STATISTICAL ANALYSIS PLAN

Final Version 1.0
17 July 2009

A Randomized Multi-Center, Double-Blind, Placebo-Controlled Study of a New Modified-Release Tablet Formulation of Prednisone (Lodotra®) in Patients with Rheumatoid Arthritis

The CAPRA-2 Study

Protocol Number: NP01-007
(17 January 2008
Amendment 1: 04 August 2008)

SPONSOR
Nitec Pharma AG
Kägenstrasse 17
4153 Reinach, Switzerland
www.nitecpharma.com

Statistical Analysis Plan Signature Page

Final Version 1.0, dated 17 July 2009

Sponsor: Nitec Pharma AG

Protocol: NP01-007

Study Title: A Randomized Multi-Center, Double-Blind, Placebo-Controlled Study of a New Modified-Release Tablet Formulation of Prednisone (Lodotra®) in Patients with Rheumatoid Arthritis

20 July 13:35 BST
Date

Author
Sandrine Cayez
Senior Statistician I

20 July 2009
Date

Reviewer
Ulrike Römer, PhD
Director Clinical Research

20 July 2009
Date

Approver
Stephan Witte, PhD
Chief Medical Officer

TABLE OF CONTENTS

LIST OF ABBREVATIONS AND DEFINITION OF TERMS

ACR	American College of Rheumatology
AE	Adverse Event
ANCOVA	Analysis of covariance
CRF	Case Report Form
CRP	C-reactive protein
DAS	Disease activity score
ESR	Erythrocyte Sedimentation Rate
EULAR	European League Against Rheumatism
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue
FACIT-G	Functional Assessment of Chronic Illness Therapy-General
HAQ-DI	Functional Disability Index of the Health Assessment Questionnaire
ICH	International Conference on Harmonisation
IL-6	Interleukin-6
LOCF	Last observation carried forward
MedDRA	Medical Dictionary for Regulatory Activities
mITT	modified Intention-to-treat
MR	Modified-Release
PP	Per-Protocol
PT	Preferred Term
RA	Rheumatoid arthritis
SAE	Serious adverse event
SAP	Statistical Analysis Plan
SD	Standard Deviation
SF-36	Short Form 36
SOC	System Organ Class
TEAE	Treatment Emergent Adverse Event
TNFα	Tumor Necrosis Factor α
VAS	Visual analogue scale

1 INTRODUCTION

The purpose of this document is to describe the statistical methods, data derivations and data summaries to be employed in this Phase III, Randomized Multi-Center, Double-Blind, Placebo-Controlled Study of a New Modified-Release Tablet Formulation of Prednisone (Lodotra®) in Patients with Rheumatoid Arthritis (RA).

The preparation of this statistical analysis plan (SAP) has been based on International Conference on Harmonisation (ICH) E3 and E9 Guidelines (1, 2) and in reference to Protocol NP01-007 (17 January 2008) and its amendment (04 August 2008)

2 STUDY OBJECTIVES

2.1 PRIMARY OBJECTIVES

The primary objective of this study is

- to evaluate if 12 weeks of treatment with 5 mg modified-release (MR) prednisone (Lodotra®) administered in the evening is superior to placebo in terms of the American College of Rheumatology (ACR)20 responder rate

2.2 SECONDARY OBJECTIVES

The key secondary objective of this study is to evaluate if 12 weeks of treatment with 5 mg MR prednisone (Lodotra®) administered in the evening is superior to placebo in terms of the relative reduction of morning stiffness from baseline

Additional secondary objectives of this study are to determine whether treatment with 5 mg Lodotra® administered in the evening is superior to placebo in terms of:

a) <u>Efficacy</u>:

- Time to Response (ACR20 criteria)
- ACR50
- ACR70
- Disease Activity Score (DAS)28 score at each visit
- European League Against Rheumatism (EULAR) response criteria
- Morning stiffness at each visit
    - ✓ Relative (to baseline) reduction of duration of morning stiffness
    - ✓ Absolute reduction of duration of morning stiffness
    - ✓ Severity of morning stiffness
    - ✓ Reoccurrence of stiffness during day
- Individual ACR20 and DAS28 criteria (ACR Core test)
    - ✓ Tender joint count
    - ✓ Swollen joint count
    - ✓ Patient’s assessment of pain - assessed using 100mm visual analogue scale (VAS)

    - ✓ Patient's global assessment of disease activity - assessed using 100mm VAS
    - ✓ Physician's global assessment of disease activity - assessed using 100mm VAS
    - ✓ Functional disability index of the Health Assessment Questionnaire (HAQ-DI)
    - ✓ Erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) as acute-phase reactants
- Requirements for additional analgesics
- Occurrence of pain in morning and evening
- Inflammatory cytokines at each visit (Interleukin-6 (IL-6) and tumor necrosis factor α (TNFα))
- Quality of life
    - ✓ Health Assessment Questionnaire (HAQ-DI, as part of ACR20)
    - ✓ Short Form 36 (Quality of Life; SF-36)
    - ✓ Fatigue (Functional Assessment of Chronic Illness Therapy-Fatigue [FACIT-F] – 13 items questionnaire)

b) Safety

- Adverse Events (AEs)
- Standard laboratory (hematology, biochemistry and urinalysis) parameters
- Physical examination findings including assessment of vital signs (blood pressure, heart rate, body weight)

3 STUDY DESIGN

This is a randomized, multi-centre, double-blind, parallel-group, placebo-controlled 13 week study comparing evening administration of 5 mg Lodotra® to placebo in patients with RA. It was planned to randomize a total of 294 patients in 70 to 80 centers in North America and Europe.

Patients must meet all inclusion and exclusion criteria at Visit 0 before receiving screening medication, and must also meet all randomization criteria at Visit 1 before receiving Lodotra® or placebo (see flow chart on page 11 and 12 of the protocol). Patients not treated with a glucocorticoid for the 6 weeks prior to the screening visit (at visit 0) will be eligible for inclusion. The single-blind screening phase will last for 1 week, and will include daily recording of duration of stiffness in the diaries prior to Visit 1 to calculate a robust baseline value (average of 7 daily values collected on days -7 to -1)

Before randomization, all patients will receive placebo on top of their standard medication for a 1 week baseline period. No medication will be withdrawn during this period, so patients will remain treated at all times during the study.

The double-blind phase of the study starts with randomized allocation of eligible patients to one of the two arms (Lodotra® or placebo) at Visit 1 (baseline; Week 0). Efficacy of

Lodotra® (5 mg daily dose [1 x 5 mg tablet], evening administration) will be derived from the comparison with placebo. Patients will be treated with blinded study medication on a fixed dose for 12 weeks. The double-blind phase will consist of four visits (Visit 1 to Visit 4; Weeks 0, 2, 6 and 12). After the double-blind treatment phase, patients should be switched to 5 mg immediate-release prednisol(lo)ne. Overall duration of the study is planned to be one and a half years.

4 CHANGES FROM PROTOCOL IN STUDY CONDUCT OR STATISTICAL ANALYSIS

- DAS28 will be analyzed using CRP value instead of ESR, as CRP is analyzed by a central laboratory.
- Age class and gender were added to the primary analysis as covariates.
- Nested effect of the pooled sites within geographical region and its interaction with the treatment was added to ACR20 sensitivity analysis.
- Duration of morning stiffness will be analyzed using Hodges Lehmann due to the non-normality of the data.
- DAS28 will be analyzed by an analysis of covariance (ANCOVA) with treatment, nested effect of the pooled sites within region [Pooled Site(Region)] (see section 8.3.1.2) as factor and a term for interaction between the nested effect and treatment; the nested effect of the pooled sites within region factor will be analyzed as a random effect.
- Urine CTX will be analyzed separately from the safety lab data as it is a biomarker of the cartilage or bone degradation.

