AstraZeneca

Statistical Analysis Plan

Study Code	D1699C00001
Edition Number	3.0
Date	23/July/2019

Study to Evaluate the Effect of Dapagliflozin on the Incidence of Worsening Heart Failure or Cardiovascular Death in Patients with Chronic Heart Failure with Reduced Ejection Fraction

Study to Evaluate the Effect of Dapagliflozin on the Incidence of Worsening Heart Failure or Cardiovascular Death in Patients with Chronic Heart Failure with Reduced Ejection Fraction

Study Statistician

______________________________ ______________

[redacted] Date

Study to Evaluate the Effect of Dapagliflozin on the Incidence of Worsening Heart Failure or Cardiovascular Death in Patients with Chronic Heart Failure with Reduced Ejection Fraction

Biometrics Team Leader

[redacted]

Date

TABLE OF CONTENTS PAGE

TITLE PAGE 1
SIGNATURE OF STUDY STATISTICIAN 2
SIGNATURE OF GLOBAL PRODUCT STATISTICIAN 3
TABLE OF CONTENTS 4
LIST OF ABBREVIATIONS 6
AMENDMENT HISTORY 8
1. STUDY DETAILS 11
1.1 Study objectives 11
1.1.1 Primary objective 11
1.1.2 Secondary objectives 11
1.1.3 Safety objectives 12
1.1.4 Explorative objectives 13
1.2 Definitions 14
1.2.1 Primary analysis censoring date 14
1.2.2 Withdrawal of informed consent 15
1.2.3 Discontinuation from study drug 15
1.2.4 Vital status 15
1.2.5 Lost to follow-up 15
1.3 Study design 15
1.3.1 Randomization 16
1.4 Number of patients 17
2. ANALYSIS SETS 17
2.1 Definition of analysis sets 17
2.1.1 Full analysis set 17
2.1.2 Safety analysis set 17
2.2 Violations and deviations 18
3. PRIMARY AND SECONDARY VARIABLES 18
3.1 Primary variable 19
3.2 Secondary variables 19
3.2.1 The composite of CV death and hospitalization for HF 19
3.2.2 Total number of (first and recurrent) hospitalizations for HF and CV death 19
3.2.3 Change from baseline at 8 months in the KCCQ total symptom score 20

3.2.4 The composite of ≥50% sustained decline in eGFR, end stage renal disease and renal death 20
3.2.5 Death from any cause 21
3.3 Safety variables 21
4. ANALYSIS METHODS 22
4.1 General principles 22
4.1.1 Estimand for primary and secondary outcomes 23
4.1.2 Hypotheses 23
4.1.3 Confirmatory testing procedure 24
4.1.4 Presentation of time-to-event analyses 24
4.1.5 Vital status and follow-up of endpoints 24
4.2 Analysis methods 25
4.2.1 Demographics and baseline characteristics 25
4.2.2 Concomitant and baseline medication 25
4.2.3 Analysis of the primary efficacy variable 26
4.2.3.1 Subgroup analysis of the primary endpoint 26
4.2.3.2 Sensitivity analysis of the primary endpoint 28
4.2.4 Analysis of the secondary efficacy variables 28
4.2.4.1 Analysis of recurrent HF events and CV death 28
4.2.4.2 Analysis of change from baseline at 8 months in the KCCQ total symptom score 30
4.2.5 Analysis of safety variables 34
4.2.5.1 Adverse events 34
4.2.5.2 Serious adverse events 35
4.2.5.3 Adverse events leading to discontinuation, interruption or dose reduction 35
4.2.5.4 Adverse events of interest 35
4.2.5.5 Laboratory Evaluation 36
4.2.5.6 Marked laboratory abnormalities 36
4.2.6 Analysis of exploratory objectives 36
5. INTERIM ANALYSES 37
6. CHANGES OF ANALYSIS FROM PROTOCOL 38
7. REFERENCES 38
8. APPENDIX 41
8.1 Appendix A Laboratory Abnormality Criteria 41
8.2 Appendix B Estimation of clinically meaningful thresholds for KCCQ total symptom score 43

LIST OF ABBREVIATIONS

Abbreviation or special term	**Explanation**
AE	Adverse event
AF	Atrial fibrillation
ATC	Anatomical Therapeutic Chemical
CEA	Clinical event adjudication
CKD-EPI	Chronic kidney disease epidemiology collaboration equation
CMH	Cochran-Mantel-Haenszel
CMWPC	Clinically meaningful within-patient change
CSP	Clinical study protocol
CSS	KCCQ clinical summary score
CV	Cardiovascular
DAE	Adverse event leading to discontinuation of investigational product
DMC	Data monitoring committee
eCDF	Empirical cumulative distribution function
eCRF	Electronic case report form
eGFR	Estimated glomerular filtration rate
ESRD	End stage renal disease
FAS	Full analysis set
HbA1c	Glycosylated haemoglobin
HF	Heart failure
HFrEF	Heart failure with reduced ejection fraction
HR	Hazard ratio
IP	Investigational Product (dapagliflozin or matching placebo)
ITT	Intention to treat
IxRS	Interactive Voice/Web Response System
KCCQ	Kansas City Cardiomyopathy Questionnaire
KM	Kaplan-Meier
LTFU	Lost to follow-up
LVEF	Left ventricular ejection fraction
MA	Marked abnormality
MAR	Missing at random

Abbreviation or special term	Explanation
MedDRA	Medical dictionary for regulatory activities
MI	Myocardial infarction
MMRM	Mixed model repeated measures
MRA	Mineralocorticoid antagonist
NT-proBNP	N-terminal pro b-type natriuretic peptide
NYHA	New York Heart Association
PACD	Primary analysis censoring date
PGIS	Patient global impression of severity
ROC	Receiver operating characteristic
SAE	Serious adverse event
SAP	Statistical analysis plan
SCV	Study closure visit
SED	Study end date
SD	Standard deviation
T2D	Type 2 diabetes
TSS	KCCQ total symptom score
WoC	Withdrawal of consent

AMENDMENT HISTORY

Date /Version	Brief description of change
1 Feb 2017 / Version 1.0	Version 1.0 signed
19 Feb 2019 /Version 2.0	Secondary objective for KCCQ changed from clinical summary score to total symptom score, following protocol update (protocol version 2.0, 26 October 2017) [Section 1.1.2, 3.2.3, 4.2.4.2] AEs leading to a risk for lower limb amputations added to safety objective, following protocol update [section 1.1.3, 3.3] Echocardiography substudy added to exploratory objective, following protocol update [section 1.1.4] Clarification of prohibited concomitant medication added [section 2.2, 4.2.2] The analysis of the KCCQ endpoint changed to a composite rank-based method, following FDA feedback regarding the initially proposed mixed model repeated measures approach. [Section 3.2.3, 4.1.1, 4.2.4.2] Adjudication requirement for potential events related to eGFR decline was removed, following protocol update [Section 3, 3.2.4, 4.2] Event types that are being withheld from reporting to health authorities limited to HF events and death (Renal events will not be withheld). [Section 3.3] Definition of baseline diabetic status for subgroup analysis added [Section 4.1] eGFR will be calculated from central laboratory creatinine values, rather than using eGFR calculated by central laboratory. [Section 4.1] Clarification that the significance level will be determined based on the exact actual proportion of primary endpoints included in the interim analysis [Section 4.2] Updates to the list of subgroups [Section 4.2.3.1, Table 1] • Diuretics removed due to small group of patients not on diuretics. • History of hypertension replaced by etiology of HF (Ischemic vs Non-ischemic/unknown). • Clarification that atrial fibrillation group will include patients with atrial flutter and be based on enrolment ECG • Definition of T2D subgroup clarified In consideration of large variability associated with few events, it has been specified that HRs with confidence intervals will not be presented if less than 15 events in asubgroup [Section 4.2.3.1] Specification that 'on+off' treatment will be considered the primary analysis approach for fractures and amputations [section 4.2.5]

Date /Version	Brief description of change
	Statement about urinalysis data removed. This was included by mistake in version 1.0. No central laboratory measurements are collected in this study [Section 4.2.5.5]
	Statement about presentation of potential events of Fournier's gangrene added [Section 4.2.5.4]
	Clarification that presentation of DKA will primarily be based on adjudicated DKA
	Statement about a primary analysis censoring date to be applied after a decision to stop the trial after interim analysis added [Section 5]
23 July 2019 / Version 3.0	Details of statistical methods, assumptions and references for the joint frailty model of recurrent HF events and CV death added based on FDA request. Details about Ghosh and Lin plot added. [section 3.2.2 and 4.2.4.1]
	Removed reference to the R package Sanon (Kawaguchi et al 2011) for derivation of the adjusted win ratio as the Mann-Whitney odds estimate and removed the intermediate step to explicitly calculate the standard error of the Mann-Whitney odds. The same statistic will be derived using SAS as described in Koch et al 1998 (reference added) [section 4.2.4.2, Estimation of treatment effect]
	Additional details regarding responder analysis of KCCQ TSS added [section 4.2.4.2, Estimation of treatment effect]. • Handling of patients whose baseline value is too high/low to make it possible to achieve the threshold for improvement/deterioration. • Details of logistic regression of the proportion of patients achieving the responder thresholds added. • Responder analysis by thresholds derived from anchor-based analysis added.
	Clarification added regarding the variable representing the number of HF events in multiple imputation of missing TSS values. [section 4.2.4.2, Handling of missing data]
	Clarification of use of imputed TSS values in responder analysis [section 4.2.4.2, Handling of missing data]
	Correction: Descriptive statistics for KCCQ scores will be presented at 4 and 8 months [section 4.2.4.2, Supportive and sensitivity analyses]
	Correction to clarify how the sensitivity analysis of change from baseline TSS is different from the main analysis in terms of handling patients who died. [section 4.2.4.2, Supportive and sensitivity analyses]
	Correction: Removed creatinine clearance from laboratory variables to be summarized. Assessment of creatinine clearance is not defined in the study protocol. [section 4.2.5.5]

Date /Version	Brief description of change
	Marked lab abnormality criteria for creatinine clearance < 45 mL/min and eGFR < 45 mL/min/1.73m^2 were removed since not considered appropriate criteria for abnormality in the enrolled study population and had been included by mistake. [Appendix A]. The criterion serum creatinine ≥1.5xbaseline was retained and serum creatinine ≥2xbaseline was added. Added Appendix B describing the estimation of clinically meaningful thresholds for KCCQ TSS.

1. STUDY DETAILS

1.1 Study objectives

1.1.1 Primary objective

Primary Objective:	**Outcome Measure:**
To determine whether dapagliflozin is superior to placebo, when added to standard of care, in reducing the incidence of CV death or a HF event (hospitalization for HF or equivalent HF event, ie an urgent HF visit)**.**	Time to the first occurrence of any of the components of this composite: 1. CV death 2. Hospitalization for HF 3. An urgent HF visit

1.1.2 Secondary objectives

Secondary Objective:	**Outcome Measure :**
To compare the effect of dapagliflozin versus placebo on CV death or hospitalization for HF.	Time to the first occurrence of either of the components of this composite: 1. CV death 2. Hospitalization for HF
To compare the effect of dapagliflozin versus placebo on total number of recurrent HF hospitalizations and CV death.	Total number of (first and recurrent) HF hospitalizations and CV death.
To compare the effect of treatment with dapagliflozin versus placebo on the Kansas City Cardiomyopathy Questionnaire (KCCQ) total symptom score for HF symptoms.	Change from baseline measured at 8 months in the total symptom score of the KCCQ, a specific HF patient reported outcome questionnaire.

To determine if dapagliflozin compared with placebo reduces the incidence of a composite endpoint of worsening renal function.	Time to the first occurrence of any of the components of this composite: 1. ≥50% sustained decline in eGFR 2. Reaching End Stage Renal Disease (ESRD) – Sustained eGFR <15 ml/min/1.73m^2 or, – Chronic dialysis treatment or, – Receiving a renal transplant 3. Renal death
To determine whether dapagliflozin, compared with placebo, reduces the incidence of all-cause mortality.	Time to death from any cause.

