Statistical Analysis Plan	
Study Code	D169EC00001
Edition Number	4.0
Date	21 September 2020

An International, Multicentre, Parallel-group, Randomised, Double-blind, Placebo-controlled, Phase III Study Evaluating the effect of Dapagliflozin on Exercise Capacity in Heart Failure Patients with Preserved Ejection Fraction (HFpEF)

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation or special term	Explanation
6MWD	6-minute walk distance
6MWT	6-minute walk test
AE	Adverse event
AF	Atrial fibrillation/flutter
ANCOVA	Analysis of covariance
ATC	Anatomical therapeutic chemical
BP	Blood pressure
CKD-EPI	Chronic kidney disease epidemiology collaboration equation
CMH	Cochran-Mantel-Haenszel test
CMWPC	Clinically meaningful within-patient change
COVID-19	Corona virus disease 2019
CSP	Clinical study protocol
CSR	Clinical study report
CV	Cardiovascular
DAE	Adverse event leading to discontinuation of investigational product
DKA	Diabetic ketoacidosis
eCRF	Electronic case report form
eGFR	Estimated glomerular filtration rate
EQ-5D-5L	EuroQol five-dimensional five-level questionnaire
ePRO	Electronic patient-reported outcome
FAS	Full analysis set
FWER	Family wise error rate
HF	Heart failure
HFpEF	Heart failure with preserved ejection fraction
IP	Investigational product (dapagliflozin or matching placebo)
ITT	Intent to treat
IxRS	Interactive voice/web response system
KCCQ	Kansas city cardiomyopathy questionnaire
LVEF	Left ventricular ejection fraction
LVPA	Light to vigorous physical activity
MedDRA	Medical dictionary for regulatory activities
METs	Metabolic Equivalent of Task units
MI	Multiple imputation
MRI	Magnetic resonance imaging
MVPA	Moderate to vigorous physical activity

Abbreviation or special term	Explanation
NC	Not calculable
NT-proBNP	N-terminal pro b-type natriuretic peptide
NYHA	New York heart association
OTB	Overall treatment benefit
PGIS	Patient global impression of severity
PGIC	Patient global impression of change
PLS	Physical limitation score
PK	Pharmacokinetic
PRO	Patient-reported outcome
PT	MedDRA preferred term
PTDV	Premature treatment discontinuation visit
Q1	First quartile
Q3	Third quartile
QoL	Quality of life
SAE	Serious adverse event
SAP	Statistical analysis plan
SOC	MedDRA system organ class
T2DM	Type 2 diabetes mellitus
TPA	Tipping point analysis
TSS	Total symptom score
VMUs	Vector magnitude units
WHO	World health organisation

AMENDMENT HISTORY

Date	Brief description of change
08 April 2019 / Version 1.0	Updated edition number in document to first full version (1.0) and updated the document date in the header. No other changes to document content.
15 April 2020 / Version 2.0	Updated according to protocol amendment. Amended to reflect the following (editorial updates are not listed): • The former primary and key secondary objectives, 6MWD and KCCQ-TSS, were combined into dual primary objectives • Sample size was increased from 400 to 500 to increase the power in evaluating the effect of dapagliflozin on KCCQ-TSS • The hypothesis testing strategy was changed from a fixed sequence approach to a weighted Bonferroni approach, allocating alpha between the primary and secondary efficacy endpoint families • Changed secondary efficacy endpoints, including vector magnitude, serum NT-proBNP, time spent in light to vigorous physical activity, and NYHA functional classification, to be exploratory endpoints • Atrial fibrillation was changed to atrial fibrillation/flutter • Exploratory endpoints, PGIS in HF symptoms, PGIC in HF symptoms, PGIC in walking ability, and Borg CR10 Scale@ for perceived dyspnea and fatigue during 6MWT, were removed. PGIS and PGIC are intentionally simple instruments designed for use in anchor-based analyses and not for comparing treatment groups • Missing data at baseline will be imputed assuming it is missing at random and clarification was added regarding which efficacy variables undergo multiple imputation • Derivation of visit-specific summaries for *MoveMonitor* measurements was clarified • Added the probability distribution function curves to anchor-based analyses

10 September 2020 / Version 3.0	Updated according to protocol amendment. Amended to reflect the following (editorial updates are not listed): • The former dual primary (6MWD and KCCQ-TSS) and exploratory (KCCQ-PLS) objectives were modified into three primary objectives • The former exploratory (total time spent in LVPA) objective was changed to secondary objective • The former secondary (movement intensity during walking) objective was changed to exploratory objective • Statistical power was re-estimated for a comparison of group-level averages of within-patient change, instead of responder analysis (as in the CSP) due to a change of main estimation method for effect size from logistic regression to the HL estimate of median difference. • Details were added regarding how alpha is distributed and propagated among the endpoints under type I error control. • In analyses of primary efficacy endpoints and secondary efficacy endpoint, ranking among the deceased patient in the primary analysis was changed to based on last value while alive, and ranking among the deceased patients in the sensitivity analysis was changed to based on time to death • Added median estimate within each treatment group and Hodges-Lehmann estimate with the corresponding distribution-free 95% confidence interval for the difference between treatment group as treatment effect in the analyses of primary efficacy endpoints and secondary efficacy endpoint • Added EQ-5D-5L question: "Usual activities" in the anchor-based analysis of KCCQ-PLS. Added the derivation of categories for this anchor variable based on EQ-5D-5L question: "Usual activities" at baseline and week 16 in Appendix 8.3. • Clarified mean change value in the endpoint (KCCQ-TSS, KCCQ-PLS, or 6MWD) corresponding to the category ‘moderate or large improvement’ for each anchor variable will be used when determining the threshold for CMWPC that will be used to define responder • Removed the anchor-based analysis and responder analysis for secondary efficacy endpoint, total time spent in LVPA

	• Described the impact of COVID-19 pandemic on trial conduct, data collection, protocol deviations, handling of missing data not due to death, and analyses • Added description of missing at random (MAR) imputation based on pre-COVID-19 data only. • Described the derivation of the categorical variable, weeks in the study impacted by COVID-19, and include the variable in the main analysis to adjust for COVID-19 impact for primary efficacy endpoints KCCQ-TSS and KCCQ-PLS • Clarified the analyses of 6MWD will remain unchanged and not impacted by COVID-19 • Added supportive analysis for KCCQ-TSS and KCCQ-PLS using placebo-based imputation dataset and not adjusting for COVID-19 impact • Added supplemental analysis using mixed-effect model for repeated measures (MMRM) for KCCQ-TSS and KCCQ-PLS • Added supportive summary statistics for KCCQ-TSS, KCCQ-PLS and 6MWD, separately for data collected prior to COVID-19 and data collected during COVID-19, based on the onset date of COVID-19 at each site • Clarified main analysis and supportive analysis for exploratory KCCQ endpoints • Clarified analyses of all *MoveMonitor* and *MoveTest* endpoints, including secondary efficacy endpoint, total time spent in LVPA, and exploratory efficacy endpoints, will use complete data, ie no patients with missing data due to reasons other than death. Only KCCQ endpoints (KCCQ-TSS, KCCQ-PLS, 6 domain scores and 1 summary score included in the exploratory efficacy endpoints) and 6MWD will use imputed datasets • Clarified subgroup analyses when subgroup category of a factor contains less than ten percent of the patients • Added Appendix 8.4 to describe the scoring algorithm of KCCQ endpoints

21 September 2020 / Version 4.0	Amended to reflect the following (editorial updates are not listed): • Added the paragraph in Section 4.2.4 to clarify the summaries based on non-missing data for KCCQ-TSS, KCCQ-PLS and 6MWD will include the medians and 1st and 3rd quartiles of change from baseline for each treatment group using complete data, ie, including patients who died prior to week 16 but excluding patients with missing data due to reasons other than death and a temporary value which is lower than all observed negative change values will be assigned to the deceased patients, and the assigned value amongst the deceased patients will be based on the last value while alive. • Clarified in Section 4.2.7 that no multiple imputation of missing data will be performed for all exploratory endpoints, except for the KCCQ endpoints. • Updated in Section 4.2.7 to clarify that the change from baseline in EQ-VAS score at week 16 will be presented. • In Appendix 8.1 for accounting for missing data – predictive mean matching, removed the last sentence "For outcome variables from the *MoveMonitor* the summarised endpoint value over the 7-day periods starting at randomisation, visit 3 (week 8) and visit 4 (week 14) will be imputed." to be consistent on the analysis method that the missing on *MoveMonitor* variables will not be imputed. • In Append 8.4, moved "(4 items)" to question 15 to clarify that social limitation domain has 4 questions.

1 STUDY DETAILS

1.1 Study objectives

1.1.1 Primary objective

Primary Objectives	Outcome Measures:
To determine whether dapagliflozin is superior to placebo in patients with chronic HF NYHA Functional Class II-IV and preserved ejection fraction (LVEF>40%) [HFpEF] in • reducing patient-reported HF symptoms • reducing patient-reported physical limitation • improving exercise capacity	Family of primary endpoints: • Change from baseline in KCCQ-TSS at week 16. • Change from baseline in KCCQ-PLS at week 16. • Change from baseline in 6MWD at week 16.

1.1.2 Secondary objective

Secondary Objective	Outcome Measure:
To determine whether dapagliflozin is superior to placebo in increasing time spent non-sedentary, evaluated in a subset of at least 100 patients	Change from baseline at the end of the study (ie, the week starting at week 14) in total time spent in light to vigorous physical activity, as assessed using a wearable activity monitor (*MoveMonitor*).

Baseline for wearable activity monitor *MoveMonitor* outcomes consists of a 7-day period measured on the week starting at enrolment visit. End of study for wearable activity monitor *MoveMonitor* outcomes consists of a 7-day period measured on the week starting at visit 4 (week 14).

1.1.3 Safety objective

Safety Objective:	Outcome Measures:
To evaluate the safety and tolerability of dapagliflozin compared to placebo in patients with HFpEF	• AEs • SAEs • DAEs • AEs leading to amputation • Potential risk factor AEs for amputations affecting lower limbs • Laboratory tests • Vital signs

1.1.4 Exploratory objectives

Exploratory Objectives:	Outcome Measures:
To determine whether dapagliflozin is superior to placebo in increasing total physical activity, evaluated in a subset of at least 100 patients	Change from baseline at end of study in total activity measured by vector magnitude units per minute, as assessed using a wearable activity monitor (*MoveMonitor*).
To determine whether dapagliflozin is superior to placebo in reducing serum NT-proBNP	Change from baseline in serum NT-proBNP at week 16.
To determine whether dapagliflozin is superior to placebo in increasing the exercise capacity during daily life, evaluated in a subset of at least 100 patients	Change from baseline at end of study in movement intensity during walking, as assessed using a wearable activity monitor (*MoveMonitor*).
To determine whether dapagliflozin is superior to placebo in reducing the proportion of patients with worsened NYHA Functional Classification	Proportion of patients with worsened NYHA Functional Classification at week 16.
To compare the effect of dapagliflozin versus placebo on physical activity, evaluated in a subset of at least 100 patients	Change from baseline at end of study for exploratory endpoints assessed using a wearable activity monitor (*MoveMonitor* or *MoveTest*), in amount, duration and intensity.
To compare the effect of dapagliflozin versus placebo on health status as assessed by EQ-5D-5L	Change from baseline in health status utilities as measured by EQ-5D-5L at week 16.

Exploratory Objectives:	**Outcome Measures:**
To compare the effect of dapagliflozin versus placebo on patient reported dyspnoea and fatigue	Change from baseline in dyspnoea at week 16. Change from baseline in fatigue at week 16.
To assess the patients' overall evaluation of net treatment benefit	Distribution of patients' assessment of benefit of IP.
To explore whether dapagliflozin compared to placebo improves symptom frequency, symptom burden, symptom stability, social limitation, and QoL	Changes from baseline in the following KCCQ domains at week 16: • TSS domains: symptom burden and symptom frequency • Overall summary score • Symptom stability domain • Self-efficacy domain • Social limitation domain • QoL domain
To assess change in oxygen saturation after 6MWT	Change from baseline in oxygen saturation difference after 6MWT at week 16.
To determine whether dapagliflozin compared with placebo has an effect on systolic BP	Change from baseline in systolic BP at week 16.
To determine whether dapagliflozin compared with placebo has an effect on body weight	Change from baseline in body weight at week 16.
To determine whether dapagliflozin compared with placebo has an effect on eGFR	Change from baseline in eGFR at week 16.
To collect and store blood samples for PK assessment	Explore dapagliflozin exposure-response relationship for efficacy and safety endpoints. The results will be analysed and reported in a separate report.
To collect and store blood samples for future exploratory genetic samples	Not applicable. Results will be analysed and reported separately.

Baseline for wearable activity monitor *MoveMonitor* outcomes consists of a 7-day period measured on the week starting at enrolment visit. End of study for wearable activity monitor *MoveMonitor* outcomes consists of a 7-day period measured on the week starting at visit 4 (week 14).

1.2 Study design

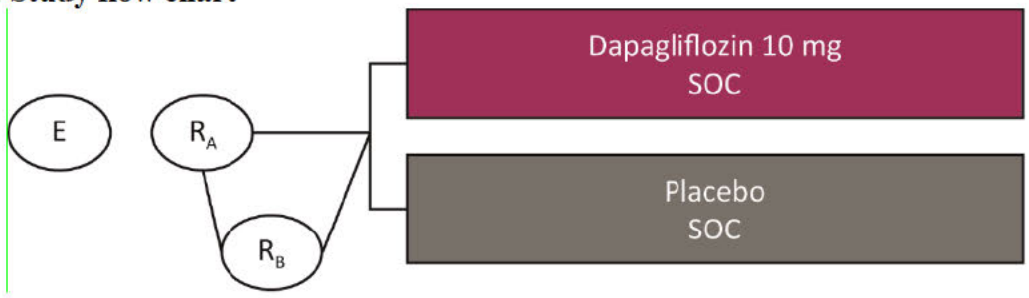
This is an international, multicentre, parallel-group, randomised, double-blind, placebo-controlled, phase III study in subjects with heart failure with preserved left ventricular ejection fraction (HFpEF), evaluating the effect of dapagliflozin 10 mg versus placebo, given once daily in addition to background regional standard of care therapy, including treatments to control co-morbidities, on change in heart failure (HF) symptoms as measured by Kansas City Cardiomyopathy Questionnaire Total Symptom Score (KCCQ-TSS), physical limitation as measured by Kansas City Cardiomyopathy Questionnaire Physical Limitation Score (KCCQ-PLS), and exercise capacity as measured by 6-minute walk distance (6MWD). The scoring algorithm is described in Appendix 8.4.

HFpEF is defined in this study as left ventricular ejection fraction (LVEF) >40% and evidence of structural heart disease. Adult subjects with HFpEF, aged ≥40 years with New York Heart Association (NYHA) Functional Class II-IV and who meet all of the inclusion criteria and none of the exclusion criteria will be randomised in a 1:1 ratio to receive either dapagliflozin 10 mg or placebo once daily.

It is estimated that approximately 1000 subjects at approximately 115 to 120 sites in 12 countries will be enrolled to reach the target of approximately 500 randomised subjects, assuming a screen failure rate of 50%. The investigational product (IP) will be added to the prescribed background therapy for HF, and background therapy for Type 2 diabetes mellitus (T2DM) when applicable, as considered appropriate by the investigator and in accordance with regional standard of care.

The anticipated total study duration is approximately 12 months. The duration of the study may be changed if the randomisation rate is different than anticipated.

Figure 1 Study flow chart



Visit	E	R_A	R_B [a]		T	FV
	1	2a	2b [a]	3	4	5
Week	-2 ±1	0	0 [a]	8 ±1	14 ±1	16 ±1
Day	-14 ±7	1	1 [a]	56 ±7	98 ±7	112 ±7

[a] Visit 2b occurs within 7 days of Visit 2a and constitutes Week 0 and Day 1 for patients who qualify for Randomisation B.

E Enrolment; FV Final visit; R_A Randomisation A; R_B Randomisation B; SOC Standard of care; T Telephone call

1.2.1 Randomisation

Subjects will be randomised 1:1 to either dapagliflozin 10 mg or placebo once daily. Randomisation will be stratified by T2DM status at randomisation (2 levels: with T2DM; without T2DM). For the purpose of stratification, T2DM is defined as established diagnosis of T2DM or HbA1c≥6.5% (48 mmol/mol) shown at the central laboratory test at enrolment (visit 1; single measure).

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

Subjects can be randomised at visit 2a [Randomisation A] or visit 2b [Randomisation B] depending on whether they fulfilled the 15% variability criterion for 6MWD. Subjects who fulfilled all other inclusion criteria and none of the exclusion criteria, with a 6MWD value at visit 1 [Enrolment] which was not within [85%, 115%] of the value at visit 2a [Randomisation A] (ie, there was >15% variability) will have a second randomisation attempt at visit 2b [Randomisation B], within 7 days of visit 2a [Randomisation A]. Such subjects can be randomised if they fulfil all inclusion criteria and none of the exclusion criteria and the 6MWD value at visit 2a [Randomisation A] is within [85%, 115%] of the value at visit 2b [Randomisation B].