5 ANALYSIS POPULATIONS

Three analysis populations will be defined for this study as outlined below.

The primary analysis population for efficacy and safety will be the safety population. The primary and key secondary endpoints will also be analyzed using the modified intention-to-treat (mITT) and Per Protocol (PP) population.

ICH E9 guideline suggests that the analysis of the primary endpoint should be analyzed according to the treatment to which patients were actually randomized (modified ITT population). However, in this study approximately 5% of patients were assigned to a treatment that was not consistent with the intended randomization schedule due to blinded errors of study personnel (i.e. distribution of medications in the wrong order). For these reasons, analyzing the primary endpoint according to the patients actually received (safety population) should not be biased and should reflect the true comparative activity of the agents.

To assess the treatment effect using different assumptions from those in the mITT and safety analyses, the primary efficacy (ACR20) and key secondary (relative change in duration of morning stiffness) variables will also be analyzed using the per-protocol (PP) population. All safety analyses will be based on the safety population.

5.1 SAFETY POPULATION

The Safety Population will include all patients who were randomized and received at least one dose of study medication. Patients will be analyzed according to the treatment which they actually received.

5.2 MODIFIED INTENTION-TO-TREAT POPULATION

The modified intention-to-treat (mITT) population will include all patients who were randomized and received at least one dose of study medication. Patients will be analyzed according to the treatment to which they were intended to be randomized to.

According to the study protocol, the treatment kits for patients were packed into "site shipper" boxes that contained one randomization block each. These were then distributed to sites. Each treatment kit is identified with a unique, predefined number (= randomization number). A list of treatment kits sent to each site will be provided by the drug distributor to determine the intended randomization schedule. The investigators then allocate the lowest kit number to the next patient eligible for randomization. If a treatment kits is damaged or out of date, the sequencing will be adjusted accordingly. All patients randomized out of order will be presented in a listing.

5.3 PER PROTOCOL POPULATION

The Per Protocol (PP) population will include all patients who were randomized, treated with study medication and did not have a major protocol deviation. Major protocol deviations leading to exclusion from the PP population will be finalized prior to unblinding during the blind data review meeting. The per-protocol population will be based on the mITT population.

6 EFFICACY ANALYSIS VARIABLES

6.1 Primary Efficacy variable

The primary efficacy variable is the ACR20 responder rate after 12 weeks of double-blind treatment with the study medication. Responders are defined as those whose improvement from baseline to endpoint (12 weeks) fulfils all three of the following criteria:

- ≥20% reduction in the tender joint count (0-28)
- ≥20% reduction in the swollen joint count (0-28)
- ≥20% reduction in 3 of 5 of the following additional measures:
    - Patient assessment of pain (VAS)
    - Patient's global assessment of disease activity (VAS)
    - Physician global assessment of disease activity (VAS)
    - HAQ-DI

- CRP or ESR as acute-phase reactant. CRP will be used. If the CRP result is not available then the ESR result will be used to calculate the ACR20 responder status

6.2 Secondary Efficacy variable

The key secondary efficacy variable is the relative change (%) in the duration of morning stiffness between baseline and endpoint (12 weeks)

Additional secondary efficacy variables are as follows:

- ACR20 response rate at week 2 (visit 2) and week 6 (visit 3)
- ACR50 response rate at week 2 (visit 2), week 6 (visit 3) and week 12 (visit 4), responders are defined as per ACR20 but using 50% reduction instead of 20%
- ACR70 response rate at week 2 (visit 2), week 6 (visit 3) and week 12 (visit 4) , responders are defined as per ACR20 but using 70% reduction instead of 20%
- Time to Response based on ACR20 criteria
- Change from baseline in DAS28 at each visit: DAS28 is a score aggregating data of 28 joints, and is calculated as:

$$\begin{aligned} DAS28 = {} & 0.56 \times \sqrt{(\text{number of tender joints})} \\ & + 0.28 \times \sqrt{(\text{number of swollen joints})} \\ & + 0.36 \times \ln(\text{CRP} + 1) \\ & + 0.014 \times \text{patient's global assessment of disease activity} \\ & + 0.96 \end{aligned}$$

  Where CRP is expressed in mg/L and patient's global assessment of disease activity (VAS) is expressed in mm. If the CRP result is missing, DAS28 will be computed using the following formula.

$$\begin{aligned} DAS28 = {} & 0.56 \times \sqrt{(\text{number of tender joints})} \\ & + 0.28 \times \sqrt{(\text{number of swollen joints})} \\ & + 0.70 \times \ln(\text{CRP}) \\ & + 0.014 \times \text{patient's global assessment of disease activity} \end{aligned}$$

- EULAR response rate at each visit: patients will be classified as patients with good, moderate or no response based on their change in DAS28[3].
- Relative reduction (%) and absolute reduction of morning stiffness between baseline and other study visits
- Relative reduction (%) and absolute reduction from baseline of morning stiffness at each week.
- Change from Baseline in Severity of morning stiffness at each visit: 100 mm VAS
- Change from baseline in terms of reoccurrence of stiffness during day (while performing routine activities). Reoccurrence of Stiffness during the day will be assessed as the percentage of days with reoccurrence of stiffness over the last 7 days prior to each visit (if 4 or more responses are missing the percentage will be set to missing)

- Change from baseline in tender and swollen joint counts at post-baseline study visit: The analysis of tender joint count and swollen joint count is based on a 28 joint assessment. For each patient, only those joints that are evaluable at baseline and endpoint will be included in the statistical analysis of joint counts.
- Change from baseline in patient assessment of the pain intensity at each visit (100 mm VAS)
- Change from baseline in physician's and patient's global assessments of disease activity at each visit (100 mm VAS)
- Change from baseline in HAQ-DI at each visit: The maximum score of all items within each of the 8 categories gives the category score for each patient. The functional disability index is the average of all 8 category scores.
- Change from baseline in inflammatory parameters at each visit: CRP, ESR, TNFα and IL-6
- Change from baseline in the occurrence of pain in morning and evening (100 mm VAS) will be computed as the change in percentage of occurrence of pain in morning/evening over the last 7 days prior to each visit (if 4 or more responses are missing the percentage will be set to missing)
- Change from baseline in use of additional analgesics (no/yes) will be computed as the change from baseline in percentage of days with the event over the last 7 days prior to each visit (if 4 or more response are missing the percentage will be set to missing)
- The use of additional analgesic will also be assessed as the number of days with additional analgesic during the first 12 weeks of double-blind treatment period.
- Change from baseline in each domain of the Short Form 36 questionnaire (Quality of Life; SF-36) and for the mental and physical component scores.
- Change from baseline in the FACIT fatigue questionnaire: this is a subset of the FACIT-F questionnaire comprising 13-items.

6.3 EFFICACY ASSESSMENTS

Efficacy data is based on:

- Individual ACR20, ACR50, ACR70 criteria (as outlined in Section 2.2)
- Individual DAS28 criteria (as outlined in Section 2.2)
- EULAR response criteria
- Laboratory assessments of acute phase reactants (ESR, CRP, IL-6 and TNFα)
- Diary entries relating to morning stiffness, stiffness during the day (while performing routine activities), and analgesics (painkillers)
- SF-36 questionnaire
- FACIT-F questionnaire (13 items fatigue section)

6.3.1 ACR20, ACR50, ACR70

The following table shows how the response for the different cut-off of the ACR criteria will be assessed based on the responses for each parameter of the ACR.

