14 April 2020
EMA/CHMP/SAWP/120610/2020
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion of clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses

Draft agreed by Scientific Advice Working Party	11-14 February 2019
Adopted by CHMP for release for consultation	23 -26 April 2019[1]
Start of public consultation	19 June 2019[2]
End of consultation (deadline for comments)	09 October 2019[3]
Adopted by CHMP	27 February 2020

Keywords	Recurrent Events, Estimand, Chronic Heart Failure, Mortality

[1] Last day of relevant Committee meeting.
[2] Date of publication on the EMA public website.
[3] Last day of the month concerned.

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000 An agency of the European Union

1. Background Information Compiled by CHMP

The objective of the submission was to seek a qualification opinion on recurrent event endpoints for clinical trials where recurrent events are clinically meaningful and where treatments are expected to impact the first as well as subsequent events. The applicant claimed that clinically interpretable treatment effect measures (estimands) based on recurrent event endpoints can be defined along with statistical analyses that are more efficient than those targeting treatment effect measures based on the first event only.

Recurrent events refer to the repeated occurrence of the same type of event over time for the same patient. They should be related to disease burden and may indicate disease progression in some instances. Evaluating a treatment effect on recurrent events may in some instances better reflect the cumulative effect of treatment over time than measuring first events (Rauch, G., et al. "Time-to-first-event versus recurrent-event analysis: points to consider for selecting a meaningful analysis strategy in clinical trials with composite endpoints." Clinical research in cardiology 2018; 107: 437-443) and recurrent and first event analyses can complement each other. Recurrent event endpoints are well established in indications where the rate of terminal events (e.g. death) is very low and reduction in mortality is not a primary goal of treatment. Examples include relapses in multiple sclerosis (CHMP, 2015), exacerbations in pulmonary diseases (e.g. chronic obstructive pulmonary disease (CHMP, 2012a) and asthma (CHMP, 2010a)), headache attacks in migraine (CHMP, 2007, 2016a), hypoglycemia episodes in diabetes mellitus (CHMP, 2012b), and seizures in epileptic disorders (CHMP, 2010b, 2016b). In these chronic diseases, time-to-first-event endpoints that focus on the treatment effect on the first event are clinically less meaningful and hence rarely used.

Regulatory interest is high and experience with recurrent event endpoints is more limited in indications where the rate of terminal events is non-negligible or even high and the clinical meaningfulness of the recurrent event is an issue of discussion if the impact of a therapeutic intervention on the terminal event is of key importance. Chronic heart failure treatment with mortality as the terminal event and recurrent events of worsening heart failure (recurrent HFE) is an indication to exemplify the need for a thorough discussion, both, from a clinical, as well as from a methodological perspective. The document focusses on chronic heart failure, but similar considerations may apply to other diseases where clinically relevant recurrent events and terminal events determine disease progression and disease burden. Aspects like time course, frequency, severity, relevance of recurrent events for prognosis and treatment decisions and association between recurrent and terminal events are disease specific as may be the impact of a therapeutic intervention. Such differences must be taken into account when transferring the results and conclusions to other diseases.

Clinical background: recurrent event analyses in chronic heart failure

The applicant emphasized the example of chronic heart failure (CHF). In the European regulatory framework the primary analysis in pivotal trials in this disease usually is based on a time-to-first-event endpoint, i.e. death alone or as a component of a composite endpoint in combination with endpoint(s) related to worsening of heart failure as e.g. time to first heart failure event (HFE) which may include hospitalisation for worsening of heart failure (HFH) but also well-defined outpatient emergency visits for worsening of heart failure. (guideline on clinical investigation of medicinal products for the treatment of chronic heart failure (CPMP/EWP/235/95, Rev.2, 20, July 2017)). Assessment of mortality in confirmatory trials should include both all-cause mortality and cardiovascular mortality. The guideline summarizes on the issue of recurrent HFE as follows: ”reoccurring hospitalisations for heart failure (HFH) are relatively common in patients with CHF and despite their significance they are rarely used as an endpoint in clinical trials compared to time to first HF hospitalisation”.

It is further stated that "the main therapeutic goals in the treatment of CHF are to reduce cardiovascular mortality and to prevent deterioration of the clinical status and hospitalisations; these goals should represent the primary aim of new agents developed for the treatment of CHF […] endpoints accounting for recurrent HFH events may under certain conditions better characterise the prognosis of patients with CHF. Recurrent events are also important as they represent a large burden to patients. The inclusion of recurrent events as co-primary endpoint may be considered, but this setting needs further justification, adjudication of the events and a clear methodological strategy". In this aspect the ability to appropriately estimate the effect of treatment on recurrent hospitalization is of importance.

The controversy on this issue relates to clinical meaningfulness of an assessment of the recurrent event, in case of no, or a negative correlation between mortality and the recurrent event, methodological issues, and the loss of information on mortality if studies become smaller when designed based on recurrent events only. These three issues are discussed here.

Mortality

Reduction of mortality is one of the main therapeutic goals in CHF. Current treatment regimens in clinical guidelines are based on robust knowledge on the effect of interventions on all-cause mortality, cardiovascular mortality and HFEs (e.g. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure, european Heart Journal doi:10.1093/eurheartj/ehw128). Robust information on all-cause and cardiovascular mortality is crucial for allocation of a new therapy in the context of other licensed medicinal products.

Although mortality rates in CHF have decreased over the last decades, all-cause mortality remains high. In the european ESC-HF pilot study, covering a period between October 2009 to May 2010, 12-month all-cause mortality rates for hospitalized (acute heart failure) and stable/ambulatory HF patients were 17% and 7%, respectively, with 12-month hospitalization rates of 44% and 32%, respectively. Similar numbers were observed in the PARADIGM HF trial (Murray et al., N Engl J Med 2014; 371: 993 and EPAR EMEA/H/C/004062/0000, run in 2009 through 2012, stopped 2014) that may serve as an example for mortality rates in present clinical heart failure studies. 17.0% and 19.8% of the patients died in the LCZ696 and the Enalapril group, respectively, after a median follow-up of 27 months. Death from cardiovascular cause was observed in 13.3 and 16.5%, first HFH in 12.8 an 15.6% in the respective arms. The statistically significant result was to a large degree based on efficacy in earlier stages of the disease (NYHA I – II). It is an example for a contemporary study (8442 patients) able to provide the data needed for assessment of effects on mortality and hospitalization for patients as included in this study.

It should be emphasized that in heart failure studies acquiring robust data on mortality is not only essential for the overall group of patients included. The SHIFT study (ivabradine, EPAR EMA/194513/2012) is an example that shows that meaningful data are also required for subgroups. In this pivotal trial, the primary endpoint (composite for cardiovascular death or first event HFH) showed a statistically significant benefit of ivabradine over placebo for the whole study population with consistent trends for mortality endpoints. However, predefined subgroup analyses by baseline heart rate (< 77 bpm, vs. ≥ 77 bpm) showed numerically increased rates of cardiovascular mortality and all-cause mortality in patients with lower baseline heart rate. These subgroup analyses contributed to the decision to restrict the indication to patients with a baseline HR ≥ 75 bpm. Reduction in variability in estimates, mainly discussed from the background of an opportunity to reduce the overall sample-size of a trial may thus limit the opportunity of the required risk-benefit assessment in an indication that suffers from high unexplained variability that should be acknowledged (Guideline on the investigation of subgroups in confirmatory clinical trials (EMA/CHMP/539146/2013)).

In general, a medicinal product can be approved based on a beneficial effect on HFEs, even if studies fail to show a mortality benefit. As a prerequisite the data must provide sufficient reassurance that mortality is not increased to a relevant degree in the overall population and in subgroups. The key example is digoxin. In a placebo controlled study including 6800 patients digoxin had no effect on all-cause mortality (RR 0.99; 95 % CI 0.91 to 1.07, The Digitalis Investigation Group (DIG), N Engl J Med 1997; 336: 525-533), but significantly improved first HFH rate (26.8 % vs. 34.7 %; RR 0.72; 95 % CI 0.66 to 0.79; P<0.001). The trial was large enough to exclude an increase in all-cause mortality by more than 7% which may be sufficient for a well-established drug. However, careful analysis of the mortality is crucial in such a case since an overall neutral effect on mortality despite a HFEs benefit may well be the result of divergent effects on mortality in subgroups. This has been discussed for the DIG trial. In a post-hoc subgroup analysis in male patients all-cause mortality was decreased at lower digoxin levels, neutral at intermediate digoxin levels and increased in patients with higher digoxin plasma levels. Similarly, in the Val-HEFT study, comparing valsartan with placebo, a beneficial effect was observed on first event HFHs (RR 0.87; 97.5 % CI, 0.77 to 0.97; p=0.009) whereas the effect on all-cause mortality was neutral (deaths during the entire trial: RR 1.02 (0.88 – 1.18)). In Val-HEFT, the neutral effect on mortality was the net result of a significantly increased mortality in patients receiving in addition ACE inhibitors and beta blockers, and a significantly decreased mortality in the other patients. However, it should be noted that this finding was not reproduced in CHARM-Added and remains controversial. This again emphasizes the need for well-populated subgroups allowing for a proper risk-benefit assessment.

Exclusion of an increase in mortality is of particular importance in CHF, considering examples of agents with a detrimental effect. E.g. in a study with 1088 patients with severe CHF Milrinone increased all-cause mortality and cardiovascular mortality by 28% and 34%, respectively. The number of patients with worsening heart failure, functional deterioration or requiring additional therapy was not different between the groups, hospitalization rate was only slightly higher in the milrinone group (44 percent vs. 39 percent; p = 0.041; Packer M et al., N Engl J Med 1991; 325:1468). Xamoterol improved breathlessness in a study with 516 patients with NYHA class III and IV heart failure but increased mortality (ITT: 32 (9.1%) vs. 6 (3.7%), p = 0.02, THE XAMOTEROL IN SEVERE HEART FAILURE STUDY GROUP, Lancet. 1990; 336:1). Exclusion of an increase in mortality is a key aspect of the assessment of chronic treatment of CHF.

Recurrent HFEs

Recurrent HFEs represent a considerable disease burden to patients. After diagnosis of heart failure 83% of patients were hospitalized at least once, 67% ≥2, 54% ≥3 and 43% ≥4 times in a US based study (period 1987–2006, Dunley SM et al., JACC 2009; 54: 1695). Most of these hospitalizations were due to non-CV reasons (61.9%), HFH made up for 16.5%, and hospitalizations for other CV reasons for 21.6%. Male sex and co-morbidities (diabetes mellitus, chronic obstructive pulmonary disease, anemia, and creatinine clearance <30 mL/min) were independent predictors of all-cause hospitalization.

After a hospitalisation for heart failure, the rate of recurrent HFHs is much higher. After discharge from a HF related hospital stay (Canada, 1999 – 2001, Chun S et al., Circ Heart Fail 2012; 5; 414) 61.3% of the patients were re-hospitalized for heart failure and 66.5% for a cardiovascular event within the first year of discharge. Differences in expected HFH rates related to whether patients have been hospitalized for HF recently or not, must be taken into account.

The study showed some peculiarities when assessing recurrent HFEs. Hospitalization rates were not uniformly distributed over time, they clustered at early post-discharge and pre-fatal time. This is expected, single or multiple pre-fatal hospitalisation events reflect deterioration of disease. However,

whereas a first event analysis of a composite of HFEs and death usually is based on a categorical understanding of how to measure deterioration of disease and does not weigh severity of events, pre-fatal clustering of HFEs introduces a weighting factor to the fatal clinical event. It can be challenged that more weight should be given to the death of a patient because he has visited a doctor for worsening heart failure shortly before deceasing. Counting pre-fatal hospitalisations increases the number of events but the clinical meaningfulness is unclear. Furthermore, HFH rate depended on the underlying disease. In ischemic heart failure, where the hospitalization rate was higher, a clear differentiation between heart failure related and ischemia related hospitalization may not be feasible in every case. Recurrent event analyses are currently not accepted in the regulatory context in cardiovascular trials aiming at the prevention of MACE related to ischemic diseases.

