-1 |
| | No Change | CFB =0 |

| Instrument | Category | Definition |
|---|---|---|
| | Worsened 1 category | CFB =+1 |
| | Worsened 2 categories | CFB =+2 |
| | Worsened 3 categories | CFB =+3 |
| | Worsened 4 categories | CFB =+4 |
| | **Collapsed Categories** | |
| | Improvement | CFB ≤ -1 point |
| | No Change | CFB=0 |
| | Worsening | CFB ≥+1 point |
| **PGIC at Visit 7** | **Un-collapsed Categories** | |
| | Very much improved | Rating "very much better" |
| | Much improved | Rating "moderately better" |
| | Minimally improved | Rating "a little better" |
| | No change | Rating "no change" |
| | Minimally worse | Rating "a little worse" |
| | Much worse | Rating "moderately worse" |
| | Very much worse | Rating "very much worse" |
| | **Collapsed Categories, 5 categories** | |
| | Very much improved | Rating "very much better" |
| | Improvement | Rating of "a little better" and "moderately better" |
| | No change | Rating of "no change" |
| | Worsening | Rating of "a little worse" or "moderately worse" |
| | Very much worse | Rating "very much worse" |
| | **Collapsed Categories, 3 categories** | |
| | Improvement | Rating of "a little better", "moderately better" or "very much better" |
| | No change | Rating of "no change" |
| | Worsening | Rating of "a little worse" or "moderately worse" or "very much worse" |

CFB=Change from baseline at Visit 7.

The change in the *CARES* total score and EFA based scales from baseline (average of Day -7 to Day -1) to the Visit 7 (average of daily measurements from day 78 to day 84) will be used in the anchor-based analyses.

The same procedure will be applied to the *PROMIS Fatigue* total score and EFA-based scales using the matching 7-day recall measurement at baseline and Visit 7.

The adequacy of the anchors proposed above is explored with the Spearman correlations between the change from baseline to Visit 7 in *CARES* and *PROMIS Fatigue* total score and EFA-based scales. If the correlation coefficient >0.30, it will be emphasized in the interpretation of the results.

The following descriptive analyses will be conducted:

- Descriptive statistics for changes in the *CARES* total score and EFA-based scales from baseline (average of Day -7 to Day -1) to the Visit 7 (average of day 78 to 84). Descriptive statistics will be conducted for all categories as described in Table .
- Descriptive statistics for changes in the *PROMIS Fatigue* total score and EFA-based scales from baseline to the Visit 7. Descriptive statistics will be conducted for all categories as described in Table .
- Empirical cumulative distribution function plots (eCDF) and smooth Probability Density Functions (PDF) will be presented for each *CARES* and *PROMIS Fatigue* total score and EFA-based scales.These curves will be generated for all categories as described in Table .

Furthermore, two descriptive tables will be generated to display change in the PRO scores from baseline at Visit 7 by baseline PGIS. This will be conducted separately for patients who achieved a 1-category and 2-category PGIS decrease at Visit 7 from baseline. Change scores will be categorized by percentiles: 10th, 25th, median, 75th, and 90th. Four baseline PGIS categories (i.e., mild, moderate, severe, very severe) will be used for the 1-category PGIS decrease analysis, while three categories for baseline PGIS (i.e., moderate, severe, very severe) will be used for the 2-category PGIS decrease analysis.

The various estimates from the different streams of evidence (distribution methods and anchor methods,) will be tabulated and will be examined for convergence in an effort to triangulate onto a single threshold value that represents meaningful within-patient worsening/improvement. However, if this is not possible, a range of thresholds will be considered.

The results of the anchor-based analyses (original, un-collapsed PGIS categories) will be considered primary when making decisions regarding the clinically meaningful worsening and clinically meaningful improvement thresholds and will be supplemented with eCDF and PDF curves. The primary anchor to establish the meaningfulness of deterioration or improvement in patient scores over time is PGIS. Specifically, our *a priori* definition of clinically meaningful deterioration threshold will be change from baseline of +1 points on the PGIS; and the *a priori* definition of clinically meaningful improvement threshold will be change from baseline -1 points on the PGIS. However, the results from other analyses will also be considered.

The eCDF and ePDF plots will be visually examined to explore the data across a range of thresholds to ensure there is separation in the curves between the anchor groups and to examine the proportion of subjects experiencing up to that change at various thresholds. The eCDF displays a continuous plot of the change from baseline on the horizontal axis and the cumulative percent of patients experiencing up to that change on the vertical axis. The eCDFs and PDFs are mathematically related, displaying a distribution of PRO change score in different but complementary ways. eCDF plots optimally display the cumulative probabilities of change across the entire distribution and provide insight into the percentiles (e.g., median) of the PRO

changes. PDF plots optimally display the moments of a distribution, such as the mean and variability.

### 7.3.5. CLASSIFICATION BY PATIENT SYMPTOMATOLOGY

One of the objectives of the interim blinded PRO data analysis of RESCUE data is to characterize patient symptomology, such that two classes of patients can be defined: low symptom burden (or non-symptomatic) and high symptom burden (or symptomatic).

The investigation will rely on quantitative approaches and will be also reviewed on the basis of qualitative considerations. This analysis will be exploratory in nature and iterative. Two methods will be considered:

- Cluster analysis (see section 7.3.5.1)
- Anchor analysis (see section 7.3.5.2)

The two symptom classes resulting from the above methods will be investigated for differential response patterns, by means of profile analysis (see section 7.3.5.3).

#### 7.3.5.1. Cluster analysis

Cluster analysis will be performed to identify subgroups of patients who differ meaningfully on symptoms at baseline. The items and/or scales to be used for cluster analysis will be determined following an examination of the results of the descriptive analyses, item-to-item correlations, and EFA. The analysis will be performed separately for *CARES* and *PROMIS* Fatigue.

Two Cluster Analysis methods may be used: hierarchical clustering and k-means (or k-medoids) clustering. The initial cluster centres may be based on prototypical (clinically meaningful) "symptomatic", "non-symptomatic" etc. cluster centres. The algorithm from the k-means cluster analysis establishes an initial set of cluster means then assigns each case (i.e., patient) to the closest cluster mean. Alternative clustering algorithms may be used as appropriate given the features of the response. A two-cluster solution will be specified to aid in interpretation.

#### 7.3.5.2. Anchor analysis

An external anchor, specifically PGIS will be used to inform the definition of a sufficiently symptomatic patient. Specifically, the median baseline score for those who reported "mild", "moderate" or "severe" severity on the PGIS at screening will used as a *threshold* for significant symptom burden. Patients will be further classified as "non-symptomatic" if their baseline score is < *threshold* and as "symptomatic" if their baseline score is ≥ *threshold*.

The analysis will be performed separately for each scale/subscale of *CARES* and *PROMIS* Fatigue.

#### 7.3.5.3. Profile analysis

The cluster analysis and the anchor analysis will identify patient profiles (symptom classes) based on shared symptoms at baseline. Differential response patterns will be assessed. Specifically, the following will be presented:

- A table with descriptive statistics for the PRO scores and a line graph with mean values and corresponding 95% confidence intervals (CI) by symptom classes and each visit (e.g., at baseline, Visit 4 and Visit 7).
- A table with descriptive statistics for the change from baseline in PRO scores by symptom classes and each visit (e.g., at baseline, Visit 4 and Visit 7).
- A cumulative distribution plot showing a continuous plot of the absolute change from baseline during the study for the PRO scores on the X-axis and the cumulative percent of patients experiencing that change on the Y-axis will be presented by symptom classes at Visit 7.

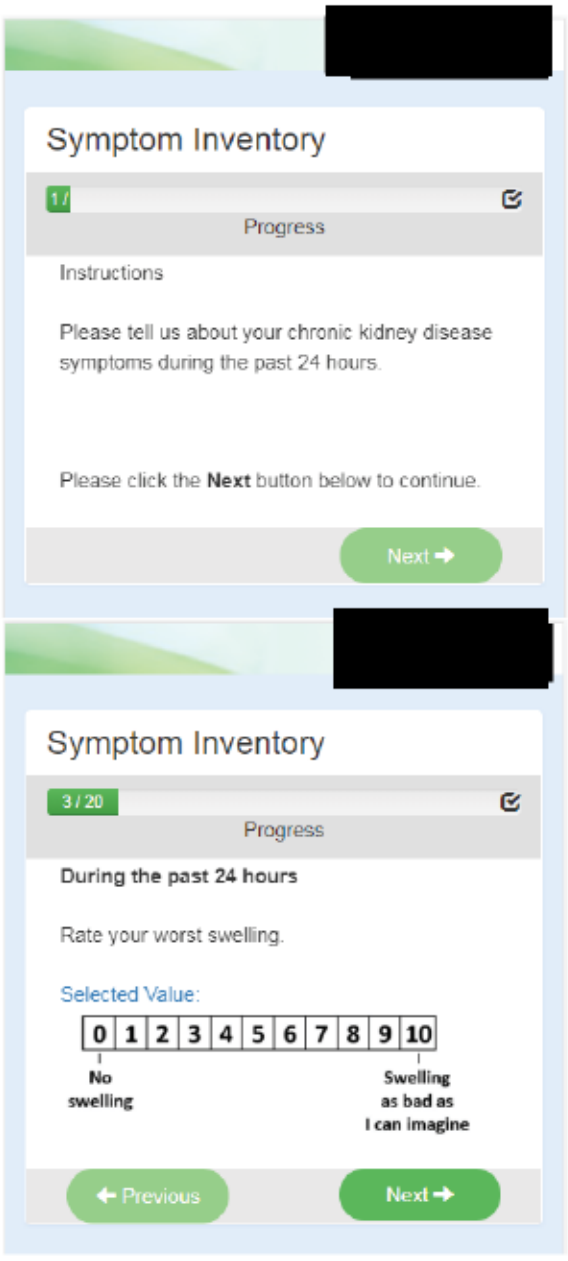
In addition, a mixed model for repeated measures (MMRM) will be used to evaluate change from baseline to Visit 7. The model will include variables for baseline hemoglobin (>= 11 or < 11 g/dL), CKD Stage (3, 4 or 5), symptom class group, visit and treatment group-by-visit interaction as categorical fixed effects, baseline value and baseline-by-visit interaction will be included as covariates. The least squares means for each symptom class group, the least squares mean differences between symptom class groups along with the associated 95% confidence intervals (CIs) and p-values will be presented.

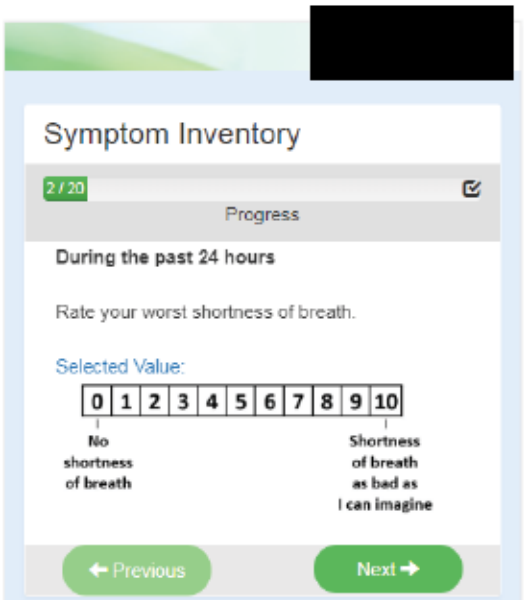
The analyses will be conducted using PROC MIXED in SAS. An unstructured covariance matrix will be used to model the within-subject correlation. The Kenward-Roger approximation will be used to adjust the denominator degrees of freedom. The analysis will be performed based on all observed post-baseline scores without any imputation of missing data. In the case when the MMRM fails to converge using an unstructured covariance matrix in any stage, a less stringent covariance matrix (e.g., autoregressive 1) will be used.

# 8. REFERENCES


| | |
|---|---|
| Cohen 1992 | Cohen J. A power primer. Psychol Bull. 1992 Jul;112(1):155-9. PubMed PMID:19565683 |
| Fabrigar et al. 1999 | Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological methods, 4(3), 272. |
| Fayers et al. 2007 | Fayers PM, Machin D. Quality of life. The assessment, analysis and interpretation of patient-reported outcomes. Second edition. Wiley 2007. |
| FDA 2009 | US Department of Health and Human Services (USDHHS). Guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. December.2009. |
| FDA 2019 | Patient-focused drug development guidance public workshop. October 15-16, 2018. https://www.fda.gov/downloads/Drugs/NewsEvents/UCM620711.pdf. Accessed 28 March 2019. |
| Liang 2000 | Liang, Matthew. (2000). Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments. Medical care. 38. II84-90. |
| Nunnally and Bernstein 1994 | Nunnally, J. C., and Bernstein, I. H. (1994). Validity. Psychometric theory, 3, 99-132. |
| QualityMetric 2011 | User's manual for the SF-36v2 Health Survey (3 ed.). Lincoln, RI. Strand, V., Schett, G., Hu, C., & Randall, M. (2013). Patient-reported Health-related Quality of Life with Apremilast for Psoriatic Arthritits: A Phase II, Randomized, Controlled Study. (Vol. 40). The Journal of Rheumatology. doi:10.3899/jrheum.121200 |
| Pett et al. 2003 | Pett M, Lackey N, Sullivan J, editors. Making Sense of Factor Analysis. Thousand Oaks, CA: Sage; 2003 |
| Shrout and Fleiss 1979 | Shrout PE and Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, Vol 86(2), Mar 1979, 420-428. |

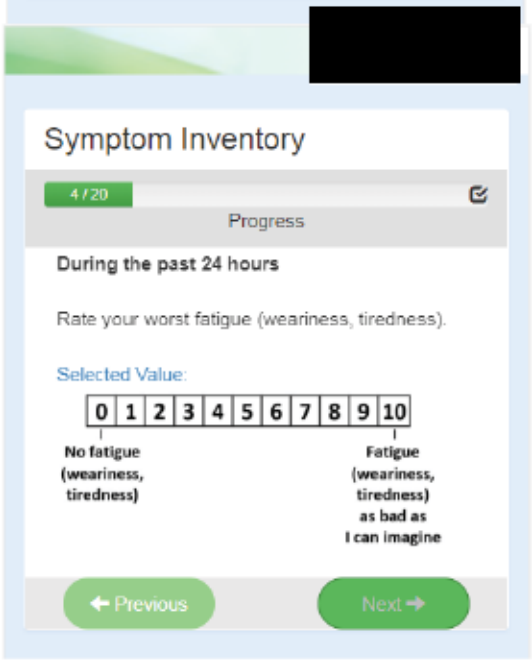
# APPENDIX A: CORVIDIA EPRO (CARES)