1.1.3 Safety objectives

Safety Objective:	**Outcome Measure :**
To evaluate the safety and tolerability of dapagliflozin in this patient population.	1. Serious Adverse Events (SAEs) 2. Discontinuation of IP due to Adverse Events (DAEs) 3. Changes in clinical chemistry/haematology parameters 4. Adverse events of interest (volume depletion, renal events, major hypoglycaemic events, fractures, diabetic ketoacidosis (DKA), AEs leading to amputation and AEs leading to a risk for lower limb amputations [“preceding events”])

1.1.4 Explorative objectives

Exploratory Objective:	**Outcome Measure :**
To compare the effect of dapagliflozin versus placebo on an expanded composite outcome reflecting worsening of HF.	Time to the first occurrence of any of the components of the expanded composite worsening HF outcome: 4. CV death 5. Hospitalization for HF 6. An urgent HF visit 7. Documented evidence of worsening HF symptoms/signs leading to initiation of a new treatment for HF sustained for at least 4 weeks or augmentation of existing oral therapy for HF (eg, increase in dose of diuretic) sustained for at least 4 weeks.
To determine whether dapagliflozin compared with placebo will have effect on New York Heart Association (NYHA) class.	Change in NYHA class from baseline.
To determine whether dapagliflozin compared with placebo will reduce the incidence of diagnosis of AF in patients without history of AF at baseline.	Proportion of patients without history of AF at baseline with a new diagnosis of AF during the study.
To determine whether dapagliflozin compared with placebo will result in a reduction of the incidence of hyper – and hypokalaemia.	Time to the first occurrence of each of any of the following central lab levels of serum potassium: • >6.0 mmol/L • >5.5 mmol/L • <3.5 mmol/L • <3.0 mmol/L
To determine whether dapagliflozin compared with placebo will affect the number of events of doubling of serum creatinine.	Number of events with doubling of serum creatinine (compared with the most recent laboratory measurement).
To determine whether dapagliflozin compared with placebo will reduce the incidence of diagnosis of T2D in patients without diabetes at baseline.	Proportion of patients without T2D at baseline with a new diagnosis of T2D during the study.
To determine whether dapagliflozin compared with placebo will have effect on HbA1c in T2D subgroup.	Changes in HbA1c from baseline.

To determine whether dapagliflozin compared with placebo will have an effect on systolic BP.	Change in systolic BP from baseline.
To determine whether dapagliflozin compared with placebo will have an effect on body weight.	Change in body weight from baseline.
To determine whether dapagliflozin compared with placebo will reduce the incidence of myocardial infarction (MI).	Time to first fatal or non-fatal MI.
To determine whether dapagliflozin compared with placebo will reduce the incidence of any stroke (ischemic, hemorrhagic, or undetermined).	Time to first fatal or non-fatal stroke of any cause.
To compare the effect of dapagliflozin versus placebo on health status assessed by Patient Global Impression of Change (PGIC) and Patient global impression of severity (PGIS) questionnaires.	Changes in health status measured by PGIC and PGIS.
To compare the effect of dapagliflozin versus placebo on health status assessed by EuroQol five-dimensional five-level questionnaire (EQ-5D-5L) to support health economic analysis and health technology assessment.	Changes in health status measured by EQ-5D-5L.
To collect and analyse pharmacokinetic (PK) samples for dapagliflozin concentration.	Not applicable. Results will be reported separately.
To assess cardiac structure and function with echocardiography at baseline and 8 months follow-up.	Not applicable. Results will be reported separately.
To collect and store samples of plasma and serum for future exploratory biomarker research.	Not applicable. Results will be reported separately.

1.2 Definitions

1.2.1 Primary analysis censoring date

The executive committee and AstraZeneca will monitor the accrual of endpoint events and when appropriate define the primary analysis censoring date (PACD) at which time at least the pre-defined target number of 844 events for the primary composite endpoint is expected to have occurred. The study sites will be instructed to plan for study closure visits to be performed after PACD.

The analyses of the endpoint events will include events with onset on or prior to PACD. Event free patients who have not been prematurely censored due to incomplete information (see

Section 3.1 and 3.2) will be censored at PACD. Endpoint events with onset after PACD will also be adjudicated.

1.2.2 Withdrawal of informed consent

Withdrawal of consent (WoC) means withdrawal from study and should only occur if the patient does not agree to any kind of further assessment at all. No data after date of WoC should be collected, with the exception of vital status (dead or alive) at the end of the study collected from public sources, which will be included in the analysis of death from any cause as a sole outcome and in patient disposition summaries. Data collected on or prior to date of WoC will be included in analyses.

1.2.3 Discontinuation from study drug

Discontinuation from study drug does not mean WoC. Optimally, patients who discontinue from study drug should continue study visits according to plan until study closure. Alternatively, if the patient does not agree to this approach, modified follow-up should be arranged. Data from patients who did not withdraw consent will be included in the intention to treat (ITT) analyses irrespective of whether the event occurred before or following discontinuation of study drug.

1.2.4 Vital status

Known vital status at the end of the study will be defined when the patient is dead or has date last know alive on or after the PACD.

For patients who have withdrawn consent, the investigator will attempt to collect vital status from publicly available sources at study closure in compliance with local privacy laws/practices.

1.2.5 Lost to follow-up

The term lost to follow-up (LTFU) will be limited to patients with unknown vital status at the end of the study as defined in Section 1.2.4. Other measures will be used to describe incomplete follow-up of the primary endpoint (Section 4.1.5)

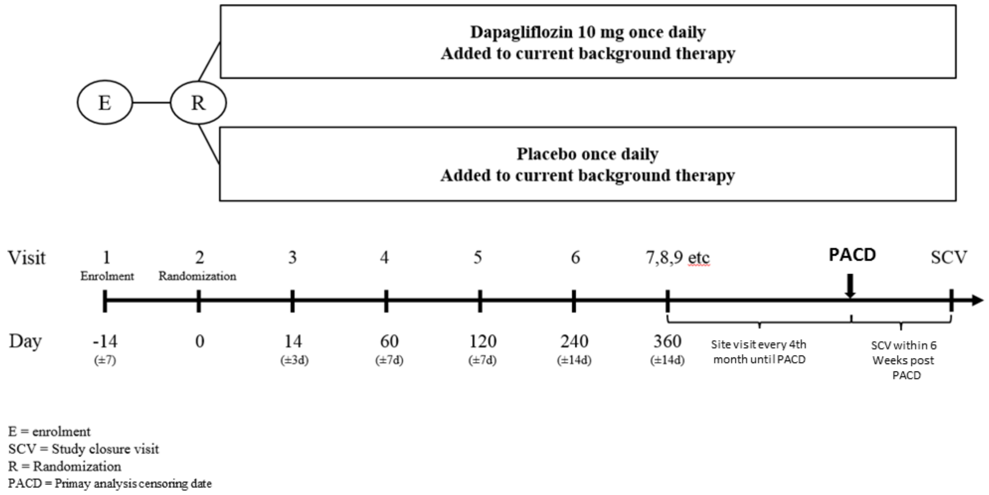
1.3 Study design

This is an international, multicentre, parallel group, event-driven, randomized, double-blind, placebo-controlled study in patients with chronic heart failure with reduced ejection fraction (HFrEF), evaluating the effect of dapagliflozin 10 mg versus placebo, given once daily in addition to background regional standard of care therapy, for the prevention of cardiovascular (CV) death or reduction of heart failure (HF) events.

It is estimated that approximately 7000 patients at approximately 500-600 sites in 20-25 countries will be enrolled to reach the target of approximately 4500 randomized patients. The investigational product (IP) will be added to the prescribed background therapy for HF (and background therapy for T2D when applicable) as considered appropriate by the investigator and in accordance with regional standards of care.

The anticipated duration of the study is approximately 33 months. The study closure procedures will be initiated when the predetermined number of adjudicated primary endpoints is predicted to have occurred (844). This date, which in the clinical study protocol (CSP) is termed study end date (SED) will be the common censoring date for efficacy time-to-event analyses. Thus, it will in this SAP be termed the primary analysis censoring date (PACD, Figure 1).

Figure 1 Study flow chart


Dapagliflozin 10 mg once daily
Added to current background therapy
E
R
Placebo once daily
Added to current background therapy
Visit
1
Enrolment
2
Randomization
3
4
5
6
7,8,9 etc
PACD
SCV
Day
-14
(±7)
0
14
(±3d)
60
(±7d)
120
(±7d)
240
(±14d)
360
(±14d)
Site visit every 4th month until PACD
SCV within 6 Weeks post PACD
E = enrolment
SCV = Study closure visit
R = Randomization
PACD = Primay analysis censoring date

1.3.1 Randomization

Patients will be randomized 1:1 to either dapagliflozin 10 mg or placebo. Randomization will be stratified by type 2 diabetes (T2D) status at randomization (2 levels: with T2D; without T2D). T2D for stratified randomization is defined as established diagnosis of T2D or HbA1c more or equal to 6.5% (48 mmol/mol) shown at central laboratory test at enrolment (visit 1). See section 4.1 for definition of baseline diabetic status for analysis.

Randomization will be performed in balanced blocks of fixed size. The randomization codes will be computer generated and loaded into the IxRS (Interactive Voice/Web Response System) database.

The number of randomized patients with and without T2D will be monitored in order to ensure a minimum of 30% in each sub-population. Randomization may be capped (ie, no more patients can be randomized in a specific sub-population) if the pre-determined limit is reached.

Randomization of patients based on geographic region will be monitored to ensure global representation. LVEF category, NYHA class and atrial fibrillation status may be capped in IxRS to avoid over- or under-representation of these patient subgroups.

1.4 Number of patients

The primary objective of the study is to determine the superiority of dapagliflozin versus placebo in reducing the incidence of the primary composite endpoint. Assuming a true hazard ratio (HR) of 0.80 between dapagliflozin and placebo, using a one-sided alpha of 2.5%, 844 primary endpoint events will provide a statistical power of 90% for the test of the primary composite endpoint. This is based on an overall 1:1 allocation between dapagliflozin and placebo. The study is event-driven. The assumed HR of 0.80 is considered as clinically relevant and has taken into account the HF outcomes in the EMPA-REG trial (Fitchett et al 2016).

With an annual event rate of 11% in the placebo treatment group, 4500 patients are estimated to provide the required number of primary events, based on an anticipated recruitment period of 18 months and an average follow-up period of approximately 24 months. The assumed placebo event rate of 11% is based on a review of recently published clinical studies in the HFrEF population, including the PARADIGM-HF trial (McMurray et al 2014). The number of patients with incomplete follow-up of endpoints is expected to be small; hence, these are not considered in the determination of the sample size.

This study is a group sequential design study with one interim analysis at 75% of the adjudicated target number of events using Haybittle-Peto boundary (a one-sided alpha=0.001), leaving a one-sided alpha of 2.496% for the final analysis.

2. ANALYSIS SETS

2.1 Definition of analysis sets

2.1.1 Full analysis set

All patients who have been randomized to study treatment will be included in the full analysis set (FAS) irrespective of their protocol adherence and continued participation in the study. Patients will be analysed according to their randomized IP assignment, irrespective of the treatment actually received. The FAS will be considered the primary analysis set for the primary and secondary variables and for the exploratory efficacy variables.

2.1.2 Safety analysis set

All randomized patients who received at least 1 dose of randomized treatment will be included in the safety population. Patients will be analysed according to the treatment actually received. For any patients given incorrect treatment, ie randomized to one of the treatment groups but actually given the other treatment, the treatment group will be allocated as follows: Patients who got both incorrect and correct treatment will be analyzed according to their randomized treatment. Patients who got only the incorrect treatment will be analyzed according to that treatment.

The safety analysis set will be considered the primary analysis set for all safety variables.

2.2 Violations and deviations

The important protocol deviations listed below will be summarised by randomized treatment group

- Patients who were randomised but did not meet inclusion and exclusion criteria
- Patients who received the wrong study treatment at any time during the study.
- Patients who received prohibited concomitant medication, which for this study is limited to open label SGLT2 inhibitors taken in combination with IP.

As the primary analysis is intention-to-treat analysis, protocol deviation will not imply exclusion from the primary analysis.

3. PRIMARY AND SECONDARY VARIABLES

Potential endpoint events and event dates will be adjudicated by an independent clinical event adjudication (CEA) committee. The committee members will not have access to the treatment codes for any patient. The CEA procedures and event definitions will be described in the CEA charter.

The primary and secondary efficacy variables and their components will only include HF events confirmed by the CEA. Deaths adjudicated as ‘cause undetermined’ with regard to CV death or non-CV death will be included as CV death in the primary efficacy analyses.

