Multi-Centre, Placebo-Controlled Study of Canakinumab for the Treatment of Adult-onset Still's disease (AOSD)

CONSIDER

(Canakinumab for treatment of adult onset Still‘s disease to achieve reduction of arthritic manifestation)

Statistical analysis plan (SAP)

Version 1: Date: 9. May 2018

Study drug:	Canakinumab.
Comparator drug:	Placebo
Indication:	Treatment of adult-onset Still’s disease with active joint involvement
Duration of the core study:	24 weeks
Clinical phase:	II
EudraCT No.	2011-001027-20
Protocol No.:	CACZ885GDE01T / NCT02204293
Duration of long-term extension study	24 months

Approved by:

Principal Investigator: Professor Dr. med. [redacted]

..... [redacted]

[redacted] Dr. [redacted]

[redacted]

[redacted]

14.5.18 .... [redacted] ....

1 General remarks

1.1 General principles

The procedures of the statistical analysis plan follow accepted guidelines especially ICH E9: Note for Guidance on Statistical Principles in Clinical Trials. If a procedure does not describe a detail and causes an open problem in the analysis then ICH E9 or other details sufficiently described here should be used to solve this problem.

1.2 Clinical parameters

The following SAP refers to names of parameters assessed in the CRF and to parameters calculated from the original CRF parameters.

All of these parameters follow one general nomenclature: var*name_t*. "*varname*" is a string in italic letters which identifies the parameter and which is valid for all time points. The suffix "_t" indicates the point in time to which the parameter refers. E.g. the parameter name of patients assessment of disease activity reported at screening is "*PATUR_01*". The corresponding names at baseline, week 4. week 8, week 12 are "*PATUR_02*", "*PATUR_03*", "*PATUR_04*", "*PATUR_05*" respectively. "var*names*" idendifying parameters assessed in the CRF are explained in the annex. Names of calculated parameters are explained in the SAP. In the case that only "varname_" is given here in the SAP this text or this formula is then valid for all time points at which this parameter is valid. Only in cases in which the *varname* refers to a specific point in time the complete *varname* is given.

Time point	Suffix of the parameters
Screening	_01
Baseline	_02
Week 4	_03
Week 8	_04
Week 12	_05
Week 16	_06
Week 20	_07
Week 24	_08

Table 1: Used suffixes of parameters (core study)

For the visits of the LTE study, the suffixes of parameters are assigned similarly in ascending order. A list is not given, since no direct reference is created in this SAP.

1.3 Abbrevations

ACR	American College of Rheumatology
ACR20	American College of Rheumatology response criterion (see p.4)
AE	adverse event
AOSD	adult onset Still's disease

BW	body weight
95% CI	95% confidence interval
CRP	C-reactive protein
DAS28	disease activity score based on 28 joint counts and ESR
DAS28(CRP)	disease activity score based on 28 joint counts and CRP
[redacted]	[redacted]
EULAR	European League Against Rheumatism
ESR	erythrocyte sedimentation rate
HAQ-DI	Health Assessment Questionnaire disabiity index
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ITT	intention-to-treat
KKS	Coordinating Center for Clinical Studies at the Charité Berlin
LOCF	last observation carried forward
LTE	long-term extension
mITT	modified ITT
PYRS	patient-years
SAE	serious adverse event

2 Rationale of the CONSIDER study

Interleukin-1 antagonists such as canakinumab have been used for the treatment of adult-onset Still's disease (AOSD) and have had a marked influence on the activity of the disease, including joint mobility. However, results from controlled clinical studies are currently not available.

3 Objective of the trial

Investigation of the efficacy of treatment with canakinumab in patients with AOSD and active joint involvement.

4 Design

4.1 Core study

The core study is performed as a multi-centre double-blinded randomized placebo controlled trial in patients with AOSD and active joint involvement.

4.2 Treatment assignment

Randomization stratified by pre-treatment status with biologic DMARDs and study centre is performed at the [redacted] in a ratio of 1:1 to the canakinumab or the placebo arm according to Atkinsons's D_A-optimal biased coin algorithm [1]. Responsible for the randomization is one statistician from the [redacted] who is not otherwise involved in the CONSIDER trial.

4.3 Long-term extension (LTE) study

Patients who participated in the core study, fulfilled the DAS28 response criterion(Δ DAS28 [*das28diff_08*] > 1.2) at week 24 and had no systemic disease manifestations of AOSD at this point in time were eligible for participation in the LTE part of the trial. All patients receive canakinumab open-label during the LTE period. The duration of the LTE part is 24 months. During the LTE period a down-titration of canakinumab from 4mg/kg BW to 2mg/kg BW is allowed in patients who fulfill DAS28 remission criteria (*das28_* < 2.6) and have no signs of systemic activity according to Yamaguchi's primary classification criteria (see study protocol for further details).

5 Primary outcome (core study)

Primary outcome: Proportion of patients with a clinically-relevant reduction in disease activity at week 12. The reduction is measured by the change in the disease activity score (DAS28). The DAS28(ESR) hereinafter referred to as DAS28 only is based on 28 joint counts, the erythrocyte sedimentation rate (mm/h) (ESR) and patient assessment of disease activity [2]. A change (Δ DAS28) > 1.2 is considered to be clinically relevant. The calculation of Δ DAS28 is shown below: it is the difference between the mean of the DAS28 scores assessed at screening and baseline and the DAS28 score assessed at week 12.