The numbers of subjects randomised to IP will be monitored, on a study level, to ensure the following characteristics are appropriately represented in the study, and caps may be applied in IxRS:

- T2DM status: the number of randomised subjects with and without T2DM will be monitored in order to ensure a minimum of 30% in each subpopulation. Randomisation may be capped (ie, no more subjects can be randomised in a specific subpopulation) if the pre-determined limit is reached.
- LVEF value: the proportion of subjects with LVEF above 40% and below 50% will be monitored to ensure a representative proportion in the study.
- Atrial fibrillation/flutter (AF) status: the proportion of subjects with AF will be monitored to ensure a representative proportion in the study.

1.2.2 Number of subjects

The study will enrol approximately 1000 subjects of which approximately 500 subjects will be randomised 1:1 to each treatment. This sample size estimate is based on supporting the primary efficacy endpoints, change from baseline at week 16 in KCCQ-TSS and 6MWD. No adjustment to sample size was made to support the addition of primary endpoints based on change from baseline at week 16 in KCCQ-PLS. Mortality over the entire study period is assumed to be 5% in each treatment group.

The sample size selection was based on the following assumptions:

[redacted]

2 ANALYSIS SETS

2.1 Definition of analysis sets

2.1.1 Full analysis set

All patients who have been randomised to study treatment will be included in the full analysis set (FAS), irrespective of their protocol adherence, addition or modification of background rescue medications, switches to alternative medications, and continued participation in the study. Patients will be analysed according to their randomised IP assignment, irrespective of the treatment they actually received. By the intent-to-treat (ITT) principle, the FAS population will be considered the main analysis set for the primary and secondary efficacy variables and for the exploratory efficacy variables, unless otherwise specified.

2.1.2 Safety analysis set

All randomised patients who received at least 1 dose of IP will be included in the safety analysis set. Patients will be analysed according to the treatment actually received. For any patients given incorrect treatment, ie, randomised to one of the treatment groups but actually given the other treatment, the treatment group will be allocated as follows: Patients who received both incorrect and correct treatment will be analysed according to their randomised treatment. Patients who received only the incorrect treatment will be analysed according to that treatment. The safety analysis set will be considered the main analysis set for all safety variables.

2.2 Violations and deviations

Only important protocol deviations (IPDs) will be listed and tabulated in the CSR, and only for randomised patients (ie, not screen failures). These are defined as protocol deviations which may significantly affect the completeness, accuracy and/or reliability of the study data, or which may significantly affect a patient's rights, safety or well-being. They will include (but are not limited to):

- 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, ie, open label SGLT2-inhibitors taken alone or in combination with IP.

All IPDs except for dosing error will be identified and documented by the study team prior to unblinding of the trial. As far as possible, the occurrence of IPDs will be monitored (blinded) during the trial, with the emphasis on their future prevention.

IPDs will not be used to exclude any patient from any analysis set, nor to exclude any data from patients included in an analysis set. Patients having IPDs will be summarised for FAS

population by randomised treatment group and overall. Patients with IPDs or any COVID-19 related protocol deviations will be listed.

2.2.1 Deviations related to COVID-19

All protocol deviations related to COVID-19 will be listed.

3 PRIMARY AND SECONDARY VARIABLES

3.1 General definitions

3.1.1 Definition of baseline

For efficacy variables, the last non-missing measurement on or prior to the date of randomisation will serve as the baseline measurement. If no such non-missing value is available, the baseline value will be imputed as described in Appendix 8.1.

The following specific rules for capturing the various efficacy data are planned:

The last value on or prior to the randomisation visit will be used as baseline for all efficacy endpoints that are intended for on-site collection, except for overall treatment benefit (OTB), which is only measured post-baseline, and also the measurements taken right before and after the 6-minute walk test (6MWT), oxygen saturation (Section 3.2.3.8) and systolic BP (Section 3.2.3.9), where the difference between the pre-6MWT value and post-6MWT value, at the randomisation visit, is used as the baseline value. Efficacy endpoints collected off-site, assessed by a wearable activity monitor (DynaPort *MoveMonitor*), are also handled differently.

Two types of accelerometers are used; the DynaPort *MoveTest* which is used by the patient in the clinic, at each visit, and the DynaPort *MoveMonitor* which is used by the patient at home (see Appendix 8.2 for a detailed description of each wearable activity monitor). At a subset of study sites, labelled "sites with wearable devices", patients will wear a wearable activity monitor during the 6MWT; this one is for use in the clinic only (DynaPort *MoveTest*), and it is different to the wearable activity monitor that will be dispensed for use at home. At the same subset of study sites, patients will be dispensed a wearable activity monitor (DynaPort *MoveMonitor*) to wear at home. The data from these two types of wearable activity monitors (DynaPort *MoveTest* and DynaPort *MoveMonitor*) will never be merged together in the analyses of the efficacy endpoints since the parameters are completely different.

The last value on or prior to the randomisation visit will be used as baseline for the exploratory endpoints assessed by the wearable activity monitor *MoveTest*. Change from baseline at week 16 in distance walked during 6MWT and change from baseline at week 16 in number of stops during 6MWT are the only two exploratory endpoints based on the data

collected by the wearable activity monitor *MoveTest*. The two endpoints based on *MoveTest* data and the corresponding parameters in *MoveTest* data are listed in Appendix 8.2, Table 4.

For efficacy endpoints assessed using the wearable activity monitor *MoveMonitor*, the summarised values (per "visit") using data collected between enrolment visit and randomisation visit, ie starting from visit 1 and collected in 7 consecutive days, will be used as baseline. For *MoveMonitor* data the measurement corresponding to a "visit" spans several days (theoretically, up to 14, as this is limited by the *MoveMonitor* device memory and settings). Therefore, in the dataset with daily summaries there is more than one record (day) associated with eg, Visit 2 (Baseline), for each combination of subject/parameter/visit. Only 7 days will be used to derive summarised values for each visit. For each pre-specified parameter of interest (efficacy endpoint) in the *MoveMonitor* data, an algorithm is applied to derive visit summaries from daily data. Rules for which days to include in the visit summary are defined in Section 3.1.3, Section 3.2.3.3, and Appendix 8.2 for *MoveMonitor*. The nine endpoints based on *MoveMonitor* data and the corresponding parameters in *MoveMonitor* data are listed in Appendix 8.2, Table 4.

For safety variables, the last non-missing measurement prior to first dose of study treatment will serve as the baseline measurement. If there is no value prior to first dose of study treatment, then the baseline value will not be imputed, and will be set to missing. This applies for safety variables including laboratory variables and vital signs.

3.1.2 Change from baseline

Change from baseline is defined as *(post-baseline value – baseline value)*.

Relative change from baseline is defined relative to the baseline value as *(post-baseline value – baseline value)/baseline value* and is only defined when the baseline value is not zero.

3.1.3 Visit windows

All summaries and analyses for efficacy and safety endpoints, except for adverse events, will be presented by time points and a visit window approach will be used to classify the data record.

It should be noted that the same approach as above will also be used to present data captured using the *MoveTest.*

MoveMonitor data, however, will be aggregated for analysis at each relevant time point by summing or averaging over up to 7 days. Details about *MoveTest* and *MoveMonitor* data are described in Section 3.2.3.3 and Appendix 8.2.

A visit window will be derived using the study day. The study day is derived from the assessment date relative to the reference start date. For safety variables, the reference start date for these measurements is the date of first dose of IP, and study day is therefore defined as *(Date of assessment – Date of first dose of IP) + 1* for post-baseline, and defined as *(Date of assessment – Date of first dose of IP)* for pre-baseline. For efficacy variables, the reference start date for these measurements is the date of randomisation, and study day is therefore defined as *(Date of assessment – Date of randomisation) + 1* for post-baseline, and defined as *(Date of assessment – Date of randomisation)* for pre-baseline.

The derivation of visit window (using study day) is described in Table 1. The window for the scheduled visits following baseline will be constructed in such a way that the upper limit of the interval falls half way between the two visits. Intervals in the table are inclusive.

Table 1 Visit windows

Assessment	Visit	Target day	Visit window for safety variables	Visit window for efficacy variables
Screening/Baseline	1 & 2	See Section 3.1.1 for baseline definitions		
Week 8	3	56	2 to 84	28 to 84
Week 16	5	112	≥85	≥85

Unless otherwise specified, if a patient has more than one measurement included within a window, the assessment closest to the target day will be used. In case of ties between observations located on different sides of the target day, the earlier assessment will be used. In case of ties located on the same side of the target day (ie, more than one value for the same day but different time), the value with the earlier entry date/time will be used. If the decision falls between two non-missing values recorded on the same day and there is no assessment time associated with at least one of them, or the same assessment time is associated with both non-missing values, the average of the two values will be selected for analysis at that visit.

3.1.4 Baseline and concomitant medication

Medications taken by any patient at any time during the study will be coded using the Anatomical Therapeutic Chemical (ATC) classification system within the World Health Organisation (WHO) Drug Dictionary.

Baseline medication is defined as medication with at least one dose taken before date of randomisation and treatment is ongoing or stop date is on or after date of randomisation.

Concomitant medication is defined as medications taken on or after date of randomisation, except for study drug.

If the start or end date for the medication is completely or partially missing, the incomplete date will be handled using the method described in Section 4.1.4.

3.2 Efficacy variable

3.2.1 Primary efficacy variables

The primary efficacy variables are change from baseline in KCCQ-TSS at week 16, change from baseline in KCCQ-PLS at week 16 and change from baseline in 6MWD at week 16. The change from baseline is selected instead of the value at week 16, to facilitate clinical interpretation of estimated differences between treatment groups.

The KCCQ is a 23-item, self-administered disease-specific instrument, which has been shown to be a valid, reliable and responsive measure for patients with HF (Green et al 2000, FDA 2020, Spertus et al 2005). The scoring algorithm is described in Appendix 8.4. The KCCQ was developed to measure the patient's perception of their health status independently, which includes HF-related symptoms (frequency, severity and recent change), impact on physical and social function, self-efficacy and knowledge, and how the patient's HF affects their quality of life. Summary scores and domain scores are transformed to a range of 0 to 100. Higher scores represent a better outcome. The clinical summary score together with its components, KCCQ-TSS and KCCQ-PLS, are qualified as a drug development tool for clinical outcome assessment by the FDA (FDA 2020).

The KCCQ-TSS incorporates the symptom frequency (4 items) and symptom burden (3 items) domains into a single summary score. The KCCQ-PLS is calculated from the average of numerical values assigned to responses to 6 items and captures how the patient's physical function is limited due to their HF. The KCCQ-TSS and KCCQ-PLS are calculated for assessments at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and premature treatment discontinuation visit (PTDV) or early withdrawal visit.

The 6MWD is the distance in meters that a patient can walk in a 6-minute period, measured by the site staff during 6MWT. The 6MWT will be conducted in accordance with the American Thoracic Society (ATS) guidelines (American Thoracic Society 2002). The 6MWD is measured at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and PTDV or early withdrawal visit. During the COVID-19 pandemic, an on-site visit may be replaced by a telephone call if allowed by local/regional guidelines to permit collection of critical study data, while preserving both patient safety and social distancing. The 6MWT will not be performed remotely and will only be possible when site is able to resume regular on-site visits and study assessments. The KCCQ-TSS and KCCQ-PLS may be collected via telephone call. The recommended approach is that the patient has the questionnaire in front of

them and reads each question aloud to the investigator, along with their response. Each response will then be entered directly into the ePRO tool by the investigator.

In addition to the impact on data collection, several aspects of the patients' lifestyle and behaviour will be impacted by the COVID-19 pandemic and associated restrictions including social distancing. These aspects are important to several of the questions in the KCCQ-TSS and KCCQ-PLS, eg, carrying groceries or hurrying as if to catch a bus, and are thus expected to influence patient responses and the resulting scores. A detailed description of how the impact of COVID-19 is accounted for in the statistical analysis is provided in Section 4.2.4.

In order to account for patients who die prior to the 16-week assessment and to accommodate non-normal distribution of the primary efficacy variables, hierarchical composite endpoints will be used. The values of change from baseline in KCCQ-TSS, KCCQ-PLS and 6MWD at week 16 in patients who survive to week 16 will be converted to ranks (across both treatment groups combined) with lower ranks attributed to worse outcomes (ie, lower ranks corresponding to negative or smaller values of change from baseline). Patients who die prior to the 16-week assessment will be assigned worse ranks than any patients surviving to 16 weeks. The ranking amongst the deceased patients will be based on the last value while alive. Details are described in Section 4.2.4.

A 6MWD responder will be defined as a patient who had a clinically meaningful improvement in 6MWD. As a starting point, a responder is defined as a patient with a ≥30 meters change from baseline at week 16 in 6MWD. Deaths are defined as non-responders. The 30-meter threshold is a pre-specified limit meant to approximate a threshold for clinically meaningful within-patient change (CMWPC), based on previous studies (Shoemaker et al 2013, O'Keeffe et al 1998, Holland et al 2014, Ferreira et al 2016). However, since most of these studies looked at group-level estimates (difference between means) and not individual-level estimates (eg, comparison of proportions of responders) a threshold, or a range of thresholds, suited to the specific target population in this study (by being based on patients included in this study), will be estimated using the anchor-based methods described in Appendix 8.3 and will replace the 30-meter threshold. Estimation of such a threshold, or a range of thresholds, will be based on blinded data and will not be in any way contingent upon treatment assignment.

A KCCQ-TSS or KCCQ-PLS responder will be defined as a patient who had clinically meaningful improvement in that score. As a starting point, in each of KCCQ-TSS and KCCQ-PLS, a responder is defined as a patient with a ≥5 point change from baseline at week 16. Deaths are defined as non-responders. The 5-point threshold is a pre-specified limit meant to approximate a threshold for CMWPC, based on previous studies (Filippatos et al 2017). However, since most earlier studies have evaluated thresholds for other summary scores in the

KCCQ (namely the overall summary score and clinical summary score), a threshold, or range of thresholds, specific to the KCCQ-TSS and KCCQ-PLS and suited to the specific target population in this study (by being based on patients included in this study) will be estimated using the same anchor-based methods as those used for 6MWD, c.f. Appendix 8.3, and these thresholds will replace the 5-point threshold. Estimation of such a threshold, or a range of thresholds, will be based on blinded data and will not be in any way contingent upon treatment assignment.

3.2.2 Secondary efficacy variable

The secondary efficacy variable is change from baseline at the end of the study (ie, the week starting at week 14) in total time spent in light to vigorous physical activity (LVPA) measured in *hours,* as assessed by the wearable activity monitor *MoveMonitor* and defined as the wear time spent with an energy expenditure ≥1.5 Metabolic Equivalents of Task (METs). Relative change from baseline at end of study will serve as a supportive variable. Change and relative change from baseline are defined in Section 3.1.2.

MoveMonitor data will be collected at sites with wearable devices and in a period of 7 days at each time point. Data collected during the 7-day period starting on the day of visit 1 (enrolment) will be retrieved at visit 2a and this data constitutes the baseline for each patient as defined in Section 3.1.1. Data collected during the 7-day period starting on the day of visit 3 (week 8) and data collected during the 7-day period starting on the day of visit 4 (week 14), will be retrieved at visit 5 (week 16), these data comprise the follow-up for each patient. End of study refers to the 7-day period starting at visit 4 (week 14).

For the secondary efficacy variable, a hierarchical composite endpoint will be derived using the same method as described in Section 3.2.1 and Section 4.2.4.

A threshold, or range of thresholds, for clinically meaningful within-patient change could theoretically be estimated using the same methods as those used for the primary efficacy endpoints, c.f. Appendix 8.3, with appropriate anchors. However, only a subset of patients are expected to provide *MoveMonitor* data and for inclusion in anchor-based analyses patients are required to also provide data on the global anchor variables. Furthermore, relevant global anchor variables for anchoring change from baseline at end of study in time spent in LVPA is not clear. Due to these factors, which severely reduce the reliability of any threshold derived in anchor-based analysis, an anchor-based analysis of the secondary endpoint will not be done for the CSR.

3.2.3 Exploratory variables

For each continuous exploratory endpoint, a hierarchical composite endpoint will be derived using the same method as described in Section 3.2.1 and Section 4.2.4, ie, NT-proBNP, dyspnoea and fatigue, KCCQ domains, oxygen saturation, systolic BP, body weight, eGFR or endpoints assessed using the wearable activity monitors.

3.2.3.1 Change from baseline at week 16 in NT-proBNP

The exploratory efficacy variable is the change from baseline in serum NT-proBNP (pg/mL) at week 16.

