August 23rd, 2018

Qualification procedure: EMEA/H/SAB/090/1/2018

Responses to the Additional Clarification Questions received on July 23rd, 2018 via email

Introduction

This preliminary response document addresses the questions raised by the Scientific Advice Working Party (SAWP) on July 23rd 2018 in the context of the Qualification procedure EMEA/H/SAB/090/1/2018 for the qualification opinion on *"Clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses".*

Please note that whenever we refer to the *original request document* we mean the document submitted on 1st February 2018.

The consortium members appreciate the opportunity to further clarify its intended qualification opinion request. As emphasized in the original request document and during the discussion meeting on July 10th, 2018 our aim is not to recommend one specific estimand. The suitability of Estimands 1 and 2 as well as alternative estimand choices will strongly depend on the specifics of the drug and therapeutic area of interest. In particular, the problem of constructing suitable estimands in chronic diseases where patients may die for disease-related reasons remains fundamentally difficult and the qualification opinion request is not meant to provide the final solution but rather to substantiate the claim that interpretable treatment effect measures based on recurrent event endpoints can be defined that may be more suitable (clinically and statistically) than traditional treatment effect measures based on the first composite event only**.** Once an appropriate estimand has been chosen, an analytical approach (main estimator and sensitivity analyses) has to be selected targeting this estimand.

The following abbreviations are used in this document in line with the original request document:

- CV cardiovascular;
- CVD cardiovascular death;
- HHF hospitalizations for heart failure;
- HR_{CV} hazard ratio for CV death;
- RR_{HHF} rate ratio for recurrent HHF.

Question 1:

It is noted from slide 29 of the presentation dated 10th July that estimates of estimand 1, the HHF rate while alive, are calculated for each treatment group by dividing the total number of HHF events for patients in that group and dividing that by the total time until death/study end for patients on that group. This would seem to provide an unbiased estimate only under assumptions that are unlikely to be valid in reality and are not valid in the simulations provide, namely that death is independent of HHR. An analysis based on first deriving the HHR for each patient and then calculating from this the average HHR for each treatment group would seem to provide unbiased estimates under less stringent assumptions.

a) Please discuss the reasons why the chosen approach to calculating the HHR was selected.
b) Please repeat the simulations looking at the performance characteristics of recurrent event analysis methods using estimates of estimand 1 and 2 calculated on a per patient basis, and discuss whether this reduces the issue of the estimated treatment effect on HHR being impacted by the treatment effect on CVD.

Response:

We first would like to clarify that the calculations presented on slide 29 of the presentation dated 10th July, 2018 do not refer to estimates of Estimand 1, rather they represent the calculations of the true value of Estimand 1 in the entire patient population of interest (i.e. millions of patients).

Furthermore, in the question a reference is made to 'unbiased estimates' but it is not clear to us which estimand this statement is referring to.

In part b) of the question, Estimand 1 and Estimand 2 are mentioned as well as the issue that the treatment effect on HHR is impacted by the treatment effect on CV death. Since Estimand 2 is a composite endpoint we want to highlight that for Estimand 2 there is no issue if its value is affected by the treatment effect on CVD. As discussed during the discussion meeting on July 10th, 2018 there would be concerns for estimands which favor a treatment with a worse effect on CVD. For Estimand 2 it has been shown that this is not the case across a range of realistic scenarios.

The two different event rate evaluations (dividing total number of events by total time versus averaging individual rates in the entire patient population of interest) result in two different estimand definitions. Both differ in their clinical interpretation and in the availability of established statistical methods that provide estimators of the particular estimand across a range of plausible scenarios. We refer to the clinical interpretation in our answer to question a) and to the availability of estimators in our answer to question b).

a) Please discuss the reasons why the chosen approach to calculating the HHR was selected.

To enhance transparency we introduce names for the two rates. Henceforth, we will use

- **Exposure-weighted rate** for the rate employed for Estimand 1 as presented on slide 29 of the presentation dated 10th July, 2018 and

- **Equal-weighted rate** for the rate which is based on first deriving the heart failure hospitalization rate for each patient and then calculating from this the average rate for each treatment group.