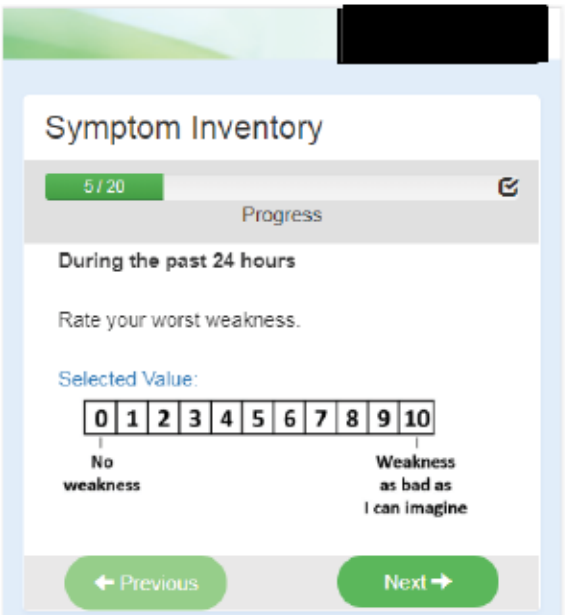


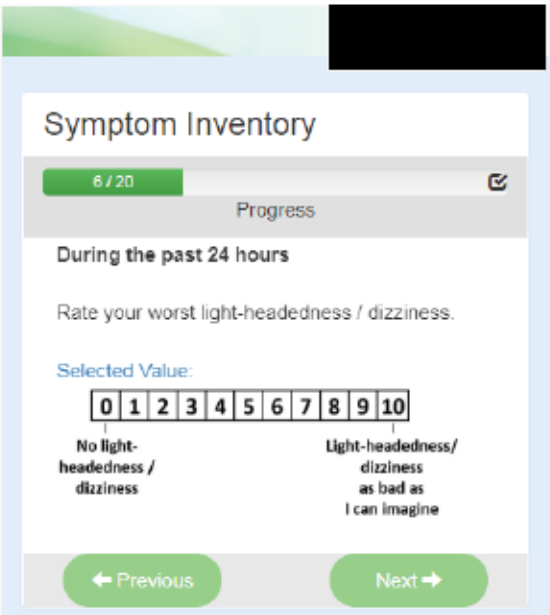



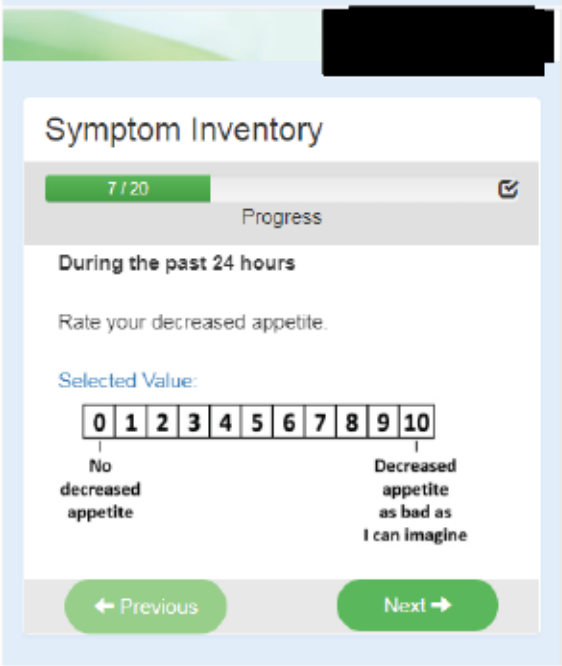


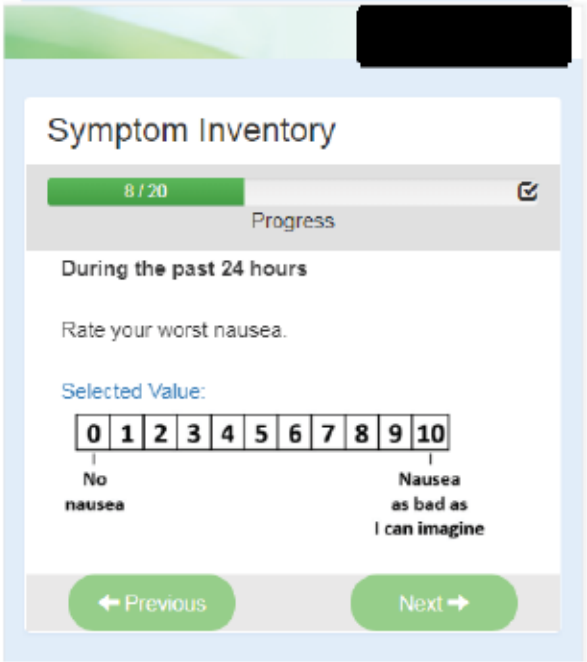


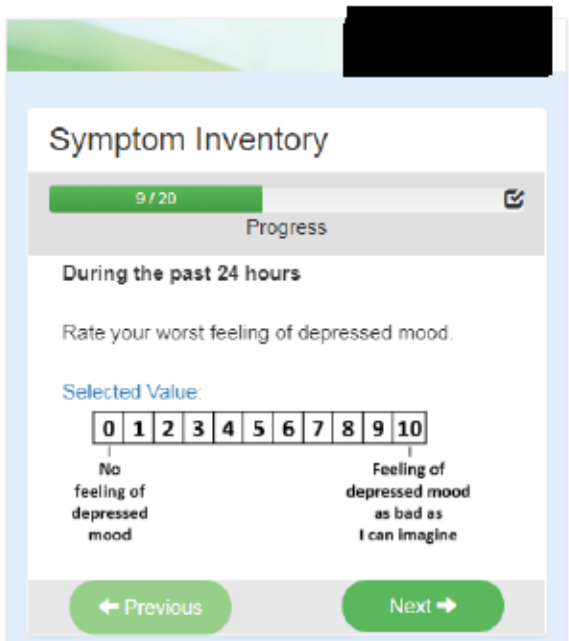


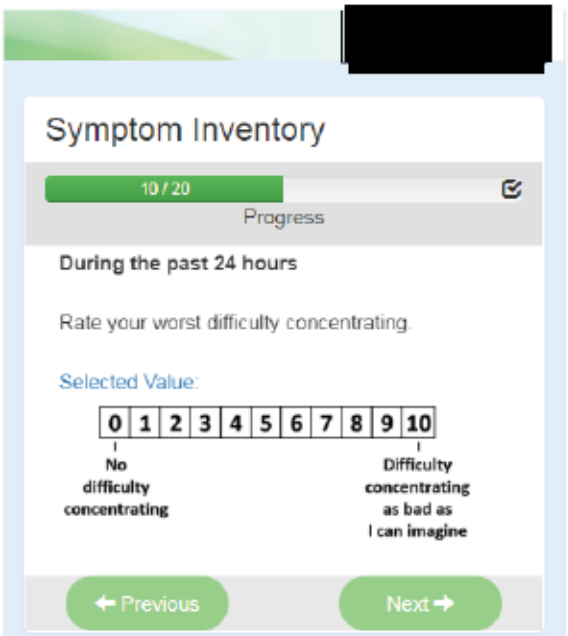


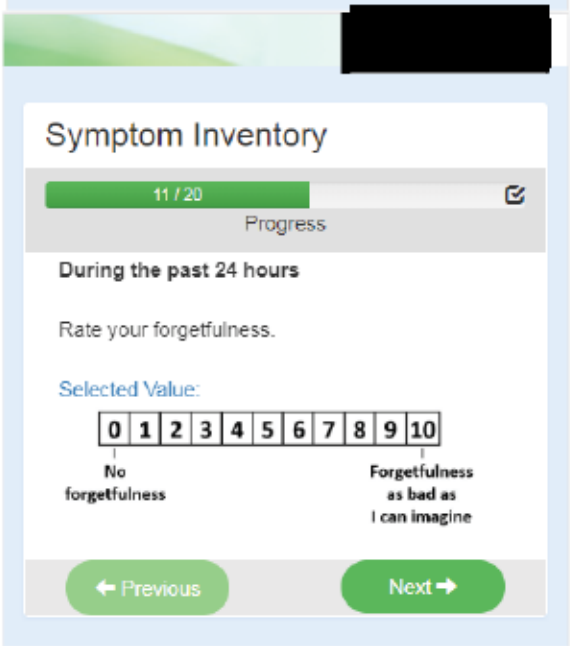



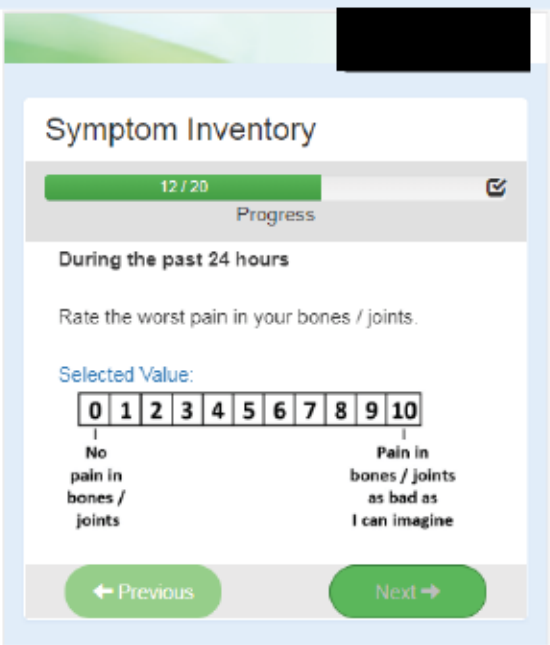



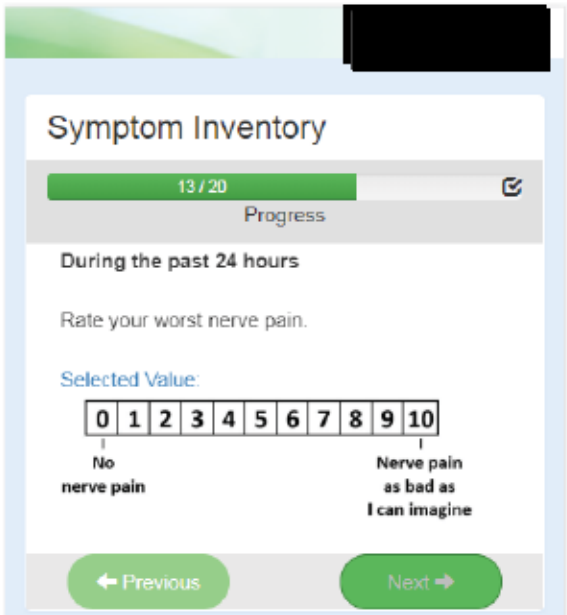



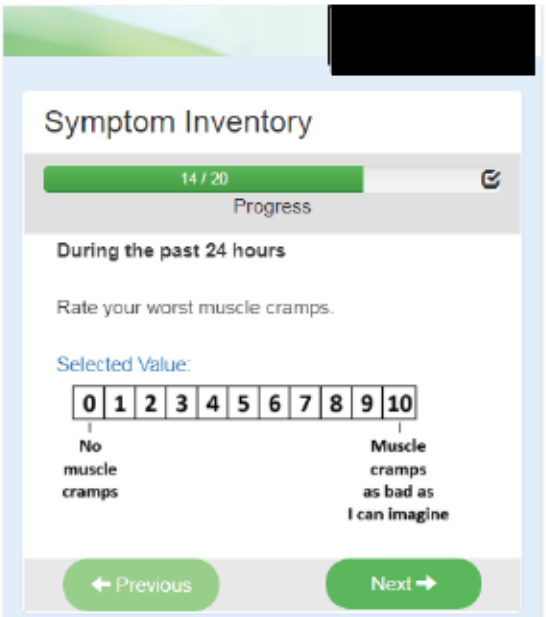


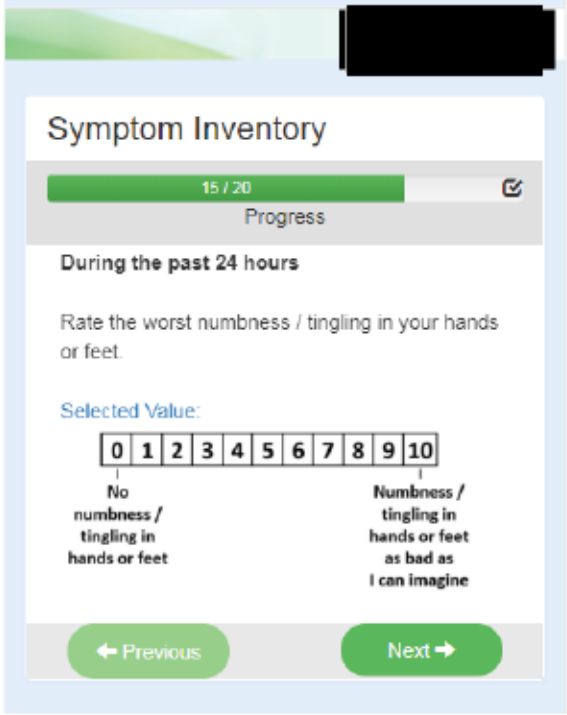



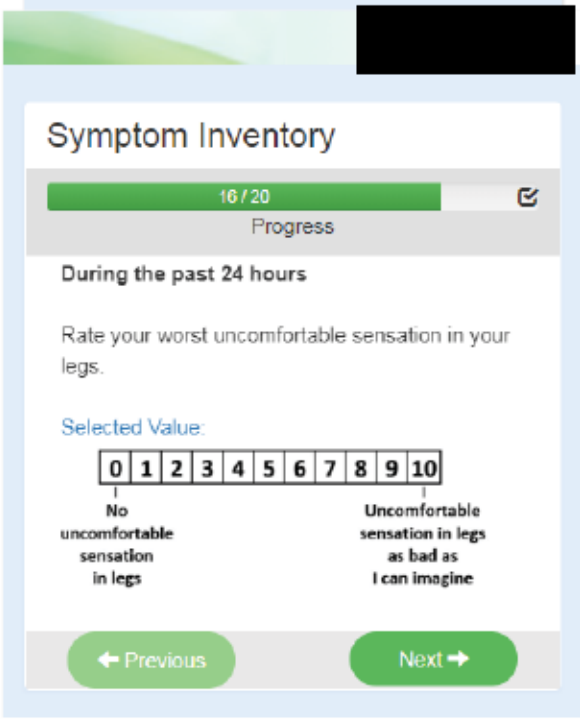



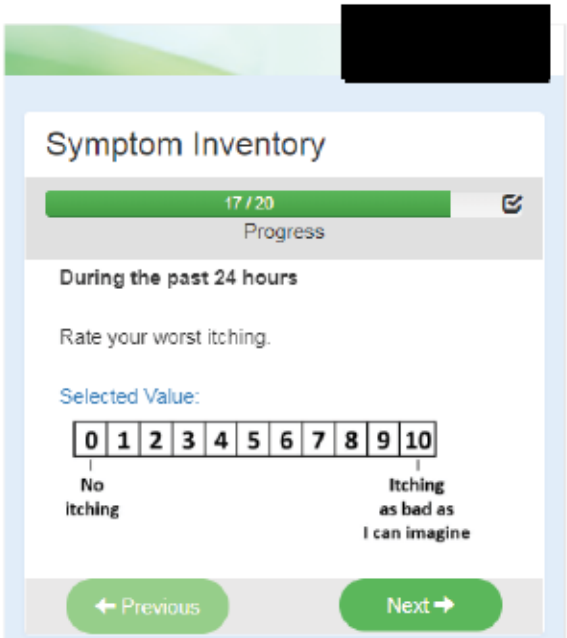



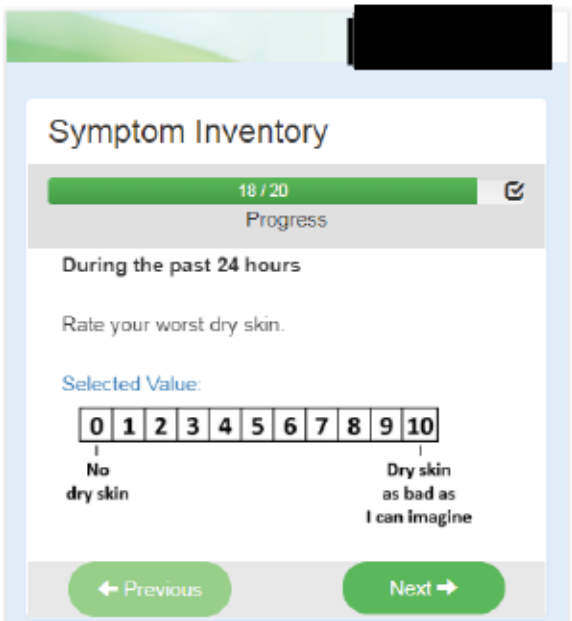


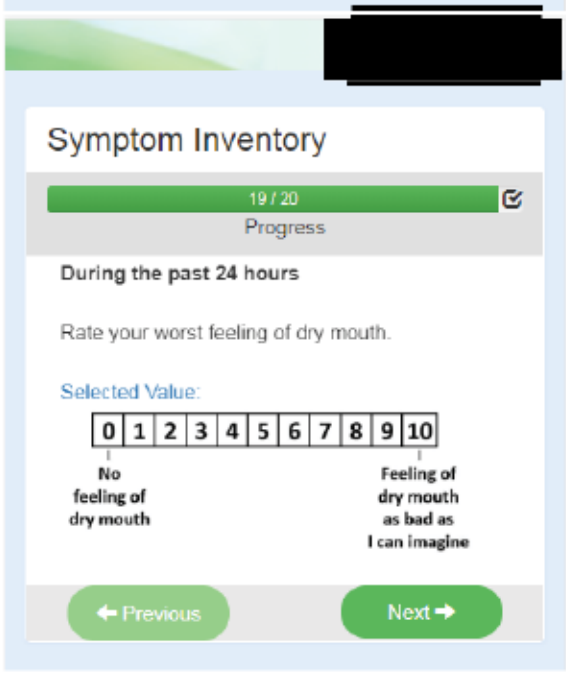



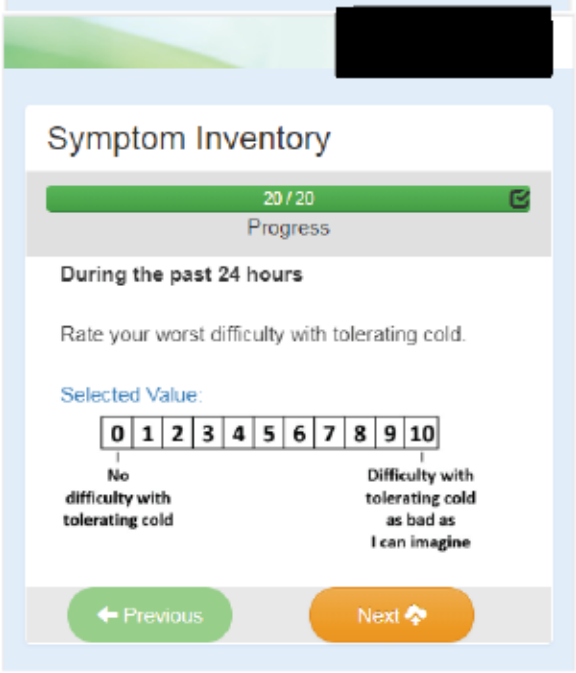



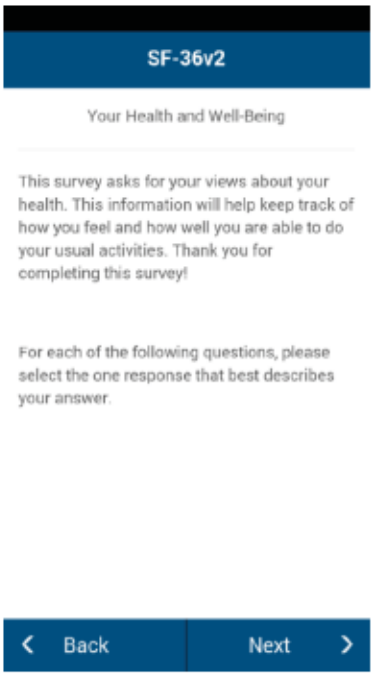


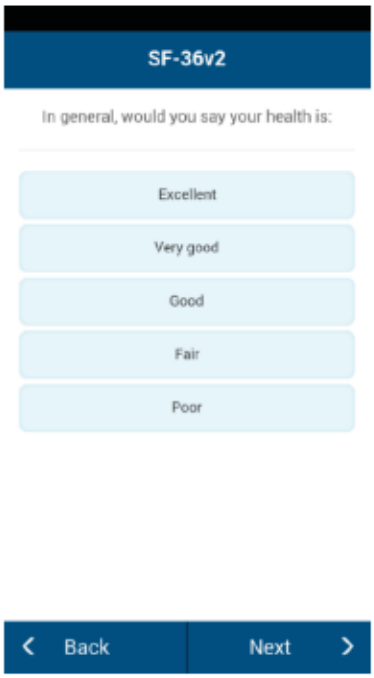



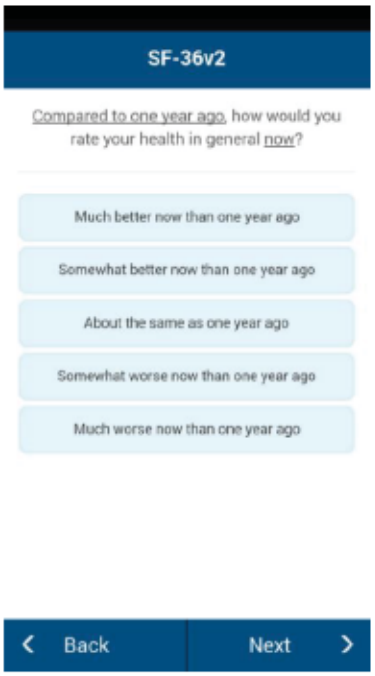



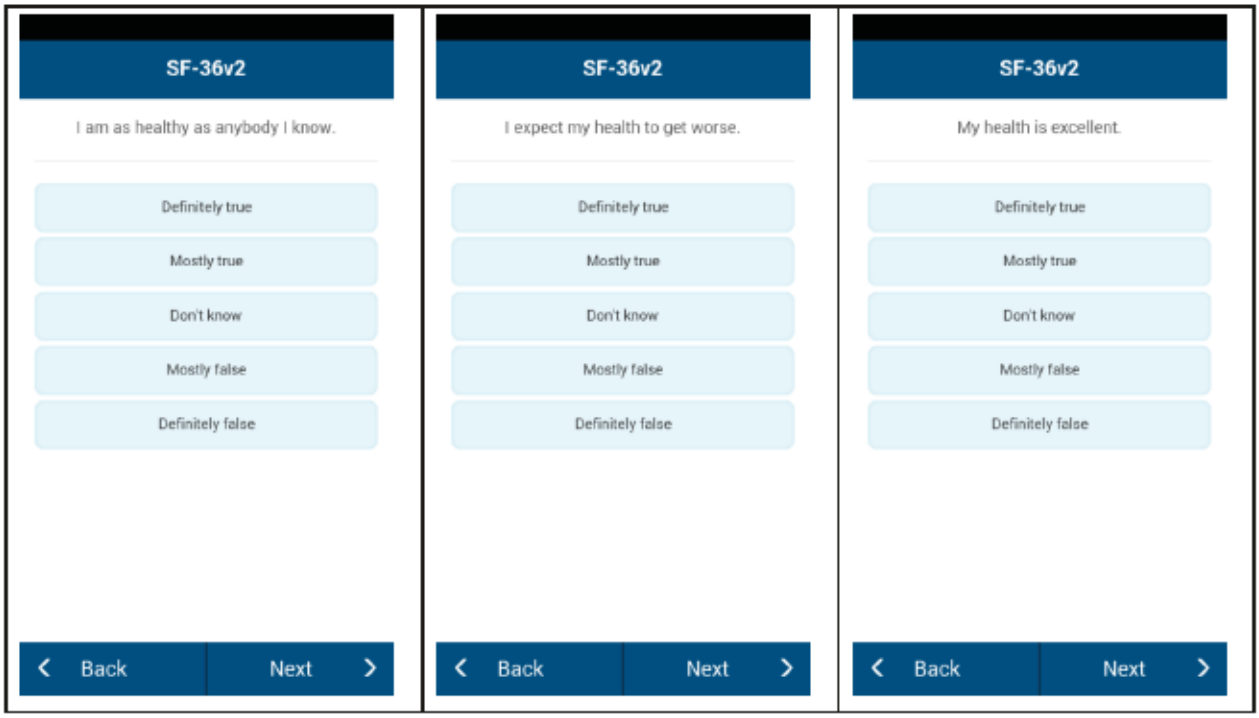



# APPENDIX C: SHORT FORM 36-ITEM HEALTH SURVEY V2 (SF-36V2)

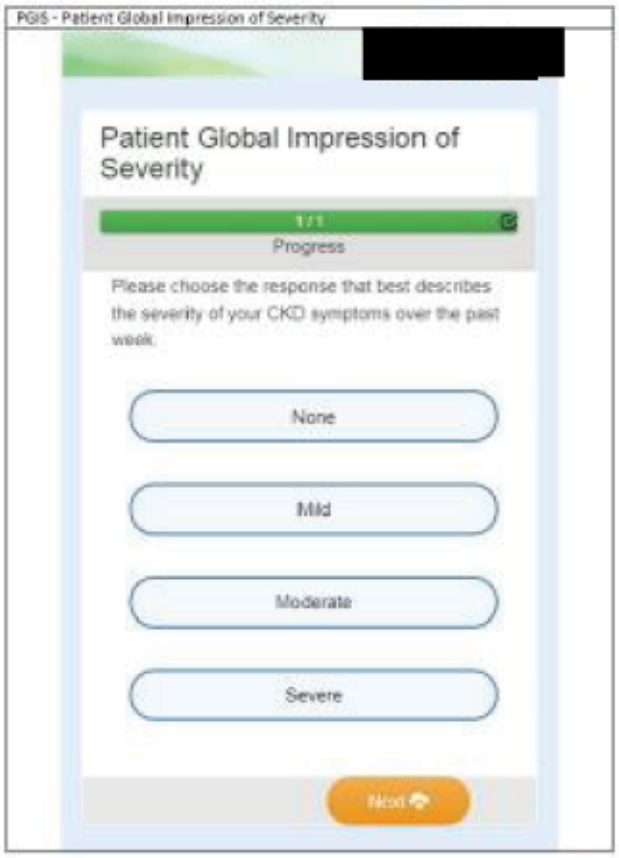




# APPENDIX D: PGIC AND PGIS




# APPENDIX D: EQ-5D-5L

| EQ-5D-5L | |
|---|---|
| EQ-5D-5L PDA version | |
| English (USA) | Country (Language) |
| **Health Questionnaire** | Health Questionnaire |
| **English version for the USA** | Version (Target Language) |
| | Version (English) |
| On the following screens please tap the statement that best describes your health TODAY. | Instruction |
| **Your mobility TODAY** | **Mobility** |
| I have no problems walking | MB1 |
| I have slight problems walking | MB2 |
| I have moderate problems walking | MB3 |
| I have severe problems walking | MB4 |
| I am unable to walk | MB5 |
| **Your self-care TODAY** | **Self-care** |
| I have no problems washing or dressing myself | SC1 |