Calculation of the DAS28 score:

$$DAS28\_ = 0.56*\sqrt{gs28sum\_} + 0.28*\sqrt{gw28sum\_} + 0.7*\ln(lbbsg\_) + 0.14*patur\_$$

$gs28sum\_$: 28 tender joint count, $gw28sum\_$: 28 swollen joint count (see CRF e.g. p.6)
$lbbsg\_$: ESR mm/h (see e.g. CRF p. 7), $patur\_$: patient global of disease activity (e.g. CRF p.9)

Mannequins which included more than 28 painful or swollen joints (e.g. p. 14) always include these 28 joints which were used for the calculation of the DAS28 scores. Patient global of disease activity is measured on a 0 to 10 numerical rating scale. Of note, DAS28 scores based on ESR and DAS28 scores based on CRP are not exchangeable. In the case of a missing ESR value the missing DAS28(ESR) will not be replaced by a DAS28(CRP)..

Calculation of Δ DAS28 at week 12

$$\Delta\text{DAS28}: das28diff\_05 = \frac{das28\_01 + das28\_02}{2} - das28\_05$$

Therefore: Primary outcome: Proportions of patients with *das28diff_05* > 1.2.

6 Secondary outcomes (core study)

6.1 Efficacy

The following secondary outcomes will be investigated at weeks 4, 8,12 and as appropriate at visits thereafter. (Varnames which are needed for the calculation are given in parenthesis).

disease activity measured by

DAS28 (*DAS28_*)

68 tender joint count (GS68sum_: sum of *GS6801_* to *GS6868_*, 0 if *GS68_* = 2)

66 swollen joint count (GW66sum_: sum of *GW6801_* to *GW6836_*, *GW6839_* to *GW6868_*, 0 if *GW68_* = 2, see e.g. CRF p.14)

28 tender joint count and 28 swollen joint count (see chapter 5.)

acute phase parameters CRP and ESR (*lbcrp_ lbbsg_*)

Ferritin (*lbfer_*)

fever episodes

Physician assessment of disease activity (*KRAKT_*, see e.g. CRF p. 14)

Patient's assessment of disease activity (*PATUR_*, see e.g. CRF p. 17)

Patient's assessment of pain (*SCHMRZ_*, see e.g. CRF p. 17)

DAS28(CRP)

$$DAS28CRP\_ = 0.56 * \sqrt{gs28sum\_} + 0.28 * \sqrt{gw28sum\_} + 0.36 * \ln(lbcrpsi\_ + 1) + 0.14 * patur\_ + 0.96$$

with *lbcrpsi_* C-reactive Protein in SI-units that is in mg/L (calculated from *lbcrp_*which is assessed in mg/L or mg/dL)

functional capacity measured by

the Health Asssessment Questionnaire disability index (HAQ-DI).[3] [4] [5] The HAQ -DI is a validated measure of physical disability and functional status. The disability dimension consists of 20 multiple choice items concerning difficulty in performing eight common activities of daily living; dressing and grooming, arising, eating, walking, reaching, personal hygiene, gripping and activities. Subjects choose from four response categories, ranging from 'without any difficulty' to 'unable to do'. (In CONSIDER the parameters *HAQ11_* to *HAQHM14_* [e.g. p. 18,19] will be used to calculate the *haq_* scores as described in [3] [4]) Details of the calculating are given in the appendix.

American College of Rheumatology (ACR) and European League Against Rheumatism (EULAR) response criteria developed for rheumatoid arthritis are applied.[6] ACR20, (ACR30, ACR50, ACR70, ACR90, ACR100) response criteria are defined as follows:

20% reduction (ACR 30: 30% reduction; ACR 50: 50% reduction ACR70 70% reduction; ACR90 90% reduction; ACR100 100% reduction) between baseline and follow-up in:

68 tender joint count

66 swollen joint count

and in three of the following five parameters

erythrocyte sedimentation rate (ESR)

functional capacity (HAQ-DI)

Physician assessment of disease activity

Patient's assessment of disease activity

Patient's assessment of pain

(varnames of the criteria: *ACR20_*, *ACR30_*, *ACR50_*, *ACR70_*, *ACR90_*, *ACR100_)*

Modified adapted ACR variable

In addition to the above ACR variable, a modified adapted ACR 30 response will be calculated using the above definition (for ACR 30) and requiring additinally no intermittent fever (i.e. oral or rectal body temperature > 38°C) in the preceding week and no more than one variable 1-7 worsening by more than 30%.

EULAR response based on Δ DAS28[7]:

Good response: *das28diff_* > 1.2 and *DAS28_* at follow-up ≤ 3.2

Moderate response: *das28diff_*> 1.2 and *DAS28_* at follow-up > 3.2

or 0.6 < *das28diff_* ≤ 1.2 and *DAS28_* at follow-up ≤ 5.1

No response: *das28diff_* ≤ 0.6 or

0.6 < *das28diff_* ≤ 1.2 and *DAS28_* at follow-up > 5.1

Note: For week 4, 8, 16 etc. the *das28diff_* parameters are calculated in the same way as described above (see 5.) for week 12. Reference is always the mean of the DAS28 parameters *DAS28_01* and *DAS28_02*.

low disease activity:

Criterion for low disease activity: *DAS28_* < 3.2

remission:

Criterion for DAS28 remission: *DAS28_* < 2.6

Criterion for extended remission: *DAS28_* < 2.6 AND no signs of systemic activity defined as any of Yamaguchi´s primary classification criteria [8] for AOSD for two consecutive visits

Change in joint mobility using the neutral zero method and will also be explored.