	≥XX% improvement achieved?							
Scenario	**SJC**	**TJC**	**C1**	**C2**	**C3**	**C4**	**C5**	**ACRXX responder**
1	**N**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**N**
2	**Y/N/M**	**N**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**N**
3	**Y**	**Y**	**Y**	**Y**	**Y**	**Y/N/M**	**Y/N/M**	**Y**
4	**Y**	**Y**	**N**	**N**	**N**	**Y/N/M**	**Y/N/M**	**N**
5	**Y**	**Y**	**M**	**M**	**M**	**Y/N/M**	**Y/N/M**	**M**
6	**Y**	**Y**	**Y**	**Y**	**N**	**M**	**M**	**M**
7	**Y**	**Y**	**Y**	**N**	**N**	**M**	**M**	**M**
8	**Y**	**Y**	**Y**	**Y**	**N**	**N**	**M**	**M**
9	**Y**	**M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**M**
10	**M**	**Y**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**M**
11	**M**	**M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**Y/N/M**	**M**

Key: **Y** is ≥XX% improvement achieved; **N** is <XX% improvement not achieved; **M** is Missing; SJC=Swelling; TLC=Tenderness; C1-C5 are the 5 remaining ACR components (Patient assessment of pain, patient's global assessment of status disease, physician's global assessment of disease activity, HAQ-DI and CRP or ESP) in any order.
XX represents the different cut-off: 20, 50 and 70.

6.3.2 Patient's and Physician's global assessment of disease activity

Disease activity is assessed by both patients and physicians using a 100 mm VAS with the endpoints 0=not active at all and 100=extremely active. Patients and physicians will mark points on the scale.

6.3.3 Patient's assessment of pain

Maximum intensity of pain is documented and intensity is assessed by marking the respective value on a 100 mm VAS (with the endpoints 0=no pain at all and 100=very intensive pain). The current state of pain assessed during the visit is used for the ACR20 assessment, pain at morning and evening is also collected using the same method in the patient's diary.

6.3.4 Tender joint count

The following 28 joints (14 left, 14 right) are assessed for tenderness: shoulder, elbow, wrist (radiocarpal, carpal and carpometacarpal are collectively designated wrist), metacarpophalangeal I-V, thumb interphalangeal, proximal interphalangeal II-V, knee. The investigator applies pressure to each joint and then moves it through a full range of motion. The tender joint count represents the number of joints in which pain is reported after either manoeuvre.

6.3.5 Swollen joint count

The same 28 joints as mentioned above in section 6.3.3 are assessed for swelling. The swollen joint count represents the number of joints in which there is synovial fluid and or soft tissue swelling, but not if bony overgrowth is found.

6.3.6 Functional disability index of HAQ-DI

The HAQ-DI includes eight blocks of questions (dressing and grooming, arising, eating, walking, hygiene, reach, grip and common daily activities) covering difficulties when

performing simple daily activities, such as personal hygiene (washing, and dressing or undressing), mobility domestic and outdoors (walking, mounting steps, going shopping, carrying things), as well as intake of food or drink and, the handling of tools used in everyday life.
The answers are to be given by marking tick-boxes at each visit to indicate the degree of difficulty on a 4 point grading system (0=none, 1=some difficulty, 2=great difficulty, 3=not able to perform at all).

The HAQ-DI score is the average across the maximum score in each category. The HAQ-DI requires at least 6 of the categories to be non-missing. Individual questions within a category are not imputed. Therefore the maximum score in each category is based on non-missing questions, and a category score is missing when all questions within a category are missing. The scoring will be adjusted with regards to the use of devices, aids and/or help from a person to perform the task. Details of computation are provided in the technical specification.

The investigator will check the plausibility and completeness of entries, without influencing the patients in their assessments.

6.3.7 EULAR

EULAR criterion is based on DAS28 and the following characterization of the disease status and its change from baseline.

	Visit		
Baseline	**Improvement > 1.2**	**0.6 < Improvement ≤ 1.2**	**Improvement ≤ 0.6**
DAS28 ≤ 3.2	Good response	Moderate response	No response
3.2 < DAS28 ≤ 5.1	Moderate response	Moderate response	No response
DAS28 > 5.1	Moderate response	No response	No response

6.3.8 Laboratory efficacy assessments (CRP, ESR, TNFα and IL-6)

Blood sampling for the assessment of the laboratory efficacy parameters must be done at the same time for all visits.

- ESR is assessed (in mm/h) by measuring the sedimentation rate in the first hour after withdrawal of blood at each visit at local laboratories using routine local standard methods and equipment. This data is used for the assessment of ACR20.
- CRP (mg/L) is analyzed from 1 mL serum by a central laboratory and is used for the determination of ACR20 and DAS28.
- IL-6 (pg/mL) and TNFα: the blood samples for the determination of these parameters are processed and stored according to protocols provided by the central laboratory
- Urine CTX

Investigators will not be notified of the final test results (CRP, IL-6 and TNFα) during the double-blind phase of the study. After database lock and unblinding of the medication, investigators will receive CRP, TNFα and IL-6 data of their study patients.

CRP, IL-6 and TNFα will be blinded by the central lab until database lock and unblinding of the treatment group.

6.3.9 SF-36

The SF-36v2 Quality of Life questionnaire consists of 36 generic health questions. There are 8 health domains of the questionnaire, each of which will be summarized (Physical functioning score (10 items), Role-physical score (4 items), Bodily pain (2 items), General health score (5 items), Vitality score (4 items), Social functioning score (2 items), Role-emotional score (3 items), and Mental health score (5 items)). Additional details of computation are provided in the technical specification.

The answers to each question (recoded as necessary) are summed for each subject at each visit, within each of the 8 domains. If an item is missing, it should be imputed as the mean of the non-missing items in its domain for the purposes of calculating the domain score. Note that this imputation applies only to the calculation of the domain scores; imputation of individual item scores will not be presented. At least 50% of the item scores in a domain must be non-missing to calculate the domain score, otherwise the domain score is set to missing.

The resulting score for each domain (after the imputation described above) is then standardized, to obtain values ranging from 0 to 100, with higher values indicating a better quality of life.

$$\text{Standardised Score} = \left(\frac{\text{Sum - Lowest possible score}}{\text{Possible raw score range}} \right) \times 100$$

In addition, the physical and mental component score will be computed using the US weighting scales. Details are provided in the technical specifications.

6.3.10 Functional assessment of chronic illness therapy-fatigue (FACIT-F)

The Fatigue questionnaire (subset of the FACIT-F questionnaire) is used to assess the impact of patient’s fatigue on their daily activity and function. It is a 13 item questionnaire which is to be completed by the patient on a 5 point grading system (0=not at all, 1=a little bit, 2=somewhat, 3=quite a bit, 4=very much). The overall score is the sum of the average of the subscales. Details of computation are provided in the technical specification.

The investigator will check the plausibility and completeness of entries, without influencing the patients in their assessments.

6.3.11 Diary

Patients were instructed to enter their data twice daily. Reoccurrence of stiffness during the day, occurrence of pain in the morning and evening and use of analgesics will be analyzed as a percentage of days with the event over the last 7 days prior to or on each visit (if 4 or more response are missing the percentage will be set to missing). Other parameters, VAS scales, duration of morning stiffness, will be computed as the mean over the last 7 days prior or on each visit. The diary data will be slotted to match the CRF visit date.

If the patient recorded that no pain was experienced and the corresponding VAS score is missing it will be set to 0 (i.e. no pain).

Duration of morning stiffness will be summarized for each week, weeks being defined from the day of first dose. All assessments within a week will be computed as per the same rules as above.

Parameters to be entered by the patients in the diary are given in the section 7.3.1.9 of the protocol.

7 SAFETY ENDPOINTS

Safety will be assessed by evaluation of the following variables: adverse events, serious adverse events (SAEs), laboratory tests (hematology, urinalysis and serum chemistry), physical examination and vital signs.

Hemoccult/guaiac tests were performed prior to randomization at Visit 0 and prior to the end of treatment at Visit 4.