Whereas it has been considered that recurrent HFEs may better characterize the prognosis of patients under certain conditions (CPMP/EWP/235/95, Rev.2, 20, July 2017) an effect on HFEs may not be predictive for an effect on mortality for every new therapeutic agent. HFH or signs and symptoms of heart failure did not exactly mirror the effect of a treatment on mortality in the above mentioned two studies with milrinone and xamoterol. The DIG study results estimated improved hospitalisation but no significant effect on mortality, possibly due to antithetic results for mortality in subgroups". Furthermore, models predicting mortality in patients with heart failure were reported to have a higher discriminative ability than those designed to predict hospitalization (Rahimi K et al., JACC heart failure 2014; 2: 440 ff; Outwerkerk W JACC heart failure 2014; 2; 429). Among the possible reasons is that hospitalization is more dependent on health care supply indicating that HFEs and mortality are not interchangeable parameters for the assessment of the outcome of a therapeutic intervention.

In summary, the main therapeutic goals in the treatment of CHF are to reduce cardiovascular mortality and to prevent deterioration of the clinical status and HFEs; these goals should represent the primary aim of new agents developed for the treatment of CHF. Recurrent events may represent a large burden to patients and endpoints accounting for recurrent HFEs may under certain conditions better characterise the prognosis of patients with CHF (c.f. CPMP/EWP/235/95, Rev.2, 20, July 2017). A particular challenge when clinically interpreting recurrent HFEs is clustering of events. Since disease specific differences and factors like health care supply have an impact on HFEs, it must be assessed whether they affect recurrent event analyses more than first event analyses. If studies become smaller when sample sizes are calculated based on recurrent HFEs data available for mortality assessment will become less robust. The impact of a new therapeutic agent on mortality, either as a measure of efficacy or at least in order to provide robust reassurance that there is no detrimental effect, is key information expected from a pivotal trial in chronic heart failure. Such data is needed not only for the overall population but also for relevant subgroups. Examples exist, where it was possible to achieve this information with a reasonably sized clinical program based on the requirements as outlined in CPMP/EWP/235/95, Rev.2. Considering requirements to rule out an excess of mortality, the number of patients needed in a study using recurrent HFEs as a component of a primary endpoint may in the end not be lower than in a study designed according to the current guideline.

Although not within the scope of this methodological qualification opinion, the application of recurrent HFE in areas, where robust data on mortality are less important (e.g. phase 2 trials, extrapolation exercises), or in rare diseases, where information on mortality primarily depends on the number of patients available and not on the study design, is endorsed by CHMP. The CHAMPION trial (Abraham WT et al., The Lancet 2011; 377: 658) may serve as an example of a small scale study for a medical device in patients where the impact of an implantable haemodynamic monitoring system of recurrent HFH was explored over a 6 month period in patients with NYHA III. These programs may substantially benefit from the development of recurrent HFE analyses in such areas.

2. Methodological issues

2.1 Calculation of HFE rate for a treatment-group:

Two different possibilities were discussed to calculate the heart failure rate per treatment group and a simplified example is presented to illustrate and discuss two different effect measures: The exposure-weighted and the patient-weighted event rate.

Patient	HFE	Follow-up (years)	HFE per year
Ann	0	3.0	0
Bill	1	3.0	0.333
Caren	3	1.5	2
Dave	0	3.0	0
Total	4	10.5	2.333
Average per patient	1	10.5 / 4 = 2.625	2.333 / 4=0.583
Average HFE per year of exposure			4/10.5=0,38

Exposure-weighted rate

The exposure (or exposure and follow-up-time) weighted annualised rate for a treatment group (the number of events per year of observation in that group) can be expressed in many ways, all of which lead to the same answer.

- It can be thought of as the total number of events observed in that group divided by the total follow-up time. In the example this gives 4/10.5, i.e. 0.38 events per year.
- It could also be thought of as the average number of HFE events per patient, divided by the average follow-up – so in the example 1/2.625, or 0.38 events per year.
- And it could also be seen as the weighted average of the event rates for each patient, with the weights being the proportion of the follow-up time contributed by that patient i.e. patients who were followed-up for longer are given more weight in the analysis. In the example this give (0x3/10.5) + (0.333x3/10.5) + (2x1.5/10.5) + (0x3/10.5) = (1/10.5) + (3/10.5) = 0.38 events per year.

This estimator is maximum likelihood if the recurrent events follow a Poisson-distribution, however, clustering of events is observed in practice which contradicts the model and the impact of this on the validity of conclusions when using the model is not clear.

Patient-weighted rate

The patient weighted annualised rate is the average of the rates observed for each patient, with each patient being given equal weight, regardless of exposure. In the example this gives (0+0.333+2+0)/4 = 0.583 events per year.

The statistical properties of this estimate are not obvious beyond assuming asymptotic normality which may hold true once the number of investigated patients is large as e.g. in chronic heart failure studies.

Comparison

The two approaches will lead to identical answers if the duration of observation is the same for all patients.

The two approaches will on average give the same answer if follow-up duration is independent of HFE e.g. the number of HFE events is no indicator of the likely duration of follow-up or survival. However, in this scenario the patient-weighted rate may be more variable, because of some very high individual patient rate-estimates from patients with one or more events, but short follow-up time.

The two approaches will give systematically different answers when the duration of follow-up is related to HFE events. An example of this would be if patients with high HFE rates are also more likely to die and therefore generally have shorter follow-up. This would lead to the patient-weighted rate being higher than the exposure-weighted rate, as the patient weighted approach would give all patients equal weight, while the exposure rated approach would generally give less weight to patients with higher HFE rates.

When interpreting these different rates, the exposure-weighted rate seems to be of some relevance to the population as a whole – e.g. if a hospital was estimating the admission rates they should expect for HF, the exposure-based approach might provide useful information in terms of events per year that they might see. However, for a patient considering what annual rate they as an individual might expect while they are alive, the patient-weighted rate would be the most informative, as every individual patient studied would have an equal chance of representing them – there is not more chance that they would be like one of the patients with long follow-up.

2.2. Calculation of the treatment effect on HFE rate

In this discussion the treatment comparison is made by taking the ratio of the events per year observed in each treatment group, the rate ratio (RR). This could be done using the exposure-weighted rate or the patient-weighted rate.

As noted above if follow-up time is the same for all patients, the estimate in each group will be the same regardless of the use of exposure or patient-weighted methodology, therefore the ratio, and hence the estimate of the treatment effect would also be the same. Similarly, the two approaches will on average give the same answer if follow-up time is independent of HFE.

However, there will be systematic differences between the two in other situations:

If a treatment, on average, delivers an x% HFE rate reduction for every patient, then the expected estimate from the patient weighted approach will be an x% reduction, regardless of follow-up time and the relationship between follow-up time and treatment and HFE.

The average estimate given by the exposure-rated analysis will vary depending on the relationships between HFE rate, treatment and follow-up duration. For example, if high HFE rates are associated with early death, and a treatment has a positive effect on HFEs, then the active treatment will manage to keep the higher HFE patients on treatment for longer than the control, making the beneficial effect seem smaller in the exposure-weighted analysis. This would be offset if the treatment had a detrimental effect on death outside the relationship between death and HFEs, meaning the effect could then seem more favourable for the exposure-related analysis.

Example:

In this example the HFE rate is halved on treatment compared to control on a per-patient basis, but because of the shorter follow-up for the patient with the highest HFE rate on control (an early death) the treatment effect estimate has a smaller magnitude than 0.5 in the exposure-weighted analysis.

Treatment

Patient	HFE	Follow-up (years)	HFE per year
Ann	0	3.0	0
Bill	1	3.0	0.33
Caren	3	3.0	1
Dave	0	3.0	0
Total	4	12	1.33
Average per patient	1	3.0	

Control

Patient	HFE	Follow-up (years)	HFE per year
Arthur	0	3.0	0
Brenda	2	3.0	0.67
Colin	3	1.5	2
Doreen	0	3.0	0
Total	5	10.5	2.67
Average per patient	1.25	2.625	

Annualised HFErates:

Exposure weighted: treatment 0.333 per year, control 0.476 per year; ratio 0.7

Patient weighted: treatment 0.333 per year, control 0.667 per year; ratio 0.5

Particularly, if the frequency of HFE was considered to be of value independently from the outcome on mortality (as is sometimes claimed) in the patient weighted approach two treatments would be considered equally effective, if all patients in treatment group A survive one year with three HFE each and those in treatment group B survive for two years with six HFE each. Interestingly the conclusion is identical if the exposure weighted approach is used. Obviously, the conclusion that both treatments may lead to the same burden for the health care system, is incorrect, as treatment B incurs higher costs for the system, but also is associated with benefit for the patient. It may also be difficult to justify to patients that treatment A should be used.

Intercurrent events, particularly if terminal / absorbing or impacting differentially (i.e. to a different degree on treated and control patients) on duration of observation by other mechanisms, cause

obvious problems with the independent interpretation of treatment effect estimates for differences in recurrent events.

2.3. Applicant's proposal

The applicant's proposals are based on the exposure-weighted rate approach. The reason for this preference is related to the drawbacks of the patient-weighted approach of the high influence of patients who die early leading to high variability and a skewed distribution of results. In addition, they state that none of the established estimators and statistical tests for recurrent events data in the literature target the patient-weighted estimate. However, high variability per se indicates lower confidence for decision making and may be an argument on its own that simply more information is needed to provide robust conclusions (i.e. regarding relevant subgroups of different risks and secondary endpoints). In addition, as mentioned above, statistical properties apply under certain assumptions regarding a model that may not apply in the current situation.

Four different methods for recurrent event analysis were looked at and compared with Cox regression analysis as the current standard – which looks at time to first event. NB refers to negative binomial regression, which targets an estimand based on the number of recurrent events. When there is complete follow-up NB provides an estimate of the RR which is the ratio of the average event numbers in the two groups. The other two methods, Wei (WLW) and Prentice (PWP) do not have such a clear interpretation. None of these directly offers an opportunity to model terminal intercurrent events.

Two main settings were considered in simulation studies, those without a terminal event (or more realistically where terminal events are rare) and those with such an event (usually death). Terminal events are events, the occurrence of which means the recurrent event can no longer be observed and obviously represent an important aspect of drug treatment and assessment of outcome on its own.

2.3.1. Scenarios without a terminal event (or where terminal event rates are low)

For the first scenario both non-informative treatment discontinuation and informative treatment discontinuation were considered. The simulated trial had a fixed 2-year follow-up for every patient. Informative discontinuation meant that patients were more likely to discontinue prematurely if they had high rates of recurrent events, with non-informative discontinuation there is no link. For both it was assumed that after discontinuation from active treatment patients were followed up and event rates went back to the control rate. It is noted that informative discontinuation does not necessarily require correlation with a higher frequency in the event of interest.

Two estimands were considered – one based on a hypothetical strategy to address discontinuation of treatment (the RR if patients remained on treatment) and the other based on the treatment policy strategy (the RR regardless of whether patients remain on treatment). Simulations were used to compare methods under different conditions. As these are simulations the model parameters were known so the true values of the estimands could also be calculated. This qualification opinion doesn't aim to address which estimand is more acceptable for regulatory decision making. However, the general concern regarding the hypothetical strategy applied to treatment discontinuation should be noted, where it is not understood why a patient who discontinued in the trial, for example because of a severe toxicity, would have continued with the medication outside the trial. In earlier phase trials where the purpose is not to gain a regulatory approval the strategy is easier to understand.

Regarding type I error, table 7A shows that in simulations as provided there is possibly a small loss of control with small sample sizes (n=50) for recurrent event methods: values generally exceed 0.025 for all methods, while Cox regression looks fine, but with larger sample sizes there are no apparent issues in the presented simulations.