6.2 Health related quality of life

SF-36 sum scores for physical and mental health will be calculated as described by the developer of the SF-36 health survey (Ware et al).[9]. The determinations and recommendations made there will followed (see annex for further details).

6.3 Safety

The occurrence of adverse events (AE) will be described in detail. An AE is an adverse medical event which occurs in a patient of the study and which is not necessarily in a causal relationship with the treatment the patient receives. AEs include symptoms of illnesses, as well as every unfavourable and unintended reaction. Clinical significant abnormal laboratory test finding are also included.

Serious adverse events (SAEs) are AEs leading to death, are life-threatening, require hospitalizations or prolongation of hospitalizations, represent an innate malformation or a congenital abnormality. (See study protocol for further details.) Adverse events will be coded using the MedDRA dictionary (version 21.0) that provides the primary system organ class and preferred term information.

6.4 Biomarker analysis

An accompanying project aimed at identifying predictive biomarkers (mRNA, protein). False positive discovery rates will be considered. Details are not laid down in this SAP.

7 Handling of missing values

The best method for handling of missing data is to prevent it. Since this is usually not achievable, methods for handling of missing data needs to be pre-specified. To minimize the possible bias and to allow an understandable interpretation of the main findings dropouts were deemed to be non-responders in the case of the primary endpoint and binary secondary endpoints. In the case of continues outcome parameters mixed linear models are the preferred method.

7.1 Missing response criteria and missings in other binary outcomes

Two cases were considered:

A) Patients who discontinued the study prior to the visit at which the response criterion is evaluated
B) Patients with missing data of the response criterion for other reasons.

7.1.1 Primary outcome

Patients with missing DAS28 scores at week 12 were considered as non-responders.

(We do not expect missing data according to case B) described above for the primary outcome at week 12. However, even if this happens we cannot exclude an information bias by the treating physician. We do not differentiate case B) from case A) for that reason.)

7.1.2 Missing secondary endpoints core study

The following is defined for all secondary endpoints (assessed at week 4 to 24)

Case A (see above): Non-response or non-fulfillment of the criterion will be imputed for the secondary endpoint.

Case B (see above): Provided at least two of the parameters which are needed to calculate the response criterion or binary outcome measure are available:

the last observation carried forward method (LOCF) will be used to impute the remaining missing components. Based on these imputed data the secondary endpoint will be calculated.

If none or only one parameter is available non-fulfillment or non-response will be imputed for the secondary endpoint.

7.2 Missing items of the HAQ-DI or the SF36

Specific rules apply for dealing with missing items in both questionnaires. These rules follow suggestions of the rheumatology and immunology department of Stanford university for the HAQ-D I[5] (and personal communication) as well as the SF36 manual. Both rules are described in detail in the annex.

7.3 Missing continuous efficacy parameters

Mixed linear models will be applied to deal with missing continuous parameters. In the case of the following heavily skewed outcome parameters CRP, ESR, Ferritin, joint counts the analysis will be based on non-parametric tests. The LOCF method will be used for that reason to replace missing data.

7.4 Safety

Event rates will be calculated as rates per 100 observed patient-years. Missing data will not be considered in this analysis.

8 Power considerations in the case of a premature termination of the CONSIDER trial

The recruitment of the patients was slower than expected. Therefore, the following power considerations were made for the case that the CONSIDER trial has to be terminated and the analysis has to be based on 36 patients randomized prior to April 1st. 2018. Under the assumptions made in the study protocol (verum response: 67%, placebo response 25%) an analysis based on n=18 patients per treatment arm would have a power of 66% to detect a significant difference between the treatment arms. However, these assumptions were possibly too cautiously. A somehow higher difference (verum response: 70% vs. placebo response 20%) would already lead to a power of 83% of the Fisher test.

9 Statistical efficacy analysis double blinded part of the core study (weeks (4,8) and 12)

9.1 Analysis sets

The statistical evaluation of the safety and efficacy of canakinumab will be conducted according to the Intention-to-treat principle (ITT). The safety analyses will be based on all patients randomized who received at least one dose of the study drug. All of these patients who did not violate one of the following important inclusion criteria:

diagnosis of adult onset Still's disease

disease activity based on DAS28 of ≥ 3.2 at screening

will be included in the efficacy analysis (mITT population).

No per-protocol analysis set will be considered.

9.2 Analysis of the primary outcome

9.2.1 Null hypothesis

Objective: To show superiority of treatment with canakinumab compared to placebo at week 12.

Null hypothesis: $p_{\text{response placebo}} = p_{\text{response canakinumab}}$

H_A: $p_{\text{response placebo}} \neq p_{\text{response canakinumab}}$

Two-sided type I error rate $\alpha = 0.05$.

9.2.2 Statistical analysis

The two-sided Fisher's exact test will be applied to compare the response rates observed in both treatment groups at week 12.

Mid -P 95% confidence intervals of the response rates will be calculated to describe the uncertainty in the estimates. Of note, the decision on the null hypothesis will be based only on the Fisher test but not on the confidence intervals.