| I have slight problems washing or dressing myself | SC2 |
| I have moderate problems washing or dressing myself | SC3 |
| I have severe problems washing or dressing myself | SC4 |
| I am unable to wash or dress myself | SC5 |
| **Your usual activities TODAY** *(e.g. work, study, housework, family or leisure activities)* | **Usual Activities** |
| I have no problems doing my usual activities | UA1 |
| I have slight problems doing my usual activities | UA2 |
| I have moderate problems doing my usual activities | UA3 |
| I have severe problems doing my usual activities | UA4 |
| I am unable to do my usual activities | UA5 |
| **Your pain / discomfort TODAY** | **Pain / Discomfort** |
| I have no pain or discomfort | PD1 |
| I have slight pain or discomfort | PD2 |
| I have moderate pain or discomfort | PD3 |
| I have severe pain or discomfort | PD4 |
| I have extreme pain or discomfort | PD5 |
| **Your anxiety / depression TODAY** | **Anxiety / Depression** |
| I am not anxious or depressed | AD1 |
| I am slightly anxious or depressed | AD2 |
| I am moderately anxious or depressed | AD3 |
| I am severely anxious or depressed | AD4 |
| I am extremely anxious or depressed | AD5 |
| We would like to know how good or bad your health is TODAY. | Vas Line 1 |
| On the next screen you will see a scale numbered 0 to 100. | Vas Line 2 |
| 100 means the best health you can imagine. | Vas Line 3 |
| 0 means the worst health you can imagine. | Vas Line 4 |
| Please tap on the scale to indicate how your health is TODAY. | Vas Line 5 |
| The best health you can imagine | Top Scale |
| The worst health you can imagine | Bottom Scale |
| YOUR HEALTH TODAY | Box Health |