Table 7A: mean treatment effects estimates (geometric mean) and type I error rates (1-sided tests, nominal significance level α=0.025) under four scenarios, with treatment effect size RR=1, baseline recurrent event rate λ_0=0.5, and dispersion parameter θ=0.25.

		$n = 50$		$n = 75$		$n = 125$	
	Method	RR	Type I error	RR	Type I error	RR	Type I error
Scenario 1: Non-informative (Hypothetical)	Cox	0.998	0.025	1	0.024	1.001	0.024
	NB	0.998	0.026	1.002	0.024	1.002	0.024
	LWYY	0.998	0.028	1.002	0.024	1.002	0.024
	WLW	0.997	0.029	1.001	0.026	1	0.025
	PWP	0.998	0.028	1.002	0.024	1.002	0.025
Scenario 2: Informative (Hypothetical)	Cox	0.994	0.025	0.999	0.024	1.001	0.022
	NB	0.995	0.028	1.002	0.025	1	0.024
	LWYY	0.995	0.029	1.003	0.026	1.001	0.024
	WLW	0.993	0.028	1.002	0.025	1.003	0.024
	PWP	0.996	0.03	1.002	0.025	1.001	0.024
Scenario 3: Non-iformative (Treatment-policy)	Cox	0.998	0.024	0.999	0.024	1.001	0.023
	NB	0.998	0.028	1	0.025	1.002	0.024
	LWYY	0.998	0.029	1	0.025	1.002	0.024
	WLW	0.997	0.028	1	0.028	1.003	0.024
	PWP	0.998	0.028	1	0.025	1.001	0.025
Scenario 4: Informative (Treatment-policy)	Cox	0.995	0.026	0.999	0.026	1.001	0.023
	NB	0.996	0.029	1.001	0.025	1.002	0.026
	LWYY	0.996	0.03	1.001	0.026	1.002	0.026
	WLW	0.994	0.029	1	0.026	1	0.026
	PWP	0.997	0.029	1.001	0.025	1.001	0.025

Tables 5 and 6 show the true value of the exposure-weighted estimand under each of the simulated scenarios, and how the estimates from each of the methods compare to this, shown by the ratio of estimate to estimand in table 5. Values of estimate/estimand greater than 1.00 in table 5 represent an on average conservative estimate i.e. estimates less favourable (or more harmful) than the true value. The true treatment effect while patients remain on treatment is 0.65 in these examples.

Table 5: Settings without terminal event (estimand vs estimate): numerical values of hypothetical estimand and treatment policy estimand under four scenarios. The ratio of the target of estimation (estimate) for each of the five analysis methods over the corresponding estimand value (estimand) is also shown. 'estimand' values are calculated analytically, 'estimate' values are calculated based on a simulated data set with 100'000 patients with RR = 0.65, θ = 0.25, and λ0 = 0.5, 1.5. Estimate/Estimand values larger (smaller) than 1 correspond to overestimation (underestimation).

	Estimand value	Estimate/Estimand		
		Method	$\lambda_0 = 0.5$	$\lambda_0 = 1.5$
Scenario 1: Non-informative (Hypothetical)	0.65	Cox	1.023	1.055
		NB	0.995	0.994
		LWYY	0.995	0.994
		WLW	0.886	0.895
		PWP	1.032	1.075
Scenario 2: Informative (Hypothetical)	0.65	Cox	1.043	1.071
		NB	1.017	1.009
		LWYY	1.020	1.014
		WLW	0.922	0.912
		PWP	1.051	1.082
Scenario 3: Non-informative (Treatment policy)	0.685	Cox	1.013	1.029
		NB	0.996	0.993
		LWYY	0.999	1.000
		WLW	0.892	0.893
		PWP	1.032	1.067
Scenario 4: Informative (Treatment policy)	0.7002	Cox	1.000	1.007
		NB	1.001	0.995
		LWYY	1.005	1.014
		WLW	0.894	0.887
		PWP	1.034	1.055

Table 6: settings without terminal event: mean treatment effect estimates under four scenarios based on 10'000 clinical trial simulations, RR = 0.65, θ = 0.25, λ0 = 0.5, 1.5.

		$\lambda_0 = 0.5$			$\lambda_0 = 1.5$		
	Method	$n = 50$	$n = 150$	$n = 250$	$n = 50$	$n = 150$	$n = 250$
Scenario 1: Non-informative	Cox	0.7	0.68	0.675	0.705	0.694	0.692
(Hypothetical)	NB	0.672	0.656	0.653	0.657	0.652	0.652
Estimand value: 0.65	LWYY	0.671	0.656	0.653	0.657	0.652	0.652
	WLW	0.615	0.591	0.586	0.602	0.591	0.59
	PWP	0.69	0.678	0.676	0.704	0.701	0.702
Scenario 2: Informative	Cox	0.705	0.687	0.681	0.709	0.698	0.696
(Hypothetical)	NB	0.679	0.666	0.661	0.665	0.659	0.658
Estimand value: 0.65	LWYY	0.681	0.668	0.663	0.668	0.663	0.661
	WLW	0.628	0.607	0.599	0.609	0.597	0.594
	PWP	0.697	0.687	0.682	0.709	0.706	0.705
Scenario 3: Non-informative	Cox	0.726	0.708	0.703	0.723	0.713	0.711
(Treatment policy)	NB	0.705	0.691	0.688	0.692	0.687	0.686
Estimand value: 0.685	LWYY	0.706	0.692	0.689	0.695	0.69	0.691
	WLW	0.646	0.624	0.619	0.631	0.62	0.619
	PWP	0.724	0.713	0.711	0.736	0.733	0.734
Scenario 4: Informative	Cox	0.729	0.713	0.709	0.724	0.714	0.712
(Treatment policy)	NB	0.718	0.706	0.702	0.707	0.702	0.701
Estimand value: 0.7002	LWYY	0.721	0.709	0.706	0.717	0.714	0.714
	WLW	0.658	0.638	0.633	0.64	0.63	0.627
	PWP	0.737	0.729	0.726	0.746	0.744	0.743

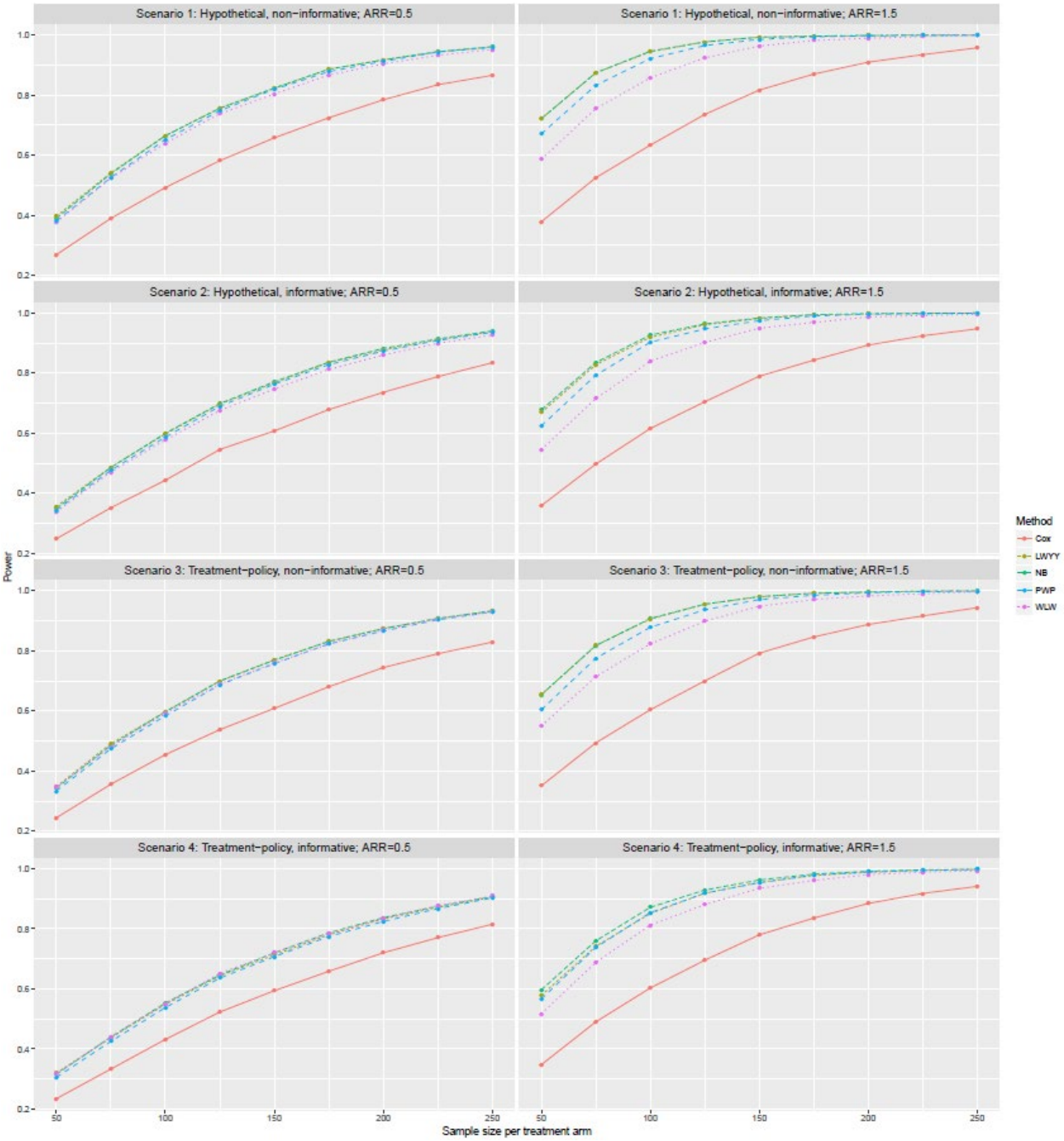
Informative discontinuation means that an effective treatment would keep the patients with a higher event rate on treatment longer allowing them to contribute more events, which explains the conservative estimation in scenario 2. Otherwise there is no suggestion of bias for NB or LWYY. WLW seems to be biased in favour of treatment while PWP is conservative.

Considering the treatment-policy approach, the treatment effect from this approach is less impressive than the 0.65 if patients would remain on treatment, as would be expected given it considers periods where patients are off-treatment. With that in mind an estimate using a treatment policy approach could be used as a conservative estimate of the hypothetical estimand when there are concerns around the assumptions that need to be made for the estimates that actually target the hypothetical estimand.

An interesting feature of the treatment policy estimand is that the true value of the estimand is dependent on the choice of design. The trials simulated here had a 2-year follow-up. If a longer follow-up was specified the true value of the treatment policy estimand would get closer to 1.0 (as the duration of follow-up increases for patients off-treatment) while for the hypothetical estimand it would remain unchanged. When such results are reported it would need to be made clear that the ratios being presented are relevant for the follow-up time specified and usually median observation times per treatment group should be reported, as well. However, this is a general feature of treatment policy estimands and the estimation of parameters of (semi-)parametric survival-functions and is not specific to the recurrent event setting.

Figure 7 shows that there is a substantial increase in power for the recurrent event methods, compared to the first event only Cox model.

Figure 7: setting without terminal event: statistical power at varied sample size under four scenarios based on 10'000 clinical trial simulations, RR = 0.65, θ = 0.25, λ0=0.5,1.5.



In the presented simulations and under the assumptions regarding the distribution of events aside from an issue with type I error control for small sample sizes, which should be investigated further, it can be agreed that methods such as negative binomial regression are more efficient than time to first events approaches in a situation where the rate of terminal events is negligibly low. The provided simulations demonstrate increased power, and the estimates of the RR reflect the true treatment effect, except for being conservative for the effect in scenario 2 where the rate of withdrawal from treatment is positively correlated with the rate of recurrent events. Obviously, type-1-error control is not given with informative censoring.