The analysis will be based on a SAS macro. The resulting table is described in 15.3 below.

9.3 Binary secondary endpoints

Fisher's exact test will be applied to compare binary secondary outcome measures at weeks 4, 8, and 12 between the treatment groups (see 6.1 above for the corresponding list). Improvement according to the criterion Δ DAS28 > 1.2 will additionally be explored at week 4 and 8. Tables with frequencies and percentages of response rates, results of the Fisher test and mid-P 95% confidence intervals (95%CI) of the response rates within both groups will be provided. Week 12 results are presented in the form shown in Table 3. See also the description of Table 4 (par. 15.4) below. Graphical presentations of the final results for lectures and publications will also be available in consultation with the PI.

9.4 Non-binary secondary endpoints

Mixed linear models (in some applications also called mixed models for repeated measures) will be used to compare treatment effects between the verum and the placebo group by taking repeated measures of the outcome parameter into account.

These models include the following fixed and random effects:

> Fixed effects: treatment group, time, treatment by time interaction
> Random effects: patient.
> Taking the rather limited number of available patients into account a restriction regarding the covariance structure will be made.

The analysis is performed using the SAS procedure PROC MIXED. The repeated statement will be used to account for the repeated measurement design and the type = TOEP to specify a Toeplitz structure of the R-side covariance matrix. The lsmeans statement of the SAS procedure PROC MIXED will be applied to calculate least square means of the outcome parameter per treatment group and time point. 95% CI of these least square means will also be calculated. Simple means and SDs of the outcome parameter (as observed) will also be provided.

In the case of heavily skewed outcome parameters (CRP, ESR, Ferritin, joint counts) Brunner's non-parametric test for repeated measures will be used. The SAS macro provided and validated by the authors will be applied. The resulting p-values of these non-parametric analyses will be used for interpretation. For description purposes the LSmeans values (and their SEMs) based on the original parameters will be provided additionally. Of note, the ranks have to be calculated based on mid-ranks of a vector which contains all observed parameter values which should be included in the analysis [10].

9.5 Dropout analyses

Usually dropouts do not differ from completers in their baseline characteristics. They however frequently differ in their disease characteristics at follow-up and especially at their last visit. It is planned to examine this possible bias for two important outcomes.

Patients of the mITT population withdrawn or lost to follow-up between screening and week 12 will be compared with their last valid DAS28 and their last valid HAQ score with the DAS28 and HAQ scores respectively of the completers at the corresponding visit. This analysis will be done separately for each treatment group. A non-parametric dropout test with a sufficient power also in small samples will be applied. The parametric version of this test will be used to estimate the mean difference and to visualize the findings [11]. In case of low drop rate, the additional analysis will not be performed.

9.6 Subgroup analyses

Patients will be stratified according to their treatment history with biologicals at enrollment. Group A consists of patients who were previously treated with biologic

DMARDs and group B consists of bio-naïve patients. Provided both subgroups consists of $n_{subgroup} \geq 6$ patients verum vs. placebo response rates will be compared within both subgroups. Strata-specific 95% CI will be provided for primary and major secondary response criteria (ACR30, adapted modified ACR30, ACR50, EULAR response, DAS28 remission). Of note, CONSIDER is not powered to detect possible differences in the verum or placebo responses in bio-naïve vs. with bDMARDs previously treated patients. If the data suggest a different response to canakinumab in bionaive and non-naive patients, this must be verified by means of an interaction test (provided sufficient data per subgroup are available). PROC LOGISTIC will be used for that purpose in the case of binary outcomes, PROC MIXED in the case of continuous outcomes. The results of these interaction tests can then be evaluated with caution in the discussion.

9.7 Sensitivity analyses

Patients will be stratified in patients with / without a previous treatment with biologic DMARDs and the Cochran-Mantel-Haenszel test will be applied to investigate the primary objective in a sensitivity analysis. In the case of a significant interaction according to 9.6 above the findings of this logistic regression analysis with no additional covariates will be used for interpretation rather than the findings of the Cochran-Mantel-Haenszel test which investigated general speaking whether the averaged odds ratio significantly differs from 1.

In dependency on the number of patients lost to follow up further sensitivity analyses will be made. In these sensitivity analyses multiple imputation methods will be used to deal with missing data. The DAS28 responses, ACR responses and remission rates will be re-calculated if more than 20% of patients were lost to follow-up in one treatment arm during the first 12 weeks or if the dropout test (described above (see 9.5)) shows a significant result in at least one treatment arm.

Missing data will be replaced 10 times. That is 10 values (10 imputations) will be calculated for each missing value. The SAS procedure PROC MI will be used in two steps. In the first step the Markov Chain Monte Carlo (MCMC) method will be applied to generate a monotone missingness pattern. In the second step linear regression in the case of continuous variables and logistic regression in the case of binary variables will be specified as methods in the PROC MI procedure to replace missing values of dropouts. All imputations will be done strata specific with strata defined by treatment group. To impute a parameter x at visit t parameter values of x preceeding t will be included as predictors of x at t (=co-variables). Pre-treatment with bDMARDs prior to enrollment will be used as another co-variable. Depending on the findings of paragraph 9.5 DAS28 and/or HAQ scores of previous visits will also be included as co-variables.