***Disclaimer:** This is a preview of the EQ-5D instrument. It demonstrates the text, questions and response options included in this version. This preview does not represent the final product and should not be used as an official EQ-5D instrument.*

# APPENDIX D: PROMIS FATIGUE 13A, FATIGUE, SEX INTEREST

## PROMIS Items

| | During the past 7 days..... | Not at all | A little bit | Somewhat | Quite a bit | Very much |
|---|---|---|---|---|---|---|
| PROMIS Short Form v1.0 – Fatigue 13a | I feel fatigued | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I feel weak all over | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I feel listless ("washed out") | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I feel tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I have trouble starting things because I am tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I have trouble finishing things because I am tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I have energy | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I am able to do my usual activities | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I need to sleep during the day | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I am too tired to eat | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I need help doing my usual activities | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I am frustrated by being too tired to do the things I want to do | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | I have to limit my social activity because I am tired | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| PROMIS items (Item identifier in bold) | **FATEXP36:** How exhausted were you on average? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | **FATEXP43:** How physically drained were you on average? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | **FATIMP52:** To what degree did your fatigue make you feel less alert? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |
| | **During the past 30 days.....** | **Not at all** | **A little bit** | **Somewhat** | **Quite a bit** | **Very much** |
| | **SFINT101:** How interested have you been in sexual activity? | 1☐ | 2☐ | 3☐ | 4☐ | 5☐ |

## Document Approval Signature(s)

This document has been signed electronically in compliance with [redacted] and [redacted] Companies' electronic signature policies and procedures. This page is a manifestation of the electronic signatures, which may be confirmed within the [redacted] and [redacted] Companies' source system in accordance with [redacted] and [redacted] Companies' records retention policies. Note: the Title field in the signature block has been intentionally left blank.

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

UserName: [redacted]
Title:
Date: [redacted]
Meaning: I Approve the Document
==========

[redacted]

# STATISTICAL ANALYSIS PLAN

| | |
|---|---|
| **Protocol Title:** | A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE) |
| **Protocol Number:** | COR-001-02 |
| **Protocol Version/Date:** | Amendment 5, 23 December 2019 |
| **Investigational Product:** | Ziltivekimab (human monoclonal antibody to IL-6) |
| **Sponsor:** | Corvidia Therapeutics<br>35 Gatehouse Drive<br>Waltham MA 02451 |
| **SAP Version/Date:** | Version 1.0, 31 March 2020 |

# SIGNATURE PAGE

**Protocol Title:** A Phase 2, Randomized, Double-Blind, Placebo-Controlled Trial to Evaluate REduction in Inflammation in PatientS with advanced Chronic Renal Disease Utilizing Antibody MEdiated IL-6 inhibition (RESCUE)

**Protocol Number:** COR-001-02

**SAP Version/Date:** Version 1.0, 31 March 2020

We, the undersigned, have reviewed and approved this Statistical Analysis Plan:

| Signature | Date |
|---|---|
| [redacted] | [redacted] |
| [redacted] | |
| DocuSigned by: [redacted] | [redacted] |
| [redacted] | |
| [redacted] | [redacted] |
| [redacted] | |

## VERSION HISTORY

| Version | Version Date | Description |
|---|---|---|
| 1.0 | 31MAR2020 | Final version |

## TABLE OF CONTENTS

## LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| ADaM | Analysis Data Model |
| ADA | Anti-Drug Antibodies |
| AE | Adverse event |
| ALT | Alanine aminotransferase |
| AST | Aspartate aminotransferase |
| ATC | Anatomical therapeutic class |
| BUN | Blood urea nitrogen |
| CDISC | Clinical Data Interchange Standards Consortium |
| CHr | Reticulocyte Hemoglobin Content |
| CRF | Case report form |
| CSR | Clinical Study Report |
| EDC | Electronic Data Capture |
| EQ-5D-5L | EuroQol-5D-5L |
| GFR | Glomerular Filtration Rate |
| hs-CRP | high-sensitivity C-Reactive Protein |
| MAR | Missing at Random |
| MCS | Mental Component Summary |
| MCV | Mean Corpuscular Volume |
| MedDRA | Medical Dictionary for Regulatory Activities |
| MNAR | Missing Not at Random |
| NDD-CKD | Non-dialysis-dependent Chronic Kidney Disease |
| NMR | Nuclear Magnetic Resonance |
| NT-pro-BNP | N-terminal prohormone-B-type natriuretic peptide |
| IWRS | Interactive Web-Response System |
| HDL | High Density Lipoprotein |
| LDL | Low Density Lipoprotein |
| PAP | Psychometric Analysis Plan |
| PCS | Physical Component Summary |
| PGIC | Patient Global Impression of Change |
| PGIS | Patient Global Impression of Severity |
| PRO | Patient Reported Outcomes |
| ePRO | electronic PRO |
| PROMIS | Patient-Reported Outcomes Measurement Information System |
| PROMIS SexFS | PROMIS Sexual Function and Satisfaction |
| RBC | Red Blood Count |
| RDW | Red cell Distribution Width |
| PK | Pharmacokinetics |
| PD | Pharmacodynamic |
| SAA | Serum Amyloid A |
| SAE | Serious adverse event |
| SAP | Statistical Analysis Plan |
| SDTM | Study Data Tabulation Model |
| SF-36 | 36-Item Short Form Survey |
| ST2 | Suppression of Tumorigenicity 2 |
| TEAE | Treatment-emergent adverse event |

| Abbreviation | Definition |
|---|---|
| TESAE | Treatment-emergent serious adverse event |
| TIBC | Total Iron Binding Capacity |
| TIMI | Thrombolysis In Myocardial Infarction |
| TSAT | Transferrin Saturation |
| UACR | Urine Albumin-to-Creatinine Ratio |
| WBC | White Blood Count |
| WHO | World Health Organization |

# 1 INTRODUCTION

The purpose of this Statistical Analysis Plan (SAP) is to provide a description of the statistical methods to be implemented for the analysis of data from the study with protocol number COR-001-02. The SAP will be finalized prior to database lock. Any deviations from the SAP after database lock will be documented in the final Clinical Study Report (CSR).

# 2 STUDY OVERVIEW

## 2.1 Study Objectives

Patients with non-dialysis-dependent chronic kidney disease (NDD-CKD), who have evidence of systemic inflammation with increased cardiovascular risk, will be enrolled into this trial. The purpose of this trial is to determine a dose to select for a potential cardiovascular outcome trial with Ziltivekimab.

### *2.1.1 Primary Objective*

The primary objective is to evaluate the effects of Ziltivekimab compared to placebo on a marker of inflammation: high-sensitivity C-reactive protein (hs-CRP).

### *2.1.2 Secondary Objectives*

The secondary objectives of the study are as follows:

- To evaluate the effects of Ziltivekimab compared to placebo on a markers of inflammation and cardiovascular risk: serum amyloid A (SAA) and fibrinogen.

### *2.1.3 Safety Objectives*

- To evaluate the safety of three dose levels of Ziltivekimab compared to placebo.

### *2.1.4 Pharmacokinetic Objectives*

- To evaluate the pharmacokinetics (PK) and PK-pharmacodynamic (PK-PD) modeling of Ziltivekimab following multiple doses at three different dose levels.

### *2.1.5 Exploratory Objectives*

The exploratory objectives of the study are as follows:

- To evaluate the effects of Ziltivekimab compared to placebo on markers of anemia (hemoglobin).
- To evaluate the effects of Ziltivekimab compared to placebo on markers of inflammation-malnutrition (albumin).
- To evaluate the effects of Ziltivekimab compared to placebo on markers of cardiovascular risk (i.e. N-terminal prohormone-B-type natriuretic peptide [NT-pro-BNP]).
- To evaluate the effects of Ziltivekimab compared to placebo on additional markers of cardiovascular risk (Suppression of Tumorigenicity 2 [ST2]).

- To evaluate the effects of Ziltivekimab compared to placebo on markers of atherosclerosis risk: LDL-C, triglycerides, ApoB, ApoA1, ApoB/ApoA1, Nuclear Magnetic Resonance (NMR) lipoprotein profile, and Lp (a).

- To evaluate the effects of Ziltivekimab compared to placebo on markers of kidney function: cystatin C, estimated Glomerular Filtration Rate (eGFR) and kidney damage: urine albumin-to-creatinine ratio (UACR).

- To determine Ziltivekemab trough drug levels following multiple Ziltivekimab doses at three different dose levels.

- To evaluate the effects of three dose levels of Ziltivekimab compared to placebo on Patient Reported Outcomes (PRO): Patient-Reported Outcomes Measurement Information System (PROMIS®) Fatigue 13a short form, selected items from the PROMIS fatigue item bank, the Optum 36-Item Short Form Survey (SF-36) v2® Health Survey, a Corvidia PRO, the PROMIS interest in sexual activity item, the Patient Global Impression of Change (PGIC), Patient Global Impression of Severity (PGIS), and the EuroQol-5D-5L (EQ-5D-5L).

- To evaluate the psychometric properties of the PROMIS Fatigue 13a short form and selected items from the PROMIS fatigue item bank, and the Corvidia electronic PRO (ePRO) items in CKD patients.

- To evaluate the effects of Ziltivekimab on systemic iron availability: transferrin saturation (TSAT), reticulocyte hemoglobin content (CHr), total iron binding capacity (TIBC), systemic iron stores (serum ferritin), serum iron, and systemic iron regulation (serum hepcidin).

- To evaluate the effects of Ziltivekimab compared to placebo on markers of cardiovascular risk (NT-pro-BNP) in patients with baseline NT-pro-BNP > 250 pg/mL.

- To evaluate the effects of Ziltivekimab compared to placebo on markers of anemia in patients with baseline hemoglobin < 11 g/dL.

- To evaluate the effects of Ziltivekimab compared to placebo on markers of inflammation malnutrition (i.e. albumin) in patients with baseline albumin < 4.0 g/dL.

## 2.2 Study Design

### *2.2.1 Overview*

This is a Phase 2, randomized, double-blind, placebo-controlled trial designed to evaluate the efficacy, safety, and pharmacokinetics of Ziltivekimab at three dose levels (7.5 mg, 15 mg or 30 mg) compared to placebo in patients with stage 3-5 CKD, not on dialysis, who have evidence of inflammation with high cardiovascular risk.

The study consists of three periods: Screening Period [Days -14 to -1], Treatment Period [Day 1 through Week 24], and Safety Follow-Up Period [Weeks 25 through 32].

Patients will undergo a Screening Period of up to 14-days during which inclusion and exclusion criteria will be evaluated. All inclusion/exclusion criteria must be met for enrollment.

Approximately 240 patients will be randomized 1:1:1:1 (60-per dose group) to one of three Ziltivekimab dose levels (7.5 mg, 15 mg or 30 mg) or Placebo in the trial. Patient randomization will be stratified by baseline hemoglobin (≥ 11 or < 11 g/dL) and CKD stage (3, 4 or 5). After the Screening Period, randomized patients will be dosed every 28 days out to Week 21. The primary, secondary, and exploratory endpoints will be analyzed after 13 weeks of treatment and then followed for additional efficacy analyses through Week 24. Selected efficacy endpoints and safety assessments will be evaluated in the Follow-up Period Week 25 through Week 32.

Patients will be randomized on Day 1 and the first dose of study drug will be administered after all assessments are conducted. Subsequent doses of study drug will be administered every 28-days for a total of 6 treatments (Weeks 1, 5, 9, 13, 17, and 21). Study visits will follow the schedule of procedures.