2.3.2. Scenarios with a terminal event

Terminal events complicate the estimation of the reduction in recurrent events, as after the terminal event occurs the patients can no longer experience the recurrent events.

Two statistics (referred to as estimands) were considered here. Firstly, a ratio based on the patient, or exposure weighted number of recurrent events (in this case hospitalisations) and secondly, a ratio based on the events when counting the terminal even (death) as an additional event.

Table 11: Settings with terminal event: mean treatment effect estimates and type I error rates for estimands 1 and 2 with non-informative treatment discontinuation based on 10'000 clinical trial simulations, RRHHF = 1 and sample size N = 4350.

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.055	0.115
		NB	1.075	0.120
		LWYY	1.124	0.254
		WLW	1.101	0.207
		PWP	1.050	0.142
	0.8	Cox	1.030	0.066
		NB	1.040	0.066
		LWYY	1.062	0.098
		WLW	1.051	0.088
		PWP	1.025	0.071
	1.0	Cox	1.004	0.048
		NB	1.006	0.050
		LWYY	1.006	0.046
		WLW	1.005	0.049
		PWP	1.002	0.050
Estimand 2 (HHF+CVD)	1.0	Cox	1.003	0.046
		NB	1.005	0.046
		LWYY	1.004	0.046
		WLW	1.004	0.050
		PWP	1.001	0.049

From table 11, looking at the rows where HRCV =1.0 we can see that the type I error control of all methods seems good under the global null-hypothesis, where there is no effect on the terminal or the recurrent event, as the type I error values are all approximately 0.05. But type I error control for the test of whether the treatment has an effect on the recurrent event can be lost when there is no effect on the recurrent event (the target of estimand 1) but there is an effect on the terminal event. (In this table that is mainly because of false-positive results in favour of the control treatment. However, if a row for HRCV values > 1.0 had been included similar results would have been seen because of false-positive results in favour of the test treatment.)

When considering the next table, we should recall that the true value of the estimand is based on the exposure-weighted approach. As noted previously, such an approach means that the magnitude of the treatment effect on HFE varies dependent on factors such as the effect of treatment on the terminal event. The results presented by the consortium confirm that assertion.

Table 8: Settings with terminal event (estimand vs estimate): True estimand values under four scenarios, as well as the treatment effects estimates from five approaches. Simulated data for 100'000 patients are generated with RRHHF = 0.7, HRCV = 0.8; 1.0; 1.25.

	Estimand value			Method	Estimates		
HR_{CV}	0.8	1.0	1.25		0.8	1.0	1.25
Scenario 1: Non-informative Estimand 1 (HHF)	0.783	0.722	0.688	Cox	0.841	0.799	0.782
				NB	0.752	0.700	0.684
				LWYY	0.784	0.722	0.687
				WLW	0.789	0.731	0.702
				PWP	0.849	0.811	0.791
Scenario 2: Informative Estimand 1 (HHF)	0.770	0.728	0.686	Cox	0.822	0.789	0.769
				NB	0.741	0.704	0.679
				LWYY	0.771	0.727	0.684
				WLW	0.774	0.731	0.692
				PWP	0.843	0.817	0.787
Scenario 3: Non-informative Estimand 2 (HHF+CVD)	0.809	0.806	0.822	Cox	0.875	0.898	0.935
				NB	0.766	0.814	0.885
				LWYY	0.809	0.806	0.821
				WLW	0.817	0.818	0.839
				PWP	0.878	0.907	0.944
Scenario 4: Informative Estimand 2 (HHF+CVD)	0.800	0.800	0.820	Cox	0.859	0.881	0.929
				NB	0.767	0.797	0.889
				LWYY	0.801	0.800	0.819
				WLW	0.807	0.806	0.831
				PWP	0.879	0.900	0.944

In table 8 the true risk ratio for hospitalization rates as used in the simulation is 0.7 for each individual treated patient but depending on the rates of terminal events the value of the estimand alters, indicating a larger beneficial effect of treatment if the treatment has an adverse effect on the terminal events. Similarly, for treatments which are reducing the rate of terminal events the effect on recurrent events seems less impressive.

In the presented simulations this pattern does not occur so markedly with estimand 2 in the above tables, but estimand 2 is a combined estimate of the effect of CVD and HFF with no clear clinical interpretation (because CVD has the same weight as one HHF).

Ideally an analysis of the data from a trial where there are recurrent and terminal events would deliver estimates of the treatment effect on both aspects; an estimate of effect of the treatment on the recurrent event, and the effect on the terminal event. The simulations show a scenario where the effect of treatment for an individual patient is that on average they would expect an reduction of 0.7 in their event rate while they are alive, yet the estimand being targeted (based on the exposure-rated approach) does not deliver this, and the value varies depending on the treatment effect on the terminal events.

In terms of the estimators being used, LWYY does well in the presented simulations, in that it produces good estimates of the true value of the exposure-weighted treatment effect, but it is questioned whether this in appropriate target for estimation.

Equal weighted (per patient) estimand

A possible alternative approach to address these issues might be to instead target a patient-weighted approach. As discussed above this would be expected to deliver on average a consistent estimate of the treatment effect on recurrent events regardless of the effect on terminal events.

Table A: Terminal event case: approximated estimand values as well as Monte Carlo standard errors (SE) under 30 scenarios. Simulated data for 200.000 patients are generated with ⟦RR⟧_HHF=0.7, ⟦HR⟧_CV=0.67;0.8;1.0;1.25;1.5.*

Endpoint	Follow-up time	HR_{CV}	Exposure-weighted rate based estimand (SE)	Equal-weighted rate based estimand (SE)
HHF	1.25	0.67	0.721(0.012)	0.703(0.013)
		0.80	0.713(0.012)	0.706(0.013)
		1.00	0.680(0.011)	0.699(0.017)
		1.25	0.690(0.011)	0.703(0.014)
		1.50	0.669(0.011)	0.703(0.015)
	3.5	0.67	0.783(0.010)	0.730(0.014)
		0.80	0.718(0.010)	0.679(0.013)
		1.00	0.704(0.009)	0.700(0.013)
		1.25	0.653(0.009)	0.682(0.013)
		1.50	0.625(0.008)	0.708(0.014)
	7	0.67	0.809(0.010)	0.698(0.015)
		0.80	0.776(0.009)	0.716(0.012)
		1.00	0.700(0.009)	0.694(0.013)
		1.25	0.642(0.008)	0.707(0.013)
		1.50	0.586(0.007)	0.708(0.013)
HHF+CVD	1.25	0.67	0.711(0.010)	0.689(0.097)
		0.80	0.742(0.010)	0.948(0.250)
		1.00	0.766(0.010)	1.099(0.167)
		1.25	0.834(0.011)	0.666(0.240)
		1.50	0.866(0.011)	3.240(2.218)
	3.5	0.67	0.764(0.009)	0.239(0.123)
		0.80	0.749(0.008)	0.856(0.103)
		1.00	0.783(0.009)	0.405(0.229)
		1.25	0.797(0.009)	1.653(0.847)
		1.50	0.816(0.009)	1.361(0.282)
	7	0.67	0.791(0.008)	0.697(0.078)
		0.80	0.797(0.008)	0.995(0.322)
		1.00	0.784(0.008)	1.621(0.630)
		1.25	0.786(0.008)	1.106(0.225)
		1.50	0.781(0.008)	1.099(0.137)

The second column of table A* shows that when this is done it does appear that the patient-weighted estimand provides estimates close to 0.7 for HHF for all values of the effect on the terminal event, irrespective of follow-up time. (This table differs from previous tables in that there are no discontinuations other than deaths – so we get a value of 0.7 for the exposure-weighted approach when there is no treatment effect on death).

The exposure-weighted estimand changes with the effect on the terminal event, but also changes with the duration of follow-up, meaning interpretation would also need to take into account changes in study design.

Whereas the exposure-weighted estimand seems to provide an estimate of the total population reduction in recurrent events that might be expected in a particular follow-up time in a certain patient population, the equal patient-weighted approach seems to target the average reduction in event rate for individual patients. While the former might have some relevance in a health economics type scenario when considering the impact on the number of hospitalisations a system might have to cope with and how this could be reduced, the latter seems more relevant when describing the impact of treatment on a particular patient.

However, there are clear limitations with the patient-weighted approach. The applicant notes that none of the investigated analysis methods targets the estimand. They also express concern over the likely increased variability of such an estimate, which would necessitate large sample sizes, and potentially lose the efficiency hoped to be gained by using a recurrent events analysis, and its skewed distribution, these issues mainly caused by the weight given to patients who have short follow-up. CHMP considers that this is evidence of population heterogeneity which needs to be understood for decision making about efficacy of the drug under consideration. Patients with short follow-up likely are

informative regarding the terminal event and should not be down-weighted with the aim to reduce variability.

The CHMP would ideally like to see an analysis which delivers an estimate which appropriately summarises the expected effect of the treatment for the average patient on their annual event rate for the recurrent event. A patient-weighted estimand would achieve that. The applicant states, however, that there are currently no methods in the literature that target this estimand, and the difficulties that exist in pursuing such an approach are clear, though more research in this direction could be fruitful. The target of estimation of the exposure-weighted estimand is not agreed to be appropriate. However, if the performance of the methods targeting this estimand were instead looked at in terms of their performance in estimating the patient-weighted estimand, it seems as if approaches that appropriately estimate this estimand are conservative in the situation where the treatment effect on the terminal event is not negative. In that context, it might be possible to support the use of approaches to analysis such as NB and LWYY, but only in situations where there is well established knowledge that the effect on the terminal event cannot be negative.

3. Conclusion- Qualification Opinion Statement

For scenarios where there are no terminal events it can be agreed that the methodology proposed provides clinically interpretable treatment effect measures that are more efficient than those targeting treatment effect measure based on the time to first event only. This conclusion is consistent with the fact that such methods are routinely used in certain disease areas, for example negative binomial analysis is used when looking at annualized relapse rate in multiple sclerosis and other indications mentioned in the introduction.

Clinical considerations regarding meaningfulness and the loss of information on mortality if studies become smaller when designed based on recurrent events are summarized in section 1 of this document. Of note, during the external consultation various comments addressed the need to have two primary objectives: First to exclude a detriment in mortality and, secondly, to properly assess the effect of treatment on the recurrent event (in the example here: recurrent rehospitalisation for worsening hear failure). This approach has also been discussed in the regulatory setting.

Methodological considerations in the scenario where there are terminal events are summarized here: The targeted effect on the recurrent event in the exposure-weighted approach alters dependent on the effect on the terminal event, meaning the effects are not clinically interpretable in the way CHMP would ideally require for an individual patient. There is also a loss of type I error for the individual assessment of the treatment effect on the recurrent event as soon as there is a non-neutral treatment effect on mortality. The proposed estimator for estimand 2 (based on adding 1 for the terminal event to the number of observed recurrent events) is maximum likelihood as soon as the individual components follow independent Poisson-distributions for the recurrent and the terminal event. As mentioned above, this independence would be difficult to assure and is likely not given in practice. Clustering of events pose additional uncertainties to the use of methodology for Poisson-distributed event rates.

The CHMP could also envisage the option to provide an analysis which delivers separate estimates which appropriately summarise the expected effect of the treatment on the (annual event rate for the) recurrent event while alive, and the effect on the terminal event for confirmatory decision making. These estimates should be unbiased from a statistical perspective and assumptions of the model should be transparent. Assessment of the treatment effect on mortality would have to precede the assessment of the treatment effect on the recurrent event.