All adjudicated events from randomization until WoC or PACD will be included in the analysis of primary and secondary endpoints, except for the analysis of all-cause death as a sole outcome, which also will include deaths (not adjudicated) after WoC, but on or before PACD.

For analysis of time to first event, data will be expressed as two variables:

- A binary variable indicating whether the event in question occurred, or the patient was censored.
- An integer variable for the number of days from randomization to the first occurrence of an event (start date of the event – randomization date + 1), or for event free patients, from randomization to censoring (censoring date – randomization date + 1).

Event free patients will be censored as described below for each respective endpoint.

3.1 Primary variable

The primary efficacy variable is time from randomization to the first occurrence of any event in the composite of CV Death, hospitalization for HF or an urgent HF visit.

The three components of the endpoint will be individually adjudicated by the CEA committee.

Patients who did not have an endpoint event will be censored at the earliest of date of WoC or non-CV death when applicable, and otherwise at the earliest of date of last clinical event assessment or PACD. Last clinical event assessment is defined as the last date when the event assessment question for a potential heart failure event was completed on the eCRF event assessment page. It is expected that patients alive and under study follow-up will have a clinical event assessment at their SCV after PACD

In analysis of hospitalization for HF and urgent HF visit to examine the contribution of each component of the composite endpoint, date of death from any cause will be an additional point of censoring.

3.2 Secondary variables

The secondary endpoints are included in a hierarchical testing sequence following the primary endpoint as ordered in Section 3.2.1 - 3.2.5.

3.2.1 The composite of CV death and hospitalization for HF

The efficacy variable is time from randomization to the first occurrence of any event in the composite of CV death and hospitalization for HF. Patients who did not have the endpoint will be censored by the same rule as for the primary endpoint.

3.2.2 Total number of (first and recurrent) hospitalizations for HF and CV death

The efficacy variable is the total number of first and recurrent hospitalizations for HF and CV death, not including urgent HF visit.

For the analysis of recurrent heart failure hospitalization and CV death, the data will be expressed in counting process style for input to the analysis as described in Section 4.2.4.1. The time from randomization to end of follow-up/censoring will be split into one or more interval with variables for start of interval, end of interval and a variable indicating if an event occurred at the end of each respective interval, or if the patient was censored. Recurrent HF hospitalizations, CV death and censoring processes all have continuous distributions so that HF hospitalization and death cannot happen at the same time. If HF hospitalization and CV death occurred at the same day, then only the CV death will be counted.

Patients who did not have the endpoint will be censored by the same rule as for the primary endpoint.

3.2.3 Change from baseline at 8 months in the KCCQ total symptom score

The efficacy variable is the change from baseline at 8 months of the Kansas City Cardiomyopathy Questionnaire (KCCQ) total symptom score (TSS). By the ITT principle, the analysis will include all data irrespective of whether the patient has discontinued study drug.

The KCCQ is a self-administered disease specific instrument for patients with HF **(**Green et al 2000, Spertus et al 2005). The KCCQ consists of 23 items measuring HF-related symptoms, physical limitations, social limitations, self-efficacy, and health-related quality of life. The TSS incorporates the symptom burden and symptom frequency domains into a single score. Scores are transformed to a range of 0-100. Higher scores represent a better outcome.

Baseline is defined as the value at randomization visit (visit 2). Change from baseline at each post-baseline analysis time point will be calculated as the value at the corresponding post-baseline analysis time point minus the baseline value. The KCCQ score is assessed by the patient at randomization, 4, 8, 12 months following randomization, thereafter every 12 months, and at the premature treatment discontinuation visit (PTDV) and the study closure visit (SCV).

In order to account for patients who die prior to the 8-month assessment and to accommodate non-normal distribution of KCCQ scores, a composite rank-based endpoint will be used. The values of change from baseline to 8 months in TSS of patients who survive to 8 months will be converted to ranks (across both treatment groups combined) with lower ranks attributed to worse outcomes (i.e., lower ranks corresponding to negative or smaller values of change from baseline). Patients who die prior to the 8-month assessment will be assigned the worst rank, i.e., worse than any patient surviving to 8 months, but among the deceased the relative ranking will be based on their last value of change from baseline in TSS while alive.

3.2.4 The composite of ≥50% sustained decline in eGFR, end stage renal disease and renal death

The efficacy variable is time from randomization to the first occurrence of any event in the composite of ≥50% sustained decline in eGFR, reaching end stage renal disease (ESRD) and renal death.

ESRD is defined as any of

- Sustained eGFR <15 ml/min/1.73m^2
- Chronic dialysis treatment
- Receiving a renal transplant

All components of the composite endpoint will be defined and described in detail in the CEA charter. The components sustained eGFR decline of ≥50% from baseline and sustained eGFR <15 ml/min/1.73m^2 will however not be adjudicated by the CEA.

Sustained eGFR decline of ≥50% from baseline and sustained eGFR <15 ml/min/1.73m^2 will be based on two consecutive central laboratory values at least 28 days apart below the respective limit. The start date of the event is the date of the first of the two qualifying consecutive central laboratory values. Thus, the analysis will include eGFR events with onset prior to PACD that are confirmed after PACD, in addition to those eGFR events confirmed prior to PACD.

Chronic dialysis will be adjudicated as treatment ongoing for at least 28 days, or when the ESRD is deemed irreversible and the dialysis treatment was stopped before day 28. Similar to the eGFR events, the onset date of chronic dialysis will be the date when the qualifying dialysis treatment started. Chronic dialysis, renal transplant and renal death will be independently adjudicated as defined in the CEA charter. Deaths adjudicated with 'undetermined' cause will not be considered as renal death.

Patients who do not have an endpoint event will be censored at the earliest of date of WoC and non-renal death when applicable, and otherwise at the earliest of date of last clinical event assessment and PACD. Clinical event assessment will be captured by the question for potential renal events on the eCRF event assessment page. The earliest assessment date among components will be used as the censoring date. For example, for the last clinical event assessment, if the central laboratory eGFR assessment is not done in conjunction with the assessment of dialysis and renal transplant, the date of last available central laboratory eGFR measurement will be used as the censoring date.

3.2.5 Death from any cause

The efficacy variable is time to from randomization to death from any cause. All deaths on or prior to PACD, including death after WoC will be included. Patients who are alive will be censored at the earliest of date last known alive and PACD.

Deaths occurring after WoC will not be adjudicated. For such events, the date of death will be collected only in eCRF.

3.3 Safety variables

The safety and tolerability of dapagliflozin will be evaluated from serious adverse events (SAEs), adverse events leading discontinuation (DAEs) of study drug, changes in clinical/haematology parameter and adverse events (AEs) of interest (volume depletion, renal events, major hypoglycaemic events, fractures, diabetic ketoacidosis, AEs leading to amputation and AEs leading to a risk for lower limb amputations ["preceding events"]).

SAEs will be collected from time of informed consent until and including the patent's last visit.

Non-serious AEs will be collected from randomization until and including the patient's last visit, only if it is a DAE, an AE of interest, an AE leading to a potential endpoint or the AE is the reason for interruption of study drug or dose reduction.

Deaths and HF events will be recorded as SAEs in the database, but will not be reported to health authorities to avoid unnecessary unblinding of efficacy endpoints that are fulfilling the SAE criteria. If it is determined by the CEA committee that a potential endpoint does not meet the endpoint criteria, but is judged by investigators to fulfil the SAE criteria, then the event will be reported to AZ patient safety data entry site and if applicable to the health authorities.

4. ANALYSIS METHODS

4.1 General principles

No multiplicity adjustment will be made to confidence intervals as they will be interpreted descriptively and used as a measure of precision. All p-values will be unadjusted. P-values for variables not included in the confirmatory testing sequence, or following a non-significant test in the sequence will be regarded as nominal.

The primary and secondary analyses include adjudicated events occurring on or prior to PACD.

Stratification of analyses for T2D status will be performed using the stratification values as entered in IxRS to determine the randomization assignment.

<u>Baseline diabetic status</u>
T2D at baseline will be defined as established diagnosis of T2D (recorded in eCRF medical history) or central laboratory HbA1c ≥6.5% at both visit 1 (enrollment) and visit 2 (randomisation).

Exploratory analyses may include “pre-diabetic” subjects not categorized as T2D above and having central laboratory HbA1c ≥5.7% at visit 1 and/or visit 2. Other subjects will be categorized to a third group, here denoted “normo-glycaemic”

In consideration of missing data, subjects without medical history of T2D and a single HbA1c ≥5.7% (including ≥6.5%) from visit 1 or visit 2 will be categorized as pre-diabetic. Subjects without medical history of T2D and no HbA1c measurement or a single measurement < 5.7% will be included in the normo-glycaemic group (thus non-diabetic).

<u>Incomplete dates</u>
All efforts should be made to obtain complete dates of clinical assessments and events. For analyses requiring complete dates, partially missing dates will be imputed based on available corroborating information. Absent of any additional corroborating information, partially missing dates will be imputed as follows: if only the year part of a date is available (YY), then the date will be set to YY0701. If only the year and month is available (YYMM), then the date will be set to YYMM15. Additional imputation rules will be defined as appropriate to ensure that eg, dates will not be imputed as prior to randomization, after death or end date before start date.

Baseline laboratory value

For all laboratory variables, with the exception of eGFR, the baseline value is defined as the last value on or prior to date of first dose of randomized study drug, or for patients who did not receive treatment, the last value on or prior to date of randomization.

eGFR

The eGFR values will be calculated (in mL/min/1.73 m^2) from the central laboratory creatinine measurements using the CKD-EPI formula (Levey at al 2009). Descriptive statistics will be presented based on laboratory data.

Baseline eGFR will be calculate as mean of visit 1 and visit 2 values. In case of re-screening, the mean of latest values at each of visit 1 and visit 2 on or prior to date of randomization will be used.

Study drug compliance

The percentage of study drug compliance for the overall treatment period will be derived for each patient based on pill counts as the number of pills taken (dispensed – returned), relative to the expected number of pills taken. The expected number of pills taken is defined as 1*(date of last dose – date of first dose +1), excluding days of interruption.

Study drug compliance will be presented descriptively, including mean, median, quartiles and 5% and 95% percentiles.

4.1.1 Estimand for primary and secondary outcomes

The primary and secondary event based objectives will be evaluated under the treatment policy estimand to reflect the effect of the initially assigned randomized study drug, irrespective of adherence to randomized study treatment. Specifically, the analysis will be performed for the full analysis set including all events that occurred on or prior to PACD, including events following premature discontinuation of study drug.

The estimand for the change from baseline in KCCQ total symptom score at 8 months will employ a combination of a treatment policy strategy and a composite strategy. For the intercurrent event of death (due to any cause) prior to the KCCQ assessment at 8 months, a composite strategy will be used, where death will be considered unfavorable and represented by a lowest (worst) rank of a combined outcome variable as described in Section 3.2.3. For all other types of intercurrent events, including but not limited to a premature discontinuation of randomized treatment, a treatment policy strategy will be used.

4.1.2 Hypotheses

To control the overall type I error rate at 2.5% one-sided, the significance level will be adjusted for one interim analysis of efficacy performed by the DMC (Section 5). For the primary endpoint the following null hypothesis will be tested at the 2.496% one-sided significance level

H0: HR [dapagliflozin:placebo] ≥1

versus the alternative hypothesis

H1: HR [dapagliflozin:placebo] <1

The secondary endpoints included in confirmatory statistical testing using a closed testing procedure (Section 4.1.3) will be based on similar one-sided alternative hypotheses for the respective treatment difference. The significance level will be re-calculated based on the exact proportion of primary endpoints included in the interim analysis and may be marginally adjusted (Section 5).

4.1.3 Confirmatory testing procedure

A closed testing procedure including a pre-specified hierarchical order of the primary and secondary endpoints will be utilized. The Type I error will be controlled at a one-sided 0.02496 level for multiplicity across primary and secondary endpoints and in consideration of the planned interim analysis. Statistical significance will be assessed in the pre-specified order of the endpoints as specified in Section 1.1.1 and 1.1.2. The testing procedure will continue down the hierarchy if the preceding endpoint is rejected at a one-sided 0.02496 level and will stop if the null hypothesis for the preceding endpoint is not rejected at a one-sided 0.02496 level.