The serum NT-proBNP is collected at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and PTDV or early withdrawal visit.

3.2.3.2 Proportion of patients with worsened NYHA Functional Classification from baseline at week 16

The exploratory efficacy variable is the proportion of patients with worsened NYHA Functional Classification from baseline at week 16.

The NYHA classification will be evaluated by the investigator and collected at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and PTDV or early withdrawal visit.

For the main analysis the data will be dichotomised into patients with worsened NYHA class at week 16 (the NYHA class is higher than baseline) and patients with improved or unchanged NYHA class. Patients who died due to any cause prior to week 16 are classified as worsened. Patients with NYHA class IV at baseline can only "worsen" if they die prior to week 16.

3.2.3.3 Change from baseline at end of study or week 16 in exploratory endpoints assessed using the wearable activity monitors

The exploratory efficacy variables based on data from the wearable activity monitor *MoveMonitor* and *MoveTest* are highlighted below, a detailed description of each endpoint is available in Appendix 8.2:

- Change from baseline at end of study in vector magnitude units (VMUs) per minute, as assessed by the *MoveMonitor*
- Change from baseline at end of study in movement intensity during walking measured in *milli-g*, as assessed by the *MoveMonitor*
- Change from baseline at end of study in movement intensity when walking for durations of >20 seconds based on *MoveMonitor* data, defined as movement intensity

measured in *milli-g*, during periods when the physical activity was classified as 'walking' and lasted for more than 20 seconds

- Change from baseline at end of study in total number of steps based on *MoveMonitor* data, defined as the total number of steps excluding steps from walking in stairs
- Change from baseline at end of study in total time spent in moderate to vigorous physical activity (MVPA) based on *MoveMonitor* data, defined as the wear time spent with an energy expenditure ≥3 METs measured in hours
- Change from baseline at week 16 in distance walked during 6MWT based on *MoveTest* data, defined as the total distance covered by the patient measured in meters
- Change from baseline at week 16 in number of stops during 6MWT based on *MoveTest* data, defined as the total number of walking interruptions
- Change from baseline at end of study in total number of sitting shifts during night rest based on *MoveMonitor* data, defined as a transition between a posture classified as 'sitting upright' and any posture classified as 'lying' (ie, 'prone', 'lying on the left', 'lying on the right' or 'supine')
- Change from baseline at end of study in total number of activity counts during worn periods based on *MoveMonitor* data, activity counts are zero when the device is not worn
- Total wear time at end of study based on *MoveMonitor* data, defined as the total time measured in hours when the number of activity counts was greater than zero

Baseline for *MoveMonitor* and *MoveTest* efficacy variables is defined in Section 3.1.1. For *MoveMonitor* efficacy variables, end of study refers to the 7-day period starting at visit 4 (week 14).

"Relative change from baseline" will serve as a supportive variable.

For each exploratory efficacy endpoint assessed by the *MoveMonitor* or the *MoveTest*, a hierarchical composite endpoint will be derived using the same method as described in Section 3.2.1 and Section 4.2.4.

3.2.3.4 Change from baseline at week 16 in EQ-5D-5L

The EQ-5D-5L is a self-reported questionnaire that is used to derive a standardised measure of health status, also referred to as a utility score.

EQ-5D-5L is collected at enrolment and randomisation visits, visit 5 (week 16), and PTDV or early withdrawal visit.

The distribution of categorical responses to EQ-5D-5L, change from baseline at week 16, and the responses on the EQ-5D visual analogue scale (VAS) scores will be summarised. The VAS score has a range of 0-100, and higher scores represent a better outcome.

3.2.3.5 Change from baseline at week 16 in dyspnoea and fatigue

The dyspnoea and fatigue scales assess shortness of breath and fatigue, respectively, on a scale from 0 to 10, where 0 indicates no dyspnoea or fatigue, and 10 indicates the worst situation. Dyspnoea and fatigue scales are collected at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and PTDV or early withdrawal visit.

The exploratory endpoint will be the change from baseline at week 16 in dyspnoea and fatigue scales. Other time points will be summarised.

3.2.3.6 Distribution of OTB at week 16

The patient OTB uses a single question that assesses the patient's impression of the benefits of the study medication relative to the negative effects. It has levels ranging from 'much greater than the negative effects', 'somewhat greater than the negative effects', 'equal to the negative effects', 'somewhat less than the negative effects', and 'much less than the negative effects'. The levels will also be collapsed into four major groups – positive ('much greater than the negative effects'), neutral ('somewhat greater than the negative effects' + 'equal to the negative effects' + 'somewhat less than the negative effects'), negative ('much less than the negative effects'), and death occurring prior to collection of the variable at the analysis time point.

OTB is collected at visit 3 (week 8), visit 5 (week 16), and PTDV or early withdrawal visit.

The exploratory endpoint will be the distribution of OTB at week 16 (collapsed class variable described above).

The second question in the patient OTB is whether the patient would choose to continue taking study medication after the end of the study, if that was an option. The response options are "Yes", "Unsure" or "No".

The distribution of responses at Week 8 and Week 16 will be reported.

3.2.3.7 Change from baseline at week 16 in KCCQ domains

The KCCQ domain and summary scores of interest include the TSS domains (symptom frequency and symptom burden), overall summary score, symptom stability domain, self-efficacy domain, social limitation domain and quality of life (QoL) domain. Domain and summary scores are transformed to a range of 0 to 100. Higher scores represent a better outcome. Scoring algorithm is defined in Appendix 8.4.

KCCQ is collected at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and PTDV or early withdrawal visit.

The exploratory endpoints will be the change from baseline at week 16 in each KCCQ domain and summary score.

3.2.3.8 Change from baseline at week 16 in oxygen saturation

Oxygen saturation forms part of the suite of assessments administered during the 6MWT and will be assessed using a standard pulse oximetry device in a sitting position right before and after the 6MWT at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and at PTDV or early withdrawal visit.

The oxygen saturation difference after 6MWT (referred to as "oxygen saturation delta") is defined as the value after 6MWT minus the value before 6MWT.

The exploratory endpoint will be change from baseline at week 16 in oxygen saturation delta.

3.2.3.9 Change from baseline at week 16 in systolic BP

Systolic BP, diastolic BP and pulse rate are collected at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and at PTDV or early withdrawal visit.

Systolic BP, diastolic BP and pulse rate are collected three times at each visit; once before conducting the 6MWT and twice as part of the suite of assessments administered during the test; right before and after the 6MWT.

The systolic BP data collected as part of the 6MWT before and after 6MWT will be used for analysis of exploratory efficacy endpoint.

At each visit, the systolic BP difference after 6MWT (referred to as "systolic BP delta") is defined as the systolic BP value after 6MWT minus the value before 6MWT.

The exploratory endpoint will be change from baseline at week 16 in systolic BP delta.

3.2.3.10 Change from baseline at week 16 in body weight

Body weight is measured at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and at PTDV or early withdrawal visit.

The exploratory endpoint will be change from baseline at week 16 in body weight.

3.2.3.11 Change from baseline at week 16 in eGFR

The eGFR values will be calculated (in mL/min/1.73 m^2) using the CKD-EPI formula (Levey et al 2009). eGFR = 141x $\min(S_{cr}/\kappa, 1)^{\alpha}$ x $\max(S_{cr}/\kappa, 1)^{-1.209}$ × 0.993^{Age} × 1.018 [if female] × 1.159 [if black]

Where

- S_{cr} is serum creatinine in mg/dL,
- κ is 0.7 for females and 0.9 for males,
- α is -0.329 for females and -0.411 for males,
- $\min(S_{cr}/\kappa, 1)$ indicates the minimum of S_{cr}/κ or 1,
- $\max(S_{cr}/\kappa, 1)$ indicates the maximum of S_{cr}/κ or 1.

Age is rounded to years.

The eGFR values will be calculated at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and at PTDV or early withdrawal visit.

The exploratory endpoint will be change from baseline at week 16 in eGFR.

3.3 Safety variables

3.3.1 Adverse events

The safety and tolerability of dapagliflozin in patients with HFpEF will be evaluated from adverse events (AEs), serious adverse events (SAEs), adverse events leading to discontinuation of IP (DAEs), AEs leading to amputation and relevant preceding AEs reflecting potential risk factors for lower limb amputations ("preceding events"). The AEs leading to amputation and preceding events are included as AEs of special interest in this study.

Amputation will be recorded in the eCRF as AE/SAE. The "preceding events", defined as non-serious and serious events potentially placing the patient at risk for a lower limb amputation regardless of whether an amputation has taken place or not, will be recorded in the eCRF as AE/SAE as well. Safety analysis of "preceding events" will be based on the predefined list of preferred terms. Additional information about amputations with underlying conditions and "preceding events" will be collected on dedicated eCRF pages.

For any AEs reported by the Investigator as Diabetic Ketoacidosis (DKA) additional information will be recorded on specific eCRF pages in addition to the AE/SAE form.

DKA definition:

A diagnosis of DKA should only be made in a clinical setting consistent with DKA (based on patient history, symptoms, and physical examination) and in the absence of more likely

alternative diagnoses and causes of acidosis (such as lactic acidosis). The following biochemical data should support diagnosis:

- Ketonaemia ≥3.0 mmol/L and/or significant ketonuria (more than 2+ on standard urinesticks).

AND

- At least 1 of the following criteria suggesting high anion gap metabolic acidosis:
  - Arterial or venous pH ≤7.3
  - Serum bicarbonate ≤18 mEq/L
  - Anion gap [Na – (Cl + HCO3)] >10

DKA events will not be adjudicated in this study. DKA events may be presented if applicable.

SAEs will be collected from time of informed consent until and including the patient's last visit. Non-serious AEs will be collected from randomisation until and including the patient's last visit.

3.3.2 Laboratory values

Blood samples will be taken at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and at PTDV or early withdrawal visit, for central laboratory assessment of the following laboratory variables:

- Sodium
- Potassium
- HbA1c
- Haematocrit
- Creatinine
- eGFR (calculated based on creatinine value, details in Section 3.2.3.11)

3.3.3 Vital signs

The following vital signs will be collected at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and at PTDV or early withdrawal visit:

- Systolic BP
- Diastolic BP
- Pulse rate
- Waist circumference
- Body weight

Systolic BP, diastolic BP and pulse rate will be measured in a sitting position, and waist circumference and body weight will be measured in a standing position. The vital signs for safety analysis will use the data collected from VS form in eCRF (details in 3.2.3.9).

3.3.4 Physical examination

Physical examinations will be performed at enrolment and randomisation visits, visit 3 (week 8), visit 5 (week 16), and at PTDV or early withdrawal visit.

Any new or aggravated clinically relevant abnormal medical finding on physical examination compared with the baseline assessment will be reported as an AE unless unequivocally related to the disease under study.

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 scheme or above the pre-specified threshold for significance, in Section 4.1.3, will be regarded as nominal.

Stratification of analyses for T2DM status will be performed using the stratification values as entered in IxRS to determine the randomisation assignment.

Summary data, including the observed value and change from baseline value at week 16 or end of study in the endpoint at each analysis time point, will be presented in tabular format by treatment group. Categorical data will be summarised by the number and percentage of patients in each category by treatment group. Continuous data will be summarised by descriptive statistics as appropriate, including N, mean, SD, minimum, first quartile (Q1), median, third quartile (Q3) and maximum.

4.1.1 Estimand for primary and secondary efficacy variables

The primary efficacy endpoints and secondary efficacy endpoint will be evaluated under a combined treatment policy (intent-to-treat) and composite variable strategy estimand including differences in outcomes at the end of the 16-week treatment period, or at end of study (ie, the week starting at week 14, for variables captured with the *MoveMonitor*) to reflect the effect of the initially assigned randomised study drug, irrespective of exposure to study drug, concomitant treatment as well as subsequent treatment after discontinuation of study drug. A composite variable strategy approach is employed to account for deaths occurring during the follow-up period. Deaths are regarded as intercurrent events and are incorporated into the hierarchical composite endpoint. The analysis will be performed for the FAS.

4.1.2 Hypotheses

For the primary efficacy endpoints KCCQ-TSS, KCCQ-PLS and 6MWD, and the secondary efficacy endpoint, total time spent in LVPA, the following hypotheses will be tested using the significance level specified in Section 4.1.3

- H0: m(r(A)) = m(r(C))

versus

- H1: m(r(A)) ≠ m(r(C))

Where H0 and H1 are the null and alternative hypotheses, respectively, and m(r(A)) and m(r(C)) represent the median of the ranked changes in each of the primary efficacy endpoints, KCCQ-TSS, KCCQ-PLS and 6MWD, from baseline at week 16, and the median of the ranked changes in secondary efficacy endpoint, total time spent in LVPA, from baseline at end of study among patients receiving dapagliflozin (Active) and placebo (Control) treatment, respectively.

When applicable, for certain exploratory efficacy endpoints, the distribution of ranked values are replaced with the proportion of responders and the analogous null and alternative hypotheses then become

- H0: OR[dapagliflozin:placebo] = 1

versus

- H1: OR[dapagliflozin:placebo] ≠ 1

Where OR[dapagliflozin:placebo] represents the odds ratio for dapagliflozin versus placebo in a logistic regression model for the outcome of observing a response at week 16 compared to baseline in the endpoint of interest.

4.1.3 Confirmatory testing procedure

To account for multiplicity when testing the primary efficacy endpoints (change in KCCQ-TSS, KCCQ-PLS and 6MWD from baseline at week 16) and secondary efficacy endpoint (change in total time spent in LVPA from baseline at end of study), a pre-specified testing strategy will be followed to control the overall type I error rate. The testing will be performed according to a gatekeeping procedure (Dmitrienko et al 2011): the three tests of KCCQ-TSS, KCCQ-PLS and 6MWD with a family wise error rate (FWER) of 0.05 (2-sided) will be conducted first. The total alpha of 0.05 will be divided among the three primary efficacy endpoints using a weighted Bonferroni method, with 0.04990 (99.8% of the total alpha) assigned to KCCQ-TSS, 0.00005 (0.1% of the total alpha) assigned to each of KCCQ-PLS and 6MWD. The secondary efficacy endpoint, total time spent in LVPA, will not be tested

unless the test of 6MWD is significant. As the testing procedure progresses, if a hypothesis test in the primary endpoint family is significant, its assigned alpha will be preserved and considered as unused and passed along fully or partially to the other tests in the primary endpoint family, as described below in Figure 2. The passed-along alpha will be added to the originally assigned alpha, before testing other endpoints in the primary family. The test of secondary endpoint will use only the unused alpha from the primary efficacy endpoints family. This testing strategy provides strong control of the overall type I error rate and as long as pre-specified alpha allocation is based on available evidence, should not inflate the type II error rate. The complete testing strategy is depicted in Figure 2.

The testing procedure will be as follows (numbering does not strictly imply a chronological order):

1 KCCQ-TSS, KCCQ-PLS and 6MWD will be tested first. The initial alpha allocated to them is 0.04990 for KCCQ-TSS, 0.00005 for KCCQ-PLS and 0.00005 for 6MWD.

2 KCCQ-TSS will be tested at the alpha level of 0.04990 (initial alpha) regardless of the outcome for the tests of KCCQ-PLS and 6MWD.