Use of an approach for the recurrent event analysis where patients are given equal weight in the analysis regardless of the duration of follow-up may have the potential to achieve this objective. There are limitations with this approach, in that it would likely lead to a higher variability which could reduce the efficiency advantages the use of recurrent event approaches hopes to obtain, but, as elaborated above, this may simply indicate that more information is needed for proper decision making. The statistical properties require further consideration. The applicant states that there are currently no established methods in the literature which target this estimand. However, based on the information provided this seems to be a possibly fruitful path to investigate and the CHMP would encourage research into devising efficient methods of estimation that target such an estimand. External comments point to the fact that in the literature further methods are available that should be included into methodological discussions to choose an appropriate way forward regarding the use of recurrent events as endpoint in clinical trials for heart failure and other indications, where a terminal event limits the observation of the recurrent event endpoint.

From both, a clinical and a statistical efficiency perspective, a better understanding of those patient characteristics that determine differences in the frequency of re-hospitalisations for worsening heart failure would be of interest and could be used for adjustment in statistical models ultimately leading as well to a reduction in variability and an increase in efficiency.

ANNEX

Background information as submitted by the applicant as a separate document

List of Issues regarding provided simulation exercises (14 June 2018)

Based on the discussion above the Scientific Advice Working Party (SAWP) determined that the applicant should discuss a list of issues, before advice can be provided. On the 21st of March 2018 the list of issues were sent to the applicant. On the 6th of April 2018 the applicant provided written responses to the list. The first list of issues and the preliminary qualification team feedback on the written responses are provided below.

For the simulations of scenarios with no terminal event

Question 1.1

For the simulations of type I error, please provide the tables using 1-sided tests at the 2.5% level rather than 2-sided tests at the 5% level. Please also include the log-rank test as part of the simulations. Please then re-discuss the issue of type I error control in studies with smaller sample sizes.

Applicant's reply: we agree that simulations using 1-sided tests at the 2.5% level may provide additional information. We present Table 7 using 1-sided tests at the 2.5% level while focusing on smaller sample sizes (n = 50, 75 and 125 per group); see Table 7A below. Table 7A shows that the 1-sided type I error inflation for smaller sample sizes (n=50) is not as pronounced as for 2-sided tests, and that the 1-sided type I error is well controlled at n $\geq$ 75 per arm. We expect the results to be similar in spirit for other scenarios. Would you thus agree that Table 7A is providing sufficient insights to the similarities of the findings based on 1-sided and 2-sided tests?

Table 7A: Mean treatment effects estimates (geometric mean) and type I error rates (1-sided tests, nominal significance level $\alpha = 0.025$) under four scenarios, with treatment effect size $\mathrm{RR} = 1$, baseline recurrent event rate $\lambda_0 = 0.5$, and dispersion parameter $\theta = 0.25$.

		$n = 50$		$n = 75$		$n = 125$	
	Method	RR	Type I error	RR	Type I error	RR	Type I error
Scenario 1: Non-informative (Hypothetical)	Cox	0.998	0.025	1	0.024	1.001	0.024
	NB	0.998	0.026	1.002	0.024	1.002	0.024
	LWYY	0.998	0.028	1.002	0.024	1.002	0.024
	WLW	0.997	0.029	1.001	0.026	1	0.025
	PWP	0.998	0.028	1.002	0.024	1.002	0.025
Scenario 2: Informative (Hypothetical)	Cox	0.994	0.025	0.999	0.024	1.001	0.022
	NB	0.995	0.028	1.002	0.025	1	0.024
	LWYY	0.995	0.029	1.003	0.026	1.001	0.024
	WLW	0.993	0.028	1.002	0.025	1.003	0.024
	PWP	0.996	0.03	1.002	0.025	1.001	0.024
Scenario 3: Non-iformative (Treatment-policy)	Cox	0.998	0.024	0.999	0.024	1.001	0.023
	NB	0.998	0.028	1	0.025	1.002	0.024
	LWYY	0.998	0.029	1	0.025	1.002	0.024
	WLW	0.997	0.028	1	0.028	1.003	0.024
	PWP	0.998	0.028	1	0.025	1.001	0.025
Scenario 4: Informative (Treatment-policy)	Cox	0.995	0.026	0.999	0.026	1.001	0.023
	NB	0.996	0.029	1.001	0.025	1.002	0.026
	LWYY	0.996	0.03	1.001	0.026	1.002	0.026
	WLW	0.994	0.029	1	0.026	1	0.026
	PWP	0.997	0.029	1.001	0.025	1.001	0.025

We also agree that the log-rank test is often the method used for the initial significance test in time-to-first-event analyses. The log-rank test is identical to the score test of the Cox regression and very

similar to the Wald test used in table 7 (as no covariates are included); see for example Andersen et al (1993), page 487. Thus we believe that the results shown in table 7 (and elsewhere in the original request document) are representative for the findings to be expected based on the log-rank test.

Qualification team comments:

No further information is required for SAWP to be able to provide an opinion. There is agreement that Cox-Regression for the purpose of these qualitative investigations is sufficient and that no additional simulation outcome needs to be provided for the log-rank test. As a general comment, wherever feasible, results from one-sided testing should be provided as decision making is clearly directional and the fact that the impact of treatment on the assessment of two endpoints is needed clearly complicates assessment.

Question 1.2

Please discuss why in settings with no terminal event where the true RR=1.0 the estimate from all methods tends to favor the control group.

Applicant's reply: as pointed out by the reviewers, the reason for all estimates being larger than one is that in the original request document the averaging across simulation runs was done on the arithmetic rather than on logarithmic scale. Table 7A (see question 1.1) shows simulation results when averaging across simulation runs is done on the logarithmic scale. The difference between arithmetic mean and geometric mean is small, and the geometric mean estimates are scattered above and below 1 as expected.

Qualification team comments:

The applicant's response is plausible, however, for the smallest investigated sample-size all estimates are still less than one. Please comment whether lacking asymptotic normality can be excluded as a reason and bias is truly absent (e.g. by providing results for an even larger sample-size n). The question should be further addressed in writing and during the discussion meeting.

Question 1.3

For the simulations of power please also include the log-rank test as this is approach more likely to be used for a significance test than Cox regression.

Applicant's reply: we believe that this question has already been addressed, therefore we kindly refer to our response to question 1.1.

Qualification team comments:

No further information is required for SAWP to be able to provide an opinion. See also comments in question 1.1.

For the situation where there is a terminal event:

Question 2.1:

Please present table 11 using 1-sided tests at the 2.5% level instead of 2-sided 5% tests. Please also add a row for HR_{CV}=1.25, add the log-rank test to the table, vary HR_{CV} for estimand 2 and provide results for varying sample size.

Applicant's reply: we present table 11 using 1-sided tests at the 2.5% level, see table 11A below. We also included varying sample sizes and added HR_{CV}=1.25 for estimand 1 and HR_{CV}=0.6, 0.8, 1.25 for estimand 2. Furthermore, the averaging across simulation runs is now done on the logarithmic scale instead of averaging on the arithmetic scale. As for the response to question 1.1 for the non-terminal event scenario, we did, however, not include the results based on the log-rank test.

For both estimands the type I error remains under control under the global null hypothesis (RR_{HHF}=1 and HR_{CV}=1) for all considered sample sizes.

With the use of 1-sided tests and including HR_{CV}=1.25 for estimand 1, we observe a type 1 error inflation in favor of the treatment that has a negative effect on CV death. The reason is that for HR_{CV}=1.25 especially the severely ill patients in the treatment group (i.e. those with high frailty) die earlier and therefore contribute fewer hospitalizations. This makes the treatment appear more effective in reducing HHF. This is in line with our previous observations for estimand 1 with RR_{HHF}=1 and HR_{CV}< 1 (see second bullet point on page 64 of the original request document). Additionally, for HR_{CV}=1.25 the type I error increases with increasing sample size, because the estimated treatment effect is below 1 (see reply to question 2.4) and a larger sample size will lead to a smaller variance of the test statistics and ultimately to more frequent rejections.

In contrast, for estimand 2 we observe in table 11A that the probability to reject in favor of the treatment with positive effect on CV death is larger than α. However, we would not refer to this as a type I error when HR_{CV}≠1. Note that the original table 11 only included results for $HR_{CV} =$ 1 because RR_{HHF}=1 and HR_{CV}=1 jointly constitute the global null hypothesis for estimand 2. When including HR_{CV}≠1, a reference to "Power" seems more appropriate. As expected, the power then increases with increasing sample size. An exception is the LWYY method, see appendix A.2.3.1 in the original request document, where similar to other simulation settings presented in the original request document the power is largely unaffected by HR_{CV}.

Table 11A: Mean treatment effects estimates (geometric mean) and type I error rates (1-sided tests, nominal significance level $\alpha = 0.025$) for estimand 1 (HHF) and estimand 2 (HHF + CVD) with non-informative treatment discontinuation and $RR_{HHF} = 1$.

Endpoint	HR_{CV}	Method	$N = 2000$		$N = 3000$		$N = 4350$		$N = 5000$	
			Estimate	Type I error	Estimate	Type I error	Estimate	Type I error	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.051	0.007	1.051	0.007	1.052	0.005	1.050	0.004
		NB	1.069	0.007	1.069	0.005	1.071	0.004	1.069	0.003
		LWYY	1.118	0.003	1.117	0.001	1.120	0.000	1.118	0.000
		WLW	1.094	0.004	1.095	0.002	1.097	0.001	1.095	0.001
		PWP	1.047	0.004	1.047	0.003	1.048	0.003	1.047	0.001
	0.8	Cox	1.025	0.014	1.023	0.014	1.027	0.010	1.024	0.009
		NB	1.033	0.012	1.032	0.016	1.035	0.010	1.034	0.009
		LWYY	1.055	0.007	1.054	0.010	1.058	0.005	1.056	0.004
		WLW	1.045	0.008	1.044	0.009	1.048	0.006	1.045	0.005
		PWP	1.023	0.010	1.023	0.013	1.024	0.008	1.023	0.007
	1.0	Cox	1.000	0.024	0.999	0.025	1.002	0.023	1.000	0.025
		NB	1.001	0.024	0.999	0.028	1.002	0.025	1.000	0.023
		LWYY	1.000	0.024	0.998	0.028	1.002	0.023	1.000	0.026
		WLW	1.000	0.023	0.999	0.028	1.002	0.024	1.000	0.024
		PWP	1.000	0.023	0.999	0.027	1.001	0.025	1.000	0.023
	1.25	Cox	0.971	0.041	0.969	0.057	0.973	0.058	0.970	0.065
		NB	0.963	0.047	0.960	0.058	0.964	0.057	0.962	0.065
		LWYY	0.940	0.069	0.937	0.091	0.942	0.100	0.939	0.113
		WLW	0.949	0.060	0.947	0.081	0.951	0.088	0.948	0.100
		PWP	0.973	0.053	0.972	0.064	0.974	0.068	0.973	0.073
Estimand 2 (HHF+CVD)	0.6	Cox	0.933	0.116	0.932	0.151	0.934	0.204	0.932	0.232
		NB	0.890	0.141	0.889	0.201	0.891	0.264	0.890	0.294
		LWYY	1.002	0.024	1.002	0.024	1.004	0.023	1.002	0.024
		WLW	0.980	0.036	0.980	0.043	0.982	0.045	0.980	0.048
		PWP	0.939	0.144	0.939	0.200	0.940	0.262	0.939	0.300
	0.8	Cox	0.967	0.053	0.966	0.069	0.969	0.074	0.967	0.083
		NB	0.945	0.059	0.944	0.082	0.946	0.088	0.945	0.100
		LWYY	1.000	0.022	0.999	0.030	1.002	0.024	1.001	0.023
		WLW	0.990	0.028	0.989	0.036	0.992	0.032	0.991	0.034
		PWP	0.970	0.060	0.970	0.084	0.970	0.096	0.970	0.103
	1.0	Cox	1.000	0.025	0.998	0.026	1.002	0.022	1.000	0.026
		NB	1.001	0.023	0.998	0.026	1.001	0.024	1.000	0.022
		LWYY	1.000	0.025	0.998	0.028	1.001	0.024	1.000	0.025
		WLW	1.000	0.026	0.999	0.027	1.001	0.025	1.000	0.024
		PWP	1.000	0.025	0.999	0.028	1.000	0.024	1.000	0.025
	1.25	Cox	1.040	0.009	1.038	0.007	1.041	0.004	1.039	0.004
		NB	1.070	0.008	1.068	0.005	1.071	0.004	1.069	0.003
		LWYY	1.002	0.024	1.000	0.028	1.004	0.023	1.002	0.023
		WLW	1.012	0.018	1.010	0.019	1.014	0.018	1.012	0.015
		PWP	1.037	0.008	1.035	0.004	1.037	0.004	1.037	0.003

Qualification team comments: No further information is required for SAWP at this stage of the procedure. Simulations including the log-rank-test are not necessarily needed and Cox-regression models should suffice to reflect the outcome of time to first event analyses that should be unbiased against negative correlation between treatment effects on mortality and on rehospitalisation for worsening hear-failure. Most challenging is the situation where a terminal event occurs in a non-negligible proportion of cases. While it is formally correct (as pointed out in response to question 2.1), that HRCV<>1 is not part of the global null-hypothesis, the question can be easily rephrased regarding the power to detect such deviations from the null-hypothesis (i.e. that there is a detriment of the new treatment on cardiovascular (or overall) mortality). Moreover it has to be noted that also the conclusions on HFH are biased. This obviously needs to be addressed once the implications for decision making are discussed from a medical perspective.