The schedule of procedures can be found in Table 1 below:

**Table 1 SCHEDULE OF PROCEDURES**

| | Screening[1] | | Treatment Period | | | | | | | | | | | | Safety F/Up | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Visit Number** | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | N/A | 7 | 8 | 9 | 10 | 11/ET | 12 | 13/ET |
| **Visit Day Window** | -14 | -7 | 1 | 8 | 22 | 29-35 | 43-49 | 57-63 | 78-84 | 85-91 | 113-119 | 141-147 | 155-161 | 162-168 | 190-196 | 218-224 |
| **Visit Window** | | ±2 | | ±2 | ±2 | | | | | | | | | | | |
| **Visit Week** | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 12 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |
| ICF signed | X | | | | | | | | | | | | | | | |
| Medical History/Update | | X | X | | | | | | | | | | | | | |
| Randomization | | | X | | | | | | | | | | | | | |
| Study drug administered (post-assessments) | | | X | | | X | | X | | X | X | X | | | | |
| Concomitant and Prior Medications | | X | X | X | X | X | X | X | | X | X | X | X | X | X | X |
| Vital Signs[2] | | X | X | | X | X | | X | | X | X | X | X | X | X | X |
| BMI | | | X | | | | | | | X | | | | X | | |
| Hematology[3] | X | X | X | X | X | X | X | X | | X | X | X | X | X | X | X |
| Lipids (fasting 8 hours)[4] | | X | X | | | | | | | X | | | X | X | X | X |
| Chemistry[5] | X | X | X | X | X | X | X | X | | X | X | X | X | X | X | X |
| Spot urine protein-creatinine ratio | | X | | | | | | | | X | | | | | | |
| Urinalysis (including albumin)[6] | | | X | | | | | | | X | | | X | X | | |
| Iron Indices[7] | X | | X | | | | | | | X | | | X | X | X | X |
| Serum Hepcidin | | | X | | | | | | | X | | | | X | | |
| TMPRSS6 Genotype test | X | | | | | | | | | | | | | | | |

**Table 1 SCHEDULE OF PROCEDURES**

| | Screening[1] | | Treatment Period | | | | | | | | | | | | Safety F/Up | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Visit Number** | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | N/A | 7 | 8 | 9 | 10 | 11/ET | 12 | 13/ET |
| **Visit Day Window** | -14 | -7 | 1 | 8 | 22 | 29-35 | 43-49 | 57-63 | 78-84 | 85-91 | 113-119 | 141-147 | 155-161 | 162-168 | 190-196 | 218-224 |
| **Visit Window** | | ±2 | | ±2 | ±2 | | | | | | | | | | | |
| **Visit Week** | -2 | -1 | 1 | 2 | 4 | 5 | 7 | 9 | 12 | 13 | 17 | 21 | 23 | 24/ET | 28 | 32/ET |
| Infectious Disease Screen[8] | | X | | | | | | | | | | | | | | |
| Special RBCs[9] | | | X | | | | | | | X | | | | X | X | X |
| IL-6[10] | | | X | | X | X | | X | | X | X | X | X | X | X | X |
| hs-CRP [11] | X | | X | | X | X | | X | | X | X | X | X | X | X | X |
| Serum Pregnancy, FSH [12] | | X | | | | | | | | | | | | | | |
| Ziltivekimab trough PK[13] | | | X | X | X | X | X | X | | X | X | X | X | X | X | X |
| ADA[14] | | | X | X | X | X | X | X | | X | X | X | X | X | X | X |
| 12-lead ECG [15] | | X | | | | | | | | | | | | X | | |
| CKD Biomarkers[16] | | X | X | | | | | | | X | | | X | X | | |
| NT-pro-BNP | | X | X | | | | | | | X | | | X | X | | |
| Lp(a), ApoB, ApoA1[4] | | X | X | | | | | X | | X | | | X | X | | |
| Lipid Profile by NMR spectroscopy (fasting 8 hours)[4] | | | X | | | | | | | X | | | X | X | | |
| Cardiac Risk Biomarker[17] | | | X | | | | | | | X | | | | X | | |
| Storage samples for Exploratory Biomarkers and RNA Testing[18] | | X | | | | | | | | X | | | | X | | |
| DNA Testing (Stored Blood Samples) | | X | | | | | | | | | | | | | | |
| Limited Physical Examination[19] | | X | | | | | | | | | | | | X | | |
| INR[20] | | | X | X | X | X | X | X | | X | X | X | X | X | | |
| Adverse Events | | | X | X | X | X | X | X | | X | X | X | X | X | X | X |
| PROMIS® 13a and Fatigue Items[21] | | | X | | | X | | | | X | | | X | X | | |
| PROMIS® Sexual Interest Items[21] | | | X | | | X | | | | X | | | | X | | |
| CARES (Corvidia ePRO)[22] | | X | | | X | | | | X | | | | X | X | | |
| PGIS[23] | | X | | | X | | | | X | | | | | X | | |
| PGIC[24] | | | | | | X | | | | X | | | | X | | |
| Optum SF-36 v2®[25] | | | X | | | X | | | | X | | | X | X | | |
| EQ-5D-5L[26] | | | X | | | X | | | | X | | | | X | | |

ET = Early Termination; F/Up = Follow-up.

1. If a Screening laboratory (excluding hematology) result is outside inclusion criteria parameters screening may be extended up to one week to allow for retesting of lab test(s) that did not meet inclusion criteria.
2. Vital signs include temperature, respiratory rate, heart rate, and blood pressure (patients may be re-evaluated within 2 weeks, at the discretion of the PI if anti-hypertensive therapy has been started or increased as a result of initial screening blood pressure). Whenever possible, vital signs will be obtained after at least 5 minutes resting in the supine position or, when necessary, in a semi-recumbent position. Vital signs will be done prior to ECG recordings.

3. Hematology: hemoglobin, hematocrit, reticulocyte count, red blood count (RBC) indices (e.g., mean corpuscular volume (MCV), red cell distribution width (RDW), platelets, white blood count (WBC), WBC differential.
4. Lipids and subfractions (fasting 8 hours): total cholesterol, low density lipoprotein (LDL) cholesterol, high density lipoprotein (HDL), and triglycerides; Lp(a), ApoB, ApoA1.
5. Chemistry: sodium, potassium, chloride, bicarbonate (or CO2), calcium, phosphate, alanine aminotransferase (ALT), aspartate aminotransferase (AST), blood urea nitrogen (BUN) creatinine, total bilirubin, direct bilirubin, alkaline phosphatase, albumin, glucose and eGFR calculated by CKD-EPI Creatinine Equation (https://www.kidney.org/content/ckd-epi-creatinine-equation-2009).
6. At Weeks 1 and 24, a portion of urine will be stored frozen for future analysis.
7. Iron Indices: TSAT, ferritin, total iron binding capacity (TIBC).
8. Infectious Disease Screen: HIV 1 and 2, Hepatitis B surface antigen, Hepatitis B surface antibody, Hepatitis C antibody, mycobacterium tuberculosis test (e.g., QuantiFERON) is preferred, but a purified protein derivative (PPD) skin test read within 48-72 hours by a qualified healthcare professional may also be performed (https://www.cdc.gov/tb/publications/factsheets/testing/skintesting.htm).
9. Special RBC: reticulocyte hemoglobin content (CHr).
10. IL-6: Interleukin-6. Pre-dose sampling.
11. hs-CRP: high-sensitivity C-reactive protein. Not using an average for inclusion, therefore test can be repeated as needed prior to randomization visit. Results after screening will be blinded.
12. Pregnancy: β-hCG; FSH: Follicle Stimulating Hormone.
13. On dosing Weeks 1, 5, 9, 13, 17, and 21, trough PK samples to be collected within -0.5 h pre-dose. Single PK blood samples will also be collected on Weeks 2, 4, 7, 23, 24/ET, 28, and 32/ET around the same time as the pre-dose samples were collected on previous dosing visits.
14. On dosing weeks 1, 5, 9, 13, 17, and 21, ADA: Anti-drug antibodies specimens to be collected within 0.5 h pre-dose.
15. Standard 12-lead ECG will be recorded in the supine position (or with the patient as flat as possible) after vital signs assessments. The ECG will be locally read by the Investigator.
16. Chronic Kidney Disease (CKD) Biomarkers: cystatin C and UACR.
17. Cardiac Risk marker: SSA, fibrinogen, and ST2.
18. Exploratory storage samples will be two 5-ml tubes of plasma, two 5-ml tubes of serum, and two 5-mL tubes of whole blood (PBMC testing at Screening, Week 13 and Week 24).
19. Limited physical exam to include skin, oropharynx, lungs, heart, abdomen, extremities (including feet), and any areas suggested by symptoms, with particular attention to signs of infection. May be performed by a physician-investigator or mid-level provider. Record abnormal findings in the source documents.
20. INR performed by a local laboratory on Warfarin patients only.
21. PROMIS, to be done on subjects' device at the clinic PRIOR to dosing.
22. CARES Corvidia ePRO, 7 day at home assessment on subject's personal device; the questionnaire will be answered daily for 7 days prior to dose (Screening Day -7 to Day -1 prior to Dose 1 (Visit 1), Dose 2 (Visit 4), Dose 4 (Visit 7), and the last 2 weeks after Dose 6 (Visit 23-24).

    **CARES: Daily recall**

    (starts with Visit-1)= baseline average of the 7 consecutive assessments (Day -7 to Day -1) prior to Dose 1

    (starts with Visit 3)= end of month of Dose 1

    no visit = end of month of Dose 3

    starts with Visit 10 and Visit 11 = starts 2 weeks after Dose 6
23. PGIS, to be done on subjects' device on the last day of CARES assessment
24. PGIC, to be done on subjects' device at the clinic PRIOR to dosing.
25. Optum SF-36 v2® Health Survey to be done on subjects' device at the clinic PRIOR to dosing.
26. EQ-5D-5L, to be done on subjects' device at the clinic PRIOR to dosing.

### *2.2.2 Randomization and Blinding*

#### *2.2.2.1 Randomization*

Patients who meet all inclusion criteria and no exclusion criteria will be randomized to one of three dose levels of Ziltivekimab (7.5 mg, 15 mg or 30 mg) or placebo for a 24 Week Treatment Period. Treatments will be assigned to randomized patients via an Interactive Web-Response System (IWRS). 240 patients will be randomized in a 1:1:1:1 ratio (60 per group) to Ziltivekimab 7.5 mg, Ziltivekimab 15 mg, Ziltivekimab 30 mg, or matching Placebo. Patients will be stratified

by baseline hemoglobin (≥ 11 or < 11 g/dL) and CKD stage (3, 4 or 5). Note that CKD stage is stratified into two categories: (1) stage 3 and (2) stages 4 and 5.

#### *2.2.2.2 Blinding*

This study is double-blind. The patients, investigators, site personnel, site monitors and Sponsor and the Contract Research Organization (CRO) clinical operational personnel will be blinded to the treatment assignment. Ziltivekimab and matching placebo will be provided as a liquid for subcutaneous injection.

The randomization list will be generated by a CRO statistician not otherwise involved in the study.

#### *2.2.2.3 Unblinding*

In case of a medical emergency or medical situation in which the treatment assignment is necessary for proper patient management, the Investigator may obtain the treatment assignment from the IWRS. The Investigator should make every reasonable attempt to contact the Medical Monitor before unblinding a patient. In all cases, the Investigator must submit a written report, including all pertinent details, to the Medical Monitor within 24 hours of the unblinding.

### *2.2.3 Study Drug*

The study drug regimens and the matching placebo regimen to be examined in this study are:

- Dose #1: Ziltivekimab, 7.5 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21 as a subcutaneous injection.
- Dose #2: Ziltivekimab, 15 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21 as a subcutaneous injection.
- Dose #3: Ziltivekimab, 30 mg per injection, administered 6 times (every 28-days) on Weeks 1, 5, 9, 13, 17, and 21 as a subcutaneous injection.
- Matched placebo injections administered subcutaneously 6 times (every 28 days) at the same frequency as the active treatment on Weeks 1, 5, 9, 13, 17, and 21.

Patients should receive 6 injections of their assigned dose during the trial. The total cumulative dosage will be 45 mg for those patients randomized to 7.5 mg per injection, 90 mg for those patients randomized to 15 mg per injection, and 180 mg for those patients randomized to 30 mg per injection.

### *2.2.4 Sample Size Determination*

Approximately 240 patients (60 per group) will be randomized.

The primary efficacy endpoint is percent change from baseline in hs-CRP (average of the hs-CRP value prior to randomization and Day 1) to Week 13 between each active group and placebo. Based on the observed treatment difference in percent change from baseline in hs-CRP of -60.74% between combined COR-001-01 active groups and placebo and the associated pooled SD of 16.893% in hs-CRP at Week 4 from the final analysis of study COR-001-SC1, a sample size of 54 per group yields more than 99% power with 2-sided alpha=0.05. Taking into consideration the dropout rate of 10% by the end of the study, a sample size of 60 per group is planned for this study.

## 2.3 Study Endpoints

### *2.3.1 Primary Endpoint*

The primary endpoint of this study is the difference in percent change in hs-CRP levels from Baseline (average of the hs-CRP value prior to randomization and Day 1) to Week 13 between each active group and placebo.

### *2.3.2 Secondary Endpoints*

The secondary endpoints are as follows:

- Difference in change in SAA from Baseline to Week 13 between each active group and placebo.
- Difference in percent change in fibrinogen from Baseline to Week 13 between each active group and placebo.

### *2.3.3 Exploratory Efficacy Endpoints*

The exploratory efficacy endpoints are listed below:

- Difference in percent change in hs-CRP levels from baseline to End of Treatment (weeks 23 through 24) between each active group and placebo.
- Difference in percent change in serum NT-pro-BNP from Baseline (average of NT-pro-BNP value prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in SAA from Baseline to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in percent change in fibrinogen from Baseline to End of Treatment (Week 24) between each active group and placebo.
- Difference in change in hemoglobin from Baseline (average of the two most recent hemoglobin values prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in serum albumin from Baseline (average of the two most recent serum albumin values prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Proportion of patients achieving hs-CRP response at Week 13 and to End of Treatment (Weeks 23 through 24), defined as hs-CRP <2.0 mg/L in each active group and placebo.
- Difference in change in ST2 from Baseline to Week 13 to End of Treatment (Week 24) between each active group and placebo.
- Difference in percent change in Lp (a) from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in percent change in LDL-C, triglycerides, ApoB, ApoA1, ApoB/ApoA1, and lipid profile by NMR spectroscopy from Baseline to Week 13 and to the End of Treatment (Weeks 23 through 24) between each active group and placebo.

- Difference in change of creatinine based eGFR and cystatin C-based eGFR from Baseline to Week 13 and to the End of Treatment (Weeks 23 through 24) between active group and placebo.
- Difference in change of UACR from Baseline to Week 13 and to the End of Treatment (Weeks 23 through 24) between active group and placebo.
- Difference in change of the total fatigue score (PROMIS Fatigue 13a short form and selected items from the PROMIS fatigue item bank) from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change of the PROMIS interest in sexual activity item from Baseline to Week 13 and to End of Treatment (Week 24) between each active group and placebo.
- Difference in change of the Corvidia ePRO items from Baseline to Week 12 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change of the PGIS index from Baseline to Week 12 and to End of Treatment (Week 24) between each active group and placebo.
- Descriptive analyses of the PGIC index at Weeks 5, 13, and End of Treatment (Week 24) in each active group and placebo.
- Difference in change of the Optum SF-36 v2® Health Survey physical component summary (PCS), mental component summary (MCS) and domain scores from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change of the EQ-5D-5L index from Baseline to Week 12 and to End of Treatment (Week 24) between each active group and placebo.
- Evaluation of the psychometric properties of the PROMIS Fatigue 13a short form and selected items from the PROMIS Fatigue item band and the Corvidia ePRO items in CKD patients – to be described in a PRO psychometric analysis plan (PAP).
- Descriptive analyses by dose and treatment may be conducted on samples stored for analysis of exploratory biomarkers, genomic and transcriptomic analysis.
- Difference in change in TSAT from Baseline to peak level, Week 13, and the End of Treatment (Weeks 23 through 24), between each active group and placebo.
- Difference in change in CHr from Baseline to peak level, Week 13, and to End of Treatment (Week 24) between each active group and placebo.
- Difference in change in TIBC from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in serum ferritin from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in serum iron from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.