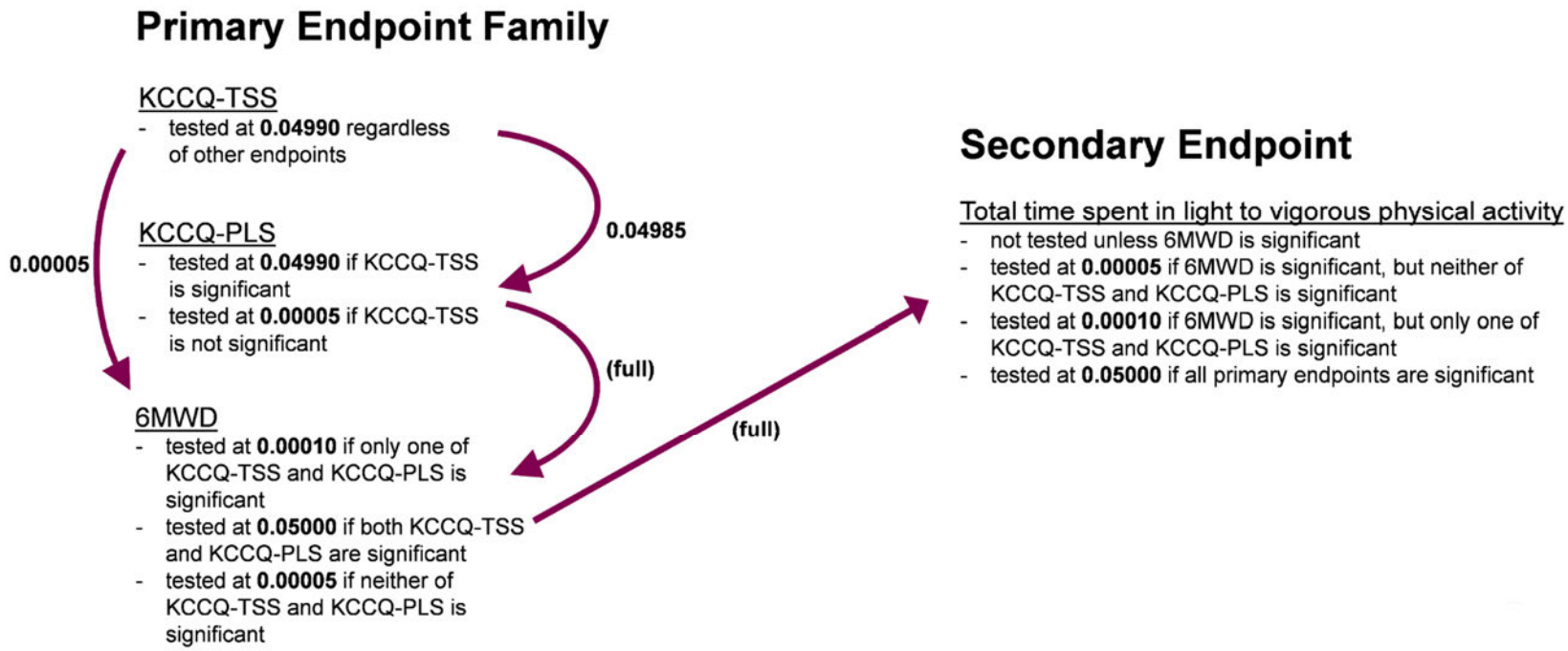
3 a. If the test of KCCQ-TSS is significant, KCCQ-PLS is tested at the alpha level of 0.04990 = 0.00005 + 0.04985 (initial alpha allocated to KCCQ-PLS plus the amount passed from KCCQ-TSS).
b. If the test of KCCQ-TSS is not significant, KCCQ-PLS is tested at 0.00005 (initial alpha allocated to KCCQ-PLS).

4 a. If the test of KCCQ-TSS is significant, but test of KCCQ-PLS is not significant, 6MWD is tested at the alpha level of 0.00010 = 0.00005 + 0.00005 (initial alpha allocated to 6MWD plus the amount passed directly from KCCQ-TSS).
b. If the test of KCCQ-TSS is not significant, but the test of KCCQ-PLS is significant, 6MWD is tested at 0.00010 = 0.00005 + 0.00005 (initial alpha allocated to 6MWD plus the amount passed directly from KCCQ-PLS).
c. If the test of KCCQ-TSS and the test of KCCQ-PLS are both significant, 6MWD is tested at 0.05000 = 0.00005 + 0.04985 + 0.00010 (initial alpha allocated to 6MWD plus the amount passed directly from KCCQ-TSS and amount passed through KCCQ-PLS).
d. If the test of KCCQ-TSS and the test of KCCQ-PLS are both not significant, 6MWD is tested at 0.00005 (initial alpha allocated to 6MWD)

5 The secondary efficacy endpoint, total time spent in LVPA, will not be tested unless the test of 6MWD is significant.
a. If the test of 6MWD is significant but neither of the test of KCCQ-TSS and the test of KCCQ-PLS is significant, the secondary efficacy endpoint of total time spent in LVPA is tested at the alpha level of 0.00005 (passed directly from 6MWD).

b. If the test of 6MWD is significant but only one of the test of KCCQ-TSS and the test of KCCQ-PLS is significant, the secondary efficacy endpoint of total time spent in LVPA is tested at the alpha level of 0.00010 (passed from 6MWD, after passing alpha from KCCQ-TSS or KCCQ-PLS to 6MWD).
c. If the test of KCCQ-TSS, the test of KCCQ-PLS and the test of 6MWD are all significant, the secondary efficacy endpoint of total time spent in LVPA is tested at the alpha level of 0.05000 (full alpha).

Figure 2 Multiple testing strategy for the three primary efficacy endpoints and secondary efficacy endpoint



Arrows indicate the amount and direction of propagated alpha which is enabled if the test was significant at the specified limit. Arrows labelled "full" indicate that all available alpha is propagated in that direction.

KCCQ Kansas City Cardiomyopathy Questionnaire; TSS Total symptom score; PLS Physical limitation score; 6MWD 6-minute walk distance

4.1.4 Incomplete dates

Dates missing the day or both the day and month of the year will adhere to the following conventions in order to classify on-study AEs and to classify baseline and concomitant medications.

In general, listings will present the actual partial or missing values rather than the imputed values that may be used in derivation. In instances where imputed values will be presented individually, imputed values will be flagged as such.

Adverse Events Dates

- The missing day of onset of an AE will be set to:
  - First day of the month that the event occurred, if the onset YYYY-MM is after the YYYY-MM of the first dose of study treatment
  - The day of the first dose of study treatment, if the onset YYYY-MM is the same as YYYY-MM of the first dose of study treatment
  - The date of informed consent, if the onset YYYY-MM is before the YYYY-MM of the first dose of study treatment.
- The missing day of resolution of an AE will be set to:
  - The last day of the month of the occurrence. If the patient died in the same month, then set the imputed date as the death date.
- If the onset date of an AE is missing both the day and month, the onset date will be set to:
  - January 1 of the year of onset, if the onset year is after the year of the first dose of study treatment
  - The date of the first dose of study treatment, if the onset year is the same as the year of the first dose of study treatment
  - The date of informed consent, if the onset year is before the year of the first dose of study treatment
- If the resolution date of an AE is missing both the day and month, the resolution date will be set to:
  - The date of the last study visit, if the patient did not die in the same year.
  - The death date, if the patient died in the same year.

Baseline and Concomitant Medication Dates

Imputation of start and end dates allows medications for patients to be classified into the categories of baseline medication or concomitant medication (or both) for tables. An assessment should be made as to the possibility that the patient's medication could fall into

each category given the information available for the dates. If it is possible given the date information that a patient's medication could fall into a given category, then the patient's medication should be included in tables for that category. If a particular category can be ruled out based on partial or full dates available, then the patient's medication should be excluded from that category.

Two variables, "Started >4 weeks prior to Visit 1" and "Treatment continues", are collected in eCRF and will be used to classify baseline and concomitant medications. In case these information variables are missing, the incomplete dates for medications will adhere to the following rules in order to classify baseline and concomitant medications.

- If both start date and end date are missing or incomplete, then start date should be imputed first.
- The missing day of start date of a medication will be set to the first day of the month of the occurrence.
- The missing day of end date of a medication will be set to the last day of the month of the occurrence.
- If the start date of a medication is missing both the day and month, the onset date will be set to January 1 of the year of occurrence.
- If the end date of a medication is missing both the day and month, the date will be set to December 31 of the year of occurrence.
- If the start date of a medication is null and the end date is not a complete date then the start date will be set to the date of the first study visit.
- If the start date of a medication is null and the end date is a complete date
  - if the end date is after the date of the first study visit then the start date will be set to the date of the first study visit.
  - otherwise the start date will be set to the end date of the medication.
- If the end date of a medication is null and the start date is not a complete date then the end date will be set to the date of the last study visit if no permanent premature discontinuation of IP occurred, otherwise discontinuation date.
- If the end date of a medication is null and the start date is a complete date
  - if the start date is prior to the date of the last study visit then the end date will be set to the date of the last study visit if no permanent premature discontinuation of IP occurred, otherwise discontinuation date.
  - otherwise, the end date will be set to the start date of the medication.

4.1.5 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. If the number of tablets dispensed or the number of tablets returned is missing for at least one observation, compliance is not calculated for that patient.

Study drug compliance will be presented descriptively, including mean, SD, median, Q1, Q3, minimum, maximum, and 5% and 95% percentiles, for safety analysis set by treatment group as defined in Section 2.1.2.

4.2 Analysis methods

4.2.1 Subject disposition

A clear accounting of the disposition of all subjects who enter the study will be provided, from screening to study completion. The number of enrolled and not randomised subjects (and reason) will be summarised. The number and percentage of subjects will be presented by treatment and overall for the following categories: randomised, received IP, did not receive IP, completed treatment, discontinued treatment (and reason), subjects who discontinued IP but completed study assessments, subjects who completed study, and subjects who discontinued study (and reason). Death will be included among reasons for study discontinuation.

4.2.1.1 Impact of COVID-19 on study visits

Number and proportion of patients and week 16 visits potentially impacted by COVID-19 will be summarised. The potential impact will be based on the onset date of COVID-19 per site, as described in Section 4.2.4 and final visit date after the onset date of COVID-19 at each site will be regarded as potentially impacted. The number and proportion of visits missed after the onset date of COVID-19, as potentially missing due to COVID-19, will also be summarised and the number and proportion of visits occurring remotely, due to COVID-19, will also be summarised.

4.2.2 Demographic and baseline characteristics

Demographic and baseline characteristics, including medical history, will be summarised, using frequency distributions and summary statistics by treatment and overall. No statistical test will be performed for comparison of any baseline measurement among treatment groups.

The demographic and baseline characteristics will also be summarised for FAS population by randomised treatment and overall for patients with LVEF>40% and <50% and LVEF≥50%, for patients with and without T2DM at randomisation, and for patients with and without AF.

4.2.3 Baseline and concomitant medication

The frequency of baseline and concomitant medication will be presented for the FAS population per ATC classification, preferred name and treatment group.

4.2.4 Analysis of the primary efficacy variables

The primary efficacy variables are the KCCQ-TSS, KCCQ-PLS and 6MWD and the endpoint for each is the change from baseline at week 16.

The objective of the study is to evaluate the efficacy of dapagliflozin compared to placebo in terms of reducing HF symptoms, reducing physical limitation and improving exercise capacity in a world where there is not an ongoing COVID-19 pandemic. Due to the way in which KCCQ-TSS and KCCQ-PLS are dependent on lifestyle and behaviour, the potential impact of COVID-19 on the analysis of these variables needs to account for this factor, to preserve the objective of the study.

Recognizing that the outbreak of COVID-19 was not simultaneous across the globe, each site in the study is elicited for a date after which their site was in some way affected by COVID-19. The assumption here is that the onset date of COVID-19 provided by each site is a good proxy for the date when the outbreak of COVID-19, together with associated restrictions and social distancing, occurred in that region. This date should also be a good proxy (better than using a global date or a date per country) for when the patients, receiving care at that site, were potentially impacted in terms of lifestyle and behavior.

The actual outbreak of COVID-19 in each region and the introduction of restrictions and social distancing, is assumed to be independent of treatment assignment. By this rationale, randomisation would balance the impact of COVID across treatment groups. On the other hand, one potential impact of COVID-19 could be a “dampening effect” on the patient-reported symptoms and physical limitation, eg due to patients not moving about as much or not having as many social interactions, due to a more restricted lifestyle or social distancing related to COVID-19.

In the hypothesis tests of KCCQ-TSS and KCCQ-PLS, a covariate variable indicating the impact of COVID-19 will be introduced in the rank ANCOVA. This enables inclusion of all data in the analysis used for the hypothesis tests, while still accounting for a potential "dampening effect" of COVID-19. This covariate variable is defined by the number of weeks where the patient was in the study after the onset date of COVID-19 at their site, calculated as (final visit date – COVID-19 start date)/7, rounded up to the closest integer. If the site is not impacted by COVID-19 or COVID-19 started after a subject had their last visit, set the value as zero. The variable is intended to address the extent to which patient lifestyle and behaviour

is affected. The number of weeks under the potential impact of COVID-19 will be categorized. By categorizing time spent in the study after the onset date of COVID-19 at their site, the impact of COVID-19 is allowed to have a non-linear impact over time. As the rank ANCOVA is a non-parametric approach based on ranks, the point estimate has no clear clinical interpretation and the inclusion of this additional covariate variable to adjust for the COVID-19 impact should help ensure that the test of the null hypothesis of no treatment difference still addresses the original objective of the study.

To preserve the objective of the study when estimating the magnitude of treatment effect, the Hodges-Lehmann (HL) estimate of the median difference and the supportive responder analysis will use data collected prior to the onset date of COVID-19 at each site. All patients are expected to have had their baseline data collected prior to the onset date of COVID-19, and for patients with follow-up data during COVID-19, that data will instead be imputed (under the original efficacy estimand) assuming missing at random, ie, missing at random (MAR) imputation based on pre-COVID-19 data only. The MAR assumption is deemed justifiable when dealing with the potential impact of COVID-19, as the timing of the onset of COVID-19 at each site can be assumed to be independent of treatment assignment for individual patients. The reason why a variable addressing the time spent under the potential impact of COVID-19 is not introduced (as was done for the rank ANCOVA used for hypothesis testing) is that the choice of functional form for such a variable could influence the point estimates for the treatment comparison, attained from this analysis. Point estimates are of explicit interest here, when estimating the magnitude of treatment effect, unlike in the non-parametric rank ANCOVA used for hypothesis testing. This multiple imputation (MI) approach comes at a cost of losing precision, as multiple imputation (MI) will generally increase the size of the standard errors, compared to using observed data. Nevertheless, this approach ensures that the analysis addresses the original objective of the study.. In a supportive analysis, all data collected during COVID-19 will be included in the HL estimate of median difference and the responder analysis.

Due to limited impact of COVID-19 pandemic expected for the primary efficacy endpoint of change from baseline at week 16 in 6MWD, the main analysis of this endpoint does not account for data collected prior to and during COVID-19.

Rank ANCOVA model

The hierarchical composite endpoint representing the patients' vital status at week 16 and the change from baseline in primary efficacy endpoints at week 16 in surviving patients, as defined in Section 3.2.1, will be analysed using the rank ANCOVA method to test the null hypothesis of no differences in the distributions of ranked outcomes between the two treatment groups.

First the change from baseline at week 16 in each of the primary efficacy endpoints and vital status at week 16, as well as values of the baseline covariate will be transformed to standardised ranks within each T2DM randomisation stratum, using fractional ranks (dividing by the denominator n+1) and the mean method for ties. Ranking for the hierarchical composite endpoint (Section 3.2.1) will be performed so that patients who die prior to the week 16 assessment are assigned the worst ranks within each stratum. This will be implemented by assigning a temporary value which is lower than all observed negative change values to patients who died prior to week 16, before deriving fractional ranks. The ranking amongst the deceased patients will be based on the last value while alive, with lower ranks assigned to smaller last values while alive.

Amongst the patients who survive to 16 weeks, the missing data on the outcome variable due to reasons other than death (eg, missing visits, early withdrawal from the study, including lost to follow-up) will be imputed prior to ranking. The missing value will be replaced by placebo-based MI using predictive mean matching. The predictive mean matching method ensures that the imputed values remain in the permissible range of the outcome values. Details about the placebo-based MI are outlined in Appendix 8.1. All data collected before and during COVID-19 will be included in the placebo-based imputation of missing data. The imputation model will include the stratification variable (T2DM at randomisation), and value of the outcome variable at the previous visits.

The patients who died will be added back to the imputed datasets, which contain the patients who survived to week 16, to construct the imputation FAS datasets. The hierarchical composite endpoint described above will be derived using the imputation FAS datasets.

Rank ANCOVA will be performed on each imputation FAS dataset. Separate ANCOVA models will be fitted to the ranked data for each randomisation stratum using a regression model with the ranked composite endpoint as the dependent variable, adjusting for the ranked baseline as a covariate. For KCCQ-TSS and KCCQ-PLS, the additional categorical variable, weeks impacted by COVID-19, will be included as covariate in the ANCOVA model to adjust for COVID-19 impact. Residuals from this regression model will be captured for testing of differences between treatment groups. The Cochran-Mantel-Haenszel (CMH) test, stratified for the T2DM status at randomisation, using the values of the residuals as scores will be used to compare treatment groups. This analysis will be repeated for each imputed dataset, and the results will be combined using Rubin's rule as implemented in the SAS Procedure MIANALYZE. The CMH tests statistic 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 imputation FAS (Ratitch et al 2013). Only the p-value from the

combined results will be presented for the confirmatory testing of the primary efficacy endpoints.

Summaries based on non-missing data for KCCQ-TSS, KCCQ-PLS and 6MWD will be presented in tabular format by treatment group. In addition, the medians and 1st and 3rd quartiles of change from baseline for each treatment group will be presented in a way which includes deaths, consistent with the primary efficacy estimand. These additional summary values will be calculated using complete data, ie including patients who died prior to week 16 but excluding patients with missing data due to reasons other than death. A temporary value which is lower than all observed negative change values will be assigned to the deceased patients, and the assigned value amongst the deceased patients will be based on the last value while alive, with lower values assigned to smaller last values while alive. Therefore, the median represents the value compared to which half of the population had a "worse" change from baseline.