Question 2.2:

Please provide additional simulations with higher mortality (~ 20%, 40% overall in the trial) to better understand the degree of type-1-error increases and behaviour of estimands 1 and 2 with varying HR_{CV} in these situations.

Applicant's reply: in an overview of published heart failure trials by Anker and McMurray (2012) the proportion of CV death events of all composite events (CV death + HHFs) was shown to be relatively stable at around 30% when considering either a time-to-first-event or a recurrent events endpoint, see the table below extracted from the article. The list of trials included in the review covers a range of overall CV mortality. For example, in the CHARM-Added trial 27.3% patients died for CV causes in the placebo arm during 41 month of median follow-up, while in the CHARM-preserved trial 11.3% patients had a CV death in the placebo arm during 36.6 months of median follow-up. As a comparison, the simulation performed in the original request document has for the base case and non-informative treatment discontinuation an overall CV mortality of 12.5% in the placebo arm during 38.5 months of median follow-up, so in this respect is similar to the CHARM-Preserved study.

Table I Number of events in 'time-to-first event' analysis and 'recurrent events' analysis of heart failure trials

Trial	Time-to-first-event (CV death or HF hospitalization): CV death as % of primary outcome ($n/n = N$)	Recurrent events (all CV deaths and all HF hospitalizations): CV death as % of all events ($n/n = N$)
CHARM-Added	316/705 = 1021 (31.0%)	649/1443 = 2092 (31.0%)
CHARM-Alternative	237/503 = 740 (32.0%)	471/1053 = 1524 (30.9%)
EMPHASIS-HF	188/417 = 605 (31.1%)	332/702 = 1034 (32.1%)
SHIFT	544/1186 = 1730 (31.4%)	940/2113 = 3053 (30.7%)
I-PRESERVE	392/661 = 1053 (37.2%)	613/1176 = 1789 (34.3%)
CHARM-Preserved	190/509 = 699 (27.2%)	340/968 = 1308 (26.0%)

n/n, CV death/HF hospitalization; *N*, CV death or HF hospitalization (time-to-first event) or total number of CV deaths plus total number of HF hospitalizations (recurrent events). CV, cardiovascular; HF, heart failure.

If the objective of additional simulations with increasing CV mortality rates is to be representative of a heart failure population, we propose to increase the HHF rate such that the proportion of CV death of all events is kept roughly at 30%. Increasing the mortality rate without changing the rate of HHF could potentially increase the generalizability of the results to other chronic indications with high terminal event rate but would no longer be representative of heart failure trials. In addition, if the rate for mortality events is as large as the rate of recurrent events or even larger, the clinical community may favor the investigation of time-to-first composite or time-to-mortality endpoints.

For the requested additional simulations to better understand the type-1-error behaviour with higher mortality, should the rate of heart failure hospitalizations be increased at the same time as increasing the mortality rate? If so, do you agree with our proposal above, i.e. to also increase the HHF rate such that the proportion of CV death of all events is kept roughly at 30%?

Preliminary qualification team comments**:** tt is agreed that the rate of heart failure hospitalizations should be increased at the same time as increasing the mortality rate. Different proportions of CV death should be investigated to further elucidate the impact of the negatively correlated effects on re-hospitalisation and mortality and particularly to understand the impact on the estimands as proposed, or requested. It is supported that different expectations regarding the terminal event may lead to different preferences regarding the choice of the analysis, but until now the discussion on the utility of recurrent hospitalisations for worsening heart failure was not perceived as specific for the situation of

heart failure with preserved ejection fraction or other selection of the patient population that are of low risk for the terminal event. The question should be further addressed in writing and during the discussion meeting.

Question 2.3:

Please discuss how it is envisaged that estimands 1 and 2 would be used in practice. Are they intended to be interpreted as an estimate of the effect on hospitalisations, or as an overall estimate of the effect of treatment combining both hospitalisations and mortality?

Applicant's reply: both estimands reflect a patient's forward-looking view of the event rate. A patient may ask: "How many events can I expect to have in the next three years, relative to how long I can expect to live in the next three years?" The proposed estimands adjust for the effect of early termination (death) by accounting for the time at risk.

Estimand 1 (HHF) could be used in settings, where it is expected that test and control treatment will not differ with respect to their effect on terminal events (deaths), based on a strong scientific rationale. In such settings, estimand 1 would measure the treatment effect on hospitalizations while alive, similar to settings without terminal events. The effect of treatments on death should be evaluated as well, and would have to be taken into account when interpreting estimand 1.

Estimand 2 (HHF+CVD) provides an overall treatment effect, including both hospitalizations and mortality, i.e. counts all disease-related "bad events" (hospitalizations for heart failure or cardiovascular deaths) while alive. It should be noted that as in other settings where composite estimands are used, the individual components would still be evaluated, in particular the treatment effect on death, and taken into account when interpreting the results. Estimand 2 weights all bad events equally, and can be seen as a natural extension of time-to-first-composite-event analyses (composite of first HHF or CVD) to the recurrent HHF setting. Other weightings are discussed in response to question 2.5 and section 3.2.1.6.2 of the original request document.

Estimand 1 and estimand 2 appear to be understandable and meaningful for patients and clinicians, have a causal interpretation, and are estimable with minimal assumptions.

We would like to illustrate this further for estimand 2, however, the following considerations also apply for estimand 1.

Using a standard causal inference framework (e.g. Hernan and Robins, 2018), we consider for each specific patient the bivariate potential outcome (number (#) of bad events, time of death/censoring) if he/she would be randomized to test treatment and control, respectively. Of note, in the actual clinical trial, the outcomes for only one of the treatments will be known, the other being missing. The table below illustrates this potential outcome framework for a trial where each patient is followed for 3.0 years (censoring) or until death. For example, patient Ann would have no bad events and would be alive at 3 years when randomized to Test; however, Ann would have 2 bad events (including death) with a death time of 1.5 years if randomized to Control.

Patient	Test		Control	
	# bad events	Time of death/censoring	# bad events	Time of death/censoring
Ann	0	3.0	2	1.5
Bill	1	3.0	1	2.5
…	…	…	…	…
AVERAGE	0.5	3.0	1.5	2.0

In the example table, the "bad event" rate while alive is 0.17=0.5/3.0 for Test and 0.75=1.5/2.0 for Control. The estimand 2 is the "bad event" rate ratio, i.e. 0.23=0.17/0.75.

Estimand 2 can simply be defined based on averages (expectations) of potential outcomes, and hence has a causal interpretation. It does not require any model assumptions for the definition. For estimation in randomized clinical trials, both semi-parametric methods (e.g. LWYY, see appendix A.2.3.1 in original request document) or parametric methods (e.g. NB, see appendix A.2.2.4 in original request document) can be considered.

As previously mentioned, the above considerations also apply for estimand 1 (HHF only).

Preliminary qualification team comments:

The applicant agrees that for estimate 1 an additional assessment of treatment effects regarding mortality is needed. Even if for estimand 2 some sort of composite endpoint thinking may be applicable, in both instances the question remains open, how a strategy for decision making should be constructed that optimizes positive conclusions regarding a treatment effect at least excluding a detrimental effect on mortality. From a drug-licensing perspective this is likely the most important question to be addressed. Whereas in end-stage disease (NYHA stage III-V and IV) some but certainly not all patients may be willing to accept a symptomatic improvement even if the treatment is associated with some uncertainty regarding the risk of dying, in earlier stages of the disease, i.e. NYHA II and III, mortality is the key aspect to be investigated. A high level of reassurance that mortality is at least not negatively affected in the whole group of these patients and in relevant subgroups is a prerequisite for designing studies using recurrent hospitalisations for worsening of heart failure analyses. Exploring recurrent hospitalisations for worsening heart failure may be of particular value in patient populations where due to the rarity of the disease information on mortality is limited as it may be the case in patients with rare forms of cardiomyopathy or with heart failure due to hereditary syndromes. In addition, recurrent heart failure event analyses may be an option in phase 2 dose finding studies or in case of an extension of an indication to a related population.

An elaborative discussion on the aspects above needs to take place in the discussion meeting.

Question 2.4:

Please discuss whether there exist alternative estimands which allows an independent evaluation of the true effect on the recurrent event independent of the terminal event (i.e. it would give 0.7 in table 8) which could then be used as a joint endpoint with a separate assessment of the RR for terminal events, and if there is one which methods could estimate it?

Applicant's reply: we would respectfully ask for some clarification on the meaning of the 'true effect' in the above question. The value of 0.7 in table 8 is certainly not the rate ratio for HHF in those alive. It is the value used in the computer simulation to generate recurrent events, both those events which in practice are observed and those events which are unobserved, i.e. those events that do not occur because the subject has died. And it is only by recovering these unobserved events and counting them together with the observed ones that one could obtain the underlying event rates and hence their ratio of 0.7. These considerations are further complicated as treatment discontinuation is an additional relevant intercurrent event in the setting of chronic heart failure studies.

In the chronic heart failure setting, evaluating the effect on the recurrent events *independent* of the terminal event based on observed data is to our knowledge not feasible. Or in other words:

disentangling the recurrent event and terminal event processes is not possible unless these processes are truly independent which would take us back to the scenarios without terminal events.

In our simulations, the association between the recurrent events and the terminal event is modelled through a shared frailty and the value 0.7 should be interpreted conditional on this subject-specific frailty and not as a marginal rate ratio. More specifically, as time progresses patient selection is taking place because the severely ill patients (i.e. those with a higher frailty) die early and patients may discontinue their study treatment. The observed recurrent event rate is thus going to change in those patients remaining alive. If the association between the recurrent events and the terminal event is positive, as simulated in section 5.2 of the original request document, then the recurrent event rate among survivors will drop as time progresses, while if the association is negative then the recurrent event rate will rise.

In table 15 (page 138) of appendix E of the original request document, see also below, we show the impact of the selection process on the true numerical value of estimand 1 which focuses on the heart failure hospitalizations only. It is these values that the estimator should be recovering rather than the 0.7 used in the simulation process.

Table 15: Numerical estimand values for two estimands with two types of treatment discontinuation. Data is generated with $RR_{HHF} = 0.7$, $HR_{CV} = 0.8, 1.0, 1.25$

	Estimand value		
HR_{CV}	0.8	1.0	1.25
Scenario 1: Estimand 1 (HHF), non-informative	0.767	0.721	0.672
Scenario 2: Estimand 1 (HHF), informative	0.767	0.719	0.669
Scenario 3: Estimand 2 (HHF+CVD), non-informative	0.812	0.815	0.820
Scenario 4: Estimand 2 (HHF+CVD), informative	0.790	0.793	0.800

In terms of data modelling, one could fit the same joint frailty model that generated the data in our simulation and recover an estimate of the parameter $\exp(\beta)$ (=0.7 in table 8). This would form an estimator under the assumption that this model is correct. But the underlying estimand is a hypothetical estimand which may not be clinically meaningful as we count both, the events which are observed and those which are unobserved, i.e. those events that do not occur because the subject has died. If the data generating process (the population) did not match the statistical model then the parameter estimate has no clear interpretation.