8 STATISTICAL EVALUATION

8.1 SAMPLE SIZE AND POWER

Superiority of an active treatment versus placebo is defined as an ACR20 response rate on active treatment that is at least 20% higher than that on placebo (e.g. 45% vs. 25%, 50% vs. 30%, or 40% vs. 20%).

The sample size calculation is based on the comparison of two proportions using the χ^2 test and a randomization ratio of 1:2 (placebo: Lodotra®).

Based on a review of selected literature and other similar studies, typical placebo response rates range between 20-30% for ACR20. Assuming an ACR20 response rate of 25% in the placebo group, a total of 294 patients (98 placebo, 196 Lodotra®) are necessary to provide 90% power to detect an ACR20 response rate of 45% in the Lodotra® group at a significance level of α=0.05.

It is estimated that a minimum of 350 patients will have to be enrolled into the study to randomize 294 patients.

Assuming a standard deviation (SD) of 89% for the key secondary efficacy variable (relative change [%] in morning stiffness) based on the SD reported in the previous Lodotra® study, the calculated sample size of 294 patients (98 placebo, 196 Lodotra®) will have 78% power to detect a difference of 30% between placebo and Lodotra® and 89% power to detect a difference of 35%.

8.2 INTERIM ANALYSIS

No interim analysis is planned for this study.

8.3 STATISTICAL METHODS

8.3.1 General Statistical Methodology

8.3.1.1 General Convention

General algorithms, imputations and conventions that will generally apply to program derivations of the data as required to perform the proposed summary tabulations, individual patient data listings, and figures will be detailed in a separate document, which will be signed off prior to unblinding.

For all variables measured during screening or at the randomization visit, the last available value prior to the first intake of study medication will be considered as the baseline value. The respective endpoint value is the first available value measured within 3 days of last intake of study medication or the visit day (whichever occurred first). Data from the diary will not be analyzed if recorded more than 3 days after the last dose of study drug.

Time points in the summaries will be the planned relative times as shown in the CRF.

All efficacy and safety variables will be summarized by treatment groups using descriptive statistics (n, mean, standard deviation [SD], median, minimum, and maximum for continuous data and absolute and relative frequencies for categorical data). Data will be summarized for baseline, endpoint and by visit (if applicable).

Descriptive statistics will be presented to assess the distribution of the baseline variables across treatment groups. No statistical test for differences between treatment groups will be applied.

Percentages will be presented to one decimal place throughout.

All statistical tests will be two-sided and performed at the 5 % significance level. 95% confidence intervals will be provided where appropriate. Data will be summarized by visit.

All dates will be displayed in DDMMMYYYY format. Visits will be referred to as shown in the protocol: “Screening (V0)”, “Baseline (V1)”, “Visit 2”, “Visit 3” and “Visit 4”

All analyses will be carried out using SAS® Version 8.02 or later on Windows 2000 or later.

8.3.1.2 Pooled Sites

As it is expected to have between 70 and 80 sites, and the sensitivity analysis for the primary endpoint is analyzed using a logistic regression with a nested effect of the sites within the region (defined as USA/Canada and Europe) and an interaction between the treatment and the nested effect, in order to estimate the difference in treatment each cell needs to have at least one observation. Therefore sites will be pooled together and country will be pooled into geographic area; when pooling sites, whenever possible sites within a country will be pooled together. Since the randomization is done at a site level, in order for the model to still be convergent after unblinding, it is expected that a minimum of 12 patients in each pooled sites (with at least 6 responders and 6 non-responders) will be sufficient. The pooling will be documented prior to the unblinding of the study.

8.3.2 Handling of Missing and Incomplete Data

8.3.2.1 Primary endpoint

Each patient will be defined as a responder or non-responder at visit 4 according to the ACR20 criteria. If the response is missing at visit 4 the following imputation schemes will be used:

For the safety population, the primary analysis will consist in imputing all missing assessments at visit 4 as non responder. As a sensitivity analysis, the missing assessments will be imputed conditionally to the completion of the study by the patient, i.e. if a patient discontinued prematurely, the ACR20 will be imputed as non-responder, while patients who completed the study but have a missing assessment for ACR20 at visit 4 will not be imputed. Analysis for the observed case only will be presented as a secondary sensitivity analysis. It is assumed that final efficacy assessments will be available and complete for all PP patients.

At any visit, if the CRP result is not available then the ESR result will be used to compute the ACR20 criteria.

8.3.2.2 Secondary endpoints

8.3.2.2.1 Diary data

Assessments of diary data will be based on the last 7 days prior to each visit. If more than 4 assessments during this 7 days period are missing, the assessment will be set to missing. In this case only, a LOCF (last observation carried forward) method will be applied for imputing missing assessments and will consist of taking the last 7 non-missing entries on the diary data prior to that visit. If strictly less than 3 assessments are available using the LOCF, the value will be set to missing.

8.3.2.2.2 CRF data

LOCF methodology will be used to impute missing assessments. If the assessment at visit 2 is missing it will not be imputed. If the visit 3 assessment is missing it will be imputed by the visit 2 assessment. If the visit 4 assessment is missing it will be imputed by the visit 3 assessment Missing subscale or item results will not be imputed, only a final score or response will be imputed if missing.

For ACR20, the worse case (as defined in section 8.3.2.1) will be applied at visit 2 and 3, in addition the worse case conditional to withdrawal will be applied at visit 3 i.e. if a patient withdrew (whatever the reason) before visit 3, he will be considered as a non-responder.

8.3.2.2.3 Time to Response (ACR20)

If a response is observed at any post-baseline visit where the assessment has been made less than 3 days after the last dose then the time to response will be computed from the date of first known as a responder.

If no response (assessment missing or “non-responder”) was observed before withdrawal or last dose (+ 3 days) or visit 4, the patient data will be censored at the date of last dose + 3 days or date of withdrawal or date of visit 4 (whichever occurred first).

8.3.2.2.4 Inflammatory parameters

The LOCF will be using the unscheduled visit result. Baseline will be defined as the last non-missing results prior or on the date of first dose. In the summary table baseline summary will be based on lab result at visit 1 and change from baseline using the baseline result (as defined above) for each visit.

8.3.2.2.5 Last dose

If the last dose is missing, it will not be imputed and all the data will be analyzed.

8.3.2.2.6 Safety laboratory parameters

If a result is reported as < lower limit of quantification, it will be analyzed as half the value of the lower limit of quantification, and the minimum value in the summary table will show the analyzed value (i.e. half the value of the lower limit of quantification).

If a result is reported as > upper limit of quantification, it will be analyzed as the value of the upper limit of quantification, and the maximum value in the summary table will show the analyzed value (i.e. the upper limit of quantification).

8.3.3 Patient Disposition

A complete accounting of patient allocation will be tabulated overall and by treatment group. The patient data will be summarized and presented for each treatment group

- All patients who signed the informed consent form
- Number of randomized patients
- Number and percentage of patients included in each population i.e. Safety, Modified ITT and PP
- Number and percentage of patients in the Modified ITT with protocol violations leading to exclusion from the PP population
- Number and percentage of patients who enrolled, who completed the study, and who prematurely withdrew. The reasons for premature withdrawal will also be summarized. Supportive listings will be provided.

Inclusion and exclusion criteria at Visit 0 before receiving screening medication, and randomization criteria at Visit 1 before receiving Lodotra® or placebo will also be listed.

8.3.4 Demography and Baseline Characteristics

All demographic variables (gender, age, race, ethnic origin and BMI) and baseline characteristic data such as medical and disease history: duration of RA, age at onset of RA, previous and concomitant illness recorded at the screening visit will be summarized by treatment group.

Supportive listings will be provided.