The effect size comparing treatment groups will be estimated using the Hodges-Lehmann (HL) estimate of the median difference between dapagliflozin and placebo, together with its 95% confidence interval. The Hodges-Lehmann estimates of the median difference between treatment groups are selected because they are not sensitive to outliers or skewed underlying distributions, and also because they can handle deaths by treating them as the worst possible outcome as done in the ranked approach. This handling of intercurrent events correspond to interpreting the median as "the change from baseline value compared to which half of the patients had a better outcome". The HL estimate is the median value of all paired differences between observations in dapagliflozin versus placebo groups, calculated using the imputed datasets. For KCCQ-TSS and KCCQ-PLS, the imputed dataset is from MAR imputation based on pre-COVID-19 data only. For 6MWD, the imputed dataset is from placebo-based imputation. The calculation will be repeated for each imputation dataset, and the results will be pooled using the central limit theorem to obtain the HL estimate and corresponding confidence interval.

Summaries based on non-missing data for KCCQ-TSS, KCCQ-PLS and 6MWD will be presented in tabular format by treatment group. In addition, the medians and 1st and 3rd quartiles of change from baseline for each treatment group will be presented in a way which includes deaths, consistent with the primary efficacy estimand. These additional summary values will be calculated using complete data, ie, including patients who died prior to week 16 but excluding patients with missing data due to reasons other than death. A temporary value which is lower than all observed negative change values will be assigned to the deceased patients, and the assigned value amongst the deceased patients will be based on the last value while alive, with lower values assigned to smaller last values while alive. Therefore, the

median represents the value compared to which half of the population had a “worse” change from baseline.

Responder analysis

For KCCQ-TSS, the number and percentage of patients by treatment group will be presented across the following categories of change from baseline:

- Death
- Deterioration from baseline (change from baseline at week 16 ≤-5)
- Stable (-5 <change from baseline at week 16 <5)
- Improvement (change from baseline at week 16 ≥5)

where "5" will be replaced by the threshold for CMWPC estimated with anchor-based analysis using change from baseline at week 16 in PGIS in HF symptoms.

For KCCQ-PLS, the number and percentage of patients by treatment group will be presented across the following categories of change from baseline:

- Death
- Deterioration from baseline (change from baseline at week 16 ≤-5)
- Stable (-5 <change from baseline at week 16 <5)
- Improvement (change from baseline at week 16 ≥5)

where "5" will be replaced by the threshold for CMWPC estimated with anchor-based analysis using change from baseline at week 16 in EQ-5D-5L question: “Usual activities”.

For KCCQ-TSS and KCCQ-PLS, the range of possible values is bounded (0-100). Therefore, unless ceiling values (near the upper end) and floor values (near the lower end) are considered, the responder definitions don't allow all patients to qualify as responders. Therefore, two modifications are made to generalize the responder definitions so that they also apply to such patients. A patient who has baseline score ≥95 is classified as having an improvement if the change from baseline at week 16 is greater than zero, and a patient who has baseline score ≤5 is classified as having a deterioration if the change from baseline at week 16 is less than zero. This will be modified according to the estimated threshold for CMWPC.

For 6MWD, number and percentage of patients in each treatment group will be summarised across the following categories:

- Death
- No improvement from baseline (change from baseline at week 16 ≤0 meters)
- Minimal improvement (0 <change from baseline at week 16 <30 meters)
- Improvement (change from baseline at week 16 ≥30 meters)

where "30 meters" will be replaced by the threshold for CMWPC estimated with anchor-based analysis using PGIC in walking ability at week 16.

The threshold for CMWPC for each of the primary efficacy endpoint will be estimated with anchor-based analysis, the details of which are described in Appendix 8.3. Additionally, for KCCQ-TSS, if the derived threshold value for CMWPC does not exceed 10 points, then a logistic regression will be constructed with 10 points improvement from baseline at week 16 to define responders, based on an exploratory analysis of the TOPCAT-HF and HF-ACTION datasets (data not published). The estimated responder threshold for CMWPC using anchor-based analyses for improvement from baseline at week 16, in the KCCQ-TSS, KCCQ PLS and 6MWD, will be indicated in the empirical cumulative distribution function curves and probability density function curves (described in Appendix 8.3). The number and percentage of patients in each category will be summarised by treatment group.

The analysis of responders based on KCCQ-TSS, KCCQ-PLS and 6MWD, respectively, will use the same datasets as used for HL estimates of the median difference between treatment groups. A responder is defined as a patient who had an improvement from baseline in the outcome. Deaths are defined as non-responders, and responder status will be determined based on the change from baseline values derived from imputed KCCQ-TSS, KCCQ-PLS and 6MWD values, respectively, for the patients who have missing data due to reasons other than death. The number and percentage of patients in the responder and non-responder categories will be presented by treatment group. The proportion of responder categories will be compared between treatment groups using a logistic regression model including treatment group, stratification variable (T2DM at randomisation), and baseline value of KCCQ-TSS, KCCQ-PLS and 6MWD, respectively. The observed proportion of 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.

Empirical cumulative distribution function and probability density function curves will be presented by treatment group to display the relative benefit of dapagliflozin over placebo across different ranges of change from baseline at week 16 in each of the primary efficacy endpoints, where patients who die prior to week 16 assessment will be represented with a value which is lower than all observed negative change values (eg, for KCCQ-TSS and KCCQ-PLS, this value will be -101). The patients who had missing outcome due to reason other than death will be excluded for the generation of the curves.

Supportive analysis due to COVID-19

A supportive analysis of KCCQ-TSS and KCCQ-PLS will be performed by not adjusting for the categorical variable weeks impacted by COVID-19 as covariate in the rank ANCOVA

model. The supportive analysis will treat data collected pre- and during COVID-19 equally and use the same imputation datasets as in the main analysis.

Supportive analysis of responders for KCCQ-TSS and KCCQ-PLS will be performed using the placebo-based imputation dataset and same logistic regression model as in the main responder analysis.

Supportive summary statistics will be presented for KCCQ-TSS, KCCQ-PLS and 6MWD, separately for data collected prior to COVID-19 and data collected during COVID-19, based on the onset date of COVID-19 at each site, as described in Section 4.2.4.

No control of Type I error is planned for the responder and supportive analyses.

4.2.4.1 Sensitivity analysis of the primary efficacy endpoints

The following sensitivity analyses will be performed to examine the robustness of main analysis results to missing data handling:

1. For FAS population, the missing due to reason other than death will be imputed in the same way as in the main analysis. Deaths prior to visit 5 (week 16) will be ranked worse than observed data. The ranking amongst the deceased patients will be based on time-to-death, with lower ranks assigned to shorter survival times. The time-to-death equals the date of death minus date of randomisation + 1. This is done to address and evaluate the impact of potential treatment differences in mortality time, on the assessment of the primary efficacy endpoints. Rank ANCOVA model and HL estimate of median difference between treatment group will be conducted in the same manner as in the main analysis.

2. For FAS population, tipping point analysis (TPA) where scenarios in terms of decreased improvement in patients in the active arm with missing data are explored to identify a 'tipping point' where statistical significance would be lost. Details about the tipping point analysis are described in Appendix 8.1.

The sensitivity analyses are only meant to be used when assessing the robustness of the main results to assumptions made in the models and are not meant to be used for labelling claims and no control of Type I error is planned.

4.2.4.2 Supplementary analysis of primary efficacy endpoints

Number and percentage of deaths up to week 16 will be presented by treatment group. To further explore the pattern of death during the study period and the difference on the pattern between the treatment groups, time to death will be presented by treatment group using a Kaplan-Meier curve (Kaplan and Meier 1958). The duration of follow-up is defined as time from the date of randomisation to date of death or date last known to be alive. If a subject dies,

the duration equals the date of death minus date of randomisation + 1. If a subject is last known to be alive, the duration equals the date subject last known to be alive minus date of randomisation + 1. Unless otherwise specified, the plot will be presented only when there are at least 5 events in one treatment group.

A supplementary analysis of KCCQ-TSS, KCCQ-PLS and 6MWD will be performed to examine the robustness of main analysis results to the influence of deaths, by using a different estimand. The supplementary analysis will use FAS population with deaths prior to visit 5 excluded (ie, only including the patients that survived to week 16), and all other missing data will be imputed in the same way as in the main analysis. ANCOVA model will be conducted with non-ranked value of change from baseline at week 16 in KCCQ-TSS, KCCQ-PLS and 6MWD as the continuous outcome. The model will include a factor for treatment group, the stratification variable (T2DM status at randomisation), and baseline KCCQ-TSS, KCCQ-PLS and 6MWD value as covariates. For KCCQ-TSS and KCCQ-PLS, this model will also include a term for the categorical variable of weeks impacted by COVID-19. The estimated mean for each treatment group, and the mean difference between treatment groups with 2-sided 95% confidence intervals and the nominal 2-sided p-values estimated from each imputed dataset will be combined using Rubin's rule, and the combined results will be presented.

For KCCQ-TSS and KCCQ-PLS, an additional supplementary analysis will be performed using a mixed-effect model for repeated measures (MMRM) method. The MMRM analysis will use FAS population with deaths prior to visit 5 excluded (ie, only including the patients that survived to week 16), and all other missing data will be imputed using placebo-based imputation as used in the main rank ANCOVA analysis. The dependent variable will be the non-ranked value of change from baseline in KCCQ-TSS or KCCQ-PLS at post-baseline visit 3 (week 8) and visit 5 (week 16). Treatment group will be fitted as the explanatory variable, visit and interaction of treatment by visit as fixed effects, the stratification variable (T2DM status at randomisation), and baseline outcome value will be fitted as a covariate. To adjust for the impact of COVID-19, the categorical variable of weeks impacted by COVID-19, together with interaction terms for that variable by visit and by treatment, respectively, will be included in the model as covariates. The variance-covariance matrix will be assumed to be unstructured. If the procedure does not converge, then a compound symmetric or Toeplitz variance-covariance matrix will be used instead. The model is:

*Change in KCCQ score = baseline value + treatment + visit + treatment*visit + (weeks impacted by COVID-19) + (weeks impacted by COVID-19)*visit + (weeks impacted by COVID-19)*treatment*

Contrasts will be used to obtain estimates of the treatment differences at week 16, given no COVID-19 impact (ie, value of the categorical variable weeks impacted by COVID-19 is set

to zero). The MMRM analysis will be conducted on each imputation dataset, and the results will be combined using Rubin's rule to present least square means (LSMEANS in SAS), treatment differences in LSMEANS, 95% confidence intervals (CI) of treatment differences, and p-values at week 16.

No control of Type I error is planned for the supplementary analyses.

4.2.4.3 Subgroup analysis of the primary efficacy endpoints

The following baseline and demographic variables are defined for the purpose of efficacy subgroup analysis to assess consistency of effects:

- T2DM status at randomisation (yes, no)
- LVEF value at baseline (>40% and <50%, ≥50%)
- AF at baseline (yes, no)
- Age (≤median, >median)
- Sex (male, female)
- Race (white, black/African American, Other)
- Geographic region (Western Europe/North America, vs rest of world)
- NYHA class at baseline (II, III/IV)
- NT-proBNP (≤median, >median)
- eGFR (<60, ≥60)
- "Site with wearable devices" (no, yes, ie, the site has wearable activity monitor devices), only for 6MWD

The geographic region of Western Europe/North American includes Denmark, Sweden, Italy, Canada, and United States. The number and percentages of patients in subgroups will be presented by treatment group in the baseline summary data. The summary of primary efficacy data will be presented by treatment group, stratified by each subgroup variable.

Since the data from *MoveMonitor* and *MoveTest* are collected in a subset of sites that have the capability of providing the wearable activity monitor to patients, a subgroup analysis of 6MWD will be performed to evaluate whether the estimated treatment effects on 6MWD are different between "site(s) with wearable devices" versus other sites.

For each of the primary efficacy endpoints, the subgroup analysis will be performed separately in each subgroup, using the same method, HL estimate of median difference between treatment groups, rank ANCOVA model and placebo-based MI, as described in Section 4.2.4 for primary efficacy endpoints. The nominal 2-sided p-values and descriptive summary measures that are not sensitive to potential outliers (medians) will be presented for each subgroup.

For KCCQ-TSS, KCCQ-PLS and 6MWD responders (as a binary outcome), respectively, the subgroup analysis will be performed to include each subgroup variable (if it's not already in the model) and the interaction of the subgroup variable and treatment group in the logistic regression model that already had treatment group, stratification variable (T2DM status at randomisation), and baseline value as covariates. For KCCQ-TSS and KCCQ-PLS, the imputed datasets by MAR imputation based on pre-COVID-19 data only and the logistic regression model adjusting for weeks impacted by COVID-19 as used in the main responder analysis will be used for subgroup responder analysis. For 6MWD, the impuated datasets by placebo-based imputation as used in the main responder analysis will be used for subgroup responder analysis. The interaction p-value will be presented in addition to the observed proportion of responders for each treatment group, odds ratio between treatment groups with 95% confidence interval and p-value for each subgroup. Odds ratios with confidence intervals for all subgroup levels will be presented in a forest plot, including observed proportions and interaction p-value. The p-values for the subgroup analyses and interaction terms will not be adjusted for multiple comparisons as the tests are exploratory and will be interpreted descriptively.

If any subgroup category of a factor contains less than ten percent of the patients, that subgroup category will be excluded from the subgroup analysis using rank ANCOVA model or logistics regression model, and only descriptive data will be summarised for that subgroup category.

4.2.5 Analysis of secondary efficacy variable

The secondary efficacy endpoint using data assessed by the wearable activity monitors *MoveMonitor*, ie, change and relative change from baseline at end of study in total time spent in LVPA, will be analysed using the same method, HL estimate of median difference and rank ANCOVA model to analyse the hierarchical composite endpoint, as described in Section 4.2.4 for the primary efficacy endpoints. The analyses use complete data, ie no patients with missing data due to reasons other than death, therefore, the HL estimate of median difference between treatment groups and its asymptotic 95% CI will be generated based on Wilcoxon's rank sum test. All models (both for change from baseline and for relative change from baseline) are adjusted for the baseline value, by including this as a continuous variable in the model specification. The expected difference in the result, between models looking at change from baseline and models looking at relative change from baseline, is the scale of the outcome variable, ie the scale on which the difference between treatment groups is presented. One scale may be more intuitive for clinical interpretation than the other due to the sometimes

uncommon units provided in endpoints assessed by the wearable activity monitors (eg, milli-*g*).

Summary data for total time spent in LVPA based on standard filters, 4 additional filter settings of interest and also without applying any filters at all (Appendix 8.2) will be presented by treatment group.

The sensitivity analysis ranking death based on time-to-death and the supplementary analyses using ANCOVA model will be performed in the same manner as described in Section 4.2.4.1 and Section 4.2.4.2 except for using complete data, ie no patients with missing data due to reasons other than death, to examine the robustness of main analysis results to the handling of death, respectively. No TPA and MMRM will be conducted for the secondary efficacy endpoint.

Subgroup analyses will be performed in the same manner as described in Section 4.2.4.3 for the primary efficacy endpoints. The summary data, stratified by subgroups, will also be presented by treatment group.

4.2.6 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 "Duration of 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 "Duration of actual exposure".

"Duration of exposure" and "Duration of actual exposure" will be presented descriptively. Patients who receive IP which is not consistent with the treatment he or she was randomised to receive will also be listed.

Treatment periods

The summaries for the "on-treatment" period will include data with onset date on or after first dose of randomised study drug and on or before 30 days after last dose of study drug, and no later than visit 5 (week 16).

The summaries for the "on and off-treatment" period will include data with onset date on or after dose of randomised study drug.

All summaries of safety endpoints described in Section 4.2.6.1 to Section 4.2.6.7 below will be presented for the on-treatment period, unless otherwise stated.

4.2.6.1 Adverse events

Adverse Events (AEs) will be classified by Primary System Organ Class (SOC) and Preferred Term (PT) according to the Medical Dictionary for Regulatory Activities (MedDRA). Summaries of AEs will use the version of MedDRA that is current at the time of database lock.

Summary table of the number and percent of patients with AEs, SAEs, DAEs, AEs leading to amputations, and AEs leading to a risk for lower limb amputations ("preceding events") (Section 3.3.1) will be presented by treatment group. The number and percent of patients with AEs will also be presented by SOC, PT and treatment group..

4.2.6.2 Serious adverse events (SAE)

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 only. The cut-off for most common SAEs will be data driven (based on blinded data) and if no appropriate cut-off is identified all SAEs will be presented by PT.

AEs with outcome death will be presented separately by SOC, PT and treatment group for the on and off treatment period.

4.2.6.3 Adverse events leading to discontinuation (DAE)

The number and percent of patients with DAEs will be presented by SOC, PT and treatment group.