10 Statistical efficacy analysis of the second part of the core study (weeks 16, 20, 24)

10.1 Description of patients included

All patients who completed the visit at week 12 and who were not withdrawn at week 12 because of non-response to canakinumab were included in the efficacy analysis at weeks 16, 20, and 24. Three groups of patients were considered:

> Group I: canakinumab responders (patients randomized to canakinumab who were primary outcome responders at week 12)
>
> Group II: placebo responders (patients randomized to placebo who were primary outcome responders at week 12),
>
> Group III: placebo non-responders at week 12 who switched to canakinumab.

Patients disposition from week 12 to week 24 will be shown by means of a flow chart. Number of dropouts and reasons for lost to follow-up will be given. Baseline and week 12 characteristics of the three groups described above as well as of the 4th group of canakinumab non-responders will be provided.

10.2 Comparison of the outcome of patients randomized to canakinumab with patients who switched from placebo to canakinumab

The outcome assessed in mITT patients randomized to canakinumab at week 4, 8, 12 will be compared with the outcome at week 16, 20, 24 of placebo patients who switched to canakinumab at week 12. In a first step the proportion of patients treated with biologic DMARDs (prior to enrollment) will be compared between both groups. Two cases are considered:

Case A: the portions differ by more than 20% or even significantly (Fisher test)

Case B: otherwise.

10.2.1 Comparison of response rates and remission rates

Taking determinations above (non-responder imputations see 7.1.2) into account number of responders and percentage of responders will be calculated for all secondary endpoints of interest (change in DAS28 > 1.2, ACR response rates, EULAR response, DAS28 remission). Case B: Mid-P 95% confidence intervals will also be calculated to allow an interpretation on the precision of the estimated rate. Case A: Logistic regression will be used to calculate adjusted rates and their corresponding 95% CI for the response rates in both groups. Adjustment will be made by the pre-treatment status with bDMARDs.

Response rates and 95% CI of group I described above (10.1) will additionally be provided.

10.2.2 Comparisons by means of disease activity parameters, HAQ-DI and SF-36

Analysis will be based on mixed linear models. Specifications made above for the double blinded period (up to week 12) are also applicable here with two exceptions: 1. In the case of the verum group, the time variable is set for baseline, week 4, 8, 12 equals 0, 4, 8, 12 as before. In the case of the placebo group, week 12 equals time=0, week 16 equals time=4 … weeks 24 equals time=12. Secondly: the pre-treatment status with bDMARDs needs to be included as an additional co-variable in case A above.

10.3 Duration of persistent response

Kaplan-Meier method will be used to estimate the duration of DAS28 response ("non-response free survival" with *das28diff_* > 1.2) in DAS28 placebo and DAS28 canakinumab responders.

10.4 Risk of deterioration in patients in remission

A simple definition of deterioration will be considered here: worsening in DAS28 > 1.2. Kaplan-Meier method will be applied to calculate the likelihood of deterioration in patients of group I and group II. No adjustment for a possible selection bias will be made.

11 Safety analysis of the core study

MedDRA will be used to code adverse events. The pt-level is used for coding and presentation of uncompiled results. The primary path is used to summarize pt terms to system organ classes. Adverse events will be summarized by presenting, for each treatment group in part I and II of the core study, the number and percentage of patients having any adverse event, having any adverse event in each primary system organ class and having each individual adverse event based on the preferred term. Since the exposure time to the randomized treatment differed between both treatment arms AE and SAE event rates per treatment exposure time will be calculated as rates per 100 patient-years (PYRS) of treatment exposure. 95% Poisson confidence intervals of these rates /100 PYRS will be calculated for events of interest and the total number of AEs, SAEs per treatment group. Events of interest are: infections, macrophage activation syndrome, anaphylactic reaction, fever episodes, AOSD typical rash (Salmon red, maculate, urticarial or maculo-papular rash), Leukocytosis of > 10 000/mm3 with > 80% neutrophils.

12 Objective of the long-term extension study

The primary objective of the LTE study is to investigate the long-term safety of canakinumab for the treatment of AOSD patients with joint pain over a period of two years.

13 Safety analysis LTE study

MedDRA (version 21.0) will be used to code adverse events. The SOC and pt-level are used for coding and presentation of uncompiled results. For each treatment group the number and percentage of patients having any adverse event, having any adverse event in each primary system organ class and having each individual adverse event based on the preferred term. Similarly as described above (chapter 11) AE and SAE event rates will be calculated as rates per 100 patient-years (PYRS) of treatment exposure from first dose to end of follow-up. 95% Exact Poisson confidence intervals of these rates /100 PYRS will be calculated for events of interest and the total number of AEs, SAEs per treatment group. Events of interest are: infections, macrophage activation syndrome, anaphylactic reaction, fever episodes, AOSD typical rash (Salmon red, maculate, urticarial or maculo-papular rash), Leukocytosis of > 10 000/mm3 with > 80% neutrophils.

14 Specifications with regard to unmasking

Rules for unmasking during the course of the study are laid down in the study protocol and are not repeated here. To perform the statistical analysis the following additional regulations are defined.

The study physicians, the PI, the KKS staff remain blind until the core study is completed. To calculate the primary outcome and important secondary outcomes at week 12 (DAS28, ACR-responses, HAQ-DI) SAS syntax and SAS macros will be developed and checked by a second SAS programmer under blind conditions.

An unmasking of the statisticians (AW, JL) will be performed after database lock and completion of validation measures. After unmasking of the statisticians the Novartis team will be unmasked on request.