8.3.5 Treatment Compliance and Exposure

Patient compliance to study medication will be calculated for each visit. Assessment will be based on tablets dispensed/returned dates recorded in the compliance page of the CRF. Compliance per visit will be calculated as follows:

$$\frac{\text{Number of tablets dispensed - Number of tablets returned}}{(\text{date of visit - date of previous visit})} \times 100$$

If the visit date occurred after the date of last dose, the date of last dose + 1 will be used in the above formula.

Overall compliance will be calculated as follows:

$$\frac{\text{Number of tablets dispensed - Number of tablets returned}}{(\text{date of last dose - date of first dose})+1}\times 100$$

Exposure per visit will be calculated as follows:

$$(\text{date of visit - date of previous visit})$$

If the visit date occurred after the date of last dose the date of last dose + 1 will be used in the above formula.

Overall Exposure to study treatment will be calculated as follows:

$$(\text{date of last dose - date of first dose})+1$$

Compliance and exposure will be summarized overall and by visit for each treatment group for the Safety population. Supportive listings will also be provided.

If the number of tablets returned is missing (or none have been collected by the investigator) it will be assumed for the compliance that all the tablets were taken by the patient.

8.3.6 Efficacy Analysis

The primary efficacy analysis will be based on the safety population. This primary analysis and key secondary variable (reduction in duration of morning stiffness) will also be repeated for the mITT and the PP populations. All efficacy variables will be summarized by treatment groups using descriptive statistics (mean, standard deviation [SD], median, minimum, and maximum for continuous data and absolute and relative frequencies for categorical data). Data will be summarized for baseline, endpoint and by visit.

8.3.6.1 Primary Efficacy Analysis

The primary efficacy analysis of the ACR20 responder status at the visit 4 (week 12) endpoint will be tested using a logistic regression model with treatment and geographic area (see section 8.3.1.2), age category (less or equal than median age or above the median) and gender as factors with a two-sided significance level of α=0.05 for the safety population. All imputation schemes will be analyzed.

In order to evaluate the consistency of results across the different study sites, the logistic regression analysis will be repeated:

- First, with treatment as factor and pooled sites as a nested effect of geographic region as factors and with a treatment-by-the nested effect interaction term included in the model.

- Secondly, with treatment and geographic region as factors and with a treatment-by-region interaction term included in the model (where region is defined as USA/Canada versus Europe).

For the evaluation of the robustness of results the primary efficacy analysis will be repeated for the mITT and PP populations. Odds ratios for the difference between treatments and the associated 95% confidence interval, as well as the difference in proportion and its associated 95% confidence interval will be presented for each population. The difference in proportion will not be adjusted.

A SAS code similar to the one below will be used for the primary endpoint analysis.

```
proc genmod data=dataset;
        class treat region agegrp gender;
        model responder = treat region agegrp gender / link=linkc dist=binomial type3;
        lsmeans treat / diff CL;
run;
```

The interaction between treatment and pooled sites nested within region will be tested as follows:

```
proc genmod data=dataset;
        class treat region pooledsite;
        model responder = treat pooledsite(region) treat* pooledsite(region) /
                          link=linkc dist=binomial type3;
        lsmeans treat / diff CL;
run;
```

The interaction between treatment and region will be tested as follows:

```
proc genmod data=dataset;
        class treat region;
        model responder = region treat* region /
                          link=linkc dist=binomial type3;
        lsmeans treat / diff CL;
run;
```

where: *treat*: treatment
region: geographic region
pooledsite: pooled sites
agegrp: age group (below or equal the to age median or above the age median)
responder: responders those who met ACR20 criteria
non-responders: those who does not meet ACR 20 criteria
linkc: logit for the odds ratio estimation and identity for the confidence interval of the difference in proportion

8.3.6.2 Secondary Efficacy Analysis

8.3.6.2.1 Duration of Morning Stiffness

As the duration of morning stiffness is expected not to be normally distributed, the difference between the treatment groups will be assessed using the median and the confidence interval of the median computed using Hodges Lehmann method (see appendix section 10.7).

The analysis will also be repeated for mean absolute and relative changes from baseline.

Duration of morning stiffness will be analyzed only on LOCF data and on the safety, mITT and PP populations.

In addition, the analysis will be presented for USA/Canada and for Europe in separate tables.

8.3.6.2.2 DAS28:

The relative change from baseline to visit 2, visit 3 and visit 4 for DAS28 will be analyzed using a mixed model with treatment, pooled sites as a nested effect of geographic region and the interaction between the nested effect and treatment. Pooled sites within geographic region will be defined as a random effect.

The following example of SAS code will be used:

```
proc mixed data= dataset;
        class treat pooledsite region;
        random pooledsite(region);
        model change = base treat pooledsite(region)*treat/ solution;
        lsmeans treat / pdiff CL;
run;
```

where: *treat:* treatment
pooledsite: pooled site
region: America or Europe
base: baseline score for each patient
change: relative change in score from baseline

The above mentioned model will also be repeated for mean absolute and relative changes from baseline. DAS28 will be analyzed on safety population and observed case as well as LOCF.

8.3.6.2.3 Time to response

The time between baseline and a patient’s first response to the ACR20 criteria will be analyzed using Kaplan-Meier methodology and treatments will be compared using a Cox model stratified by geographic region. Time to first response is defined as the date when all the assessment leading to the ACR20 response has been collected.

The following example SAS code will be used:

```
proc lifetest data=dataset;
        time onset*censor(0);
        test treat;
run;
```

where: *treat*: treatment
onset: the time between baseline and a patient’s first response to the ACR20 criteria
censor: censor flag

Hazard ratio using Cox model stratified by region will also be presented.

A SAS code similar to the following will be used.

```
proc phreg data=dataset;
        model onset*censor(0) = treat / risklimits;
        strata region;
        ods output parameterestimates=_paramsA;
run;
```

where: *treat*: treatment
onset: the time between baseline and a patient’s first response to the ACR20 criteria
censor: censor flag
region: America or Europe

8.3.6.2.4 ACR50 and ACR70

ACR50 and ACR70 responder status at each visit endpoint will be tested separately using a logistic regression model with treatment as factor with a two-sided significance level of α=0.05 for the safety population. Observed data and LOCF imputation scheme will be analyzed.

Odds ratios for the difference between treatments and the associated 95% confidence interval, as well as the difference in proportion and its associated 95% confidence interval will be presented for each population.

A SAS code similar to the one below will be used for the primary endpoint analysis.

```
proc genmod data=dataset;
        class treat;
        model responder = treat / link=linkc dist=binomial type3;
        lsmeans treat / diff CL;
run;
```

where: *treat*: treatment
responder: responders those who met ACRX criteria (X = 50, 70)
non-responders those who did not meet ACRX criteria (X = 50, 70)
linkc: logit for the odds ratio estimation and identity for the confidence interval of the difference in proportion

8.3.6.2.5 EULAR Response:

EULAR response will be analyzed using logistic regression with treatment and geographic region as factors.

Odds ratios for the difference between treatments and the associated 95% confidence interval will be presented for the Safety population.

A similar code as the following will be used.

```
proc genmod data=dataset;
        class treat region;
        model responder = treat region / link= clogit dist=mult type3;
        lsmeans treat / diff CL;
run;
```

where: *treat*: treatment
region: America or Europe
responder: responders those who met ACR20 criteria
non-responders those who did not meet ACR 20 criteria

8.3.6.2.6 Usage of Additional Analgesics:

The number of days with additional analgesic use will be summarized and will be based on the number of days when additional analgesic were used over the treatment period (from first dose to last dose date). Wilcoxon rank sum test will be used to compare the number of days of analgesic use among treatment groups. P-value based on the normal approximation will be used.

The following example of SAS code will be used:

```
proc npar1way data=dataset WILCOXON;
        class treat;
        var analg;
run;
```

where: *treat*: treatment
analg: number of days additional analgesics used

8.3.6.2.7 Other parameters:

The following parameters will be analyzed using an ANCOVA.