- Difference in change in serum hepcidin from Baseline to Weeks 13 and to End of Treatment (Week 24) between each active group and placebo.
- Difference in percent change in serum NT-pro-BNP from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo in patients with baseline NT-pro-BNP > 250 pg/mL.
- Difference in change in hemoglobin from Baseline (average of the two most recent hemoglobin values prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo in patients with baseline hemoglobin < 11 g/dL.
- Difference in change in serum albumin from Baseline (average of the two most recent serum albumin values prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo in patients with baseline albumin < 4.0 g/dL.

### *2.3.4 Calculation/Derivation of Endpoint Variables*

#### *2.3.4.1 PROMIS Fatigue 13a Short Form*

The PROMIS Fatigue 13a short form contains 13 items that assess symptoms and impacts of fatigue. The PROMIS Fatigue 13a short form will be scored using item-level analysis according to the scoring manual. The PROMIS fatigue 13a short form and selected PROMIS items can be found in Appendix B of the protocol. The scoring manual for PROMIS Fatigue can be found at:http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Fatigue_Scoring_Manual.pdf.

Each question from the PROMIS Fatigue 13a short form has five response options (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much). To use the scoring tables from the PROMIS fatigue scoring manual, a summed score is calculated. The lowest possible summed score is 13 and the highest possible summed score is 65. All questions must be answered in order to produce a valid score using the scoring tables. Once calculated, the total summed raw score is translated to a T-score. The T-score rescales the raw score into a standardized score with a mean of 50 and a standard deviation (SD) of 10. The T-score and standard error are then used to calculate the 95% confidence interval around the observed score. The 95% confidence interval is calculated as T-Score ±1.96*SE. The T-score will not be calculated if a response is missing for one or more question from the PROMIS Fatigue 13a short form.

A higher PROMIS T-score represents more fatigue.

#### *2.3.4.2 PROMIS Sexual Function and Satisfaction (PROMIS SexFS)*

The PROMIS SexFS is a customizable self-reported set of measures that includes 79 items in 11 domains. For the purpose of this protocol, only one item will be used: interest in sexual activity. This item has five response options (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much). The PROMIS evaluations will be done on a subject's device at the clinic PRIOR to dosing.

#### *2.3.4.3 Optum SF-36 v2® Health Survey*

The Optum SF-36 v2® Health Survey is perhaps the most widely used health-related quality of life (HRQoL) survey instrument in the world today. It is comprised of 36 items that assess eight health concepts: physical functioning, role limitations caused by physical health problems, role

limitations caused by emotional problems, social functioning, emotional well-being, energy/fatigue, pain, and general health perceptions (see Appendix C of the protocol). Physical and mental health summary scores are also derived from the eight Optum SD-36 scales.The Optum evaluations will be done on a subject's device at the clinic PRIOR to dosing. Optum's PRO CoRE sotware version 1.5 will be used to score SF-36 data.

#### *2.3.4.4 CARES Corvidia ePRO*

The CARES Corvidia ePRO is a new PRO instrument under development (see Appendix D of the protocol). It consists of 19 symptom items, with a 24-hour recall, asking patients to report their worst level of that symptom in the past 24 hours on a numeric rating scale from 0 (no symptom) to 10 (symptom as bad as I can imagine).

Data will be collected at home on subject's personal device or one provided by Corvidia; the questionnaire will be answered daily for 7 days prior to dose (Screening prior to Dose 1 (Visit 1), Dose 2 (Visit 4), Dose 4 (Visit 7), and the last 2 weeks after Dose 6 (Visit 23-24). For each item, weekly scores will be derived as the average of the 7 consecutive daily scores as follows

Baseline = average of Day -7 to Day -1

Visit 4= average of Day 22 to Day 28

Visit 7 = average of Day 78 to Day 84.

If more than 4 daily scores out of the 7 days (>50%) within the weekly period are missing, then the score is set to missing.

A bi-weekly score (the End of Treatment) will also be derived as the average of Day 155 to Day 168. If more than 8 daily scores out of the 14 days (>50%) within the bi-weekly period are missing, then the score is set to missing.

#### *2.3.4.5 PGIC and PGIS scales*

The Patients' Global Impression of Change (PGIC) scale and the Patients' Global Impression of Severity (PGIS) scale represent clinically relevant tools to assess perceived impact of disease management. The PGIC evaluates overall health status as perceived by the patient in a seven-point, single-item scale including 'very much better', 'moderately better', 'a little better', 'no change', 'a little worse', 'moderately worse', and 'very much worse'. The PGIS includes 4 severity scales (none, mild, moderate, and severe) and patients are asked to rate their overall severity of symptoms in the present (e.g., at each visit). The PGIC and PGIS are the most commonly used anchor based method of assessing clinically important change and severity in which the external judgment of meaningful change is made by the patient (see Appendix E of the protocol). The PGIC and PGIS evaluations will be done on subjects' device at the clinic PRIOR to dosing.

#### *2.3.4.6 EQ-5D-5L*

The EurQol-5D-5L (EQ-5D-5L) is a standardized measure of health status developed by the EuroQol Group in order to provide a simple, generic measure of health for clinical and economic appraisal (see Appendix F of the protocol). The EQ-5D-5L is to be done on a subject's device at the clinic PRIOR to dosing. EQ-5D-5L consists of two parts:

- EQ-5D-5L descriptive system

- Visual Analog Scale (VAS) with instant recall of "how you feel today"; scale goes from 0-100, where 100 means the best health you can imagine and 0, the worst one.

The EQ-5D-5L descriptive system includes the basic domains common to each generic health status measure: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The instrument includes one single question per domain. For each question, there are 5 levels of responses, corresponding to increasing levels of impairment (no problems, slight problems, moderate problems, severe problems, and extreme problems), and coded 1 to 5

#### *2.3.5 Safety Endpoints*

- Proportion of patients with AEs, serious AEs (SAEs), severe hematologic AEs, severe non-hematologic AEs, and AEs leading to discontinuation.
- Proportion of subjects with thrombolysis in myocardial infarction (TIMI) major bleeding event.
- Description and frequency of events of special interest by
  - Serious infections.
  - Severe injection-related reactions.
  - Gastrointestinal perforations.
  - Hypersensitivity reaction during study drug administration.
  - Anaphylaxis occurring at any time, even if considered unrelated to the study drug.
  - Neutrophil < 500/mm$^3$ (severe) or neutrophil < 1000/mm$^3$ (severe) with evidence of concurrent infection. These events will be separately summarized by treatment group and dose.
  - Thrombocytopenia (platelet count < 50,000/mm$^3$ [severe]) or platelet count < 75,000/mm$^3$ (moderate) with evidence of concurrent TIMI major bleeding. These events will be separately summarized by treatment group and dose.
  - Malignancies
- Description of additional safety assessments by treatment group and dose: vital signs, ECG, clinical laboratory, and anti-drug antibodies (binding and neutralizing).

# 3 STATISTICAL METHODOLOGY

## 3.1 General Considerations

### *3.1.1 Analysis Day*

Analysis day will be calculated from the date of first dose of study drug. The day of the first dose of study drug will be Day 1, and the day immediately before Day 1 will be Day -1. There will be no Day 0.

### *3.1.2 Analysis Visits*

Scheduled visits will be assigned to analysis visits as recorded in the electronic data capture (EDC) system. If a scheduled visit is not available, unscheduled and early termination visits will be assigned to analysis visits using analysis visit windows based on the actual date the assessment took place. The low analysis day of the analysis window will be calculated as the midpoint between the scheduled assessment and previously scheduled assessment for that parameter. The high analysis day of the analysis window will be calculated as the midpoint between the scheduled assessment and the next scheduled assessment for that parameter.

Where multiple measurements for a particular parameter appear within an analysis window, the scheduled visit will be used. If no scheduled visit appears in the analysis window, the result closest to the target day will be used. If equidistant and both are unscheduled and/or early termination visits, the later result will be used for the summary measure. For applicable efficacy parameters, "End of Treatment" visit will be derived where the average of assessments at Week 23 and Week 24/ET visits from the EDC will be calculated. Otherwise Week 24/ET visit will be used as the derived "End of Treatment" visit.

| Analysis Visit | Target Day | Low Analysis Day | High Analysis Day |
|---|---|---|---|
| Baseline | 1 | | 1 |
| Week 2 | 8 | 2 | 14 |
| Week 4 | 22 | 15 | 28 |
| Week 5 | 35 | 29 | 42 |
| Week 7 | 49 | 43 | 55 |
| Week 9 | 63 | 56 | 70 |
| Week 13 | 85 | 71 | 99 |
| Week 17 | 114 | 100 | 128 |
| Week 21 | 141 | 129 | 153 |
| Week 23 | 157 | 154 | 161 |
| Week 24 | 165 | 162 | 168 |
| End of Treatment | 161 | 154 | 168 |
| Week 28 | 190 | 169 | 211 |
| Week 32 | 224 | 211 | 236 |

The observational period for the study will start from informed consent and end with study completion. Any event occurring after the defined observational period, even if collected on the case report form (CRF), may not be included in the planned statistical analyses. However, all data, including that reported after the defined observational period, will be included in the patient data listings.

### *3.1.3 Definition of Baseline*

For the primary endpoint hs-CRP, baseline will be the average of the hs-CRP value prior to randomization and Day 1. For SAA, baseline will be calculated as the average of the values at Visit Week -1 and Day 1. For laboratory parameters hemoglobin and serum albumin, baseline will be calculated as the average of two most recent lab values prior to randomization and Day 1. For NT-pro-BNP, baseline will be calculated as the average NT-pro-BNP value prior to randomization and Day 1. For the CARES Corvidia ePRO items, baseline will be the average of 7 consecutive measurements (Day -7 to Day -1) prior to first dose initiation. For all other evaluations that are collected at multiple occasions prior to initiation of Study Drug, the latest evaluation will be considered the baseline evaluation for analysis, unless otherwise specified. Similarly, end of treatment values will be calculated by taking the average of assessments taken at Week 23 and Week 24.

#### *3.1.4 Summary Statistics*

All study-collected data will be summarized by treatment group for the appropriate analysis population, using descriptive statistics, graphs, and/or raw data listings. Descriptive statistics for continuous variables will include number of patients (n), mean, standard deviation (SD), median, quartiles (Q1 and Q3), minimum (min) and maximum (max) values. Analysis of categorical variables will include frequency and percentage.

#### *3.1.5 Handling of Dropouts and Missing Data*

Missing data will be imputed only in the context of sensitivity analysis as described in section 3.4.1.

### 3.2 Analysis Populations

#### *3.2.1 Intent-to-Treat (ITT) Analysis Population*

The ITT analysis population includes all randomized patients and will be the primary population for analyses of disposition, baseline, and efficacy data.

#### *3.2.2 Per-Protocol (PP) Analysis Population*

The PP analysis population will be the secondary population for analyses of disposition, baseline data, and efficacy data. For PP analysis population, treatment classification will be based on the randomized treatment. PP analysis population is defined as all randomized patients who completed the study and did not incur a significant protocol violation. Significant protocol violations may include but are not limited to:

- Failed to meet eligibility criteria
- Took the wrong study drug (i.e., did not take the randomized study drug)
- Had study drug compliance <80%
- Took a restricted concomitant medication
- Failed to complete the primary efficacy assessment

A list of subjects with significant protocol violations leading to exclusion from PP analysis population will be finalized prior to unblinding the randomized treatment assignments.

#### *3.2.3 Safety Analysis Population*

The Safety analysis population is defined as all randomized patients who received at least one dose of study drug. All safety data will be analyzed using the Safety analysis population. For the Safety analysis population, treatment classification will be based on the actual treatment received.

#### *3.2.4 Pharmacokinetics (PK) Analysis Population*

The PK analysis population is defined as all randomized patients who received at least one dose of study drug and had at least one post-dose PK blood sample. For the PK analysis population, treatment classification will be based on the actual treatment received.

## 3.3 Subject Data and Study Conduct

### *3.3.1 Subject Disposition*

Counts and percentages of subjects who were randomized, discontinued early from the study, and completed the study will be summarized by treatment and in total based on all randomized subjects (ITT analysis population). Reasons for early discontinuation will also be summarized.

### *3.3.2 Protocol Violations*

Protocol deviations are defined in the Protocol Deviation Plan and classified as CSR reportable or non-CSR reportable. All protocol deviations will be listed.

### *3.3.3 Demographic and Baseline Characteristics*

The following demographic and baseline characteristics will be summarized:

- Age (years)
- Sex
- Race
- Ethnicity
- Height (cm)
- Weight (kg)
- Body mass index (BMI) ($kg/m^2$)
- Baseline Hemoglobin (>=11 or <11 g/dL)
- CKD Stage (Stage 3, Stage 4 or Stage 5)

Demographic and baseline characteristics will be summarized with descriptive statistics or counts and percentages of subjects as appropriate by treatment and in total for the ITT, per-protocol, and safety analysis populations.

### *3.3.4 Medical History*

Medical history will be coded to system organ class and preferred term using the Medical Dictionary for Regulatory Activities (MedDRA) Version 21.0. Counts and percentages of subjects with medical history by system organ class and preferred term will be summarized by treatment and in total based on the ITT analysis populations.

### *3.3.5 Prior and Concomitant Medications*

Prior and concomitant medications will be coded to anatomical therapeutic chemical (ATC) class and preferred term using the World Health Organization (WHO) Drug Dictionary. For summary purposes, medications will be considered prior medications if they were given before the first dose of study drug and concomitant medications if they were taken before and were continuing after the first dose of study drug or initiated after the first dose of study drug.

If a medication has incomplete start or stop dates, dates will be imputed to determine whether a medication should be considered prior or concomitant. If a medication start date is incomplete, the first day of the month will be imputed for missing day and January will be imputed for missing month. If a medication stop date is incomplete, the last day of the month will be imputed for missing day and December will be imputed for missing month. Incomplete start and stop dates will be listed as collected without imputation.