Scientists Novartis SGS Cephac Europe (www.sgs.com) who perform the pk/pD analysis of blood samples of the patients will be unmasked on request.

15 Description of tables, flow charts, figures

15.1 Flow chart of patients enrolled and analyzed

According to the CONSORT statement a flow chart of patients screened, randomized, withdrawn or lost to follow-up and analyzed at week 12 and 24 will be provided. A similar flow chart including the visits 9 to 33 will be shown in the case of the LTE analysis.

15.2 Baseline characteristics

Baseline characteristics of both treatment groups and the total group will be shown in the baseline table for the following parameters

Number (percent) of female patients and mean (SD) values for:

age, disease duration, 68 tender joint count, 66 swollen joint count, ESR, CRP, DAS28, physician global, patient global, pain, HAQ, SF36 physical health, SF36 mental health.

15.3 Primary outcome

In the statistical analysis table 2 below will be calculated by means of a SAS macro.

Group	n of patients	n of responders	Response (%)	Response 95%CI	two-sided p-value Fisher test
Placebo	xx	xx	xx.x	[xx.x; xx.x]	
Canakinumab	xx	xx	xx.x	[xx.x; xx.x]	x.xxxxx

Table 2: Statistical analysis of the primary outcome at week 12

15.4 Binary secondary outcomes at week 4, 8, and 12

Similar to table 2 above findings of binary secondary outcomes at week 12 will be summarized in table 3. Again number of responders or patients in remission, percentage of response/of remission its 95% CI and the p-value of the Fisher test will be shown. This means table 3 summarizes the findings at week 12.

Outcome parameter	Group	n of patients	n of responders	Response (%)	Response 95%CI(%)	two-sided p-value Fisher test
ACR 20	Placebo	.	.	.	.	
ACR 20	Canakinumab	.	.	.	.	
....	....	....	....	....	....	....
DAS28 remission	Placebo	.	.	.	.	
DAS28 remission	Canakinumab	.	.	.	.	

Table 3: Secondary outcomes at week 12.

Table 4 is an extension of table 3 showing now outcomes at week 4, 8, 12. Table 4 describes the course in response for the different criteria.

15.5 Non-binary secondary outcomes at week 4, 8, and 12

Similar to the binary outcomes two types of tables will be provided. Table 5 with week 12 results only and table 6 with findings at week 4, 8, 12.

For each outcome and time point the following data are provided: n of patients with valid data, mean (SD) of the observed data, least square means (LSmeans) and their 95% CI. For the calculation of the LSmeans see 9.4 above.

15.6 Safety results core study

A list of all MedDRA coded (SOC and pt-term level) SAEs with corresponding event rates (n, n/100 PYRS) will be provided for each exposure group. MedDRA terms of non-serious AEs will also be listed in a detailed (SOC and pt-term) level and a summarized level.

16 References

Bibliography

1. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika. 1982; 69(1):61-67.

2. Prevoo ML, van't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum. 1995 1995; 38:44-48.

3. Lautenschlager J, Mau W, Kohlmann T, Raspe HH, Struve F, Bruckle W, et al. [Comparative evaluation of a German version of the Health Assessment Questionnaire and the Hannover Functional Capacity Questionnaire]. Z Rheumatol. 1997 May-Jun; 56(3):144-155.

4. Fries JF. The Assessment of Disability - from 1St to Future Principles. Br J Rheumatol. 1983 1983; 22(3):48-58.

5. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: dimensions and practical applications. Health Qual Life Outcomes. 2003 Jun 9; 1:20.

6. Felson DT, Anderson J, Boers M, et al. American College of Rheumatology preliminary criteria for improvement in rheumatoid arthritis patients. Arthritis Rheum. 1995 1995; 38:727-735.

7. van Gestel AM, Haagsma CJ, van Riel PL. Validation of rheumatoid arthritis improvement criteria that include simplified joint counts. Arthritis Rheum. 1998 Oct; 41(10):1845-1850.

8. Yamaguchi M, Ohta A, Tsunematsu T, Kasukawa R, Mizushima Y, Kashiwagi H, et al. Preliminary criteria for classification of adult Still's disease. J Rheumatol. 1992 Mar; 19(3):424-430.

9. Ware JE, New England Medical Center H, Health I. SF-36 physical and mental health summary scales : a user's manual. Boston: Health Institute, New England Medical Center, 1994.

10. Brunner E, Domhof S, Langer F. Nonparametric analysis of longitudinal data in factorial experiments: J. Wiley, 2002.

11. Listing J, Schlittgen R. A Nonparametric Test for Random Dropouts. Biom J. 2003 2003; 45(1):113-127.

Annex B: Calculation of the HAQ-DI

Within each of the 8 categories only the item indicating the most severe impairment contributes to the category score.

In the first step the HAQ-DI Items have to be recoded.

- Without ANY Difficulty will be recoded in 0
- With SOME Difficulty will be recoded in 1
- With MUCH Difficulty will be recoded in 2
- UNABLE to do will be recoded in 3

In the second step the maximum of the items within each of the eight categories will be taken. SAS-code:

```
haq1_ = MAX(haq11_,haq12_);
haq2_ = MAX(haq21_,haq22_) ;
haq3_ = MAX(haq31_,haq32_, haq33_);
haq4_ = MAX(haq41_,haq42_);
haq5_ = MAX(haq51_,haq52_, haq53_);
haq6_ = MAX(haq61_,haq62_);
haq7_ = MAX(haq71_,haq72_, haq73_);
haq8_ = MAX(haq81_,haq82_, haq83_);
```

There is no condition on the completeness of the items. The maximum of several variables will result in a non-missing value if at least one valid value is available.