4.2.6.4 Amputations and preceding events

The number and percent of patients with AEs leading to amputation and AEs reflecting potential risk factors for lower limb amputations ("preceding events") (Section 3.3.1) will be presented by SOC, PT and treatment group. Amputations are presented for the on and off treatment period.

4.2.6.5 Laboratory evaluation

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

Laboratory variables, including the absolute values at each scheduled visit and the corresponding change from baseline, will be summarised by treatment group.

4.2.6.6 Marked laboratory abnormalities

The number and percent of patients with a marked abnormality in clinical laboratory tests listed in Table 2 will be summarised over time by treatment group.

Laboratory abnormalities will be evaluated based on marked abnormality (MA) criteria for the laboratory variables listed in Table 2. When there is more than one limit for a variable, summaries will be provided for each limit.

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.

If value at baseline is missing, MA is determined by the laboratory test results at each post baseline visit.

Table 2 Marked abnormality criteria for safety laboratory variables

		Marked Abnormality Criteria	
Clinical laboratory variables	SI units	Low	High
Haematocrit	RATIO	< 0. 20	> 0. 55
Haematocrit	RATIO		> 0.60
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

BL is the baseline measurement

Laboratory MAs occurring during the on-treatment period will be summarised 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.

4.2.6.7 Vital signs

For vital signs, including body weight and waist circumference, the absolute values at each scheduled visit and the corresponding change from baseline will be summarised by treatment group.

Absolute values and changes from baseline will also be listed and compared to the AZ-defined reference ranges (Table 3), and classified as low (below range), normal (within range or on the limits) or high (above range). All values falling outside the reference ranges will be flagged.

Table 3 Vital signs reference ranges

Parameter	Standard Units	Lower Limit	Upper Limit	Change from Baseline Criteria
Diastolic Blood Pressure (sitting)	mmHg	60	100	±15
Systolic Blood Pressure (sitting)	mmHg	90	160	±30
Pulse Rate (sitting)	Beats/min	50	100	±20

4.2.7 Analysis of exploratory efficacy endpoints

The analysis of exploratory efficacy endpoints will use same approach as for the main analysis of primary and secondary efficacy endpoints, that is, use rank ANCOVA model to analyse the hierarchical composite endpoints for continuous outcome variables, and logistic regression model for binary outcome variables. For continuous outcomes, hierarchical composite endpoints will be derived, with a ranking scheme assigning deaths ranks which are lower than observed data ranks and ranking among deaths will be based on the last value while alive. For binary outcomes, deaths will be considered as not improved or as worsened, as appropriate. All exploratory endpoints, except for the KCCQ endpoints, will be analysed using complete data, ie, no multiple imputation (MI) of missing data is performed for patients with missing data due to reasons other than death. Summary data for each exploratory endpoint will be presented by treatment group.

For KCCQ domains including TSS domains (symptom frequency and symptom burden), overall summary score, symptom stability domain, self-efficacy domain, social limitation domain and QoL domain, a placebo-based MI with predictive mean matching will be used to impute the missing values on the outcomes at week 16 due to reasons other than death (Appendix 8.1). HL estimate of median difference between treatment groups and rank ANCOVA model will be performed in the same manner as in the main analysis, ie, including the categorical variable, weeks impacted by COVID-19 as covariate in the model. A supportive analysis of rank ANCOVA model not including the weeks impacted by COVID-19 variable as covariate will be performed. In the supportive analysis, HL estimate of median difference between treatment groups will use the placebo-based MI imputation dataset.

Analysis of change or relative change from baseline at week 16 in the exploratory endpoints assessed by *MoveTest*, and change or relative change from baseline at end of study in the exploratory endpoints assessed by *MoveMonitor*, including VMUs per minute and movement intensity during walking, will be performed using the same method; rank ANCOVA model to

analyse the hierarchical composite rank-based endpoint, as described in Section 4.2.5, for the secondary efficacy endpoint assessed by *MoveMonitor*. These data will be analysed using complete data, ie no patients with missing data due to reasons other than death.

Summary data for NYHA classification at baseline and each analysis time point, and across the categories for worsened NYHA classification from baseline at week 16 will be presented by treatment group. Analysis of the proportions of patients with worsened NYHA classification from baseline at week 16 will be performed using logistic regression, classifying all deaths prior to week 16 assessment as worsened NYHA classification. The logistic regression model will include treatment group, stratification variable (T2DM status at randomisation), and baseline NYHA class (a categorical variable) as covariates. The observed proportion, odds ratio between treatment groups and its 95% confidence interval and corresponding 2-sided p-value will be presented.

Summary data for OTB at week 8 and week 16 will be presented by treatment group. Logistic regression will be performed to analyse the proportion of patients with positive OTB at week 16 (versus all other collapsed categories and death). The logistic regression model will include treatment group and stratification variable (T2DM at randomisation) as covariates.

The responses to the individual questions of EQ-5D-5L and EQ-VAS score will be summarised at baseline and week 16. Change from baseline at week 16 will be summarised to the following categories: question completed at baseline and the time point, and 3 derived categories (No change from baseline, Deteriorated from baseline and Improved from baseline) for each individual question, and change from baseline in EQ-VAS score at week 16 will be summarised, with mean, SD, median, Q1 and Q3.

Analyses of change from baseline at week 16 in NT-proBNP, oxygen saturation difference after 6MWT, dyspnoea and fatigue, systolic BP difference after 6MWT, body weight, and eGFR will be performed using same method, rank ANCOVA model to analyse the hierarchical composite endpoint, as described in Section 4.2.4 for the primary efficacy endpoints. The comparison using rank ANCOVA model will adjust for baseline value as covariate and is stratified by the stratification variable (T2DM at randomisation).

5 INTERIM ANALYSES (NOT APPLICABLE)

6 CHANGES OF ANALYSIS FROM PROTOCOL

Statistical power was re-estimated for a comparison of group-level averages of within-patient change, instead of responder analysis (as in the CSP) due to a change of main estimation method for effect size from logistic regression to the HL estimate of median difference. The sample size estimation stays unchanged.

7 REFERENCES

American Thoracic Society 2002
American Thoracic Society. ATS Statement: Guidelines for the Six-Minute Walk Test. Am J Respir Crit Care Med 2002;166:111-7.

Brandes et al 2012
Brandes M, Van Hees V, Hannöver V, Brage S. Estimating energy expenditure from raw accelerometry in three types of locomotion. Med Sci Sports Exerc 2012 Nov; 44(11):2235-2242.

Colley et al 2010
Colley RC, Connor Gorber S, Tremblay MS. Quality control and data reduction procedures for accelerometry-derived measures of physical activity. Health Reports 2010 Mar; 21(1):63-69.

CHMP/EMA/SAWP 2018
Committee for Medicinal Products for Human Use, European Medicines Agency, Scientific Advice Working Party. Qualification of opinion on Proactive in COPD. 2018 April 12. Date last accessed: March 18, 2019.

Dijkstra et al 2010
Dijkstra B, Kamsma YP, Zijlstra W. Detection of gait and postures using a miniaturized triaxial accelerometer-based system: accuracy in patients with mild to moderate Parkinson's disease. Arch Phys Med Rehabil 2010 Aug; 91(8):1272-1277.

Dmitrienko et al 2011
Dmitrienko A, Millen BA, Brechenmacher T, Paux G. Development of gatekeeping strategies in confirmatory clinical trials. Biom J 2011 Nov;53(6):875-893.

FDA 2020
Food and Drug Administration Center for Drug Evaluation and Research. Qualification of the Kansas City Cardiomyopathy Questionnaire Clinical Summary Score and its Component Scores: A Patient-Reported Outcome Instrument for Use in Clinical Investigations in Heart Failure. April 2020. Clinical Outcome Assessment DDTCOA-000084.

Ferreira et al 2016
Ferreira JP, Duarte K, Graves TL, Zile MR, Abraham WT, Weaver FA, et al. Natriuretic Peptides, 6-Min Walk Test, and Quality-of-Life Questionnaires as Clinically Meaningful Endpoints in HF Trials. J Am Coll Cardiol 2016;68(24):2690-707.

Filippatos et al 2017
Filippatos G, Maggioni AP, Lam CSP, Pieske-Kraigher E, Butler J, Spertus J, et al. Patient-reported outcomes in the SOluble guanylate Cyclase stimulatoR in heArt failurE patientS with PRESERVED ejection fraction (SOCRATES-PRESERVED) study. Eur J Heart Fail. 2017;19(6):782-91.

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

Holland et al 2014
Holland AE, Spruit MA, Troosters T, Puhan MA, Pepin V, Saey D, et al. An Official European Respiratory Society/American Thoracic Society technical standard: field walking tests in chronic respiratory disease. Eur. Respir J. 2014;44(6):1428-46.

Kaplan and Meier 1958
Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of American Statistical Association. 1958; 53:457-481.

Kosiborod et al 2020
Kosiborod et al. Effects of Dapagliflozin on Symptoms, Function, and Quality of Life in Patients With Heart Failure and Reduced Ejection Fraction: Results From the DAPA-HF Trial. Circulation. 2020 Jan 14;141(2):90-99.

Langer et al 2009
Langer D, Gosselink R, Sena R, Burtin C, Decramer M, Troosters T. Validation of two activity monitors in patients with COPD. Thorax 2009 Jul; 64(7):641-642.

Levey at al 2009
Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 5; 150(9):604-12.

Lewis et al 2017
Lewis EF, Claggett BL, McMurray JJV, Packer M, Lefkowitz MP, Rouleau JL, et al. Health-Related Quality of Life Outcomes in PARADIGM-HF. Circ Heart Fail 2017; 10(8):e003430.

Nassif et al 2019
Nassif et al. Dapagliflozin Effects on Biomarkers, Symptoms, and Functional Status in

Patients With Heart Failure With Reduced Ejection Fraction: The DEFINE-HF Trial. Circulation. 2019 Oct 29;140(18):1463-1476.

O'Keeffe et al 1998
O'Keeffe ST, Lye M, Donnnellan C, Carmichael DN. Reproducibility and responsiveness of quality of life assessment and six-minute walk test in elderly heart failure patients. Heart 1998;80:377-82.

Rabinovich et al 2013
Rabinovich RA, Louvaris Z, Raste Y, Van Remoortel H, Giavedoni S, et al. Validity of physical activity monitors during daily life in patients with COPD. Eur Respir J 2013 Nov; 42(5):1205-1215.

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

Shoemaker et al 2013
Shoemaker MJ, Curtis AB, Vangsnes E, Dickinson MG. Clinically meaningful change estimates for the six-minute walk test and daily activity in individuals with chronic heart failure. Cardiopulm Phys Ther J 2013 Sep; 24(3):21-29.

Spertus et al 2005
Spertus J, Peterson E, Conard MW, Heidenreich PA, Krumholz HM, Jones P, et al. Monitoring clinical changes in patients with heart failure: A comparison of methods. Am Heart J 2005;150:707-15.

Storm et al 2015
Storm FA, Heller BW, Mazzà C. Step detection and activity recognition accuracy of Seven physical activity monitors. PLoS One 2015 Mar 19; 10(3):e0118723.

Van Hees et al 2009
Van Hees V, van Lummel RC, Westerterp KR. Estimating activity-related energy expenditure under sedentary conditions using a tri-axial seismic accelerometer. Obesity 2009 Jun; 17(6):1287-1292.

Van Remoortel et al 2012
Van Remoortel H, Giabedoni S, Raste Y, Burtin C, Louvaris Z, Gimeno-Santos E, et al. Validity of activity monitors in health and chronic disease: a systematic review. Int J Behav Nutr Phys Act 2012 Jul 9; 9:84.

Van Roie et al 2019
Van Roie E, Van Driessche S, Hujiben B, Baggen R, van Lullem RC, Delecluse C. A body-fixed-sensor-based analysis of stair ascent and sit-to-stand to detect age-related differences in leg-extensor power. PLoS One 2019 Jan 17; 14(1):e0210653.

Zakeri et al 2015
Zakeri R, Levine JA, Koepp GA, Borlaug BA, Chirinos JA, LeWinter M, et al. Nitrate's effect on activity tolerance in heart failure with preserved ejection fraction trial: rationale and design. Circ Heart Fail 2015;8(1):221-8.

8 APPENDIX

8.1 Accounting for missing data

Subject 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. Because of the short study duration, the number of deaths observed during the study is expected to be low. Deaths will be included in each estimand as described in Section 4.1.1.

An increased frequency of missing data is expected after the onset date of COVID-19 at each site, as described in Section 4.2.4. This missingness is assumed to be independent of treatment assignment and if a data point is missing due to reasons other than death, it will be imputed using the standard approach.

For all analyses of 6MWD, for the main hypothesis tests of KCCQ-TSS, KCCQ-PLS, and other KCCQ endpoints, for all handling of data prior to the onset date of COVID-19, and for the supportive analysis using all data without accounting for the potential impact of COVID-19, a placebo-based MI will be undertaken.

In the main HL estimate of median difference and responder analysis for KCCQ-TSS and KCCQ-PLS, all data collected during COVID-19 (ie, after the onset date at each site) is imputed assuming missing at random (not placebo-based MI) as described below.

Note that MI of missing data is not performed for safety variables and is not performed for all efficacy variables. Only the following efficacy variables are affected:

- KCCQ (KCCQ-TSS, KCCQ-PLS, 6 domain scores and 1 summary score included in the exploratory efficacy endpoints, no explicit imputation of the 23 individual items/questions), TPA is only performed for KCCQ-TSS and KCCQ-PLS
- 6MWD

Furthermore, imputation is always done for the values at each visit. This means the change values (or relative change values) are never directly imputed. Only the components of the change (and relative change) are imputed for the efficacy variables listed above. The exploratory efficacy variables which are directly imputed for the values at each visit using MI are KCCQ domain and summary scores., the data collected during COVID-19 will be treated equally as the observed data pre-COVID-19 and used for the imputation.

Placebo-based MI

In the main analysis, the underlying assumption is that the trajectory for subjects in the dapagliflozin group who dropped out for a treatment-related reason is similar to that of the

placebo subjects, and that subjects who dropped out for other reason also have a similar trajectory as the completers in the placebo group and are imputed assuming missing not at random. This placebo-based MI will be done by constructing the imputation of missing observations in the treatment groups using the observed data in the placebo group only, that is, one imputation model of placebo outcomes will be used to impute missing values for all discontinued subjects in both treatment groups. For ITT analyses, this approach is considered more conservative than assuming data is missing completely at random (which would be the case if missing data was simply excluded) or assuming that data is missing at random (which would be the case if a mixed model with repeated measures was used directly) because the assumptions mean that as soon as subjects withdraw for a treatment related reason, they begin to worsen immediately.

Predictive mean matching approach

The first step in placebo-based predictive mean matching consists of estimating regression coefficients from a linear regression model among subjects in the placebo arm without missing data. New regression coefficients are then drawn at random from the posterior predictive distribution of the linear model and the new coefficients are then used to generate predicted values for all subjects, including both those with missing data and those without. The set of closest subjects is then identified for each subject with missing data and a replacement for each missing value is drawn at random from the observed data among the closest subjects. This process, except for the first deterministic linear regression step, is repeated for each imputation dataset. The missing data on the outcome variable at baseline, visit 3 and 5 will be imputed in a sequential manner. Missing data at baseline will be imputed making the most liberal assumptions, because the likelihood of baseline data being missing should not be influenced by randomised treatment assignment. Then, the missing value at visit 3 will be imputed (treated as monotone missing data) including the potentially imputed baseline value as a covariate in the imputation model, and finally the missing value at visit 5 will be imputed in a similar way except that the (potentially imputed) value at baseline and at visit 3 will be used as a covariate in the imputation model for visit 5.

MAR imputation based on pre-COVID-19 data only

In the main HL estimate of median difference for KCCQ endpoints and responder analysis for KCCQ-TSS and KCCQ-PLS, all data collected during COVID-19 (ie, after the onset date at each site) will be treated as missing and imputed under the assumption of MAR. The purpose of this MI assuming MAR is to only use data collected prior to COVID-19 to inform imputation of data during COVID-19, thus preserving the original efficacy estimand. The imputation will be done in three steps. First, identify the visit that is potentially impacted by COVID-19, ie, visit date after the onset date of COVID-19 at the site. If visit date is missing,

visit target date (Table 1) will be compared to the onset date of COVID-19 to determine whether the visit is potentially impacted by COVID-19 or not. Then set the value of endpoint at this visit as missing, regardless of whether the visit has observed data of the endpoint, and treat them as if they had died (ie, exclude them), in the second step of the sequential imputations. In the second step, impute the missing data (not potentially impacted by COVID-19) on the outcome variable at baseline, visit 3 and 5 in a sequential manner using placebo-based imputation with predictive mean matching approach and including stratification variable (T2DM at randomisation) and value of the outcome variable at the previous visits in the imputation model. The missing potentially impacted by COVID-19 will be excluded in each sequential imputation. In third and final step, impute the missing potentially impacted by COVID-19 at visit 3 and 5 using the imputed data from step 2 and MAR imputation, including treatment group, stratification variable (T2DM at randomisation) and value of the outcome variable at the previous visits in the imputation model.