Counts and percentages of subjects taking prior and concomitant medications by Anatomical therapeutic class (ATC) and preferred term will be summarized by treatment and in total based on the Safety analysis population. Patients will be counted only once by medication class or name.

### *3.3.6 Study Drug Exposure*

Weeks of exposure to study drug will be calculated as: (date of last dose of study drug – date of first dose of study drug + 1)/7. Note that the exposure calculation is intended to describe the length of time a subject was exposed to study drug and therefore does not take study drug interruptions into account. Weeks of exposure to study drug will be summarized by treatment based on the Safety analysis population with descriptive statistics.

Study drug is planned to be administered every 28 days for a total of 6 treatments on Weeks 1, 5, 9, 13, 17, and 21.

Counts and percentages of number of injections received (1 to 6) along with summary statistics will be summarized by treatment based on the Safety analysis population.

## 3.4 Efficacy Assessment

For all efficacy variables, the number of patients (n), mean, standard deviation (SD), median, quartiles (Q1 and Q3), minimum (min) and maximum (max) values for the observed value, change from baseline, and percent change from baseline (as applicable) will be reported. The difference (Ziltivekimab - Placebo) in changes or percent changes (as applicable) will be presented.

A mixed model for repeated measures (MMRM) will be used to evaluate change (or percent change) from baseline. The model will include variables for baseline hemoglobin (>= 11 or < 11 g/dL), CKD Stage (3, 4 or 5), treatment group, visit and treatment group-by-visit interaction as categorical fixed effects, baseline value and baseline-by-visit interaction will be included as covariates. The least squares means for each dose, the least squares mean differences from placebo along with the associated 95% confidence intervals (CIs) and p-values will be presented. If the normality assumption is not met, a nonparameteric test will be selected for efficacy analysis.

An unstructured covariance matrix will be used to model the within-subject correlation. The Kenward-Roger approximation will be used to adjust the denominator degrees of freedom. The analysis will be performed based on all observed post-baseline scores without any imputation of missing data. In the case when the MMRM fails to converge using an unstructured covariance matrix in any stage, a less stringent covariance matrix (e.g., autoregressive 1) will be used

Sensitivity analyses will be conducted to explore the robustness of the result for primary efficacy endpoint based on MMRM. Efficacy data will be summarized by randomized treatment group based on the ITT and PP analysis populations.

The differences in proportions and 95% CIs will be based on the normal approximation to binomials.

Missing data will be imputed only in the context of sensitivity analysis as described in section 3.4.1.

The pharmacodynamic efficacy of three Ziltivekimab doses compared to placebo will be assessed by evaluating the following:

### *3.4.1 Primary Efficacy Endpoint*

Primary Analysis

The primary analysis variable is the difference in percent change in hs-CRP levels from baseline (average of the hs-CRP value prior to randomization and Day 1) to the Week 13 between each active group and placebo.

The primary efficacy analysis will be carried out using a mixed model for repeated measures (MMRM) to evaluate percent change in hs-CRP levels from baseline in the ITT populations. The MMRM model will include percent change in hs-CRP from baseline as the dependent or response variable, variables for baseline hemoglobin (>= 11 or < 11 g/dL), CKD Stage (3, 4 or 5), treatment group, visit and treatment group-by-visit interaction as categorical fixed effects, baseline value and baseline-by-visit interaction will be included as covariates. Restricted Maximum Likelihood (REML) method will be used.

The sample SAS code can be found below:

```
*************************************************************************
*Note: PCHG = Percent change in hs-CRP from baseline
*      HGB = Baseline Hemoglobin category (>=11 or <11g/dL)
*      CKD = CKD Stage (3, or 4/5)
*      TRT = Treatment group
*      VISIT = Analysis visit
*      TRT*VISIT = Treatment group by visit interaction
*      BASE = Baseline value
*      BASE*VISIT = Baseline by visit interaction
*      USUBJID = Unique subject identifier
*************************************************************************;
proc mixed data=efficacy method=reml covtest;
   class HGB CKD TRT VISIT;
   model PCHG = HGB CKD TRT VISIT TRT*VISIT BASE BASE*VISIT/
                                                DDFM=KENWARDROGER;
   repeated / type=un subject=USUBJID;
   ODS OUTPUT LSMeans=lsmeantable Diffs=LSDiffs;
run;
```

The treatment difference in terms of mean percent change from baseline will be estimated and tested in this MMRM model. The least squares means for each dose, the least squares mean differences from placebo along with the 95% confidence intervals (CIs) and p-values will be presented from the above results.

Normality assessment for the primary efficacy variable will be performed using the Shapiro-Wilk test. Normality test will be performed on the model residuals. If substantial deviation from assumption of normality is encountered (Shapiro-Wilk p-value < 0.01), a nonparametric analysis will be performed. Median and quartiles will be presented for hs-CRP values at baseline, Week 13, and percent change in hs-CRP from baseline to Week 13. The Hodges-Lehmann estimator of the location shift between each treatment arm and placebo will be presented. The asymptotic 95% confidence limits for the location shift and p-value from the Wilcoxon test will also be presented. The nonparametric analysis will account for the covariates baseline hemoglobin (>= 11 or < 11 g/dL) and CKD Stage (3, 4 or 5) by aligning responses within each stratum defined by the covariates prior to analysis.

The SAS code for Hodges-Lehmann estimates is listed below:

```
proc npar1way hl alpha=.05 align=strata(hl);
        class TRT;
        strata HGB CKD;
        var PCHG;
        ods output wilcoxontest=WILT hodgeslehmann=HL1;
        run;
```

Sensitivity Analyses

Sensitivity analyses will be performed on the PP analysis population with the same method used to analyze the primary efficacy endpoint. In addition, a sensitivity analysis will be performed on the ITT analysis population using a pattern-mixture model using sequential modeling with multiple imputation to assess the robustness of the primary MMRM results to the possible violation of the missing-at-random (MAR) assumption. The sensitivity analysis will consider dropout reasons while imputing missing values after the discontinuation. Subjects who discontinued due to lack of efficacy or adverse events in the active arms are assumed to have no treatment effect after the discontinuation. These subjects are assumed to copy the profile of placebo arm and missing values are imputed based on the distribution estimated from the placebo group under the missing not at random (MNAR) using copy-reference approach. The rest of missing values in the placebo arm and active arms will be imputed using the observed data in their respective group under the MAR assumption. The multiple imputation model will include factos such as treatment arm, baseline hemoglobin category and CKD stage in addition to the data outcomes at each visit.

The following steps will be performed:
Step 1 Intermittent (non-monotone) missing data will be imputed using the MCMC option of SAS PROC MI (using seed=5414).
Step 2 For monotone missing data, the values for each pattern will be imputed via the chained equation method with range restriction, using SAS PROC MI option MONOTONE REG ().

Under step 2, missing values for subjects who discontinued due to lack of efficacy or adverse events will be imputed under MNAR assumption using copy-reference approach, and missing values for subjects who discontinued due to reasons other than lack of efficacy or adverse events will be imputed under the MAR assumption.

These two steps will be carried out sequentially to construct 100 hypothetical complete data sets.

An ANCOVA model (not repeated measures, and with no random effects) with the following covariates will be performed for each imputation using the SAS Mixed procedure:

- HGB = Baseline Hemoglobin category (>=11 or <11g/dL)
- CKD = CKD Stage (3, or 4/5)
- TRT = Treatment group
- BASE = Baseline hemoglobin value

The SAS MIAnalyze procedure will be used to combine the results of these analyses for the imputations. The overall least square means, standard errors, 95% confidence intervals, and p-values will be reported.

Reasons for subject discontinuation that are identified as lack of efficacy will be discussed and finalized prior to database lock.

### *3.4.2 Secondary Efficacy Endpoints*

The secondary analysis variables will be summarized and analyzed in the same manner as the primary analysis variable (Section 3.4.1). Each MMRM model will use change or percent change from baseline variable as applicable as the dependent or response variable.

The following secondary variables will be analyzed:

- Difference in percent change in SAA from Baseline to Week 13 between each active group and placebo.
- Difference in percent change in fibrinogen from baseline to Week 13 between each active group and placebo.

### *3.4.3 Exploratory Efficacy Endpoints*

The following exploratory analysis variables will be summarized in the same manner as the primary analysis of the primary efficacy endpoint for the continuous efficacy endpoints.

For binary endpoints (proportions), differences in proportions and 95% CIs will be based on the normal approximation to binomials.

The sample SAS code for normal approximation for binary endpoints can be found below:

```
*******************************************************************************;
*Note: ACHIEVE = status of patients achieving endpoint (1 if Yes, 2 if No)
*       TRT = Treatment Group
*******************************************************************************;
proc freq data=endpoint;
      by TRT;
      tables ACHIEVE / binomial (Wald exact);
      ods output BinomialCLs=CL BinomialProp=Proportion;
run;
```

The following exploratory endpoints will be analyzed:

- Difference in percent change in hs-CRP levels from baseline to End of Treatment (weeks 23 through 24) between each active group and placebo.
- Difference in percent change in serum NT-pro-BNP from Baseline (average of NT-pro-BNP value prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in SAA from Baseline to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in percent change in fibrinogen from Baseline to End of Treatment (Week 24) between each active group and placebo.
- Difference in change in hemoglobin from Baseline (average of the two most recent hemoglobin values prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.

- Difference in change in serum albumin from Baseline (average of the two most recent serum albumin values prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Proportion of patients achieving hs-CRP response at Week 13 and to End of Treatment (Weeks 23 through 24), defined as hs-CRP <2.0 mg/L in each active group and placebo.
- Difference in change in ST2 from Baseline to Week 13 to End of Treatment (Week 24) between each active group and placebo.
- Difference in percent change in Lp (a) from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in percent change in LDL-C, triglycerides, ApoB, ApoA1, ApoB/ApoA1, and lipid profile by NMR spectroscopy from Baseline to Week 13 and to the End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change of creatinine based eGFR and cystatin C-based eGFR from Baseline to Week 13 and to the End of Treatment (Weeks 23 through 24) between active group and placebo.
- Difference in change of UACR from Baseline to Week 13 and to the End of Treatment (Weeks 23 through 24) between active group and placebo.
- Difference in change of the total fatigue score (PROMIS Fatigue 13a short form and selected items from the PROMIS fatigue item bank) from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change of the PROMIS interest in sexual activity item from Baseline to Week 13 and to End of Treatment (Week 24) between each active group and placebo.
- Difference in change of the Corvidia ePRO items from Baseline to Week 12 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change of the PGIS index from Baseline to Week 12 and to End of Treatment (Week 24) between each active group and placebo.
- Descriptive analyses of the PGIC index at Weeks 5, 13, and End of Treatment (Week 24) in each active group and placebo.
- Difference in change of the Optum SF-36 v2® Health Survey physical component summary (PCS), mental component summary (MCS) and domain scores from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change of the EQ-5D-5L index from Baseline to Week 12 and to End of Treatment (Week 24) between each active group and placebo.
- Evaluation of the psychometric properties of the PROMIS Fatigue 13a short form and selected items from the PROMIS Fatigue item band and the Corvidia ePRO items in CKD patients – to be described in a PRO psychometric analysis plan.
- Descriptive analyses by dose and treatment may be conducted on samples stored for analysis of exploratory biomarkers, genomic and transcriptomic analysis.
- Difference in change in TSAT from Baseline to peak level, Week 13, and the End of Treatment (Weeks 23 through 24), between each active group and placebo.

- Difference in change in CHr from Baseline to peak level, Week 13, and to End of Treatment (Week 24) between each active group and placebo.
- Difference in change in TIBC from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in serum ferritin from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in serum iron from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo.
- Difference in change in serum hepcidin from Baseline to Weeks 13 and to End of Treatment (Week 24) between each active group and placebo.
- Difference in percent change in serum NT-pro-BNP from Baseline to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo in patients with baseline NT-pro-BNP > 250 pg/mL.
- Difference in change in hemoglobin from Baseline (average of the two most recent hemoglobin values prior to randomization and Day 1) to Week 13 and to End of Treatment (Weeks 23 through 24) between each active group and placebo in patients with baseline hemoglobin < 11 g/dL.
- Difference in change in serum albumin from Baseline (average of the two most recent serum albumin values prior to randomization and Day 1) to Week 13 and tos End of Treatment (Weeks 23 through 24) between each active group and placebo in patients with baseline albumin < 4.0 g/dL.

For the above efficacy variables, the observed values, changes from baseline, and percent changes from baseline (as applicable) will be summarized by treatment group using descriptive statistics using ITT.

For PROMIS fatigue 13a short form, the observed value and change from baseline in total fatigue score will be summarized by treatment at each visit.

For PROMIS SexFS, the observed value and the change from baseline will be summarized by treatment at each visit..

The PCS, MCS, and domain scores from the Optum SF-36 v2® Health Survey will only be analyzed descriptively. The observed value and change from baseline in total Optum SF-36 v2® Health Survey score will be summarized by treatment at each visit.

Psychometric analyses will be conducted using the trial data to assess the psychometric properties of the CARES Corvidia ePRO instrument. The structural validity of CARES Corvidia ePRO instrument will be evaluated by means of exploratory factor analysis. A total summary scale or set of scales will be proposed in light of the factor loadings as well as clinical and conceptual considerations. Additional details will be described in the PAP. The score(s) developed as psychometrically sound will be used to evaluate efficacy.

Descriptive analyses of the EQ-5D-5L results will be presented. Number and percentage of patients with each of five levels (no problems, slight problems, moderate problems, severe problems, and extreme problems) of response in each domain will be presented by treatment and visit. The observed value and change from baseline on the EQ-5D-5L VAS score will be summarized by treatment at each analysis visit.

Additional exploratory analyses including descriptive summaries and inferential statistics for the following will be described in a separate PRO Statistical Analysis Plan (PRO SAP): CARES Corvidia ePRO, SF-36 v2 Health Survey and derived scores, EQ-5D-5L, PROMIS 13a and Fatigue Items, PROMIS interest in sexual activity, PGIS, and PGIC.