Missing scores and scores of 0 or 1 in a category are increased to 2 if the patient uses devices or help from another person to carry out the corresponding activities.

If at least six out of the eight parameters are valid (not missing) the HAQ-DI parameter will be calculated

SAS-code: haq_scor_ = Mean(haq1,haq2,haq3,haq4,haq5,haq6,haq7,haq8);

Annex C: Calculation of the SF36 scales

The scoring of the SF36 scales is clearly and unambiguously described in the manual of Ware JE (1994). In the following only some key issues are repeated.

The following varnames of the items are used *SF3601_ to SF3636_*. In the first step the items 1 (*SF3601_*), *6* (*SF3620_*), 7 (*SF3621_*), 8 (*SF3622_*), 9a (*SF3623_*), 9d (*SF3626_*), 9e (*SF3627_*), 9h (*SF3630_*), 11b (*SF3634_*), 11d (*SF3636_*) have to be recoded as described in the manual.

Item	Response	Precoded Value	Final value
1 (*SF3601_*)	Excellent	1	5
	Very good	2	4.4
	Good	3	3.4
	Fair	4	2
	Poor	5	1
6 (*SF3620_*)	Not at all	1	5
	Slightly	2	4
	Moderately	3	3
	Quite a bit	4	2
	Extremely	5	1
7 (*SF3621_*)	None	1	6
	Very mild	2	5.4
	Mild	3	4.2
	Moderate	4	3.1
	Severe	5	2.2
	Very severe	6	1
8 (*SF3622_*)if item 7 (*SF3621_*) is answered (response to item 7 is given in parentheses)	Not at all	1 (7: 1)	6
	Not at all	1 (7: 2 - 6)	5
	A little bit	2 (7: 1 - 6)	4
	Moderately	3 (7: 1 - 6)	3
	Quite a bit	4 (7: 1 - 6)	2
	Extremely	5 (7: 1 - 6)	1
8 (*SF3622_*) if item 7 (*SF3621_*) is not answered	Not at all	1	6
	A little bit	2	4.75
	Moderately	3	3.5
	Quite a bit	4	2.25
	Extremely	5	1
9a (*SF3623_*) & 9e (*SF3627_*)	Almost of the time	1	6
	Most oft he time	2	5
	A good bit of the time	3	4
	Some oft he time	4	3
	A little bit of the time	5	2
	None of the time	6	1
9d (*SF3626_*)& 9h (*SF3630_*)	All of the time	1	6
	Most of the time	2	5
	A good bit of the time	3	4
	Some of the time	4	3
	A little of the time	5	2
	None of the time	6	1
11b (*SF3634_*) & 11d (*SF3636_*)	Definitely true	1	5
	Mostly true	2	4
	Don't know	3	3
	Mostly false	4	2
	Definitely false	5	1

Table: Recoding values of SF36 items

In the second step the eight subscales physical functioning (*SF36_PF_*), role physical (*SF36_RP_*), bodily pain (*SF36_BP_*), general health (*SF36_GH_*), vitality (*SF36_VT_*), social functioning (*SF36_SF_*), role emotional (*SF36_RE_*), mental health (*SF36_MH_*) will be calculated as described in the manual. Of note, Ware JE (1994) recommended to calculate the score of the subscale if a "respondent answered at least half of the items ... or half plus one in the case of scales with an odd number of items". Ware JE (1994) further recommended imputing missing items by the average score of the remaining items of the subscale. These recommendations will be followed here. In the case of missing data the raw score of the corresponding subscale needs to be re-calculated after the handling of the missing items.

In the next step the corresponding standardized subscales will be calculated. This standardization is done by using means and standard deviations observed in the US general population.

```
SF36_PF_Z_ = (SF36_PF_ - 84.5240452)/22.8948992;
SF36_RP_Z_ = (SF36_RP_ - 81.1990721)/33.7972923;
SF36_BP_Z_ = (SF36_BP_ - 75.4919631)/23.5587879;
SF36_GH_Z_ = (SF36_GH_ - 72.2131559)/20.1696447;
SF36_VT _Z_ = (SF36_VT_ - 61.0545296)/20.8694255;
SF36_SF_Z_ = (SF36_SF_ - 83.5975259)/22.3764186;
SF36_RE_Z_ = (SF36_RE_ - 81.2946729)/33.0271727;
SF36_MH_Z_ = (SF36_MH_ - 74.8421239)/18.0118961;
```

In the last step the Physical and Mental Component Summary Scores will be calculated as weighted and T-transformed sums of the eight subscales.

```
agg_phys_ = (SF36_PF_Z_ * .42402)+( SF36_RP_Z_ * .35119)+( SF36_BP_Z_ *
.31754)+( SF36_SF_Z_ * -.00753)+ (SF36_MH_Z_ * -.22069)+( SF36_RE_Z_ * -
.19206)+( SF36_VT_Z_ * .02877)+( SF36_GH_Z * .24954);
agg_ment_ = (SF36_PF_Z_ * -.22999)+( SF36_RP_Z_ * -.12329)+( SF36_BP_Z_ * -
.09731)+( SF36_SF_Z * .26876)+ (SF36_MH_Z_ * .48581)+( SF36_RE_Z_ * .43407)+(
SF36_VT_Z_ * .23534)+( SF36_GH_Z_ * -.01571);
```

```
SF36_PCS_ = 50 + (agg_phys_*10);
```

```
SF36_MCS_ = 50 + (agg_ment_*10);
```

SF36_PCS_ and SF36_MCS_ Scores are not necessarily missing if single items are missing (see above) but only if at least the score of one subscale is missing.