If the study is stopped in the efficacy interim analysis (Section 5), testing of secondary endpoints will continue down the hierarchy at one-sided significance level 0.001.

4.1.4 Presentation of time-to-event analyses

In general, summary tables of time-to-event analyses will include the number and percent of patients with event per treatment group, event rate, hazard ratio with 95% confidence interval and p-value. The event rate will be derived as the number of patients with event divided by the total duration of follow-up across all patients in a given group, presented as patients with event per 100 patient years.

Kaplan- Meier (KM) estimates of the cumulative proportion of patients with events will be calculated and plotted per treatment group, with the number of patients at risk indicated below the plot at specific time points. The KM plots will be presented for all time to event analyses, including the individual components of the composite endpoints.

4.1.5 Vital status and follow-up of endpoints

Potential endpoints will be collected and adjudicated from randomization throughout the study until and including the patient’s last visit. The investigator will attempt to collect vital status (dead or alive) at the end of the study for all patients, including vital status from publicly available sources for patients who have withdrawn consent, in compliance with local privacy laws/practices.

Known vital status at the end of the study will be defined when the patient is dead or has date last know alive (derived from the eCRF final status form) on or after the PACD. In patient disposition the number of patients who are dead, alive or with unknown vital status will be reported separately for patients who did/did not withdraw consent. The term lost to follow-up (LTFU) will be limited to only patients with unknown vital status.

Follow-up of the primary endpoint will be defined in terms of completion of the event assessment question for a potential HF event as described for censoring in Section 3.1. Thus a patient that is not LTFU, ie with known vital status, may have incomplete follow-up of endpoints.

Complete follow up of the primary endpoint will be defined as the patient had a primary endpoint event, died from non-CV death or had complete event assessment on or after the PACD.

In addition to the number and percent of patients with complete follow-up, the proportion of total patient time with complete follow-up will be reported per treatment group.
Patient time with complete follow-up will be defined as time from randomization until the earliest of first primary endpoint event, death, WoC, censoring due to incomplete event assessment (in cases where last complete event assessment is prior to PACD) or PACD. The denominator, representing maximum complete follow-up, will be the time to first primary endpoint event, death or PACD.

4.2 Analysis methods

4.2.1 Demographics and baseline characteristics

Demographic and baseline characteristics, including medical history, will be summarized, using frequency distributions and summary statistics based on the FAS data set, for each treatment group as well as for all patients combined. No statistical test will be performed for comparison of any baseline measurement among treatment groups.

4.2.2 Concomitant and baseline medication

Baseline medication is defined as medication with at least one dose taken before date of randomization and with no stop date before date of randomization.

Concomitant medication is defined as medications taken post randomization, irrespective of study drug.

The frequency of baseline and concomitant medication will be presented for the FAS per ATC class and treatment group.

Summaries of prohibited medication (as defined in CSP Section 7.7.1) will be presented. In this study prohibited medication is limited to open label SGLT2 inhibitors taken in combination with IP.

4.2.3 Analysis of the primary efficacy variable

The primary variable is the time to first event included in the primary composite endpoint. The primary analysis will be based on the ITT principle using the FAS, including events with onset on or prior to PACD, adjudicated and confirmed by the CEA committee.

In the analysis of the primary composite endpoint, treatments (dapagliflozin versus placebo) will be compared using a Cox proportional hazards model with a factor for treatment group, stratified by T2D status at randomization, and adjusting for history of hospitalization for heart failure. The analysis will use WoC, non-CV death, last clinical event assessment and PACD for censoring of patients without any primary event as described in Section 3.1. The Efron method for ties and p-value based on the score statistic will be used. The event rates, p-value, HR, and 95% confidence interval will be reported.

The contribution of each component of the primary composite endpoint to the overall treatment effect will be examined. In the analysis of the components, all first event of the given type will be included irrespective of any preceding non-fatal composite event of a different type. Consequently, the sum of the number of patients with events in the component analysis will be larger than the number of patients with composite events. Methods similar to those described for the primary analysis will be used to separately analyze the time from randomization to the first occurrence of each component of the primary composite endpoint.

Kaplan-Meier estimates of the cumulative proportion of patients with event will be calculated and plotted, for the composite endpoint and for the individual components.

4.2.3.1 Subgroup analysis of the primary endpoint

Exploratory subgroup analyses of the primary composite endpoint will be performed for the characteristics listed in Table 1. A test of interaction between randomized treatment group and the subgroup variable will be performed in each Cox model, stratified by T2D and including as covariates history of hospitalization for HF, the relevant subgroup variable and the interaction between treatment and the subgroup variable. In addition to the number and percent of patients with event, event rate estimate, HR with 95% confidence interval and p-value for each subgroup, the interaction p-value will be presented. HRs with confidence interval will be presented in a forest plot, also including the event rate and interaction p-value. The p-values for the subgroup analyses and interaction will not be adjusted for multiple comparisons as the tests are exploratory and will be interpreted descriptively.

Table 1 Characteristics and categories for sub group analysis of the primary endpoint

Characteristic	Categories
Age (years)	<= 65, >65
Sex	Male, female
Race	White, Black or African, Asian, Other
Geographic region	Asia (China, India, Japan, Taiwan, Vietnam) Europe (Bulgaria, Czech Republic, Denmark, Germany, Hungary, Netherlands, Poland, Russia, Slovakia, Sweden, UK) North America (Canada, US) South America (Argentina, Brazil)
NYHA class at enrollment	II, III/IV
LVEF at enrollment	<= median, > median
NT-proBNP at enrollment	<= median, > median
Prior Hospitalization for HF	Yes, No
MRAs at baseline	Yes, No
Type 2 diabetes at baseline *	Yes, No
Atrial fibrillation or flutter at enrolment ECG	Yes, No
Etiology of HF	Ischemic, Non-ischemic/unknown
BMI (kg/m2) at enrollment	<30, ≥30
Baseline eGFR (ml/min/1.73m^2)	<60, ≥60

* Defined as history of T2D or HbA1c ≥6.5% at both visit 1 and visit 2 (Section 4.1) The subgroup analysis by T2D status will exclude T2D as a stratification factor from the model.

The subgroup analyses will be repeated for the CV death component of the primary composite endpoint and the secondary composite endpoint of CV death or HF hospitalization.

No hazard ratio estimates with confidence interval and p-values will be given for subgroups with less than 15 events in total, both treatment groups combined.

4.2.3.2 Sensitivity analysis of the primary endpoint

Undetermined cause of death

A sensitivity analysis of the primary analysis where deaths adjudicated as 'undetermined' cause are not included as endpoint events, but treated as censoring events, will be performed.

Missing data and informative censoring

The time-to-event analysis using the Cox regression depends on the assumption of non-informative or ignorable censoring, corresponding to the missing-at-random assumption. The missing data in this context are patients who are prematurely censored due to WoC, LTFU or otherwise incomplete follow-up of endpoints. The amount of missing data will be described eg, in terms of the number of patients and patient time with incomplete follow-up as described in Section 4.1.5.

Patient retention and follow-up are at the forefront of study planning and conduct, and the amount of incomplete follow-up is expected to be small. To assess the impact of missing data and the robustness of the results with regard to the assumption of non-informative censoring, sensitivity analysis will be planned based on the evaluation of the missing follow-up and discussed in relation to the observed efficacy signal. This may include analysis where scenarios in terms of increased risk in censored patients are explored to identify a 'tipping point' where statistical significance would be lost.

4.2.4 Analysis of the secondary efficacy variables

The time-to-event secondary variables include

- time to the first occurrence of hospitalization for HF or CV death
- time to the first occurrence of any of the components of the renal composite endpoint
- time to death from any cause.

These endpoints will be analysed in the same manner as the primary variable, including stratification by T2D status at randomization. The composite of hospitalization for HF or CV death will be adjusted for history of hospitalization for heart failure. The renal composite endpoint will be adjusted for baseline eGFR. Censoring is described in Section 3.2 for each endpoint.

Additionally, exploratory subgroup analysis will be conducted for the secondary composite endpoint of CV death or HF hospitalization in the same way as the subgroup analysis for the primary endpoint.

4.2.4.1 Analysis of recurrent HF events and CV death

The composite outcome of recurrent HF hospitalizations or CV death will be analysed by the semi-parametric proportional rates model (Lin et al 2000) known as the LWYY (Lin-Wei-Yang-Ying) method, to test the treatment effect and to quantify the treatment difference in terms of the rate ratio with 95% confidence interval and p-value. Recurrent HF hospitalizations, CV death and censoring processes all have continuous distributions so that HF hospitalization and death cannot happen at the same time. If a HF hospitalization and CV death occurred at the same day, then only CV death will be counted. Moreover, CV death is a terminating event which, unlike censoring, prevents the occurrence of new HF hospitalizations.

In addition, the two components in the composite endpoint (total HF hospitalizations and CV death) will be analysed separately to quantify the respective treatment effects and check the consistency between the composite and the components. For the analysis of total HF hospitalizations component, occurrence of CV death can be regarded as semi-competing risk (informative censoring) and may introduce a bias in the treatment effect estimate for HF hospitalizations (dilution of effect size if the drug has a positive effect on both components). In order to address this concern and to account for the correlation between the two components, the joint modelling (frailty model) approach (Rogers et al 2016) will be used for the component analyses.

For joint frailty analysis the following assumptions are made: (Liu et al 2004)

- As noted above, HF hospitalization and death cannot happen at the same day (only the CV death will be counted) and CV death is a terminating event which prevents the occurrence of new HF hospitalizations.
- The hazard function of CV death (the instantaneous risk of death given that the subject is still alive) follows a proportional hazards model with constant baseline hazard, stratified by T2D status at randomization with treatment and history of hospitalization for heart failure as covariates.
- The hazard function of recurrent HF hospitalizations (the instantaneous risk of having another HF hospitalization given that the subject is still alive) follows a proportional hazards model as well, with constant baseline hazard, stratified by T2D status at randomization with treatment and history of hospitalization for heart failure as covariates.
- Within subject recurrent HF hospitalizations and CV death are correlated events. This correlation varies from subject to subject. To account for this correlation, the hazard function of recurrent HF hospitalizations and hazard function of CV death are modeled by a joint a subject specific random frailty term which follows a gamma distribution with unit mean and unknown variance.
- Censoring (including non-CV death) is noninformative, in particular it does not depend on random frailty.

Estimation in the frailty proportional hazards model will be done using Gaussian quadrature approach (Liu and Huang 2008, Lu and Liu 2014, SAS Institute Inc. 2018). Since SAS PROC NLMIXED supports only normal distributions, a probability integral transformation to generate numbers from a gamma distribution will be used (Nelson et al 2006) following Liu and Huang 2008.

Additionally, non-parametric estimates of the marginal mean of the cumulative number of recurrent HF hospitalization rates over time will be calculated allowing for death as terminal event, and the estimates will be plotted (Ghosh and Lin 2000). The estimation will be done in SAS using the PROC PHREG and by treating the situation as competing risks with several records per subject (Andersen et al 2019)

4.2.4.2 Analysis of change from baseline at 8 months in the KCCQ total symptom score

Hypothesis testing
The composite rank-based endpoint representing the patients' vital status at 8 months and the change from baseline to 8 months in TSS in surviving patients, as defined in Section 3.2.3, will be analysed using the rank ANCOVA method (Stokes et al 2012) to test the null hypothesis of no differences in the distributions of ranked outcomes between the two treatment groups. Analysis will be stratified by T2D status at randomisation, and adjusted for the baseline TSS value as follows.

First the change from baseline to 8 months in TSS and vital status at 8 months, as well as values of the baseline TSS covariate will be transformed to standardized ranks, using fractional ranks and mean method for ties. Ranking for the composite endpoint will be done so that patients who died prior to the 8-month assessment are assigned the worst ranks within each stratum. Among the deceased the relative ranking will be based on their last value of change from baseline in TSS while alive before deriving fractional ranks. Then, separate regression models will be fit to the ranked data for each randomization stratum using a regression model for the ranked composite variable as dependent variable, adjusting for the ranked baseline covariate. Residuals from this regression model will be captured for further testing of differences between treatment groups. The Cochran-Mantel-Haenszel (CMH) test, stratified for the T2D status at randomization, using the values of the residuals as scores will be used to compare treatment groups.

Responses missing for reasons other than death will be imputed as described in section 'Handling of missing data".