- Tender joint count,
- Swollen joint count,
- Patient assessment of pain (VAS) - CRF data,
- Patient assessment of pain (VAS) morning and evening - diary data
- Patient’s global assessment of disease activity (VAS),
- Physician global assessment of disease activity (VAS),
- Quality of life questionnaire: HAQ-DI, FACIT-Fatigue, FACIT-G and SF-36
- Inflammatory parameters: CRP, ESR, IL-6, TNFα
- Urine CTX
- Severity of morning stiffness
- Reoccurrence of morning stiffness
- Additional analgesics (proportion of days with use of analgesics within the last 7 days prior to visit)

FACIT-G questionnaire will only be summarized; no analysis of change from baseline will be analyzed.
The mean absolute change and relative change from baseline to endpoint will be analyzed using ANCOVA with treatment and geographic region as the factors.

If the change from baseline in CRP, IL-6 and TNFα, is not normally distributed (normality tested using Kolmogorov-Smirnov), the data will be log transformed before the analysis and then the estimates will be back transformed, in addition differences between the treatment groups will be assessed using the confidence interval and not the p-value (p-value corresponding to the log-transformed data).

The following example of SAS code will be used:

```
proc mixed data= dataset;
        class treat region;
        model change = base treat region / solution;
        lsmeans treat / pdiff CL;
run;
```

where: *treat*: treatment
region: America or Europe
base: baseline score for each patient
change: absolute change in score from baseline

8.3.7 Safety Analysis

All summaries will be performed on the Safety population. All safety variables will be summarized by treatment groups using descriptive statistics. For categorical data, the number and percentage of patients will be presented and for continuous data the number of patients (n), mean, standard deviation (SD), median, minimum and maximum will be presented. Data will be summarized for baseline, endpoint and by visit.

8.3.7.1 Adverse Events

Absolute and relative frequencies of treatment emergent adverse events (TEAEs) will be calculated by system organ class and preferred term for all AEs, possibly related AEs, SAEs, and AEs leading to withdrawal.
All AEs count data will be summarized for the number of patients in each treatment group in whom the events occurred, and the rate of occurrence of the event. Incidence rates of TEAEs will be summarized by System Organ Class (SOC) and preferred term (PT) with respect to the Medical Dictionary for Regulatory Activities (MedDRA).

In addition, TEAEs will be summarized by seriousness, relationship to the study drug (‘Yes, reasonable causal relationship, ‘No causal relationship) for each treatment group.

If a patient has more than one occurrence of the same AE, the patient will be counted only once within that preferred term in the summary tables. The most severe occurrence of an AE, as well as the most extreme relationship of the AE to the study procedures, will be indicated in cases of multiple occurrences of the same AE.

AEs in the tables will be sorted by decreasing frequencies of SOC and PT. Supportive listings will be provided.

8.3.7.2 Laboratory Evaluation

The laboratory parameters include Hematology, Clinical Chemistry, and Urinalysis. Hematology and clinical chemistry will be analyzed for differential patterns of changes between treatment groups. Summary of laboratory results and shift table (between visit 4 and Baseline) will be presented.

Supportive individual listings will be provided.

8.3.7.3 Physical Examination

Physical Examination findings will be summarized by body system for each treatment group at Visit 0 and Visit 4. Shifts from normal to abnormal between baseline and endpoint will also be displayed.

Supportive individual listings will be provided.

8.3.7.4 Vital Signs

Vital signs (Blood pressure (mmHg) (systolic/diastolic), Pulse (beats/minute), Weight (kg) and Height (m)) will be summarized descriptively (value and absolute change from baseline) by visit (all visits) and treatment group.

8.3.7.5 Pregnancy Test

Findings from the HCG pregnancy test will only be listed.

9 REFERENCES

1. Guidelines for Industry: Structure and Content of Clinical Study Reports (E3), International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, July 1996.

2. Guidelines for Industry: Statistical Principles for Clinical Trials (E9), International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, September 1998.

3. Van Gestel AM, Anderson JJ, Van Riel PLCM et al.: ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. J Rheumatol 1999; 26: 705–711

10 APPENDIXES

The following section describes the table and listings templates, the algorithms, imputations and conventions that will generally apply to program derivations of the data as required to perform the proposed summary tabulations, individual patient data listings, and figures.

10.1 Tables template

See attached document (appendix II).

10.2 Listings template

See attached document (appendix III)

10.3 Layout

All computer-generated tables should be produces in landscape mode. The output area is restricted to 23.9 cm x 15.49 cm to allow printing both on letter and on A4 size paper with suitable margins. To achieve a readable output using SAS Monospace with font size 8 the following SAS option may not be exceed:

- Linesize=140
- Pagesize=46

The number of decimal places will be displayed as follows.

- Mean and median: one more than the number of decimal places allotted in the CRF.
- Standard deviation [SD]: two more than the number of decimal places allotted in the CRF.
- Minimum and maximum: equal to the number of decimal places allotted in the CRF.

Percentages will be presented with 1 decimal place.

The following number of decimal place for the derived variables will be used:

- Duration of RA will be presented with 1 decimal place.
- BMI will be presented with 1 decimal place.
- DAS28 will be presented with 1 decimal place.
- Diary data will be presented with 1 decimal place.
- HAQ-DI Score, 2 decimal places.
- SF-36 Scores, 1 decimal place.
- Fatigue Score and FACIT-G 1 decimal place.

10.4 Categorization

The following categorization will be used where applicable.

a. Age Classes

The patient will be summarized based on the following age categories.

- Young, if age is 45 years or less
- Middle-aged, if age is between > 45 and 65 years
- Elderly, if age is between > 65 and 75 years
- Very elderly, if age is > 75 years

b. Duration of Rheumatoid Arthritis

With regards to the duration of rheumatoid arthritis, the following categories are defined.

- < 2 years
- >= 2 to < 5 years
- >= 5 to < 10 years
- >= 10 years

c. Extent of Exposure

With regards to the total number of treatments days during the double blind treatment phase, the different categories are defined as follow.

- < 14 days
- >= 14 to < 28 days
- >= 28 to < 42 days
- >= 42 to < 56 days
- >= 56 to < 70 days
- >= 70 to < 84 days
- >= 84 days

d. Compliance with study medication

With regards to the percentage intake of study medication during the double blind treatment phase, the different categories are defined as follow.

- < 80%
- >= 80% to < 95%
- >= 85% to < 105%
- >= 105% to < 120%
- >= 120%

e. Disease Activity Score

With regards to the disease activity score, the different categories are defined as follow.

- Inactive, if DAS28 is 3.2 or less
- Moderate, if DAS28 is > 3.2 and =< 5.1
- Very Active, if DAS28 is > 5.1
- Not available if DAS28 is missing

10.5 Derivations

The following section provides details on the derivation of the variables used in the Tables, Listings and Figures.

a. Age

Age [years] is the integer of time from the date of birth [DOB] to date of informed consent [DOIC].

$$\text{Age} = \text{INT}(\text{DOIC} - \text{DOB})$$

b. Duration of an event

Duration [days] is the difference between the end date [ENDT] and the start date [STDT] plus one day.

$$\text{Duration} = (\text{ENDT} - \text{STDT}) + 1$$

Conversion from days to years will be done by dividing the number of days by 365.25.

c. Duration of morning Stiffness

Daily Duration in morning stiffness [min] is the difference between the time of resolution of morning stiffness [ENTM] and the time of wake up [STTM] (both times expressed in minutes).

$$\text{Duration\_Stiffness} = (\text{ENTM} - \text{STTM})$$

Due to missing or inconsistent subject diary entries (with regard to wake-up time, end of stiffness time, or stiffness yes/no marker) or due to entries which indicate that the morning stiffness on a particular day did not end, the above formula is not always applicable. The following table defines the rules to be applied depending on the relevant combination.