Official Title:	**A Multi-Centre, Placebo-Controlled Phase II Study of Canakinumab for the Treatment of Adult-Onset Still's Disease (AOSD)-core study, including an open-label long-term extension – LTE**
NCT Number:	NCT02204293
Document Date:	Amendment : 2 October 2018

Amendment

Statistical analysis plan (SAP)

Multi-Centre, Placebo-Controlled Study of Canakinumab for the Treatment of Adult-onset Still's disease (AOSD)

CONSIDER

Rationale: Considering the limited sample size (n=35) and taking the low numbers of patients with missing week 12 data (n=4) into account the following changes will be made to improve the power of the analysis of continuous secondary endpoints. These changes are made prior to unblinding of the statistical analysis team [redacted] but after being aware of the findings of a recent RCT in adult-onset Still's disease (Kaneko Y, et al. Ann Rheum Dis 2018;0:1–10. doi:10.1136). Taking advantage of the low number of dropouts the application of the last observation carried forward (LOCF) method can be regarded as justified. This allows the application of the analysis of covariance (ANCOVA), which increases the power not only compared to the t-test but in general also in comparison the time*group interaction test of a mixed linear model. For this reason the following changes are made:

Paragraph 7.3 Missing continuous efficacy parameters (page 10)

For CRP values where only <2mg/L or < 1mg/L is specified, the center of the range (for < 2mg/L i.e. 1mg/L) is imputed.

In applications described in paragraph 9.4 below the last observation carried forward method (LOCF) will be used to replace missing values of continuous efficacy parameter. In addition ~~M~~mixed linear models will be applied to deal with missing continuous parameters (see 9.4 below). ~~In the case of the following heavily skewed outcome parameters CRP, ESR, Ferritin, joint counts the analysis will be based on non-parametric tests. The LOCF method will be used for that reason to replace missing data.~~

Paragraph 9.4 Non-binary secondary endpoints (page 12-13)

Analysis of covariance (ANCOVA) will be applied to compare the outcome of continuous or non-binary parameters at week 12 between the verum and the placebo group after adjustment for the baseline status (covariable). For parameters assessed at screening and at the baseline visit the average of both values will be used as covariable. The last observation carried forward method (LOCF) will be used to impute missing values of the outcome at week 12. The statistical comparison of the groups is carried out using type 3 test of the ANCOVA model. In addition baseline (covariable) adjusted least square means (LSmeans) of the outcome parameter and their corresponding 95% CI will be calculated.

In order to be able to compare the course of the secondary efficacy parameter in both groups over the period of 12 weeks, a ~~M~~mixed linear model~~s~~ (in some applications also called mixed model~~s~~ for repeated measures) will be calculated. These models allow comparisons of ~~used to compare~~ treatment effects between the verum and the placebo group by taking repeated measures of the outcome parameter into account.

These models include the following fixed and random effects:

Fixed effects: treatment group, time, treatment by time interaction
Random effects: patient.
Taking the rather limited number of available patients into account a
restriction regarding the covariance structure will be made.

The analysis is performed using the SAS procedure PROC MIXED. The repeated statement will be used to account for the repeated measurement design and the type = TOEP to specify a Toeplitz structure of the R-side covariance matrix. The lsmeans statement of the SAS procedure PROC MIXED will be applied to calculate least square means of the outcome parameter per treatment group and time point. 95% CI of these least square means will also be calculated. The LSmeans and their 95% CI calculated for baseline and for week 12 can then be compared with the LSmeans and their 95% CI calculated by

ANCOVA to discuss a possible impact of LOCF. To discuss this impact in the case of noticeable differences in the baseline status between the groups a further mixed linear model with the baseline status as covariable can be applied. In this case missing week 4 data of dropouts have to be replaced by LOCF to include those patients in the analysis and to calculate the corresponding LSmeans. Simple means and SDs of the outcome parameter (as observed) will also be provided.

In the case of heavily skewed outcome parameters (CRP, ESR, Ferritin, joint counts) a non-parametric ANCOVA proposed by Brunner and colleagues will be used. ~~Brunner's non-parametric test for repeated measures will be used.~~ The SAS macro npar provided and validated by the authors will be applied. The resulting p-values of these non-parametric analyses will be used for interpretation. For description purposes the LSmeans values (and their ~~SEMs~~95% CI) based on ~~the original parameters~~parametric ANCOVA will be provided additionally. Of note, the ranks have to be calculated based on mid-ranks of a vector which contains all observed parameter values which should be included in the analysis.

Berlin, October 2, 2018

Approved by:

Principal Investigator: Prof. Dr. med. [redacted] [redacted]

....

[redacted]: Dr. [redacted] [redacted]

[redacted]

....