The p-value from the CMH test of treatment effect at 8 months will be used for the confirmatory statistical testing of the 3rd secondary endpoint in the multiple testing procedure described in section 4.1.3

Estimation of treatment effect

Win ratio:
For a summary statistic that uses the same ranking as that used in the hypothesis test, but has a clinical interpretation, the win ratio (WR) and the corresponding 95% confidence interval (Wang and Pocock 2016) will be reported. It is noted that the WR differs from the statistic used for hypothesis testing, so that exact consistency is not expected between these two analyses, e.g. on rare occasions, the confidence interval for WR could exclude unity while the pre-planned hypothesis test could be non-significant, or the hypothesis test could be

significant with the confidence interval for WR including unity. Formal inference for the superiority of the treatment over control will be made only from the preplanned hypothesis test.

The win ratio represents the odds of having a more favourable outcome versus a less favourable outcome when assigned to the dapagliflozin 10 mg treatment group as opposed to placebo. More specifically, each patient in the dapagliflozin group is compared with each patient in the placebo group and each pair is labelled as “winner”, “loser”, or “tie”, depending on whether the patient on dapagliflozin has a more favourable, less favourable, or the same outcome, respectively, with respect to the composite ranked endpoint compared to the patient on placebo. Win ratio is defined as the ratio of the number of “winner” pairs to the number of “loser” pairs for the dapagliflozin arm. If the estimated win ratio is greater than 1 then the treatment effect is estimated to be in favour of dapagliflozin.

The win ratio statistic adjusted for the randomization stratification factor and baseline TSS will be obtained using the methodology in

Koch et al 1998, Kawaguchi et al 2011 for the stratified Mann-Whitney estimators for the comparison of two treatments with randomization-based covariance adjustment. The win ratio statistic will be calculated as Mann-Whitney odds, i.e., WR=MW⁄(1-MW), where MW is the adjusted Mann-Whitney estimate. This transformation is monotonous in the domain of the Mann-Whitney estimate. The 95% confidence interval for the win ratio will be obtained by transforming the bounds of the confidence interval (

Koch et al 1998) for the Mann-Whitney estimate, using the same transformation as for the win ratio.

<u>Responder analysis:</u>

Number and percentage of patients in each treatment group will be summarized across the following categories:

5 point improvement from baseline to 8 months in TSS vs no significant improvement:
- Change from baseline in TSS ≥ 5 points, vs
- Death prior to the 8 months assessment or change from baseline in TSS < 5 points.

5 point deterioration from baseline to 8 months in TSS vs no significant deterioration:
- Death prior to the 8 months assessment or change from baseline in TSS ≤ -5 points, vs
- Change from baseline to 8 months in TSS > -5 points.

If a patient had a baseline value of TSS ≥ 95 points, ie near the “ceiling”, they will be defined as having a 5 point improvement only if they had TSS ≥ 95 points at 8 months. Similarly, if a patient had a baseline value of TSS ≤ 5 points, ie near the “floor”, they will be defined as having a 5 point deterioration only if they had TSS ≤ 5 points at 8 months.

The proportion of TSS responder categories will be compared between treatment groups using a logistic regression model including treatment group, stratification variable (T2D at randomisation) and baseline TSS value. The observed number of and proportion of TSS responders, odds ratio between treatment groups and its 95% confidence interval and corresponding 2-sided p-value estimated from each imputed dataset will be combined using Rubin's rule, and the combined results will be presented.

Additional responder analysis will be performed in the same way as for 5 points improvement and deterioration described above, using the thresholds of clinically meaningful within-patient change from baseline TSS derived from anchor-based analyses (10 point improvement, 10 point deterioration, 15 point improvement; Appendix B Section 8.2), with "ceiling" and "floor" values handled consistently.

Empirical cumulative distribution function (eCDF) plots will be presented by treatment group to summarize the distribution of change from baseline to 8 months in TSS values, where patients who die prior to the 8-month assessment will be represented with the value of -101 (a value below the worst possible change from baseline).

Handling of missing data

The number of patients with missing vital status at 8 months is expected to be negligible. If some patients are LTFU or patients who withdrew consent have unknown vital status, the main analysis will be done with these patients assigned the worst ranks (same as deaths).

In the context of analysing the composite ranked endpoint as described above, missing data may arise when patients miss the 8-month KCCQ assessment while remaining in the study during the 8-month assessment window, or when patients withdraw consent from the study prior to 8 months. If a patient is known to have died prior to the 8-month assessment, the patient is considered to have a non-missing composite outcome and will be handled as described above (assigned the worst rank). Otherwise, patients who are alive at 8 months and have missing baseline or 8-month KCCQ assessments will have their missing TSS imputed using the multiple imputation (MI) methodology as follows.

Missing TSS values at baseline or at 8 months will be imputed under the Missing at Random (MAR) assumption. The imputation will be done using a predictive mean matching multiple imputation model and a method of Fully Conditional Specification as implemented in the SAS Procedure MI (FCS statement). The predictive mean matching method ensures that the imputed values remain in the permissible range of the TSS values. The imputation model will include the treatment group, T2D randomization stratum, TSS at baseline, month 4, and 8, and a categorical variable representing the number of HF events (0, 1, ≥2 events) in the interval from randomization to 4 month and in the interval from 4 to 8 months. Occurrences of HF events will be determined based on the investigator-reported potential HF events. The auxiliary variable related to HF events is included in the imputation model to improve the imputation accuracy, because the occurrence of HF events is associated with quality of life assessed by KCCQ.

The number of closest observations used to sample an imputed value by the predictive mean matching method will be 5 (SAS default setting).

Each imputed dataset will be analysed using the methods described in the “Hypothesis testing” and “Estimation of treatment effect” sub-sections above. The results from multiple imputed datasets will be combined using Rubin’s rule as implemented in the SAS Procedure MIANALYZE.

- In the analysis of rank ANCOVA, the CMH test statistic used for the hypothesis test has a chi-square distribution. In order to apply Rubin’s combination rule, which assumes approximate normal distribution of the statistics being combined, a normalizing Wilson-Hilferty transformation will be applied to the CMH test statistics from each imputed dataset (Ratitch et al 2013). The standardized transformed statistic will be computed as follows:

$$st_{wh\_cmh}{}^{(m)} = \frac{\sqrt[3]{\frac{cmh^{(m)}}{df}} - \left(1 - \frac{2}{9 \times df}\right)}{\sqrt[2]{\frac{2}{9 \times df}}}$$

  where $cmh^{(m)}$ is the CMH statistic from the m^{th} imputed dataset and df is the number of degrees of freedom associated with the statistic (in this case equal to 1). The transformed statistics are approximately normally distributed with a mean of 0 and variance of 1 and can be combined using Rubin’s rule.

- For the estimation of the win ratio, a combined Mann-Whitney estimate (MW) and its standard error ($SE(MW)$) will first be obtained by applying Rubin’s rule to the corresponding estimates from multiple imputed datasets. Then the win ratio and its 95% confidence interval will be obtained based on the combined Mann-Whitney estimate and its standard error as previously described.

- For the summaries of number and percentage of subjects in the categories of significant improvement and deterioration from baseline, as discussed in the “Estimation of treatment effect” sub-section above, the average number and percent of subjects in each category across all multiple imputed datasets will be reported. The TSS responder analyses will use the imputation datasets created for the main analysis. Therefore, deaths will be defined as non-responders, and responder status will be determined based on the imputed TSS values for the patients who have missing TSS due to reasons other than death.

Supportive analyses and sensitivity analyses

The number and percent of patients who die prior to the 8-month assessment will be summarized by treatment group.

Descriptive statistics of scores and change from baseline at 4, and 8 months will be presented for total symptom score, overall summary score, clinical summary score and domains (Physical limitation, symptom stability, symptom frequency, symptom burden, quality of life, self-efficacy and social limitation).

The testing and estimation described for change from baseline at 8 months in TSS, will be repeated in an exploratory fashion for change from baseline in TSS at 4 months, and for the overall summary score and clinical summary scores at 4 and 8 months.

To assess the impact on TSS change from baseline of a treatment effect on mortality, an alternative ranking will be applied, where patients deceased prior to the 8-month assessment will be assigned the same worst rank regardless of their last value while alive . This is done to reduce the impact of treatment differences in time to CV death on the assessment of this KCCQ secondary endpoint.

4.2.5 Analysis of safety variables

Analysis set
For safety analyses, all summaries will be based on the safety analysis set (Section 2.1.2).

Exposure
The total exposure to study drug will be defined as the length of period on study drug, calculated for each patient as date of last dose – date of first dose +1.

An alternative measure where days of interruption are removed will be calculated and termed actual exposure.

Total and actual exposure will be presented descriptively.

Treatment periods
The summaries for the on-treatment period will include events with an onset date on or after first dose of randomized study drug and on or before 30 days after last dose of study drug. Additional presentations will include all events with onset on or after first dose of study drug regardless of whether patients are on or off study treatment at the time of the event (the 'on +off ' treatment period).

All summaries of AEs and safety lab data described in Section 4.2.5.1 to 4.2.5.6 below will be presented for the on-treatment period. Additional summaries based on the on+off treatment period will be presented for SAEs and AEs of special interest. For fractures and amputations on+off treatment will be considered the primary analysis approach, while the on-treatment period will be the primary analysis approach for other AEs of special interest.

4.2.5.1 Adverse events

Summaries of AEs will primarily be based on the on-treatment period, except for fractures and amputations as described in section 4.2.5

In addition to SAEs, the collection of AEs that are not serious is limited to DAEs, AEs leading to interruption of IP or dose reduction and AEs of interest. Thus, summaries of AEs will be limited to these categories and general summaries of all AEs are not planned.

AEs will be classified according to MedDRA by the medical coding team at AstraZeneca data management centre, using the most current version of MedDRA possible.

Summaries by system organ class (SOC) and preferred term (PT) will be sorted by international order for SOC and by descending order of PT in the dapagliflozin treatment group.

No statistical tests to compare crude AE frequencies between treatment groups will be performed.

A summary table of the total number and percent of patients with SAE, DAE, AE leading to dose reduction and temporary interruption, and for each category of AE of interest per treatment group will be provided.

4.2.5.2 Serious adverse events

SAEs will be presented as described below both on treatment and on+off treatment.

The number and percent of patients with SAEs will be presented by SOC, PT and treatment group. The most common SAEs will also be presented by PT and treatment group only.

AEs with outcome death will be presented separately by SOC and PT.

4.2.5.3 Adverse events leading to discontinuation, interruption or dose reduction

The number and percent of patients will be presented by SOC and PT for AEs leading discontinuation, AEs leading to temporary interruption and dose reduction (separately for each of the three categories based action taken "Drug Permanently Discontinued", "Drug Interrupted" and "Drug Reduced" respectively).

4.2.5.4 Adverse events of interest

Each category of AEs of interest will be presented separately. AEs of interest fall in the categories volume depletion, renal events, major hypoglycaemic events, fractures, diabetic ketoacidosis (DKA), AEs leading to amputation and AEs leading to a risk for lower limb amputations ["preceding events"].

In addition to the pre-specified events of interest, analysis and presentation of potential events of Fournier's gangrene will also be included.

Potential DKA events will be adjudicated. The adjudicated outcome will be considered the main analysis for DKA.

For each AE of interest, a summary table including the number and percent of patients with any event in the AE category, SAE, DAE, AE leading to interruption and dose reduction. Each AE of interest category will also be tabulated with frequency by PT.

In addition to presentations of the number of patients with event, the total number of events counting multiple events per patient will be presented for AE of special interest, where relevant eg for major hypoglycaemic events and amputation.

In addition to the presentation of on-treatment events, on+off presentations will be provided for all AEs of interest.

4.2.5.5 Laboratory Evaluation

All summaries of clinical chemistry/haematology parameters will be based on samples analyzed at the central laboratory, and presented in SI units.

The result and the change from baseline of each clinical chemistry/haematology tests, including estimated GFR , will be summarized by treatment group at each scheduled visit using descriptive statistics, including n, mean, SD, median and quartiles.

4.2.5.6 Marked laboratory abnormalities

The number and percent of patients with a marked abnormality in clinical laboratory tests will be summarized over time by treatment group.