Wake-up time	End of morning stiffness	Stiffness (yes/no)	Calculated duration of stiffness
Missing	Missing or Documented	Missing or Documented	Missing
Documented	Missing	Missing	Missing
Documented	Missing	No	Set to 0
Documented before 12:00	Missing	Yes	12:00 – Wake up time
Documented after 12:00	Missing or Documented	Yes	Set to missing
Documented before 12:00	Documented before 12:00	Missing or No or Yes	End of stiffness – Wake up time
Documented before 12:00	Documented after 12:00 or reported as 00:00	Missing or No or Yes	12:00 – wake up time

If duration of stiffness is negative due to wrong entries of the wake up time or the time of end of stiffness, the duration will be set to missing. Different pages with the same date for the same patient will be analyzed as per section 10.6.
If the patient stated on the CRF diary that the morning stiffness did not subdue it was entered as 00:00 in the diary.

d. *Duration, Age of onset of rheumatoid arthritis in case of partial date*

If the date of diagnosis of rheumatoid arthritis is a partial date the following rules will be implemented to compute the duration of the rheumatoid arthritis and the onset age of rheumatoid arthritis.

Duration of RA:

- If the date of diagnosis is not missing or partial, then the duration of RA (in year) will be computed as: date of informed consent – date of diagnosis divided by 365.25.
- If only the year and month of diagnosis are recorded, the number of months between the date of informed consent and the date of diagnosis will be computed and converted to years by dividing by 30.4375.
- If only the year is recorded, the number of years between the two dates will be computed (for example, if both events occurred the same year, the duration of RA will be set to 0).

Onset Age of RA

- If the date of diagnosis is not missing or partial, then the onset age (year) will be computed as: integer of [(date of diagnosis – date of birth) divided by 365.25].
- If only the year and month of diagnosis are recorded, the number of months between the date of birth and the date of diagnosis will be computed and converted to years by taking the integer value of the number of months between the two events converted to years i.e. the integer value of the quotient of the difference in months and 30.4275.
- If only the year is recorded, the number of years between the two dates will be computed.

e. Prior/Concomitant medication and TEAEs partial date

Treatment emergent Adverse Events:

- If the start date is not partial and is on or after the first dose of medication then the AE is a treatment emergent adverse event.
- If the start date of the AE is a partial date, the AE will be classified as TEAE only in the following scenario:
  - if the year of the start of the AE event is after the year of the first dose
  - if the year is the same for both the first dose and the start of the AE and the month of the AE start date is on or after the month of the first dose or if the month is missing
  - if the AE start date is missing

Concomitant medications:

- If the start date is not partial and is on or after the first dose of medication then the medication is concomitant.
- If the start date of the medication is a partial date, the medication will be classified as concomitant only in the following scenario:
  - if the year of the start date of the medication is after the year of the first dose
  - if the year is the same for both the first dose and the start of the medication and the month of the medication start date is on or after the month of the first dose or if the month is missing
  - if the medication start date is missing and the medication is ongoing

Prior medications:

- If the start date is not partial and is strictly before the first dose of medication then the medication is a prior medication.
- If the start date of the medication is a partial date, the medication will be classified as prior only in the following scenario:
  - if the year of the start date of the medication is before the year of the first dose
  - if the year is the same for both the first dose and the start of the medication and the month of the medication start date is on or before the month of the first dose or if the month is missing

f. HAQ-DI Score

The subject must have a score for at least 6 of the 8 categories. If there are less than 6 categories completed, the HAQ-DI score cannot be computed.

- The highest score reported for any component question of the eight categories determines the score for that category
- If either devices and/or help from another person are checked for a category and the highest score for this category is 0 or 1 then the score is set to 2. The other categories will be ignored.
- A global score is calculated by summing the scores for each of the categories and dividing by the number of categories answered

The aids or devices and help from another person are linked to the 8 categories as follows.

Category	Aids or devices	Help from another person
Hygiene	Raised toilet seat Bathtub seat Bathtub bar Long-handled appliances in bathroom	Hygiene
Reach	Long-handled appliances for reach	Reach
Grip	Jar opener (for jars previously opened)	Gripping and opening things
Activities		Errands and chores
Dressing and grooming	Devices used for dressing (button hook, zipper pull, long-handled shoe horn, etc.)	Dressing an d grooming
Arising	Special or built up chair	Arising
Eating	Built up or special utensils	Eating
Walking	Cane Walker Crutches	Walking

g. SF-36 Score

The SF-36v2 scoring system requires 2 assumptions: (i) a higher score indicates a better health state and (ii) there is a linear relationship between the item scores and the underlying health concepts defined by their scales. As not all the raw item scores recorded for the SF-36v2 satisfy these assumptions, some recoding is required. All questions will be scored as per the raw data values collected on the eCRF with the following exceptions:

- Seven questions will have their coding inversed so that 5=1, 4=2, 3=3, 2=4 and 1=5. These questions are: 6, 9a, 9d, 9e, 9h, 11b and 11d.

The SF-36v2 scoring system relies on an assumption of linearity among the responses. However, for 3 of the 36 questions, it was found that the intervals were not evenly spaced among some of the qualitative responses so the values were recoded to preserve the linearity assumption. The questions affected are question 1 (General Health) and questions 7 and 8 (Bodily Pain).

Question 1 (General Health)

Q1 (verbatim responses)	Q1 (raw value)	Q1 (recoded)
Excellent	1	5.0
Very good	2	4.4
Good	3	3.4
Fair	4	2.0
Poor	5	1.0

Question 7 (Bodily Pain)

Q7 (verbatim responses)	Q7 (raw value)	Q7 (recoded)
None	1	6.0
Very mild	2	5.4
Mild	3	4.2
Moderate	4	3.1
Severe	5	2.2
Very severe	6	1.0

Question 8 will have its score inversed too, but also depends on the response given for question 7, in the following manner:

Q7 (raw value)	Q8 (verbatim responses)	Q8 (raw value)	Q8 (recoded)
1	Not at all	1	6
2, 3, 4, 5 or 6	Not at all	1	5
any	A little bit	2	4
any	Moderately	3	3
any	Quite a bit	4	2
any	Extremely	5	1

If question 7 is not answered then question 8 will have its score recoded to preserve linearity, in the following manner:

Q7 (raw value)	Q8 (verbatim responses)	Q8 (raw value)	Q8 (recoded)
missing	Not at all	1	6.0
missing	A little bit	2	4.75
missing	Moderately	3	3.5
missing	Quite a bit	4	2.25
missing	Extremely	5	1.0

Domain Scores

The 8 health domains are comprised of the individual items as follows:

- Physical Functioning Score => (Q3A Q3B Q3C Q3D Q3E Q3F Q3G Q3H Q3I Q3J)
- Role-Physical Score => (Q4A Q4B Q4C Q4D)
- Bodily Pain => (Q7 Q8)
- General Health Score => (Q1 Q11A Q11B Q11C Q11D)
- Vitality Score => (Q9A Q9E Q9G Q9I)
- Social Functioning Score => (Q6 Q10)
- Role-Emotional Score => (Q5A Q5B Q5C)
- Mental Health Score => (Q9B Q9C Q9D Q9F Q9H)

Note that Q2 is a general question and is not contained in any of the scales.

The answers to each question (recoded as necessary) are summed for each subject at each visit, within each of the 8 domains. If an item is missing, it should be imputed as the mean of the non-missing items in its domain for the purposes of calculating the domain score. Note that this imputation applies only to the calculation of the domain scores; imputation of individual item scores will not be presented. At least 50% of the item scores in a domain must be non-missing to calculate the domain score, otherwise the domain score is set to missing.