Preliminary qualification team comments:

This point needs to be kept for the discussion meeting. As outlined in response to question 2.3 both estimands do have a causal interpretation, but (as explained above) it is not fully clear what can be estimated, once the terminal event occurs.

Question 2.5:

Please explore further the power and type I error of rank-based approaches such as win-ratio in various scenarios, and those using weighted composites (of which estimand 2 in your example was a specific case with weight of 1 given to the terminal event).

Applicant's reply: we split our response to your question into two parts. In the first part (see 1. below), we discuss the win ratio as one example of a rank based approach. In the second part (see 2. below), we discuss weighted composites.

1. Win ratio approach

Next to other prioritized outcome measures (e.g. Buyse, 2010), the win ratio has been proposed (Pocock et al., 2012) as an effect measure that considers different outcomes according to their clinical relevance. To the best of our knowledge, the literature about the win ratio focuses on the estimation of the win ratio, distributional properties of these estimators, and on the calculation of confidence intervals for the win ratio. However, win ratio estimands as well as their clinical interpretability and relevance have not been discussed in the literature yet.

Before considering the value of additional simulations, we would like to seek advice from the SAWP on the win ratio approach:

- How would the estimand targeted by the win ratio approach (e.g. according to Pocock et al., 2012) be described using the framework and language suggested by the ICH E9(R1) draft addendum (ICH, 2017)?
- The interpretation of the win ratio critically depends on the follow-up time T (Oakes 2016). To our knowledge this fact has received little to no attention in the medical literature. For illustration, if the follow-up time T converges to infinity in a heart failure trial, every subject will experience a death and the HF hospitalization will have no effect on the win ratio. In other words, the larger the follow-up time T, the less weight we assign to HF hospitalizations. How should the win ratio approach be used in clinical research given that the interpretation of any results will critically depend on the follow-up time, i.e. results can generally not be generalized to other follow-up schemes?
- How should recurrent hospitalizations for heart failure be included into the win ratio approach? For example, the comparison could be based on the time-to-first HHF (Pocock et al., 2012) or the rate of HHFs (Rogers et al. 2016).
- Is the matched or the unmatched version of the win ratio approach more clinically relevant? In case of the matched approach, how should patients be matched?
- In practice, the interpretability and efficiency of the win ratio approach seems to heavily depend on the censoring distribution. The findings of any simulation study will thus also strongly depend on the assumed censoring distribution, which might attenuate the usefulness of simulation results. Would you agree?

2. Weighted composites

In a weighted composite endpoint, the individual components of the endpoint are assigned weights. The weights are chosen to reflect the clinical importance of the individual components of the composite endpoint. A number of statistical methods considering weighting of endpoints have been considered in recent years, such as the Mao and Lin (2016) or Luo et al. (2017). However, as highlighted by Anker et al. (2016): "[Statistical methods to weight outcomes] are limited by lack of consensus on the relative weighting of events and inconsistency across studies." Thus, while from a statistical perspective weighted outcomes might be appealing, the definition of weights in a manner that is scientifically justified and agreed upon within the clinical community is not feasible from a clinical perspective. A more detailed discussion of these aspects is given in section 3.2.1.6.2 of the original request document.

As pointed out by the reviewers, estimand 2 also constitutes a weighted endpoint in the sense that a cardiovascular hospitalization is weighted the same as a cardiovascular death.

Since we have already considered a weighted endpoint (estimand 2) and the lack of consensus in the clinical community on an appropriate weighted composite endpoint, would you agree that additional simulation focusing on weighted composite endpoints would be of limited value unless the weighted composite is informed by a clinical rational/consensus?

Preliminary qualification team comments:

This point needs to be kept for the discussion meeting.

Question 2.6:

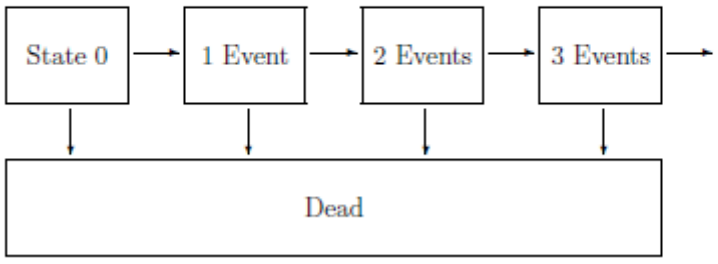
Discuss the utility of multi-stage models to simulate and estimate both, the effect of treatment on mortality and, the effect on HFH. These estimates should be investigated in simulations regarding their statistical properties, interpretability, and yardsticks to their utilization.

Applicant's reply: the utility of multi-state models in the presence of terminal events was discussed in the appendix of the original request document; see for example A.1.5 and A.2.5.

Moreover, some of the models, which were explored in the simulation study, are in fact specific examples of multi-state models, e.g. the PWP and the Negative Binomial models, see also appendix A.2.2 of the original request document. The simulation results for the associated models can thus be considered as multi-state model results, i.e. the modeling assumptions and partial likelihoods correspond to particular multi-state models.

Besides these models, one could consider more general multi-state models as depicted in Figure 14 in appendix A.1.5 of the original request document.

Figure 14: Recurrent events considered as a multi-state model with a terminal event.



This would allow the estimation of various hazard functions as well as their dependence on the number of previous events, the treatment and other covariates. Such general models, however, have several challenges:

- Treatment effects within particular higher-order transitions are difficult to interpret as they represent effects patients will benefit from only if they have entered that particular state before. In particular, these comparisons are no longer protected by randomization.
- Estimating the various transition hazards sounds attractive at first glance; however, it is not obvious how to combine this information into interpretable and clinically meaningful overall treatment effect measures. Therefore, the key challenges mentioned in section 3.2 of the original request document still remain.

- Estimation of specific transition hazards to and from certain higher event numbers might not be feasible due to the sparseness of data.

Would you agree that we have already provided a discussion on multi-state models – including simulations for specific multi-state models under a broad range of scenarios?

If simulations for additional multi-state models are required, we would like to seek advice from the SAWP on the multi-state models of interest, e.g.

- Which overall treatment effect measure (estimand) should we target?
- How many states should the model have?
- Should the treatment effects for the different transitions all vary?

Preliminary qualification team comments:

A large number of different estimates for transition rates can be estimated from the data, but simplifications may be possible, as well (e.g. why not assume that some of the parameters are constant?). For simplicity, at the moment multi-stage models may be dropped from the discussion to first precisely clarify the appropriate regulatory question.

The scientific advice working party (SAWP) determined that the applicant should discuss a list of issues, before advice can be provided.

Pending issues (from the first round of discussions)

Question 1.2: please discuss why in settings with no terminal event where the true RR=1.0 the estimate from all methods tends to favor the control group. Please comment whether lacking asymptotic normality can be excluded as a reason and bias is truly absent (e.g. by providing results for an even larger sample-size n).

Question 2.2: please provide additional simulations with higher mortality (~ 20%, 40% overall in the trial) to better understand the degree of type-1-error increases and behaviour of estimands 1 and 2 with varying HR_{CV} in these situations.

Question 2.3: please discuss how it is envisaged that estimands 1 and 2 would be used in practice. Are they intended to be interpreted as an estimate of the effect on hospitalisations, or as an overall estimate of the effect of treatment combining both hospitalisations and mortality?

Question 2.4: please discuss whether there exist alternative estimands which allows an independent evaluation of the true effect on the recurrent event independent of the terminal event (i.e. it would give 0.7 in table 8) which could then be used as a joint endpoint with a separate assessment of the RR for terminal events, and if there is one which methods could estimate it?

Question 2.5: Please explore further the power and type I error of rank-based approaches such as win-ratio in various scenarios, and those using weighted composites (of which estimand 2 in your example was a specific case with weight of 1 given to the terminal event).

Additional list of issues

For the situation where there is a terminal event:

1. The use of the frailty model requires further justification because preference would always be given to not add unstructured variability to the model:
    a. Is it impossible to explain the high variability in the frequency of rehospitalisation by means of co-variates?
    b. If there have been attempts to explain this high variability, which models have been investigated?
    c. Please discuss examples, where modelling of the high variability in rehospitalisation-rates has been attempted and in how far this has been successful / not successful.
2. ValHeft is not considered a useful example to discuss the application of recurrent events of worsening of heart failure for decision making. The key result in ValHeft was an increased mortality in patients on background ACE-inhibitor and beta blocker therapy (n = 1610), which was considered a robust result, and a decreased mortality in the other patients (n = 3400). Overall, this led to an apparent neutral effect in mortality in the study. The applicant is asked to comment on how such different results in subgroups in mortality can be detected if studies are designed based mainly on recurrent hospitalisation events and how such heterogeneity is accounted for in the modelling approaches.
3. Please discuss examples of clinical trials, where an analysis of rates of rehospitalisation for worsening heart-failure was helpful for decision making about the efficacy of a drug, or where results on HFH and mortality led to different conclusions. Please discuss this also in the context of an overall assessment of benefit and risks.

Written responses from applicant (12 October 2018)

We agree that simulations for settings without treatment discontinuation are helpful for the interpretation of the results presented in table A, see [3]. The new simulations are provided in table A* below. For ease of reference, we also include table A.

In the simulations provided below, a joint frailty model was used, where the rate ratio parameter for recurrent HHF was fixed at RR_{HHF} = 0.7. It should be noted that RR_{HHF} is just a parameter in a specific (joint frailty) model used to simulate recurrent HHF data and it is not clear whether this parameter has a meaningful interpretation to patients, see also [2, reply to question 2.4].

Estimand values were approximated by simulating data from 200.000 patients, and hence are subject to Monte Carlo error. The corresponding Monte Carlo standard errors (SE) were added to table A* below to better assess the uncertainty in these approximated estimand values.

Table A: Terminal event case: true estimand values for four scenarios, as well as the treatment effect estimates based on five established approaches. Simulated data for 100.000 patients are generated with $RR_{HHF} = 0.7$, $HR_{CV} = 0.8;1.0;1.25$.

	Exposure-weighted rate based estimand*			Equal-weighted rate based estimand			Method	Estimates		
HR_{CV}	0.8	1.0	1.25	0.8	1.0	1.25		0.8	1.0	1.25
Scenario 1: Non-informative HHF	0.783	0.722	0.688	0.752	0.727	0.72	Cox	0.841	0.799	0.782
							NB	0.752	0.700	0.684
							LWYY	0.784	0.722	0.687
							WLW	0.789	0.731	0.702
							PWP	0.849	0.811	0.791
Scenario 2: Informative HHF	0.770	0.728	0.686	0.745	0.794	0.728	Cox	0.822	0.789	0.769
							NB	0.741	0.704	0.679
							LWYY	0.771	0.727	0.684
							WLW	0.774	0.731	0.692
							PWP	0.843	0.817	0.787
Scenario 3: Non-informative HHF+CVD	0.809	0.806	0.822	0.93	1.759	3.737	Cox	0.875	0.898	0.935
							NB	0.766	0.814	0.885
							LWYY	0.809	0.806	0.821
							WLW	0.817	0.818	0.839
							PWP	0.878	0.907	0.944
Scenario 4: Informative HHF+CVD	0.800	0.800	0.820	0.799	1.498	1.737	Cox	0.859	0.881	0.929
							NB	0.767	0.797	0.889
							LWYY	0.801	0.800	0.819
							WLW	0.807	0.806	0.831
							PWP	0.879	0.900	0.944

*In the original request document, this estimand was called Estimand 1 (HHF) and Estimand 2 (HHF+CVD), respectively.