Laboratory abnormalities will be evaluated based on marked abnormality (MA) criteria. The list of MAs is provided in Appendix A (Section 8.1)

An on-treatment value will be considered an MA if either

- the on-treatment value is beyond an MA limit AND the baseline value is not beyond the same limit,
  OR
- both the baseline and on-treatment value are beyond the same MA limit AND the on-treatment values is more extreme (farther from the limit) than was the baseline

Laboratory MAs occurring during the on-treatment period will be summarized by treatment group. The directions of changes (high or low) in MAs will be indicated in the tables. Additionally, for each patient with a MA for a parameter, all the patient’s values of that parameter over the treatment period will be listed.

4.2.6 Analysis of exploratory objectives

The analysis of the exploratory variables will in the same fashion as the primary and secondary efficacy variables be based on the ITT principle, including data irrespective of whether the patient has discontinued study drug.

The time to event endpoints including

- MI
- stroke
- first occurrences of serum potassium falling above/below defined thresholds
- the wider composite reflecting worsening HF (CV death, hospitalization for HF, urgent HF visit or worsening HF symptoms/signs leading to initiation of new treatment or augmentation of existing oral treatment)

will be analysed with the same methods as the primary endpoint.

Change from baseline to each visit for HbA1c, body weight, blood pressure and EQ-5D will be analysed with a repeated measures method. The analysis of HbA1c will be limited to patients with T2D at baseline as defined in Section 4.1. All non-missing visit data will be used, including measurements after discontinuation of study drug. The model will include terms for treatment group, visit, visit*treatment group and baseline measurement as a covariate. The model will be used to derive a least squares estimate of the treatment difference with 95% confidence interval and corresponding two-sided p-value. The missing data will not be imputed.

The proportion of patients with no worsening NYHA classification will be analyzed by a logistic regression by visit with treatment group, baseline NYHA and T2D at randomization. The odds ratio between treatment groups and its 95% confidence interval and corresponding two-sided p-value will be presented.

The proportion of patients with new diagnosis of T2D in non-diabetic patients (as defined in Section 4.1) will be analyzed by a logistic regression with treatment group and baseline HbA1c. The odds ratio between treatment groups and its 95% confidence interval and corresponding two-sided p-value will be presented. In addition to analysing patients who are non-diabetic at baseline, the subgroup of pre-diabetic patients as defined in Section 4.1 will be analysed.

The proportion of patients with new diagnosis of AF in patients without history of AF at baseline will be analysed with logistic regression with treatment group and T2D at randomization.

5. INTERIM ANALYSES

An interim analysis is planned to be performed when approximately 75% of the primary events have been adjudicated, using a Haybittle-Peto rule. There will in principle be one planned interim analysis, with the possibility for the DMC to do subsequent interim analysis if they deem necessary. The significance level for the final analysis will be determined by the Haybittle-Peto function based on the actual number and timing of interim analyses.

The interim analysis will assess superiority of dapagliflozin to placebo. The interim analysis will use a one-sided alpha level of 0.001. At the interim analysis, the primary composite endpoint will be first tested at the specified alpha level. If superiority is achieved for the primary endpoint, then the superiority of dapagliflozin to placebo on CV death will be tested at a one-sided alpha level of 0.001. If CV death is significant, then an action is triggered whereby the DMC will evaluate the totality of the available efficacy data and safety data, to determine if benefit is unequivocal and overwhelming such that the DMC recommends ending the study.

If the interim analysis leads to a decision to terminate the study early based on the pre-defined stopping guidelines for superiority, the executive committee will define a PACD, on or after which study closure visits will commence. The study report will be based on all events occurring on prior to the PACD.

If the study is stopped for superiority at an interim analysis, then testing of secondary endpoints will continue down the hierarchy at a one-sided alpha level 0.001.

6. CHANGES OF ANALYSIS FROM PROTOCOL

The analysis of the KCCQ total symptom score was in the protocol (section 8.5.4) based on a repeated measures analysis. This endpoint will be analysed as a composite rank-based endpoint as described in section 3.2.3 and 4.2.4.2.

7. REFERENCES

Andersen et al 2019
Andersen PK, Angst J, Ravn H. Modeling marginal features in studies of recurrent events in the presence of a terminal event. Lifetime data anal 2019 Jan: 1-15.

Coon and Cook 2018
Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 2018 Jan; 27(1):33-40.

De Vet et al 2010
De Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, Hendriks EJ, Tertwee CB. Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol 2010 Jan; 63(1):37-45.

Fitchett et al 2016
Fitchett D, Zinman B, Wanner C, Lachin JM, Hantel S et al. Heart failure outcomes with empagliflozin in patients with type 2 diabetes at high cardiovascular risk: results of the EMPA-REG OUTCOME® trial. Eur Heart J 2016; 37 (19): 1526-1534

Froud and Abel 2014
Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of pythagoras. Theoretical considerations and an example application of change in health status. 2014 PloS ONE 9(12): e114468.

Ghosh and Lin 2000
Ghosh D, Lin DY. Nonparametric analysis of recurrent events and death. Biometrics 2000; 56:554–562.

Green et al 2000
C. Patrick Green, MD, Charles B. Porter, MD, FACC, Dennis R. Bresnahan, MD, FACC, John A.Spertus, MD, MPH, FACC Development and Evaluation of the Kansas City Cardiomyopathy Questionnaire: A New Health Status Measure for Heart Failure. JACC, Vol.35 No 5, April 2000:1245-55

Kawaguchi et al 2011
Kawaguchi A, Koch G, Wang X. Stratified Multivariate Mann-Whitney Estimators for the Comparison of Two Treatments with Randomization Based Covariance Adjustment. Statistics in Biopharmaceutical Research, 2011;3(2), 217–231.

Koch et al 1998
Koch GG, Tangen CM, Jung JW, Amara IA. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. Statist. Med 1998; 17(15-16), 1863-1892.

Levey at al 2009
Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, Kusek JW, Eggers P, VanLente F, Greene T, Coresh J; CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009 May 5; 150(9):604-12

Lin et al 2000
Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2000; 62(4):711–730.

Liu et al 2004
Liu L, Wolfe RA, Huang X. Shared Frailty Models for Recurrent Events and a Terminal Event. Biometrics, 2004;60(3);747-756

Liu and Huang 2008
Liu L, Huang X. The use of Gaussian quadrature for estimation in frailty proportional hazards models. Statist. Med. 2008; 27(14):2665–2683

Lu and Liu 2014
Lu L, Liu C. Analysis of Correlated Recurrent and Terminal Events Data in SAS. Proceedings of the Northeast SAS Users Group Conference 2008

McMurray et al 2014
McMurray JJ, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR, Rouleau JL, Shi VC, Solomon SD, Swedberg K, Zile MR; PARADIGM-HF Investigators and Committees. Angiotensin-neprilysin inhibition versus enalapril in heart failure. N Engl J Med 2014; 371 (11):993-1004

Nelson et al 2006
Nelson KP, Lipsitz SR, Fitzmaurice GM, Ibrahim J, Parzen M, Strawderman R. Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. J Comp Graph Stat 2006; 15 39-57

Ratitch et al 2013
Ratitch B, Lipkovich I, O’Kelly M. Combining Analysis Results from Multiply Imputed Categorical Data. PharmaSUG 2013 - Paper SP03

Rogers et al 2016
Jennifer K Rogers, Alex Yaroshinsky, Stuart J Pocock, David Stokar and Janice Pogoda. Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model. Statist. Med. 2016, 35 2195–2205

SAS Institute Inc. 2018
SAS/STAT® 15.1 User’s Guide. Cary, NC: SAS Institute Inc. 2018. Example 86.5: Failure Time and Frailty Model, page 7097

Spertus et al 2005
Spertus J Am Heart J 2005. Monitoring clinical changes in patients with heart failure: A comparison of methods. Multicenter Study

Stokes et al 2012
Stokes ME, Davis CS, Koch GG.Categorical Data Analysis Using the SAS System. Carry, NC: SAS Institute Inc.

Wang and Pocock 2016
Wang D, Pocock S. A win ratio approach to comparing continuous non-normal outcomes in clinical trials. Pharmaceut. Statist. 2016, 15 238–245

8. APPENDIX

8.1 Appendix A Laboratory Abnormality Criteria

Table 2 provides the criteria for assessing marked abnormalities in safety laboratory parameters. When there is more than one limit for a variable, summaries will be provided for each limit.

If both the baseline and on-treatment values of a parameter are beyond the same MA limit for that parameter, then the on-treatment value will be considered a MA only if it is more extreme (farther from the limit) than was the baseline value.

The following three criteria will also be summarized by treatment group in examination of the elevated AT (ALT and/or AST) and total bilirubin:

- (AST or ALT > 3XULN) and (Total Bilirubin > 1.5XULN within 14 days on or after AT elevation)
- (AST or ALT > 3XULN) and (Total Bilirubin > 2XULN within 14 days on or after AT elevation)

(AST or ALT > 3XULN) and {(Total Bilirubin > 2XULN and no ALP ≥ 2XULN) within 14 days on or after AT elevation}

Table 2 Marked abnormality criteria for safety laboratory variables and elevated AT (ALT and/or AST) and total bilirubin

		Marked Abnormality Criteria	
Clinical laboratory variables	**Units**	**Low**	**High**
Hematology			
HCT	vol	< 0.20	> 0.55
HCT	vol		> 0.60
Hemoglobin	g/L	< 60 g/L	> 180 g/L
Hemoglobin	g/L		> 200 g/L
Blood Chemistry			
ALP	U/L		> 1.5X ULN
ALP	U/L		> 3X ULN
ALT	U/L		> 3X ULN
AST	U/L		> 3X ULN

		Marked Abnormality Criteria	
Clinical laboratory variables	**Units**	**Low**	**High**
AST or ALT	U/L		> 3X ULN
ALT	U/L		> 5X ULN
AST	U/L		> 5X ULN
AST or ALT	U/L		> 5X ULN
ALT	U/L		> 10X ULN
AST	U/L		> 10X ULN
AST or ALT	U/L		> 10X ULN
ALT	U/L		> 20X ULN
AST	U/L		> 20X ULN
AST or ALT	U/L		> 20X ULN
Total Bilirubin	µmol/L		> 1.5X ULN > 2X ULN
Na (Sodium)	mmol/L	< 130 mmol/L	> 150 mmol/L
Na (Sodium)	mmol/L	< 120 mmol/L	
K (Potassium)	mmol/L	≤ 2.5 mmol/L	≥ 6.0 mmol/L
Creatinine	µmol/L		≥1.5X BL CREAT
Creatinine	µmol/L		≥2X BL CREAT

BL is the baseline measurement

8.2 Appendix B Estimation of clinically meaningful thresholds for KCCQ total symptom score

B1 Methods

Thresholds for clinically meaningful within-patient change (CMWPC) will be estimated according to predefined algorithms using an anchor-based approach, supplemented with graphical visualizations of the distribution across anchor categories. Clinically meaningful thresholds will be estimated for change from baseline TSS at 8 months.

Section B1 describes the methods which were applied to blinded study data, with results and derived thresholds described in section B2 below. The analysis was performed on the FAS population, on blinded study data across both treatment arms only including patients with complete data at baseline and 8 months.

Anchor-based approaches

Anchor-based approaches estimate a threshold by 'anchoring' the results on a separate variable, often a patient-reported outcome. The anchor-based analysis will employ the patient global impression of severity (PGIS) in HF symptoms. Meaningful change will be evaluated using observed scores according to a predefined algorithm. The responses to PGIS at baseline and 8 months will be used in the analysis.

Categorization of anchors

The change from baseline PGIS at 8 months will be categorized and categories will be collapsed in different ways, to provide a clearer distinction between patients who have and have not experienced a meaningful change according to this anchor.

The ordinal responses to PGIS at baseline and 8 months will be assigned the following numeric values:

- 1 ('no symptoms')
- 2 ('very mild')
- 3 ('mild')
- 4 ('moderate')
- 5 ('severe')
- 6 ('very severe')

Change from baseline PGIS at 8 months will be categorized as small, moderate or large improvement/deterioration or stable as defined in Table 3.