The resulting score for each domain (after the imputation described above) is then standardised, to obtain values ranging from 0 to 100, with higher values indicating a better quality of life.

$$\text{Standardised Score} = \left[\ \left(\frac{\text{sum} - \text{lowest possible score}}{\text{possible raw score range}} \right)\ \right] \text{ x } 100$$

- Physical and Mental Component Summary Scale.

Physical and Mental Summary Scale are computed based on the US based population standardization. The scoring of these two component summary scale involved the three following steps:

- Standardization of the 8 domains of the SF-36 as computed previously, as per following formula.
  - PFZ = (Physical Functioning Score - 84.52404) / 22.89490
  - RPZ = (Role-Physical Score - 81.19907) / 33.79729
  - BPZ = (Bodily Pain Score - 75.49196) / 23.55879
  - GHZ = (General Health Score – 72.21316) / 20.16964
  - VTZ = (Vitality Score – 61.05453) / 20.86942
  - SFZ = (Social Functioning Score – 83.59753) / 22.37642
  - REZ = (Role-Emotional Score – 81.29467) / 33.02717
  - MHZ = (Mental Health Score – 74.84212) / 18.01189
- Weighting and aggregation of the 8 domains scores
  - AGG_PHYS = 0.42402×PFZ + 0.35119×RPZ + 0.31754×BPZ + 0.24954×GHZ + 0.02877×VTZ – 0.00753×SFZ - 0.19206×REZ - 0.22069×MHZ
  - AGG_MENT = -0.22999×PFZ – 0.12329×RPZ - 0.09731×BPZ - 0.01571×GHZ + 0.23534×VTZ+ 0.26876×SFZ + 0.43407×REZ + 0.48581×MHZ
- Transforming the aggregate scale score to a T-score
  - Physical Component Score = 50 + 10×AGG_PHYS
  - Mental Component Score = 50 + 10×AGG_MENT

h. FACIT-F and its subscales scores

The following derivation will be performed on the items and subscale scores in order to compute the FACIT-F Score.

Subscale	**Items**	**Score**
Physical Well-being (PWB)	All items	4 – raw score
Social/family Well-Being (SWB)	All items	Raw score
Emotional Well-Being (EWB)	Q1 and Q3 to Q6	4 – raw score
Emotional Well-Being (EWB)	Q2 (I am satisfied with how I am coping with my illness)	Raw score
Functional Well-Being (FWB)	All items	Raw Score
Fatigue Subscale (FS)	Q1 to Q6 and Q9 to Q13	4 – raw score
Fatigue Subscale (FS)	Q7 (I have energy)	Raw Score
Fatigue Subscale (FS)	Q8 (I am able to do my usual activities)	Raw Score

Subscale score is computed as follow:

$$\text{Subscale Score} = \frac{\text{Sum of the Item} \times \text{Number of items}}{\text{Number of items answered}}$$

Subscale	Number of items
Physical Well-being (PWB)	7
Social/family Well-Being (SWB)	7
Emotional Well-Being (EWB)	6
Functional Well-Being (FWB)	7
Fatigue Subscale (FS)	13

Total Score of FACIT-F is the sum of the subscale scores. Total score ranges from 0 to 160.
Total Score of FACIT-G is the sum of the subscale scores (PB, SWB, EWB and FWB). Total score ranges from 0 to 108.

If more than 50% of the items (e.g., a minimum of 4 of 7 items, 4 of 6 items, etc) within a subscale are missing the subscale score cannot be computed. The FACIT scale is considered to be an acceptable indicator of patient quality of life as long as overall item response rate is greater than 80% (e.g., at least 32 of 40 FACT-F items completed, 22 of 27 for the FACT-G). If the total number of items answered is less than 80% the FACIT-F and FACIT-G score cannot be computed and will be set to missing.

10.6 Collapsing the diary information

Twice daily the patients were asked to fill the diary (one page for the morning and one page for the afternoon). At each visit a new diary was provided to the patient and the completed one was retrieved by the investigator during the visit. On the day of the visit the afternoon diary page was not completed by the patient as it was kept with the investigator. The afternoon data of the visit day is recorded on the diary using different methods.

- Scenario 1: the afternoon page was removed from the diary during the visit and stapled to the new diary.
- Scenario 2: the afternoon was completed on the afternoon of the day 1 of the new diary
- Scenario 3: the afternoon was completed on the afternoon of one of the additional pages
- Scenario 4: the afternoon was completed on the afternoon of one of the additional pages and the morning was recopied from the morning data.

However, as the date corresponding to the diary information is recorded only on the morning page, in order to link morning and afternoon data across the same visit the following rules have been applied to the programming of the diary data.

If two pages have the same date but the data were reported on consecutive visits, the two dates will be collapsed using the worse case scenario (see below).
If the first page (day 1) on the next diary has a missing date and the following day entry correspond to the day after the visit, then day 1 will be linked to the last day of the previous diary.
If the afternoon page (scenario 3) has a day out of sequence, the afternoon page will be linked to the last day of the previous visit.
If the morning and afternoon pages (scenario 4) has a day out of sequence, the morning and afternoon pages will be linked to the last day of the previous visit.

If for a diary visit a patient provided only one page with an afternoon data (whatever the day entered), it will be associated with the last day of the previous visit.

In order to implement the above rules, the page number, date and day of the diary were used to impute missing information and be able to collapse the data.
It is expected that it will not always be possible to link morning and afternoon data for the following reasons:

- missing date and day on the diary
- wrong day entered

Therefore these pages, for which it is not possible to associate a date, will be removed from the diary data and not listed.

It is assumed that day 0 entries corresponds to the day of the visit n the date of the diary entry will be imputed using that assumption.

In addition, it is also expected that some of pages of the diary will not be recorded with the correct date. If it is not possible to self evidently correct the date (done by data management), the diary information were collapsed with the other entries recorded on the same date by using the worse case scenario.

Worse case is defined as follows. If one or more values for the same date is missing then the missing will be disregarded, if all values are missing then the variable will be set to missing.

- Time of wake up: minimum of the different times available for the same date
- Presence of stiffness at wake up: set to yes if at least one of the values is set to yes for the same date, set to the no or missing otherwise
- Severity of morning stiffness: maximum of the different values of the VAS for the same date
- Time of morning stiffness subdue: maximum of the different times available for the same date
- Intensity of pain at wake up: maximum of the different values of the VAS for the same date
- Time of medication intake: the time which represents the greatest deviation from 22:00 then the latest.

- Reoccurrence of pain: set to yes if at least one of the values is set to yes for the same date, set to the no or missing otherwise
- Intensity of pain during the day:
- Additional painkiller taken: set to yes if at least one of the values is set to yes for the same date, set to the no or missing otherwise. If painkiller information has been recorded on the same page set to yes.
- Painkiller dose/time information: take all medications recorded on the different pages.

10.7 Hodges-Lehmann Estimate of Between Treatment Difference in Medians

Step 1: Create 2 separate datasets, one for each treatment group, and create a separate variable for the response.
Step 2: Calculate all possible differences between the 2 treatment groups.
Step 3: Calculate the median of these differences

Corresponding distribution-free CI (based on the Wilcoxon Rank Sum test) to be calculated as follows:

Lower limit: $0^{C\alpha}$
Upper limit: $0^{(XY+1-C\alpha)}$

where:

X = sample size for first treatment group
Y = sample size for second treatment group
C_α is an integer that approximates the ordered value of the lower confidence interval

For large samples C_α is an integer approximated by the following:

$C_\alpha \approx XY/2 - Z_{\alpha/2} [XY(X+Y+1)/12]^{1/2}$

Note: $\alpha = 0.05$ for the calculation of a 95% CI