Tipping point analysis

A TPA will be employed, to assess the sensitivity of the results of the main analysis to the handling of missing data. The TPA will only be performed for the primary efficacy endpoints and only when a main analysis result is statistically significant. In TPA, missing data is first replaced using the MI procedure described above. The imputed post-baseline values in the active arm are then subsequently shifted toward a null hypothesis, until the significance of model estimates is overturned. The plausibility of the magnitude in the shift parameter required to overturn the significance is then interpreted from a clinical perspective, to evaluate the robustness of the results to the approach used to handle missing data. The upper boundary for shift parameter for 6MWD is set to 100 meters, since if a shift parameter of 100 meters could not overturn the significance, it indicates that significance is not sensitive to the handling of missing data. For KCCQ-TSS and KCCQ-PLS, the upper limit of the shift parameter is set to 40 points. Values will not be shifted outside of the range of the outcome variable.

8.2 Wearable activity monitors

Data from wearable activity monitors

Two types of wearable activity monitors (DynaPort *MoveTest* and DynaPort *MoveMonitor*; McRoberts BV, The Hague, The Netherlands) are available for all patients who are randomised at a subset of sites where the devices are available. The two different wearable activity monitors are identical in dimensions and appearance but serve different purposes; the *MoveTest* is intended for supervised use at the site for a few minutes at a time and the *MoveMonitor* is intended for unsupervised use at home over several days. The DynaPort device (both types) contains a tri-axial accelerometer, a rechargeable battery, raw data storage

on a Micro-SD card. The device is worn on a belt around the waist and the dimensions are 85 mm x 58 mm x 11.5 mm, it weighs 55 grams, the sample frequency is 100 Hz and the battery lifetime can sustain at least 7 days of continuous measurement. Both wearable activity monitors only need to be placed around the waist (like a belt) and worn by the patient, as instructed. No additional manipulation is required by the patient (eg, turning the device on/off or starting/stopping the measurement). The setup of the wearable activity monitors is handled entirely by the site staff.

The first wearable activity monitor is the *MoveTest*, used at the site during the 6MWT (Van Roie et al 2019). The *MoveTest* assesses performance during the 6MWT and can potentially provide different, more granular data on physical ability than the manual 6MWT. In addition to an estimate of the distance walked during the 6MWT, for use in an exploratory analysis, the *MoveTest* also provides other parameters for use in exploratory endpoints to assess physical ability (Table 4). The analysis time points for variables collected using the *MoveTest* are the same as for the primary efficacy endpoints. As the *MoveTest* is used during in-clinic visits, more than one measurement can be available for a given visit, in the same way as for other variables measured in-clinic and such cases will be handled in the same way as specified in Section 3.1.3. The two endpoints measured using the *MoveTest* are designated as exploratory.

The second wearable activity monitor is the *MoveMonitor*, used by the patient at home, to assess physical ability during day-to-day activities. The *MoveMonitor* has a validated classification algorithm by which activities such as sitting, standing and walking can be identified, along with postures such as prone, lying, sitting and standing, and different measures of physical activity can be assessed using these classifications (Langer et al 2009, Dijkstra et al 2010, Storm et al 2015). As the *MoveMonitor* is used by the patient at home, more than one measurement cannot be available for a given time period unless the use of the wearable activity monitor was not according to the instructions. Any such violation of instructions will result in a missing per "visit" summary value for that time period, as the integrity of the collected data cannot be ascertained.

Data from the two different types of wearable activity monitors (in-clinic versus at home) will never be aggregated.

Data collected from the *MoveMonitor* will span a period of 7 days at each time point for collection. Data collected during the 7-day period starting on the day of the enrolment visit 1 will be retrieved at visit 2a and this data constitutes the baseline for each patient. Data collected during the 7-day period starting on the day of visit 3 (week 8) and data collected during the 7-day period starting on the day of visit 4 (week 14), will be retrieved at visit 5 (week 16) and these data comprise the follow-up for each patient.

Daily summaries for each of the *MoveMonitor* parameters that are specified as secondary or exploratory efficacy endpoints are aggregated, using pre-specified algorithms, into per-visit summaries at baseline, the week starting on the day of visit 3, and the week starting on the day of visit 4 (end of study). The change from baseline (or relative change from baseline) at end of study will constitute the endpoint for all variables based on the *MoveMonitor* data.

The secondary and exploratory endpoints using *MoveMonitor* and *MoveTest* data are summarised in Table 4.

Table 4 Endpoints based on data from wearable activity monitors

Endpoint	Secondary/ Exploratory	*MoveMonitor*/ *MoveTest*
Change from baseline at end of study in total time spent in LVPA	Secondary	*MoveMonitor*
Change from baseline at end of study in VMUs per minute	Exploratory	*MoveMonitor*
Change from baseline at end of study in movement intensity during walking	Exploratory	*MoveMonitor*
Change from baseline at end of study in movement intensity when walking for durations of >20 seconds	Exploratory	*MoveMonitor*
Change from baseline at end of study in number of steps, excluding steps from walking in stairs	Exploratory	*MoveMonitor*
Change from baseline at end of study in total time spent in MVPA	Exploratory	*MoveMonitor*
Change from baseline at Week 16 in distance walked during 6MWT	Exploratory	*MoveTest*
Change from baseline at Week 16 in number of stops during 6MWT	Exploratory	*MoveTest*
Change from baseline at end of study in number of sitting shifts during night rest	Exploratory	*MoveMonitor*
Change from baseline at end of study in total number of activity counts during worn periods (activity counts are zero when the device is not worn)	Exploratory	*MoveMonitor*
Total wear time at end of study	Exploratory	*MoveMonitor*

Baseline for wearable activity monitor *MoveMonitor* outcomes consists of a 7-day period measured on the week starting at enrolment visit. End of study for wearable activity monitor *MoveMonitor* outcomes consists of a 7-day period measured on the week starting at visit 4 (week 14).

For the main analysis of endpoints based on the *MoveMonitor*, a set of predefined data reduction rules, referred to as 'filters', are employed to ensure that the *MoveMonitor* data collected corresponds to instructed use and is representative of the physical ability intended to be captured. This is common practice in studies using wearable activity monitors and has been shown to positively influence the reliability of the data (Colley et al 2010). Filters will define the data considered available for the analysis of each endpoint. The same set of predefined filters will be applied to all endpoints measured using the *MoveMonitor*. To assess the

sensitivity of the results to the filter definitions, additional analyses will be performed with predefined variations of the filter definitions and also without applying any filters at all.

Filter definitions

Filters are comprised of data reduction rules, used to ensure that the *MoveMonitor* data contributing to the analysis of endpoints come from periods when the device was used according to instructions, that the data sufficiently reflects the activity during each day and that the period of use is long enough to be representative of physical ability during day-to-day activities.

The standard filter, applied at each analysis time point, has the following rules:

- A minimum of 10 hours of wear time, between 6:00 AM and 10:00 PM, will be required for the data during a day to be considered 'sufficient'.
- A minimum of 3 days with 'sufficient' data, out of the total 7 days, will be required for a patient to be considered to have a non-missing value in the analysis.

The standard filter is used for all efficacy endpoints based on *MoveMonitor* data. To assess the sensitivity of the results to the filter definitions, the variations of the filter rules will be applied to the secondary efficacy endpoints, ie, total time spent in LVPA.

Variations of the filters rules consist of:

- Setting the minimum no. of hours of wear time, between 6:00 AM and 10:00 PM, to 6 hours and 14 hours, respectively, for data to be classified as 'sufficient'.
- Setting the minimum no. of days with 'sufficient' data to 1 day and 5 days, out of the total 7 days, for the patient to be classified as having a non-missing outcome in the analysis.

The variations of the filter rules yield 4 additional filter settings of interest:

- The first variation consists of requiring a minimum of 3 days, out of the total 7 days, with at least 6 hours of wear time, between 6:00 AM and 10:00 PM.
- The second variation consists of requiring a minimum of 3 days, out of the total 7 days, with at least 14 hours of wear time, between 6:00 AM and 10:00 PM.
- The third variation consists of requiring a minimum of 1 day, out of the total 7 days, with at least 6 hours of wear time, between 6:00 AM and 10:00 PM.
- The fourth variation consists of requiring a minimum of 5 days, out of the total 7 days, with at least 14 hours of wear time, between 6:00 AM and 10:00 PM.

In addition of 4 additional filter settings of interest listed above, total time spent in LVPA will also be derived based on all available *MoveMonitor* data collected during the 7-day period at baseline and at end of study, regardless of the number of days or number of hours per day

during which the device was worn (as long as it is possible to derive per-visit summaries). This analysis does not apply any filters at all.

Summary data for total time spent in LVPA based on standard filters, 4 additional filter settings of interest and also without applying any filters at all will be presented by treatment group.

Sensitivity analyses and supplemental analyses defined for the primary efficacy endpoints (except for TPA and analyses related to COVID-19, including the MMRM analysis) will also be performed for the secondary efficacy endpoint based on *MoveMonitor* data, ie total time spent in LVPA. Additional analysis of total time spent in LVPA may be performed if deemed appropriated.

No sensitivity analyses and supplemental analyses will be performed for exploratory objectives.

Handling of missing data due to death
Deaths are included in the definition of the estimand and handled by the ranking scheme, analogously to the primary efficacy endpoints.

Handling of missing data not due to death
Missing data which is not due to death will be excluded from analysis of all *MoveMonitor* and *MoveTest* endpoints

The secondary efficacy variable based on data from the *MoveMonitor*
The secondary efficacy variable based on data from the *MoveMonitor* is change from baseline at end of study in total time spent in LVPA, as assessed by the *MoveMonitor*.

Secondary endpoint: change from baseline at end of study in total time spent in LVPA
Total time spent in LVPA is defined as the total wear time spent with an energy expenditure ≥1.5 METs. These calculations rely on the estimation of energy expenditure, which has been validated previously (Van Hees et al 2009, Van Remoortel et al 2012, Brandes et al 2012, Rabinovich et al 2013). Total time spent in LVPA is measured in hours. Time spent in LVPA per day is calculated as the sum of all time spent with an energy expenditure ≥1.5 METs during each day. After applying the filters for *MoveMonitor* data, the time spent in LVPA per day will be summed over the days that have "sufficient" wear time at baseline, week 8 and end of study, respectively. The estimand for change from baseline at end of study in total time spent in LVPA is defined analogously to the primary efficacy endpoints, with intercurrent events handled in the same way.

Change in total time spent in LVPA from baseline at end of study will then be calculated as

$$LVPA_{EOS} - LVPA_{BL}$$

where $LVPA_{BL}$ represents the total time spent in LVPA at baseline and $LVPA_{EOS}$ represents the time spent in LVPA at end of study.

Analysis of the secondary efficacy variable

The main estimator for the secondary efficacy variable from the *MoveMonitor* is the rank that will be used for rank ANCOVA model, with deaths handled analogously to the primary efficacy endpoints.

Exploratory endpoint: change from baseline at end of study in VMUs per minute

VMUs per minute is defined as the root mean square of the values along the y-signal (y-axis: caudal-cranial) and are calculated according to the norm. VMUs per minute represent a proxy for the intensity of effort (CHMP/EMA/SAWP 2018). The unit for VMUs is activity counts. The VMUs per minute, per day, is calculated as the sum of the VMUs per minute measured during all worn periods during each day, independent of activity (Shoemaker et al 2013). After applying the filters for *MoveMonitor* data, the VMUs per minute per day will be summed over the days that have "sufficient" wear time at baseline, week 8 and end of study, respectively. The estimand for change from baseline at end of study in VMUs per minute is defined analogously to the primary efficacy endpoints, with intercurrent events handled in the same way.

Change in VMUs per minute from baseline at end of study will then be calculated as

$$VMU_{EOS} - VMU_{BL}$$

where VMU_{BL} represents the mean VMUs per minute at baseline and VMU_{EOS} represents the mean VMUs per minute at end of study.

Exploratory endpoint: change from baseline at end of study in movement intensity during walking

Movement intensity during walking will be assessed by *MoveMonitor* data as mean movement intensity during periods classified as 'walking' (at least 3 consecutive steps during a 'standing' posture are required for the classification as 'walking'). Movement intensity is defined as the root mean square of the sum of the values along the x-, y-, and z-signals (x-axis: posterior-anterior; y-axis: caudal-cranial; z-axis: medial-lateral). The unit for movement intensity is milli-*g* (1 milli-$g = 9.81 \times 10^{-3}$ ms^{-2}). The daily mean movement intensity is calculated as the sum of mean movement intensity measured during each walking period divided by the sum of the lengths of the walking periods. After applying filters to the *MoveMonitor* data, the mean movement intensity per day will be summed over the days that

have "sufficient" wear time at baseline, week 8 and end of study, respectively. The estimand for change from baseline at end of study in movement intensity during walking is defined analogously to the primary efficacy endpoints, with intercurrent events handled in the same way.

Change in movement intensity during walking from baseline at end of study will then be calculated as

$$MI_{EOS} - MI_{BL}$$

where MI_{BL} represents the mean movement intensity during walking at baseline and MI_{EOS} represents the mean movement intensity during walking at end of study.

Other exploratory endpoints based on data from *MoveMonitor* and *MoveTest*

- Change from baseline at end of study in movement intensity when walking for durations of >20 seconds based on *MoveMonitor* data, defined as movement intensity measured in *milli-g*, during periods when the physical activity was classified as 'walking' and lasted for more than 20 seconds
- Change from baseline at end of study in total number of steps based on *MoveMonitor* data, defined as the total number of steps excluding steps from walking in stairs
- Change from baseline at end of study in total time spent in MVPA based on *MoveMonitor* data, defined as the wear time spent with an energy expenditure ≥3 METs measured in hours
- Change from baseline at week 16 in distance walked during 6MWT based on *MoveTest* data, defined as the total distance covered by the patient measured in meters
- Change from baseline at week 16 in number of stops during 6MWT based on *MoveTest* data, defined as the total number of walking interruptions
- Change from baseline at end of study in total number of sitting shifts during night rest based on *MoveMonitor* data, defined as a transition between a posture classified as 'sitting upright' and any posture classified as 'lying' (ie, 'prone', 'lying on the left', 'lying on the right' or 'supine'), with night rest defined as the longest period >3 hours without interruptions of more than 30 minutes (eg, walking) identified through a process where, between 12:00 PM on the day of interest and 12:00 PM the following day, all lying periods ≥10 minutes are stitched together with sitting and not wearing periods <15 minutes
- Change from baseline at end of study in total number of activity counts during worn periods based on *MoveMonitor* data, activity counts are zero when the device is not worn

- Total wear time at end of study based on *MoveMonitor* data, defined as the total time measured in hours when the number of activity counts was greater than zero

Analysis of the exploratory efficacy endpoints

The estimand for each exploratory endpoint is defined analogously to the primary efficacy endpoints, with intercurrent events handled in the same way. The estimator and estimation of treatment effect for each exploratory endpoint, are analogous to those for the primary efficacy endpoints. Each exploratory endpoint based on *MoveMonitor* data will be analysed by both change from baseline and relative change from baseline.

8.3 Anchor-based analyses

Clinically meaningful threshold

Thresholds for meaningful within-subject change will be estimated according to predefined algorithms using anchor-based approaches, supplemented with empirical cumulative distribution function curves and probability density function curves. Clinically meaningful thresholds will be estimated for the following efficacy variables:

- Change from baseline in KCCQ-TSS at week 16
- Change from baseline in KCCQ-PLS at week 16
- Change from baseline in 6MWD at week 16

The analysis will be performed on the FAS population, across both treatment groups. Anchor-based analysis will only include patients who survived to week 16 with complete data, ie no patients with missing change value at week 16 due to any reason, for variables included in that analysis.

Anchor-based approaches

Anchor-based approaches estimate a threshold by ‘anchoring’ the results on a separate variable, often a patient-reported outcome (PRO) with a simpler scale, referred to as an anchor variable or a global anchor. The anchor-based analyses in this study will employ the following PROs as global anchors:

- PGIS in HF symptoms (used for KCCQ-TSS)
- PGIC in HF symptoms (used as a supportive global anchor for KCCQ-TSS)
- PGIC in walking ability (used for 6MWD and as a supportive global anchor for KCCQ-PLS)
- EQ-5D-5L question: "Usual activities" (used for KCCQ-PLS)

For each anchor, meaningful change will be evaluated using observed scores according to a predefined algorithm. The responses to the PGIS and PGIC scales, and the EQ-5D-5L question, at visit 5 (week 16) will be used in the analysis. For PGIS and EQ-5D-5L, the baseline value (randomisation) will also be used in the analysis.