#### *3.4.4 Multiple Comparisons Procedures*

The overall familywise error rate (FWER) will be controlled at $\alpha = 0.05$ using the sequential testing procedure for the primary efficacy endpoint comparing between the three dose levels of Ziltivekimab vs. placebo. The high dose (30 mg) will be compared with placebo at the beginning. If significant then the middle dose (15 mg) will be compared with placebo. If the middle dose is significant, then the low dose (7.5 mg) will be compared with placebo. The procedure stops whenever a nonsignificant test occurs.

#### *3.4.5 Data Plots*

Graphical data display for the efficacy parameters will be employed to present observed data and changes from baseline. The Ziltivekimab 7.5 mg, Ziltivekimab 15 mg, and Ziltivekimab 30 mg active dose groups as well the placebo group will be superimposed for visual comparison.

## 3.5 Pharmacokinetic Analysis

Pharmacokinetic analysis will be described in a separate pharmacokinetic analysis plan.

## 3.6 Pharmacokinetic/Pharmacodynamic Assessment

PK-PD analysis will be described in a separate PK-PD analysis plan.

## 3.7 Safety Assessment

Safety data will be summarized by actual treatment received (and in total for selected analyses) based on the Safety analysis population.

#### *3.7.1 Adverse Events (AEs)*

AEs will be captured from the date of informed consent through study completion. All AEs will be coded to system organ class and preferred term using MedDRA version 21.0.

Treatment-emergent adverse events (TEAEs) are defined as AEs that initiated or worsened on or after the date of first dose of study drug up to the end of safety-follow-up. For AEs occurring on the first dosing day, if the start time cannot be ascertained, the event will be counted as treatment-emergent. These events will be identified in the data by coded terms.

Adverse events of special interest (Section 2.3.4) will be summarized by dose group and treatment assignment.

An overview of AEs will be provided including counts and percentages of subjects with the following:

- Any TEAEs (overall and by maximum severity)
- Any study drug related TEAEs (overall and by maximum severity)
- Any TEAEs of special interest (overall and by maximum severity)
- Any serious AEs (SAEs)

- Any treatment-emergent serious AEs (TESAEs)
- Any TEAEs leading to discontinuation of study drug
- Any TEAEs leading to discontinuation of study
- Any AEs leading to death
- Any severe hematologic AEs
- Any severe non-hematologic AEs

The number and percentage of patients reporting TEAEs and SAEs for each preferred term will be tabulated by system organ class, by system organ class and severity, and by system organ class and relationship to Study Drug. If more than one event occurred with the same preferred term for the same patient, the patient will be counted only once for that preferred term using the most severe or related occurrence for the summary by severity, or relationship to study drug, respectively.

Listings will be presented specifically for SAEs and TEAEs leading to discontinuation of study drug.

### *3.7.2 Clinical Laboratory Tests*

Blood and urine samples for clinical laboratory tests will be collected and processed by a central laboratory.

Values and changes from baseline for lab parameters (excluding efficacy lab parameters) will be presented at each scheduled visit and baseline by laboratory test. The incidence of abnormalities (as defined by normal ranges) prior to the first dose of study drug and after the first dose of study drug will be summarized with counts and percentages of subjects.

All laboratory measurements will be listed.

### *3.7.3 Vital Signs*

Vital signs and change from baseline in vital signs will be summarized descriptively at each visit by treatment group. The number and percentage of patients with exceeding pre-defined absolute and relative threshold values will be summarized. These threshold values are presented in Table 2.

When calculating the percentages for the criterion related to a threshold, the numerator will be the number of subjects with normal baseline with respect to the specific criterion and at least 1 post-baseline outlier value within the analysis period; and the denominator will be the number of subjects with a baseline and at least 1 post-baseline assessment within the analysis period.

Table 2 Pre-defined Threshold Value for Vital Signs

| Parameter | Criteria |
|---|---|
| Systolic Blood Pressure (SBP) | >25 mmHg increased or decreased from baseline |
| SBP | >160 mmHg |
| SBP | <90 mmHg |
| Heart Rate (HR) | >100 beats per minute |
| HR | <50 beats per minute |
| Respiration Rate | >24 breaths per minute |

| | |
|---|---|
| BMI | >10% increased from baseline |
| BMI | >10% decreased from baseline |

All vital signs measurements will be listed.

### *3.7.4 Electrocardiograms*

ECG interpretation (normal vs. abnormal) will be summarized using frequency and percentage at each visit by treatment group. ECG intervals (PR, QT, HR, and QTcF) will be summarized descriptively at each visit. The number and percentage of patients with exceeding pre-defined absolute and relative threshold values will be summarized. These threshold values are presented in Table 3.

When calculating the percentages for the criterion related to a threshold, the numerator will be the number of subjects with normal baseline with respect to the specific criterion and at least 1 post-baseline outlier value within the analysis period; and the denominator will be the number of subjects with a baseline and at least 1 post-baseline assessment within the analysis period.

Table 3 Pre-defined Threshold Value for ECG

| **Parameter** | **Criteria** |
|---|---|
| PR Interval | >200 msec |
| QTcF | >450 msec |
| QTcF | >480 msec |
| QTcF | >500 msec |
| QTcF | Increase from baseline>30 msec |
| QTcF | Increase from baseline>60 msec |

### *3.7.5 Physical Examinations*

Physical examination clinically significant new or worsening findings will be reported as adverse events and will therefore be summarized as described for adverse events.

### *3.7.6 Antibodies to Ziltivekimab*

The immunogenic potential of Ziltivekimab will be assessed by summarizing the number and percentage of patients who develop detectable anti-drug antibodies (ADA). Anti-drug antibody titers will be summarized descriptively for ADA positive samples and the impact of ADA on PK will be assessed if data allows.

### *3.7.7 Bleeding Events*

Bleeding events will be classified using the Thrombolysis In Myocardial Infarction (TIMI) Bleeding Classification

Table 4 TIMI Bleeding Classification

| Parameter | Criteria |
|---|---|
| Major | Intracranial hemorrhage or a >=5 g/dL decrease in the hemoglobin concentration or a >=15% absolute decrease in the hematocrit |

| | |
|---|---|
| Minor | Observed blood loss: >=3 g/dL decrease in the hemoglobin concentration or >=10% decrease in the hematocrit. No observed blood loss: >=4 g/dL decrease in the hemoglobin concentration or >=12% decrease in the hematocrit |
| Minimal | Any clinically overt sign of hemorrhage (including imaging) that is associated with a <3 g/dL decrease in the hemoglobin concentration or <9% decrease in the hematocrit |
| All TIMI definitions take into account blood transfusions, so that hemoglobin and hematocrit values are adjusted by 1 g/dL or 3%, respectively, for each unit of blood transfused. Therefore, the true change in hemoglobin or hematocrit if there has been an intervening transfusion between two blood measurements is calculated as follows: Δ Hemoglobin = [baseline Hgb – post-transfusion Hgb] + [number of transfused units]; Δ Hematocrit = [baseline Hct – post-transfusion Hct] + [number of transfused units × 3]. | |

Tables will be generated for each of the TIMI Bleeding Classification parameters by treatment and in total. Bleeding events will also be listed.

#### *3.7.8 Exploratory Analyses of Efficacy and Safety Endpoints Based on Baseline Hemoglobin (>=11 or <11 g/dL)*

Exploratory analyses may be performed for the primary and secondary efficacy endpoints with a stratified analysis by baseline hemoglobin (>=11 g/dL or <11g/dL) and CKD stage (3, 4 or 5).

# 4 ANALYSIS TIMING

## 4.1 Blinding and Database Lock

All investigators and patients will remain blinded throughout the study. Additionally, members of the study team including operations team, data management, and statistical team will also remain blinded until the final database lock.

Once all subjects complete Visit 7 (Week 13), the database will be cleaned and locked for the purpose of analyzing the efficacy and safety data for primary analysis. An independent, unblinded statistician will perform this analysis and provide results to a team at Corvidia that will use this information to plan future studies. No subject-level data will be provided to Corvidia or any members of the project team so that the treatment assignments of all subjects remains blinded until completion of the entire study. A formal Blinding Plan will be approved for implementation in the study, prior to the Visit 7 (Week 13) lock and statistical analysis.

After the last subject completes the last visit, the final database will be cleaned, locked and analysis will be performed. At this time, members of the study team will become unblinded to the subject treatment assignments.

Prior to the each database lock, the blinded team may perform a blinded data review using tables, figures, and listings.

## 4.2 Interim Analysis

Interim analysis will not be performed for this study,

## 4.3 Pre-Final Analysis

After the final database is locked and exclusions from analysis populations have been finalized, the randomized treatment assignments will be unblinded and the pre-final analysis will be generated. Pre-final TFLs will be provided approximately 4 weeks after final database lock.

## 4.4 Final Analysis

After all comments on the pre-final analysis have been resolved and the study database is declared final, the final analysis will be generated. Final TFLs will be provided approximately 1 week after the study database is declared final. If there were no changes to the pre-final analysis or the study database, the pre-final TFLs may be considered final. In addition to TFLs, Study Data Tabulation Model (SDTM) data and Analysis Data Model (ADaM) data along with associated files will be provided. Associated files may include: annotated CRFs, SDTM specifications, SDTM programs, ADaM specifications, ADaM programs, TFL programs, and Clinical Data Interchange Standards Consortium (CDISC) Define packages for both SDTM and ADaM data.

# 5 CHANGES FROM PROTOCOL-SPECIFIED STATISTICAL ANALYSES

The schedule of procedures defined in the protocol is listed as Table 1 in this SAP. The schedule of precedures states that cardiac risk markers (SAA, fibrinogen, and ST2) will be assessed at Visit Weeks 1, 13, and 23/ET resepectively. The SAP clarifies that SAA will be analyzed for percentage change from baseline where the baseline for SAA will be evaluated as the average of values at Visit Week -1 and Day 1, and the End of Treatment will be evaluated as the average of values at Visit Week 23 and Week 24/ET.

Due to the impact of the Coronavirus Disease 2019 (COVID19) pandemic on the RESCUE clinical trial, an immediate discontinuation of dosing was announced on March 18, 2020 to ensure the safety of trial patients. The announcement required the early termination visit to be scheduled as early as possible for ongoing patients followed by their entry into the safety follow-up period.

At the time of discontinuation of the study, approximately 140 patients had completed the 24-week treatment period and more than 120 patients were non-completers of the treatment period. As a result, the methods for primary, secondary, or exploratory endpoints that required analysis of change or percent change from baseline to end of treatment were reconsidered as follows:

- For patients who were early terminated from the RESCUE trial due to COVID19 and had completed Week 13 efficacy assessments, the End of Treatment visit will be derived as the average of their last 2 visits after Week 13 including Week 13. The End of Treatment value for such patients will be defined as the average of assessments at Week 13 and Week 17 or Week 17 and Week 21, or Week 21 and Week 23 depending upon the last two completed visits for each patient. The rationale behind taking the average of the last 2 visits as End of Treatment visit is that Ziltivekimab remains effective in the same manner after three months of dosing which is equivalent to the Week 13 endpoint.

MMRM analysis performed for baseline to End of Treatment is based on (1) the assumption of missing at random with small amount of missing data and (2) End of Treatment (the average of Week 23 and Week 24) with the same time window for all patients. Given (1) the missing data

for a large number of early terminated patients due to COVID19 and (2) the time window for End of Treatment visit (the average of the last 2 visits including and after Week 13) varying from patient to patient, MMRM results may not be appropriate to analyze change or percentage change from baseline to End of Treatment. Therefore, analysis for change or percentage change from baseline to End of Treatment will be performed using an analysis of covariance model (ANCOVA), where the percent change or change from baseline to End of Treatment described above will be used as the outcome variable. The ANCOVA model will also include baseline hemoglobin (>=11 or <11 mg/dL), CKD Stage (3, 4 or 5), and treatment group as factors; baseline hs-CRP and the baseline value (for efficacy parameters other than hs-CRP) as covariates.

The sample SAS code for ANCOVA analysis is provided below:

```
*********************************************************************
*Note: PCHG = Percent change from baseline for efficacy parameter
*      HGB = Baseline Hemoglobin category (>=11 or <11g/dL)
*      CKD = CKD Stage (3, or 4/5)
*      TRT = Treatment group
*      CRP = Baseline hs-CRP value
*      BASE = Baseline value for efficacy parameter
*********************************************************************;

proc mixed data=efficacy ;
   class  HGB CKD TRT;
   model pchg=HGB CKD TRT BASE CRP;
   lsmeans TRT/  pdiff=control('0') cl ;
   ods output  diffs=Diff;
run;
```

Analysis of change or percent change from baseline to Week 13 for the primary, secondary, and exploratory endpoints will be carried out using MMRM analysis as originally planned and will include scheduled post-baseline visits up to and including Week 13. Details of MMRM analysis can be found in section 3.4.1 of this SAP.

# 6 PROGRAMMING SPECIFICATIONS

Analyses will be performed using SAS® version 9.3 or higher. All available data will be presented in subject data listings which will be sorted by subject and visit date as applicable. Detailed Programming Specifications will be provided in a separate document.

## APPENDIX A: REFERENCES

Babitt JL, Lin HY. Mechanisms of Anemia in CKD. J Am Soc Nephrol. 2012 Sep 28; 23(10): 1631–1634.

Bretz F, Maurer W, Hommel G. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. Stat Med. 2011 Jun 15;30(13):1489-501.

Go AS, Chertow GM, Fan D, McCulloch CE, Hsu CY. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. N Engl J Med. 2004 Sep 23;351(13):1296-305.

Hays RD, Morales LS. The RAND-36 measure of health-related quality of life. Ann Med. 2001 Jul;33(5):350-7.

Li L, SAS® v9.4 MNAR statement for multiple imputations for missing not at random in longitudinal clinical trials. PharmaSUG2019 – Paper ST-103, 2019.

Rao SV, O'Grady K, Pieper KS, Granger CB, Newby KL, Mahaffey KW. A Comparison of the Clinical Impact of Bleeding Measured by Two Different Classifications Among Patients With Acute Coronary Syndromes. J Am Coll Cardiol 2006;47:809 –16.

Ratitch B, and O'Kelly M. Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures. PharmaSUG2011 – Paper SP04, 2011.

Toblli, Jorge Eduardo, et al. "Intravenous iron reduces NT-pro-brain natriuretic peptide in anemic patients with chronic heart failure and renal insufficiency." *Journal of the American College of Cardiology* 50.17 (2007): 1657-1665.