Table 3 Categories of change from baseline PGIS at 8 months

		PGIS at 8 months					
		No symptoms	**Very mild**	**Mild**	**Moderate**	**Severe**	**Very Severe**
PGIS at baseline		**1**	**2**	**3**	**4**	**5**	**6**
No symptoms	**1**	0 Stable	+1 SD	+2 MD	+3 LD	+4 LD	+5 LD
Very mild	**2**	-1 SI	0 Stable	+1 SD	+2 MD	+3 LD	+4 LD
Mild	**3**	-2 MI	-1 SI	0 Stable	+1 SD	+2 MD	+3 LD
Moderate	**4**	-3 LI	-2 MI	-1 SI	0 Stable	+1 SD	+2 MD
Severe	**5**	-4 LI	-3 LI	-2 MI	-1 SI	0 Stable	+1 SD
Very severe	**6**	-5 LI	-4 LI	-3 LI	-2 MI	-1 SI	0 Stable

LD Large deterioration. MD Moderate deterioration. SD Small deterioration. SI Small improvement. Moderate improvement. LI Large improvement.

The categories in Table 3 will be further collapsed as

- 'moderate or large deterioration' in the categorization with 5 categories (version A)
- 'small or moderate deterioration' in the categorization with 5 categories (version B)
- 'small or moderate improvement' in the categorization with 5 categories (version B)
- 'moderate or large improvement' in the categorization with 5 categories (version A)

The change from baseline TSS at 8 months, will be used repeatedly in the anchor-based analyses. To explore the adequateness of each anchor categorization, the Spearman correlation coefficient between change from baseline TSS and change from baseline PGIS at 8 months will be assessed.

The larger the correlation coefficient between an anchor and the endpoint, the greater the confidence in the classifications. An anchor is considered adequate if it has a correlation coefficient of 0.3 or greater (Coon and Cook 2018).

Descriptive statistics (mean, SD, median, quartiles, minimum and maximum), empirical cumulative distribution function (eCDF) will be presented for each categorization. The eCDF curves display a continuous plot of the change from baseline on the horizontal axis, and the cumulative proportion of patients experiencing changes from baseline up to that level, on the vertical axis. If the eCDF curves show very poor distinction between categories, they may be complemented with curves illustrating the probability density function for that categorization.

Additionally, to better characterize the association between change from baseline in each endpoint, and CMWPC according to the anchors, the sensitivity and specificity will be calculated based on receiver operating characteristic (ROC) curve analysis using logistic regression analyses for each cut-off in change from baseline TSS. The following categories will be examined:

- Large improvement
- Moderate or large improvement
- Moderate or large deterioration
- Large deterioration

The CMWPC thresholds derived from analysis of ROC curves, for each anchor category, will be defined by the change value corresponding to the cut-point in the ROC space which minimizes the sum of squares of 1-sensitivity and 1-specificity, ie, which is closest to the top-left corner in the ROC space (Froud and Abel 2014). The certainty with which an estimated threshold can be used will depend on its corresponding sensitivity and specificity values. It is commonly not recommended to apply thresholds to individual patients when sensitivity and 1-specificity are lower than 75% (De Vet et al 2010).

Establishing the clinically meaningful threshold

The various estimates from the different streams of evidence (tables and plots of the distribution, ROC curve analyses) will be examined for convergence in an effort to triangulate onto a single threshold value which represents CMWPC (for improvement and deterioration, respectively) and the TSS responder analysis will be performed for this threshold. However, if the values are too disparate, a range of clinically relevant thresholds may be identified. CMWPC thresholds identified will be indicated in the eCDF for change from baseline TSS by treatment, in the unblinded results, and responder analysis will be performed for the thresholds.

B2 Summary of results of anchor-based analysis on blinded study data

The descriptive statistics for change from baseline KCCQ-TSS at 8 months in different categories of change from baseline PGIS at 8 months in Table 4 suggest that a large improvement in PGIS corresponds to a median increase in TSS of 15 points. A moderate or large improvement in PGIS (5 categories, version A) corresponds to a median increase in KCCQ-TSS of about 10 points, whereas a small or moderate improvement (5 categories, version B) corresponds to a median increase in KCCQ-TSS of about 8 points. Similarly, a moderate or large deterioration in PGIS corresponds to a median decrease in KCCQ-TSS of about 6 points, whereas a small or moderate deterioration corresponds to a median decrease in KCCQ-TSS of about 2 points. The mean values are generally more extreme, as would be expected, but as this is a complete-case analysis (excluding deaths and not accounting for baseline values which are near the "ceiling" and "floor" of the KCCQ-TSS scale) the median values represent more reliable estimates of average change.

Table 4 Distribution of change from baseline KCCQ-TSS at 8 months by change from baseline PGIS at 8 months

	N	(%)	Mean	SD	Min	Q1	Median	Q3	Max	Correlation*
PGIS at 8 Months: 7 Categories										0.35
Large Improvement	255	(7)	20.0	22.7	-22.9	2.1	15.6	34.4	90.6	
Moderate Improvement	402	(10)	11.9	18.7	-35.4	0.0	9.4	20.8	72.9	
Small Improvement	831	(21)	9.6	17.4	-49.0	0.0	8.3	17.7	80.2	
Stable	1522	(39)	3.2	15.5	-63.5	-4.2	1.0	10.4	82.3	
Small Deterioration	561	(14)	-3.0	18.2	-71.9	-11.5	-2.1	7.3	50.0	
Moderate Deterioration	223	(6)	-6.9	20.8	-100.0	-17.7	-4.2	6.3	43.7	
Large Deterioration	89	(2)	-11.9	21.6	-65.6	-25.0	-8.3	0.0	54.2	
PGIS at 8 Months: 5 Categories (Version A)										0.34
Moderate or Large Improvement	657	(17)	15.0	20.7	-35.4	0.0	10.4	26.0	90.6	
Small Improvement	831	(21)	9.6	17.4	-49.0	0.0	8.3	17.7	80.2	
Stable	1522	(39)	3.2	15.5	-63.5	-4.2	1.0	10.4	82.3	
Small Deterioration	561	(14)	-3.0	18.2	-71.9	-11.5	-2.1	7.3	50.0	
Moderate or Large Deterioration	312	(8)	-8.4	21.1	-100.0	-19.3	-6.3	4.2	54.2	
PGIS at 8 Months: 5 Categories (Version B)										0.34
Large Improvement	255	(7)	20.0	22.7	-22.9	2.1	15.6	34.4	90.6	
Small or Moderate Improvement	1233	(32)	10.3	17.9	-49.0	0.0	8.3	19.8	80.2	
Stable	1522	(39)	3.2	15.5	-63.5	-4.2	1.0	10.4	82.3	
Small or Moderate Deterioration	784	(20)	-4.1	19.0	-100.0	-12.5	-2.1	6.8	50.0	

Large Deterioration	89	(2)	-11.9	21.6	-65.6	-25.0	-8.3	0.0	54.2	

[a] Absolute value of the Spearman correlation coefficient for change from baseline KCCQ-TSS at 8 months and change from baseline PGIS at 8 months with each categorization
Categories of change from baseline PGIS at 8 months as defined in Table 3

The eCDF curves in Figure 2 demonstrate a clear separation between all categories of deterioration and the no change category, in the interval between 5 and 20 points decrease in KCCQ-TSS at 8 months. For improvement, the separation is clear between improvement and no change in the interval between 5 and 20 points increase in KCCQ-TSS at 8 months, but the distinction is not as clear between small and moderate improvement. From

Figure 3 it’s evident that the combined moderate or large categories of deterioration and improvement are clearly separated from the no change category. This is also observed for combined small or moderate categories of deterioration and improvement in Figure 4.

Figure 2 Empirical cumulative distribution function for change from baseline KCCQ-TSS at 8 months versus change from baseline PGIS at 8 months with 7 categories

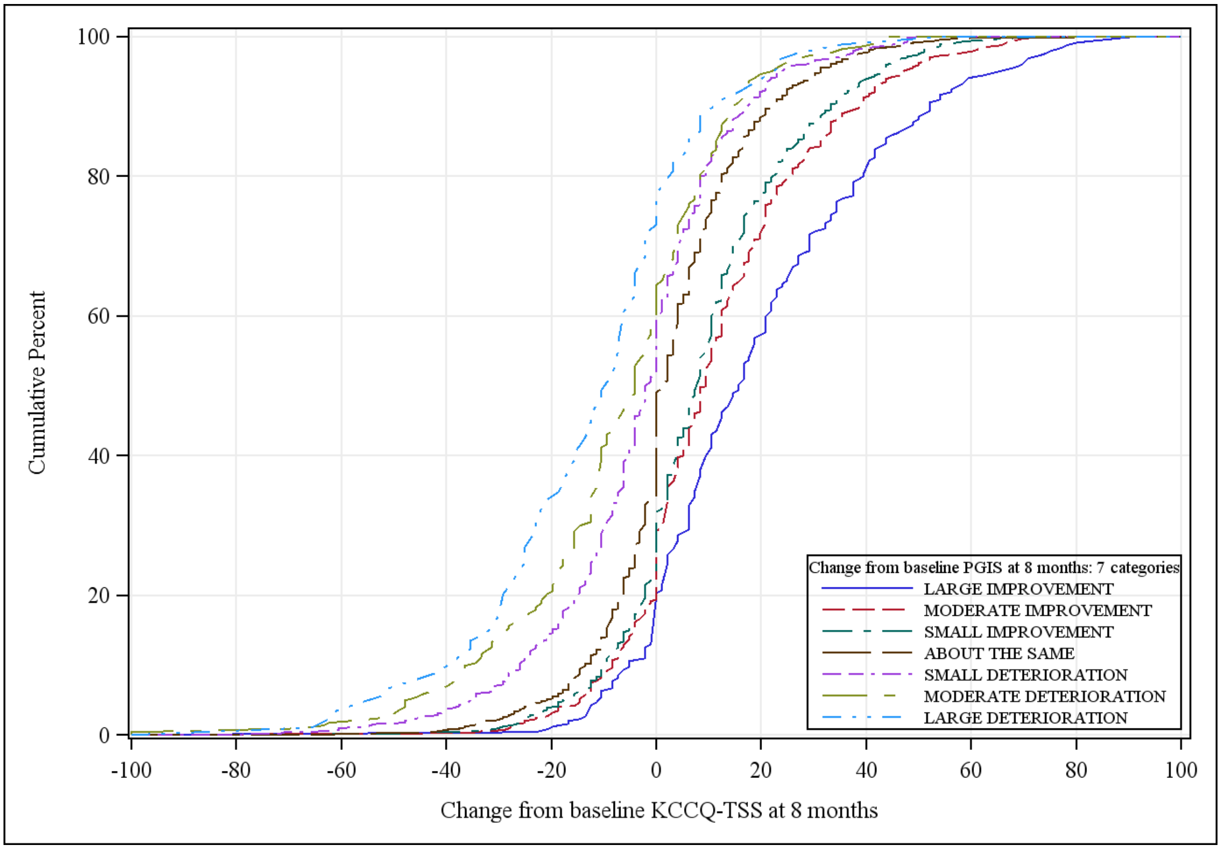


Figure 3 Empirical cumulative distribution function for change from baseline KCCQ-TSS at 8 months versus change from baseline PGIS at 8 months with 5 categories (version A)