Categorisation of anchors

The PGIS and PGIC scales, and the EQ-5D-5L question will be categorised to provide a clearer difference between patients who have and have not experienced different degrees of change according to the anchors.

The PGIC in walking ability and PGIC in HF symptoms assess how a patient perceives his or her overall change in walking ability or HF symptoms since the start of the study. These scales have levels ranging from 'much worse' to 'much better'. The values are supplied on an ordinal scale and should never be analysed as continuous variables. The ordinal responses to PGIC at visit 5 (week 16) will be assigned the following numeric values to allow categorisation:

- -3 ('much worse')
- -2 ('moderately worse')
- -1 ('a little worse')
- 0 ('about the same')
- +1 ('a little better')
- +2 ('moderately better')
- +3 ('much better')

The numeric values above are arbitrary. They may seem to imply equidistant categories, but this is merely an illusion (we could have assigned logarithmic values and the categorisation process would have worked equally well). There is nothing continuous about the ordinal verbal response options in a PGIC scale. The numeric values are simply a representation, to aid in explaining the collapsing of certain categories. Having noted that, the following categories will be defined, using the numeric values assigned above, for PGIC at visit 5 (week 16):

- 'moderate or large deterioration' (-2 to -3)
- 'large deterioration' (-3)
- 'moderate deterioration' (-2)
- 'small deterioration' (-1)
- 'stable' (0)
- 'small improvement' (+1)
- 'moderate improvement' (+2)
- 'large improvement' (+3)
- 'moderate or large improvement' (+2 to +3)

The PGIS in HF symptoms instrument assesses how a patient perceives his or her overall current severity of HF symptoms. Patients will choose from 6 response options ranging from 'no symptoms', 'very mild', 'mild', 'moderate', to 'severe', and 'very severe'. The values are

supplied on an ordinal scale and should never be analysed as continuous variables. The ordinal responses to PGIS at baseline and visit 5 (week 16) will be assigned the following numeric values to allow categorisation:

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

Again, similar to what was done for the PGIC, the PGIS contains ordinal verbal response options. Numeric values are somewhat arbitrarily assigned to illustrate the "number of steps" a patient has moved in a shift table. Having noted that, the following numerical change values are defined, in Table 5, based on the numeric values assigned to each response option in PGIS. Collapsed or “transformed” categories corresponding to ‘moderate or large deterioration’ (+3 to +5) and ‘moderate or large improvement’ (-3 to -5) will also be defined in accordance with the categories for PGIC.

Table 5 Definition of transformed and raw numeric change from baseline values for PGIS in HF symptoms

	PGIS in HF symptoms at baseline					
PGIS in HF symptoms at week 16	**No symptoms (1)**	**Very mild (2)**	**Mild (3)**	**Moderate (4)**	**Severe (5)**	**Very severe (6)**
No symptoms (1)	Stable (0)	SIm (-1)	MIm (-2)	LIm (-3)	LIm (-4)	LIm (-5)
Very mild (2)	SDt (+1)	Stable (0)	SIm (-1)	MIm (-2)	LIm (-3)	LIm (-4)
Mild (3)	MDt (+2)	SDt (+1)	Stable (0)	SIm (-1)	MIm (-2)	LIm (-3)
Moderate (4)	LDt (+3)	MDt (+2)	SDt (+1)	Stable (0)	SIm (-1)	MIm (-2)
Severe (5)	LDt (+4)	LDt (+3)	MDt (+2)	SDt (+1)	Stable (0)	SIm (-1)
Very severe (6)	LDt (+5)	LDt (+4)	LDt (+3)	MDt (+2)	SDt (+1)	Stable (0)

Transformed change from baseline scores are given as LDt, MDt, etc. and raw change from baseline scores are given in parentheses as (+5), (+4), (+3), (+2), (+1), (0), (-1), (-2), (-3), (-4) or (-5).
LDt Large deterioration. MDt Moderate deterioration. SDt Small deterioration. SIm Small improvement. MIm Moderate improvement. LIm Large improvement. HF Heart failure. PGIS Patient global impression of severity.

EQ-5D-5L question: "Usual activities" asks how the patient is limited in their usual activities (eg, work, study, housework, family or leisure activities). This is about as close of an anchor as we can muster up for the KCCQ-PLS. Patients will choose from 5 response options ranging

from ‘I have no problems doing my usual activities’, ‘I have slight problems doing my usual activities’, ‘I have moderate problems doing my usual activities’, ‘I have severe problems doing my usual activities’, to ‘I am unable to do my usual activities’. The values are supplied on an ordinal scale and should never be analysed as continuous variables. The ordinal responses to EQ-5D-5L question: "Usual activities" at baseline and visit 5 (week 16) will be assigned the following numeric values to allow categorisation:

- 1 (‘I have no problems doing my usual activities’)
- 2 (‘I have slight problems doing my usual activities’)
- 3 (‘I have moderate problems doing my usual activities)
- 4 (‘I have severe problems doing my usual activities’)
- 5 (‘I am unable to do my usual activities’)

Based on the numeric values assigned to each response option in EQ-5D-5L question: "Usual activities", numerical change values are defined, as in Table 6. Collapsed or “transformed” categories corresponding to ‘moderate or large deterioration’ (+2 to +4) and ‘moderate or large improvement’ (-2 to -4) will also be defined in accordance with the categories for PGIS.

Table 6 Definition of transformed and raw numeric change from baseline values for EQ-5D-5L question: "Usual activities"

	EQ-5D-5L question: "Usual activities" at baseline				
EQ-5D-5L question: "Usual activities" at week 16	**I have no problems doing usual activities (1)**	**I have slight problems doing usual activities (2)**	**I have moderate problems doing usual activities (3)**	**I have severe problems doing usual activities (4)**	**I am unable to do usual activities (5)**
I have no problems doing usual activities (1)	Stable (0)	SIm (-1)	MIm (-2)	LIm (-3)	LIm (-4)
I have slight problems doing usual activities (2)	SDt (+1)	Stable (0)	SIm (-1)	MIm (-2)	LIm (-3)
I have moderate problems doing usual activities (3)	MDt (+2)	SDt (+1)	Stable (0)	SIm (-1)	MIm (-2)

Table 6 Definition of transformed and raw numeric change from baseline values for EQ-5D-5L question: "Usual activities"					
	EQ-5D-5L question: "Usual activities" at baseline				
EQ-5D-5L question: "Usual activities" at week 16	**I have no problems doing usual activities (1)**	**I have slight problems doing usual activities (2)**	**I have moderate problems doing usual activities (3)**	**I have severe problems doing usual activities (4)**	**I am unable to do usual activities (5)**
I have severe problems doing usual activities (4)	LDt (+3)	MDt (+2)	SDt (+1)	Stable (0)	SIm (-1)
I am unable to do usual activities (5)	LDt (+4)	LDt (+3)	MDt (+2)	SDt (+1)	Stable (0)

Transformed change from baseline scores are given as LDt, MDt, etc. and raw change from baseline scores are given in parentheses as (+4), (+3), (+2), (+1), (0), (-1), (-2), (-3) or (-4).
LDt Large deterioration. MDt Moderate deterioration. SDt Small deterioration. SIm Small improvement. MIm Moderate improvement. LIm Large improvement. EQ-5D-5L EuroQol five-dimensional five-level questionnaire.

Utilisation of anchors

The main anchor used to establish the threshold for CMWPC from baseline at week 16 in KCCQ-TSS is change from baseline PGIS in HF symptoms at week 16, using the category ‘moderate or large improvement’ (defined by combining the categories ‘moderate improvement’ and ‘large improvement’). The analyses of other categories for PGIS in HF symptoms and the analysis of PGIC in HF symptoms will be regarded as supportive.

The main anchor used to establish the threshold for CMWPC from baseline at week 16 in KCCQ-PLS is change from baseline in EQ-5D-5L question: "Usual activities" at week 16, using the category ‘moderate or large improvement’ (defined by combining the categories ‘moderate improvement’ and ‘large improvement’). The analyses of other categories for EQ-5D-5L question: "Usual activities" and the analysis of PGIC in walking ability will be regarded as supportive.

The main anchor used to establish the threshold for CMWPC from baseline at week 16 in 6MWD is PGIC in walking ability at week 16, using the category ‘moderate or large improvement’. The analyses of other categories of PGIC in walking ability will be regarded as supportive.

Anchor-based analysis

The change from baseline at week 16 in KCCQ-TSS, KCCQ-PLS, and 6MWD will be used repeatedly in the anchor-based analyses.

Descriptive statistics (mean, SD, median, Q1, Q3, minimum and maximum), empirical cumulative distribution function curves and probability density function curves will be presented for each combination of anchor, category and endpoint. The empirical cumulative distribution function curves display a continuous plot of the change from baseline on the horizontal axis, and the cumulative percent of patients experiencing changes from baseline up to that level on the vertical axis. The probability distribution function curves will display kernel density curves for all the categories of change from baseline overlaying the histogram with all categories combined. The percent of patients with change from baseline within each category in the histogram is displayed on the vertical axis.

Establishing the clinically meaningful threshold

To have a completely prespecified decision algorithm for defining the thresholds for CMWPC, threshold value will be defined by the mean change from baseline at week 16 value in the endpoint (KCCQ-TSS, KCCQ-PLS, or 6MWD) corresponding to the category 'moderate or large improvement', for change from baseline at week 16 in PGIS in HF symptoms (for KCCQ-TSS), for change from baseline at week 16 EQ-5D-5L question "Usual activities" (for KCCQ-PLS), or in PGIC in walking ability at week 16 (for 6MWD). Rounding up to the nearest integer will be done when determining threshold values. The reason why the collapsed category is selected is because small sample sizes are expected in the extreme category ('large improvement') and there will be a more balanced number of patients compared to the 'stable' category.

8.4 KCCQ scoring algorithm

The KCCQ is a 23-item, self-administered disease specific instrument, which has been shown to be a valid, reliable and responsive measure for patients with HF (Green et al 2000, FDA 2020, Spertus et al 2005). The KCCQ was developed to measure the patient's perception of their health status independently, which includes HF-related symptoms (frequency, severity and recent change), impact on physical and social function, self-efficacy and knowledge, and how the patient's HF affects their quality of life.

The 23 items and corresponding 15 questions are listed in the Appendix H of the CSP. The 6 items in question 1 constitute the physical limitations score. The question 2 is for the symptom stability domain. The question 3, 5, 7 and 9 constitute the symptom frequency domain, and the questions 4, 6 and 8 constitute the symptom burden domain. The total symptom score incorporates the symptom frequency (4 items) and symptom burden (3 items). The questions 10 and 11 constitute the self-efficacy domain. The questions 12, 13 and 14 constitute quality

of life (QoL) domain. The question 15 (4 items) is for the social limitation domain. Overall summary score is the average of the physical limitation score, total symptom score, quality of life score, and the social limitation score.

Each KCCQ item or question is scored by assigning each response an ordinal value, beginning with 1 for the response that implies the lowest level of functioning. If at least half of the components within the domain are not missing, then the domain score can be calculated by summing the responses of the questions actually answered within the domain. Scale scores are transformed to a 0 to 100 range by subtracting the lowest possible scale score, dividing by the range of the scale and multiplying by 100. If the domain has more than one component, the domain score will be the mean value of the transformed score over the actually answered components. Higher scores represent a better outcome. The scoring algorithm of each domain and summary score is described in detail below.

Physical Limitation

Code responses to each of Questions 1a-f as follows:

Extremely limited = 1
Quite a bit limited = 2
Moderately limited = 3
Slightly limited = 4
Not at all limited = 5
Limited for other reasons or did not do = <missing value>

If at least three of Questions 1a-f are not missing, then compute
Physical Limitation Score = 100*[(mean of Questions 1a-f actually answered) – 1]/4

Symptom Stability

Code the response to Question 2 as follows:

Much worse = 1
Slightly worse = 2
Not changed = 3
Slightly better = 4
Much better = 5
I've had no symptoms over the last 2 weeks = 3

If Question 2 is not missing, then compute
Symptom Stability Score = 100*[(Question 2) – 1]/4

Symptom Frequency

Code responses to Questions 3, 5, 7 and 9 as follows:
Question 3

Every morning = 1
3 or more times a week but not every day = 2
1-2 times a week = 3
Less than once a week = 4
Never over the past 2 weeks = 5

Questions 5 and 7

All of the time = 1
Several times a day = 2
At least once a day = 3
3 or more times a week but not every day = 4
1-2 times a week = 5
Less than once a week = 6
Never over the past 2 weeks = 7

Question 9

Every night = 1
3 or more times a week but not every day = 2
1-2 times a week = 3
Less than once a week = 4
Never over the past 2 weeks = 5

If at least two of Questions 3, 5, 7 and 9 are not missing, then compute:

S3 = [(Question 3) – 1]/4
S5 = [(Question 5) – 1]/6
S7 = [(Question 7) – 1]/6
S9 = [(Question 9) – 1]/4

Symptom Frequency Score = 100*(mean of S3, S5, S7 and S9)

Symptom Burden

Code responses to each of Questions 4, 6 and 8 as follows:

Extremely bothersome = 1
Quite a bit bothersome = 2
Moderately bothersome = 3
Slightly bothersome = 4
Not at all bothersome = 5
I've had no swelling/fatigue/shortness of breath = 5

If at least one of Questions 4, 6 and 8 is not missing, then compute

Symptom Burden Score = 100*[(mean of Questions 4, 6 and 8 actually answered) – 1]/4

Total Symptom Score

Total Symptom Score = mean of Symptom Frequency Score and Symptom Burden Score

Self-Efficacy

Code responses to Questions 10 and 11 as follows:
Question 10
- Not at all sure = 1
- Not very sure = 2
- Somewhat sure = 3
- Mostly sure = 4
- Completely sure = 5

Question 11
- Do not understand at all = 1
- Do not understand very well = 2
- Somewhat understand = 3
- Mostly understand = 4
- Completely understand = 5

If at least one of Questions 10 and 11 is not missing, then compute
Self-Efficacy Score = 100*[(mean of Questions 10 and 11 actually answered) – 1]/4

Quality of Life

Code responses to Questions 12, 13 and 14 as follows:
Question 12
- It has extremely limited my enjoyment of life = 1
- It has limited my enjoyment of life quite a bit = 2
- It has moderately limited my enjoyment of life = 3
- It has slightly limited my enjoyment of life = 4
- It has not limited my enjoyment of life at all = 5

Question 13
- Not at all satisfied = 1
- Mostly dissatisfied = 2
- Somewhat satisfied = 3
- Mostly satisfied = 4
- Completely satisfied = 5

Question 14
- I felt that way all of the time = 1
- I felt that way most of the time = 2
- I occasionally felt that way = 3
- I rarely felt that way = 4
- I never felt that way = 5

If at least one of Questions 12, 13 and 14 is not missing, then compute
Quality of Life Score = 100*[(mean of Questions 12, 13 and 14 actually answered) – 1]/4

Social Limitation

Code responses to each of Questions 15a-d as follows:
- Severely limited = 1

Limited quite a bit = 2
Moderately limited = 3
Slightly limited = 4
Did not limit at all = 5
Does not apply or did not do for other reasons = <missing value>
If at least two of Questions 15a-d are not missing, then compute
Social Limitation Score = 100*[(mean of Questions 15a-d actually answered) – 1]/4

Overall Summary Score

Overall Summary Score = mean of the following available summary scores:
Physical Limitation Score
Total Symptom Score
Quality of Life Score
Social Limitation Score

Clinical Summary Score

Clinical Summary Score = mean of Physical Limitation Score and Total Symptom Score

For PLS and social limitation score, the response to the items/questions could be "Limited for other reasons or did not do" and "Does not apply or did not do for other reasons", respectively. The response represents the scenario that the question doesn't apply and the score is not calculable (NC). For example, if a patient stays at home, the item "Hurrying or jogging (as if to catch a bus)" is not answerable, and the corresponding scale score will be not calculable. If at a time point, at least 4 items in PLS selected "Limited for other reasons or did not do", or at least 3 items in social limitation score selected "Does not apply or did not do for other reasons", the corresponding PLS or social limitation score at the time point will be considered as NC. If PLS or social limitation score is not calculable at baseline or visit 5, the patient will be excluded from the corresponding analysis. NC at visit 3 is rare and only plays a role in the sequential imputations, therefore, it will be treated as missing.

SIGNATURE PAGE

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

[redacted]		
Document Title:	Statistical Analysis Plan Edition 4	
[redacted]	[redacted]	
[redacted]	[redacted]	
Server Date **(dd-MMM-yyyy HH:mm 'UTC'Z)**	**Signed by**	**Meaning of Signature**
25-Sep-2020 04:15 UTC	[redacted]	Author Approval
22-Sep-2020 13:25 UTC	[redacted]	Content Approval
22-Sep-2020 13:01 UTC	[redacted] FMD K&L	Author Approval

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