Table A*: terminal event case: approximated estimand values as well as Monte Carlo standard errors (SE) under 30 scenarios. Simulated data for 200.000 patients are generated with $RR_{HHF} = 0.7$, $HR_{CV} = 0.67;0.8;1.0;1.25;1.5$.

Endpoint	Follow-up time	HR_{CV}	Exposure-weighted rate based estimand (SE)	Equal-weighted rate based estimand (SE)
HHF	1.25	0.67	0.721(0.012)	0.703(0.013)
		0.80	0.713(0.012)	0.706(0.013)
		1.00	0.680(0.011)	0.699(0.017)
		1.25	0.690(0.011)	0.703(0.014)
		1.50	0.669(0.011)	0.703(0.015)
	3.5	0.67	0.783(0.010)	0.730(0.014)
		0.80	0.718(0.010)	0.679(0.013)
		1.00	0.704(0.009)	0.700(0.013)
		1.25	0.653(0.009)	0.682(0.013)
		1.50	0.625(0.008)	0.708(0.014)
	7	0.67	0.809(0.010)	0.698(0.015)
		0.80	0.776(0.009)	0.716(0.012)
		1.00	0.700(0.009)	0.694(0.013)
		1.25	0.642(0.008)	0.707(0.013)
		1.50	0.586(0.007)	0.708(0.013)
HHF+CVD	1.25	0.67	0.711(0.010)	0.689(0.097)
		0.80	0.742(0.010)	0.948(0.250)
		1.00	0.766(0.010)	1.099(0.167)
		1.25	0.834(0.011)	0.666(0.240)
		1.50	0.866(0.011)	3.240(2.218)
	3.5	0.67	0.764(0.009)	0.239(0.123)
		0.80	0.749(0.008)	0.856(0.103)
		1.00	0.783(0.009)	0.405(0.229)
		1.25	0.797(0.009)	1.653(0.847)
		1.50	0.816(0.009)	1.361(0.282)
	7	0.67	0.791(0.008)	0.697(0.078)
		0.80	0.797(0.008)	0.995(0.322)
		1.00	0.784(0.008)	1.621(0.630)
		1.25	0.786(0.008)	1.106(0.225)
		1.50	0.781(0.008)	1.099(0.137)

We now discuss the patterns seen in table A*, first for rate based estimands which focus on the HHF endpoint, and second for rate based estimands which focus on the composite endpoint (HHF+CVD).

Estimands on HHF

Table A* shows that both the exposure-weighted and the equal-weighted rate based estimand have an estimand value of about 0.7, corresponding to the value of the parameter RR_{HHF} in the joint frailty model, when there is no treatment effect on CVD (HR_{CV}=1) and no treatment discontinuation. For the exposure-weighted rate based estimand, the estimand value can also be derived analytically (rather than by simulation), and gives an exact value of 0.7 in this setting, see [1, section E.3.2]. As seen in table A, estimand values are larger than 0.7 in case of treatment discontinuation. This is expected since a patient who discontinues from an effective treatment loses the beneficial effect, and hence the treatment effect is attenuated.

Table A* also shows that if there is a treatment effect on CVD, the values for the equal-weighted rate based estimand are still close to 0.7 for all considered study durations and treatment effects on CVD. In other words, this estimand is not affected by different treatment effects on CVD. However, it is unclear whether this holds in general and whether the model parameter of a joint frailty model itself is a meaningful quantity for a patient, see [2, response to question 2.4].

As seen in table A*, the exposure-weighted rate based estimand gives estimand values below 0.7 for treatments with detrimental treatment effect on CVD. This is more pronounced the longer the follow-up times, since selection effects due to CVD are becoming increasingly pronounced. In case of positive correlation between CVD risk and HHF risk, the survivors tend to have lower HHF rates, which for $HR_{CV} > 1$ suggests an increasing treatment effect on HHF. Hence, as discussed in [1, Section 5.2.3.1.1] and [3, reply to question 1], the exposure-weighted rate based estimand focusing on only HHF is sensitive to a potential treatment effect on CVD (positive or negative) in cases where there is a dependence between any unobserved risk factors for HHF and CVD events.

Estimands on HHF+CVD

Table A* shows that for the exposure-weighted rate based estimand on HHF+CVD, the estimand values are larger than 0.7 if there is no treatment effect on CVD. This is expected, as this composite estimand combines treatment effects on both HHF and CVD, and hence the overall effect is attenuated compared to the effect on HHF alone for all scenarios considered. If there is a worsening treatment effect on CVD, this is captured by this estimand, as estimand values are getting closer to 1. This effect is more pronounced for shorter follow-up times (1.25 to 3.5 years), and less the case for longer follow-up times (7 years). The reason is for long follow-up times, selection effects due to CVD are becoming increasingly important. In case of positive correlation between CVD risk and HHF risk, the survivors tend to have lower HHF rates, which for $HR_{CV} > 1$ suggests an increasing treatment effect on HHF. For the composite endpoint, where the treatment effects on HHF and CVD are combined and the HHF and CVD events occur at different rates, we still see estimand values that decrease with study duration.

Table A* indicates that calculation of a reliable estimand value for the equal-weighted rate based estimand on HHF+CVD is difficult, as seen by the very large Monte Carlo standard errors. Hence, interpretation of these approximated estimand values in table A* (and table A) remains inconclusive. This estimand appears to be very sensitive to patients who die quickly after randomization, and hence provide large event rates; see also [3, reply to question 1]. It should be noted that the distribution of individual HHF+CVD rates is extremely skewed. For heavily skewed distributions, the mean may not be the most appropriate summary measure. The median may also not be appropriate in this setting, as often the majority of patients have zero events. Other robust summary measures such as trimmed means could be used, however, to our knowledge experience in a clinical trial context is limited.

Estimation

The exposure-weighted rate based estimand may be estimated by the LWYY approach. In all considered scenarios, this analysis method targets the estimand (table A, table A*, and table A** provided in the appendix).

For the equal-weighted rate based estimand, none of the investigated analysis methods (LWYY, NB, WLW, PWP) targets the estimand (table A, table A*, table A**). One possibility would be to use the plug-in estimator, see [3, reply to question 1]. However, the properties of this estimator have not been investigated in the scientific literature, and it is unclear whether there are more appropriate methods for estimating the equal-weighted rate based estimand.

Summary and conclusions

The exposure-weighted rate based estimand for the composite (HHF+CVD) endpoint has the desirable property of appropriately capturing treatment effects on CVD. However this is not the case when the HHF endpoint rather than the composite (HHF+CVD) endpoint is used. The exposure-weighted rate based estimand is targeted by the LWYY estimator for all considered scenarios.

The equal-weighted rate based estimand for the HHF endpoint appears to be not affected by varying treatment effects on CVD. The interpretation of the approximated estimand values of the equal-weighted rate based estimand for the composite (HHF+CVD) endpoint remains inconclusive due to the heavily skewed distribution of the individual event rates. The meaningfulness of using the mean as a summary measure for the equal-weighted based estimand remains also debatable. In terms of estimation, the plug-in estimator targets the equal-weighted rate based estimand, but whether this estimator is appropriate for inference is an open research question.

As highlighted in [1] and [3], finding suitable estimands in chronic diseases where patients may die for disease-related reasons remains fundamentally difficult both for time-to-event and recurrent event endpoints. Our aim is to substantiate the claim that interpretable treatment effect measures based on recurrent event endpoints can be defined in a way that may be more suitable (clinically and statistically) than traditional treatment effect measures based on the first composite event only. We do not seek to recommend a specific estimand choice, but rather to discuss the value and limitations of different treatment effect measures and their associated statistical analyses.

Appendix

Table A**: terminal event case: treatment effect estimates as well as standard errors (SE) based on five established approaches under 30 scenarios. Simulated data for 200.000 patients are generated with $RR_{HHF} = 0.7$, $HR_{CV} = 0.67;0.8;1.0;1.25;1.5$.

Endpoint	Follow-up time	HR_{CV}	Cox(SE)	NB(SE)	LWYY(SE)	WLW(SE)	PWP(SE)
HHF	1.25	0.67	0.779(0.014)	0.713(0.016)	0.721(0.016)	0.721(0.016)	0.780(0.012)
		0.8	0.779(0.014)	0.707(0.016)	0.713(0.016)	0.715(0.016)	0.778(0.012)
		1	0.742(0.015)	0.674(0.017)	0.679(0.016)	0.679(0.016)	0.748(0.012)
		1.25	0.752(0.015)	0.689(0.017)	0.690(0.016)	0.689(0.016)	0.755(0.012)
		1.5	0.735(0.015)	0.671(0.017)	0.669(0.016)	0.666(0.016)	0.738(0.013)
	3.5	0.67	0.845(0.011)	0.748(0.014)	0.785(0.013)	0.796(0.013)	0.853(0.008)
		0.8	0.798(0.011)	0.690(0.014)	0.719(0.013)	0.734(0.013)	0.809(0.009)
		1	0.790(0.011)	0.688(0.014)	0.703(0.013)	0.722(0.013)	0.804(0.009)
		1.25	0.758(0.011)	0.647(0.014)	0.653(0.013)	0.670(0.013)	0.766(0.009)
		1.5	0.746(0.011)	0.632(0.015)	0.623(0.013)	0.646(0.013)	0.751(0.009)
	7	0.67	0.862(0.010)	0.736(0.014)	0.813(0.012)	0.832(0.011)	0.873(0.007)
		0.8	0.847(0.010)	0.721(0.014)	0.778(0.012)	0.799(0.011)	0.855(0.007)
		1	0.814(0.010)	0.671(0.014)	0.699(0.012)	0.734(0.011)	0.818(0.007)
		1.25	0.778(0.010)	0.638(0.014)	0.639(0.012)	0.679(0.011)	0.787(0.008)
		1.5	0.738(0.010)	0.600(0.015)	0.582(0.012)	0.632(0.012)	0.756(0.008)
HHF+CVD	1.25	0.67	0.763(0.012)	0.683(0.015)	0.712(0.014)	0.711(0.014)	0.766(0.011)
		0.8	0.806(0.012)	0.728(0.015)	0.742(0.014)	0.743(0.014)	0.805(0.010)
		1	0.833(0.012)	0.769(0.015)	0.766(0.013)	0.764(0.014)	0.837(0.010)
		1.25	0.912(0.012)	0.868(0.015)	0.834(0.013)	0.832(0.013)	0.908(0.010)
		1.5	0.956(0.012)	0.923(0.015)	0.865(0.013)	0.862(0.013)	0.948(0.010)
	3.5	0.67	0.817(0.009)	0.694(0.013)	0.765(0.011)	0.775(0.011)	0.829(0.007)
		0.8	0.830(0.009)	0.708(0.013)	0.750(0.011)	0.764(0.011)	0.838(0.007)
		1	0.882(0.009)	0.786(0.013)	0.783(0.011)	0.801(0.011)	0.887(0.007)
		1.25	0.933(0.009)	0.852(0.013)	0.796(0.011)	0.816(0.010)	0.928(0.007)
		1.5	0.977(0.009)	0.925(0.013)	0.814(0.011)	0.842(0.010)	0.971(0.007)
	7	0.67	0.840(0.008)	0.688(0.013)	0.795(0.010)	0.809(0.009)	0.847(0.006)
		0.8	0.870(0.008)	0.736(0.013)	0.799(0.010)	0.820(0.009)	0.875(0.006)
		1	0.916(0.008)	0.790(0.013)	0.784(0.010)	0.820(0.009)	0.909(0.006)
		1.25	0.955(0.008)	0.867(0.013)	0.783(0.010)	0.829(0.009)	0.955(0.006)
		1.5	0.993(0.008)	0.934(0.013)	0.777(0.010)	0.837(0.009)	0.992(0.006)