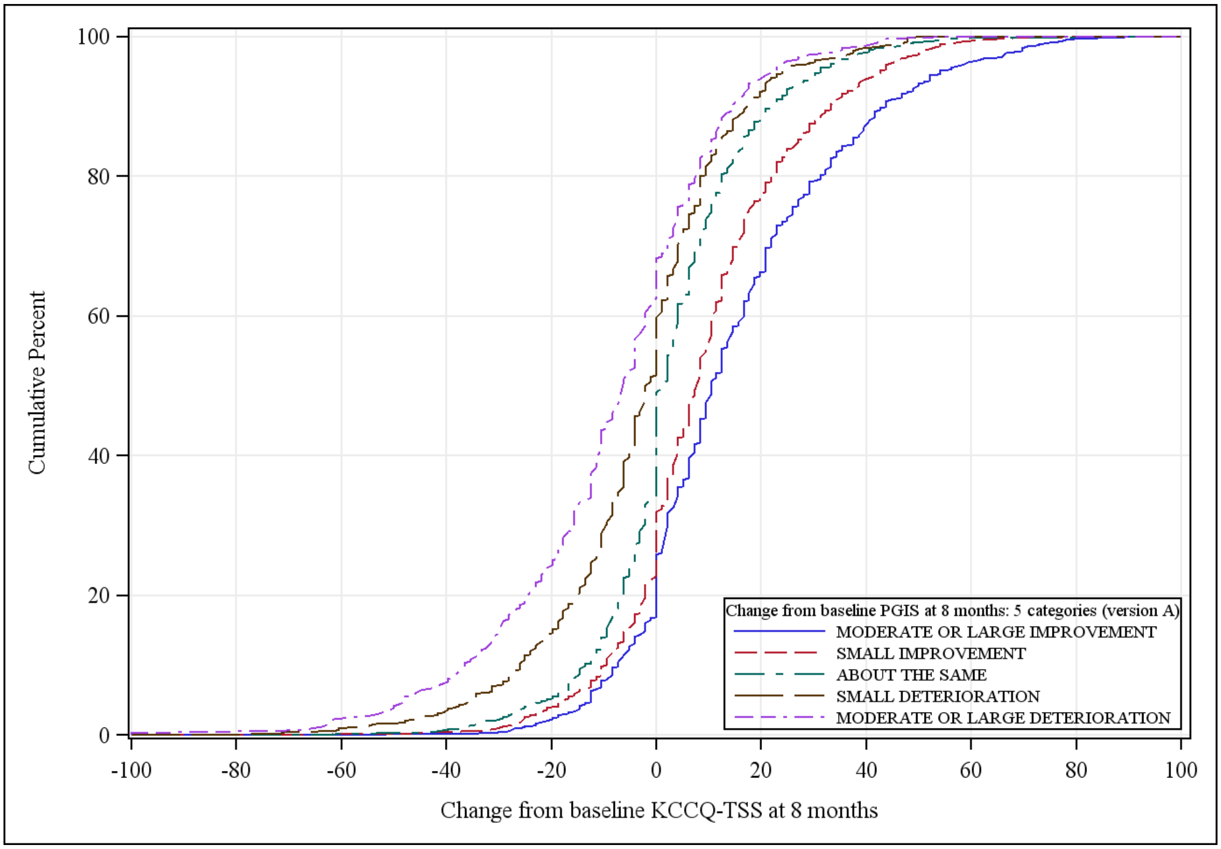
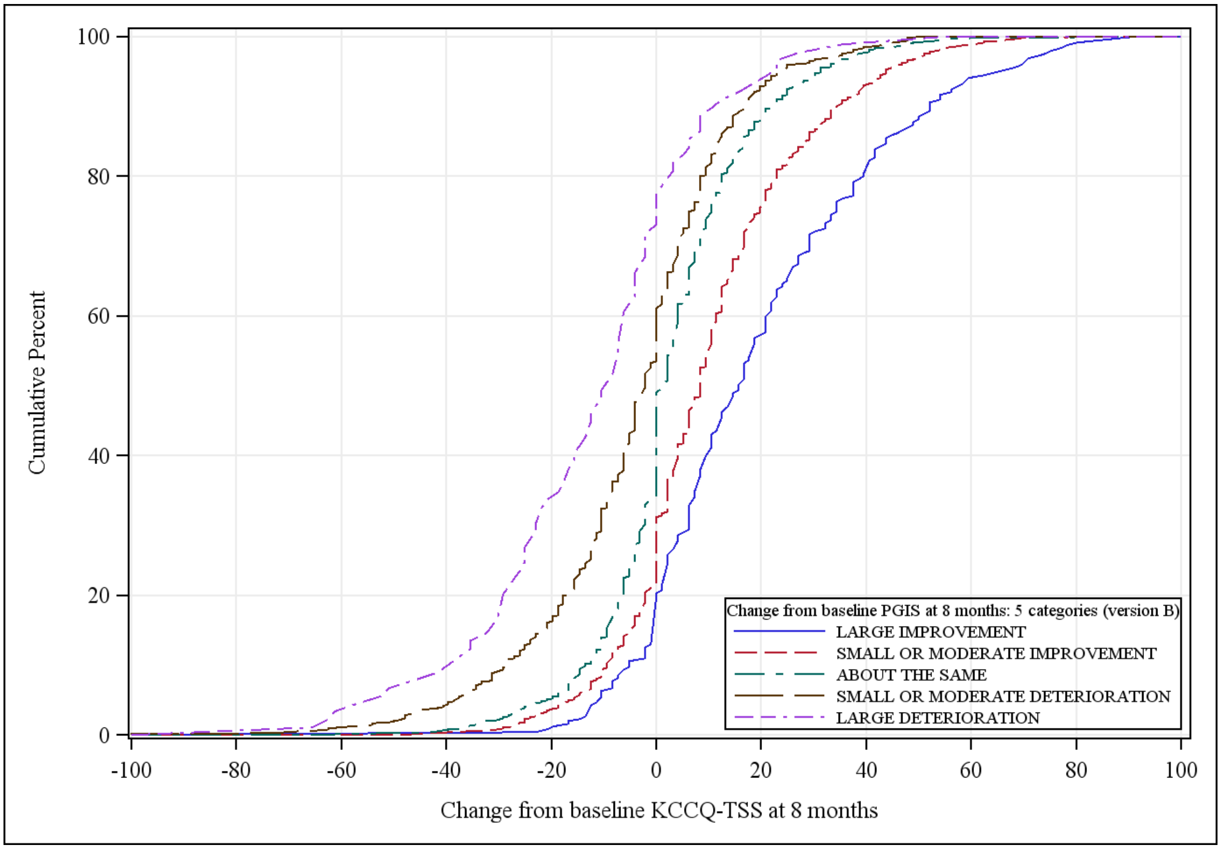


Figure 4 Empirical cumulative distribution function for change from baseline KCCQ-TSS at 8 months versus change from baseline PGIS at 8 months with 5 categories (version B)



The thresholds from analysis of ROC curves, in

Table 5 suggest that an increase in KCCQ-TSS at 8 months of 9 points may be a sufficient cut-off to identify a moderate or large improvement, or even a large improvement, based on ROC curve analysis. For deterioration, the suggested cut-off for a large deterioration is a decrease in KCCQ-TSS at 8 months of 2 points. The sensitivity and specificity of these cut-offs are not very high.

Table 5 Characteristics of thresholds for change from baseline in KCCQ-TSS at 8 months as assessed by PGIS at 8 months based on ROC analysis

	Estimated threshold[a]	**Sensitivity**[b]	**Specificity**[c]
Change from baseline PGIS at 8 months			
Large improvement	9.4	0.62	0.68
Moderate or large improvement	7.3	0.60	0.65
Moderate or large deterioration	-1.0	0.62	0.69
Large deterioration	-2.1	0.72	0.69

[a] Anchor-based threshold for clinically meaningful within-patient change is defined as the point on the ROC curve which minimizes the sum of squares of (1-sensitivity) and (1-specificity).

[b] Sensitivity is defined as the true positive rate for each response threshold, i.e. the number identified as responders by both KCCQ-TSS and change from baseline PGIS at 8 months divided by the total number identified as responders by change from baseline PGIS at 8 months.

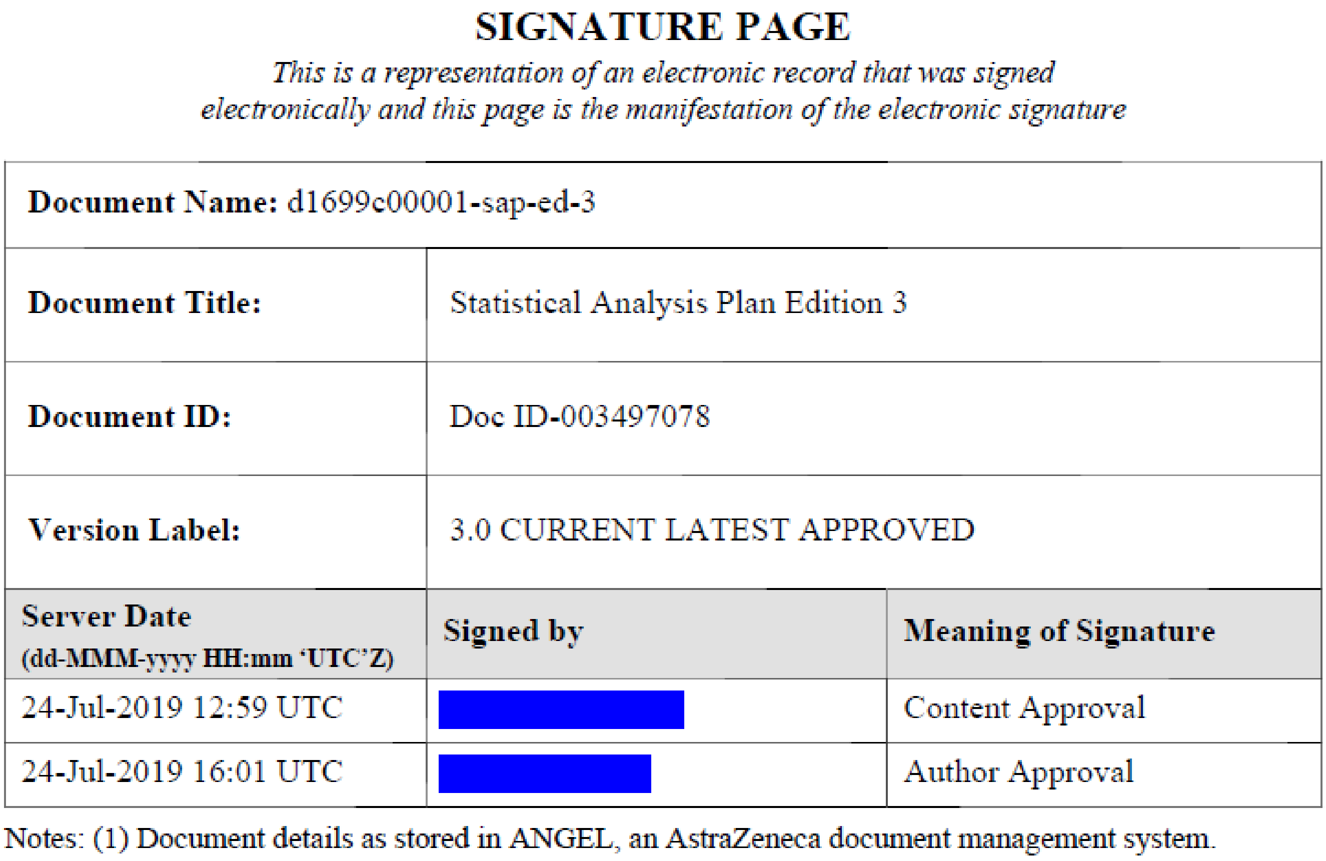
[c] Specificity is defined as the true negative rate for each response threshold, i.e. the number identified as non-responders by both KCCQ-TSS and change from baseline PGIS at 8 months divided by the total number identified as non-responders by change from baseline PGIS at 8 months.

Selection of thresholds for supportive responder analysis

Based on the results of the threshold analysis summarized in this appendix and clinical discussions regarding CMWPC it was decided that a clinically relevant and achievable threshold for improvement, in the study population, is estimated to correspond to an increase in KCCQ-TSS at 8 months of between 5 and 10 points, with a large improvement possibly corresponding to an increase of 15 points. A clinically relevant deterioration is estimated to correspond to a decrease in KCCQ-TSS at 8 months of between 5 and 10 points.

In the analyses (see section 4.2.4.2) of the third secondary efficacy endpoint, change from baseline measured at 8 months in the total symptom score of the KCCQ, decreases in KCCQ-TSS at 8 months of 5 and 10 points and increases of 5, 10 and 15 points will be indicated in the eCDF for change from baseline TSS by treatment. In addition to the responder analyses of

5 points improvement and deterioration, further supportive analyses will be performed for the 10 point deterioration, 10 point improvement and 15 point improvement CMWPC thresholds.



SIGNATURE PAGE

This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature

Document Name: d1699c00001-sap-ed-3		
Document Title:	Statistical Analysis Plan Edition 3	
Document ID:	Doc ID-003497078	
Version Label:	3.0 CURRENT LATEST APPROVED	
Server Date **(dd-MMM-yyyy HH:mm 'UTC'Z)**	**Signed by**	**Meaning of Signature**
24-Jul-2019 12:59 UTC	[redacted]	Content Approval
24-Jul-2019 16:01 UTC	[redacted]	Author Approval

Notes: (1) Document details as stored in ANGEL, an AstraZeneca document management system.