Note that these rates form the basis for the summary measures which are used in the corresponding estimand definitions.

In the following we will provide

- a **precise definition** of the different rates;
- a discussion of the **clinical interpretability** of the rates.

Definition of the rates:

For the purpose of clarity, we consider only one intercurrent event: disease-related death, i.e. CVD. The population of interest is well-defined through some inclusion/exclusion criteria and the variable of interest is the number of hospitalizations for heart failure (HHFs) while the patient is alive.

For patient i from the entire population of interest with size m let

- $\boldsymbol{N_i}$ = number of HHFs by time of CVD or end of study;
- $\boldsymbol{T_i}$ = time that patient i is in the study and alive, i.e. time of CVD or end of study (say 3 years);
- $\boldsymbol{R_i = N_i/T_i}$ = individual rate while alive, i.e. number of HHFs divided by the time that the patient is in the study and alive.

For a population of size $m = 4$ (rather than millions of patients) this set-up is illustrated through **toy example A**, see Figure 1.

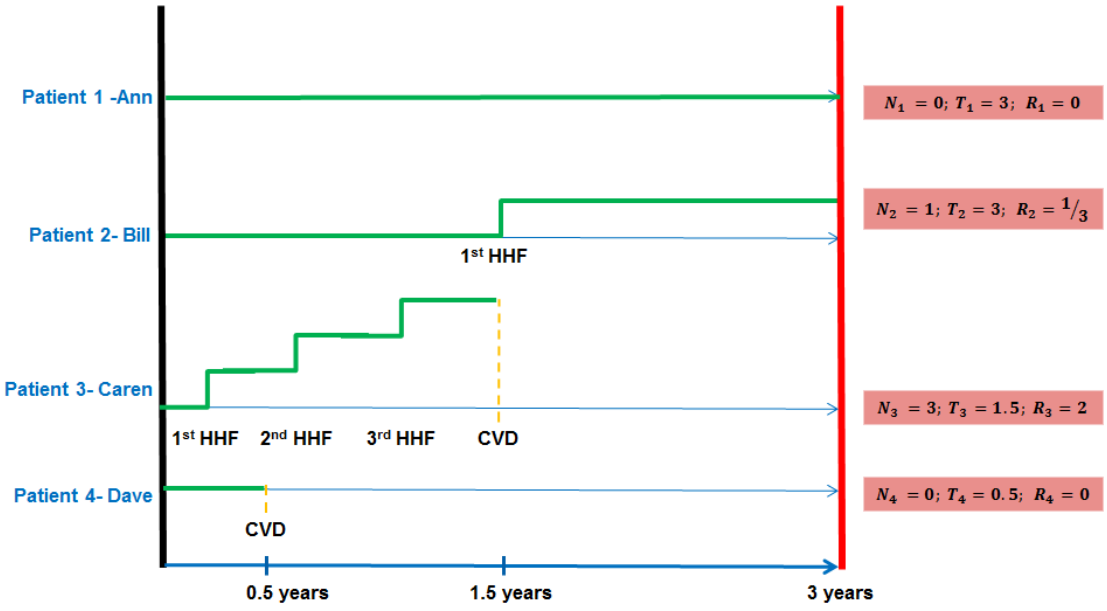

Figure 1 Toy Example A: Schematic illustration of the HHF and CVD experience of 4 patients.

The **exposure-weighted rate** can then be derived by dividing the total number of HHFs events by the total time until CVD or end of study:

$$\begin{aligned}\frac{\mathbb{E}(\boldsymbol{N})}{\mathbb{E}(\boldsymbol{T})} &= \frac{\sum_{i=1}^{m} \boldsymbol{N_i}}{\sum_{i=1}^{m} \boldsymbol{T_i}} \\ &= \frac{\boldsymbol{N_1}}{\boldsymbol{T_1}}\frac{\boldsymbol{T_1}}{\sum_{i=1}^{m} \boldsymbol{T_i}} + \cdots + \frac{\boldsymbol{N_m}}{\boldsymbol{T_m}}\frac{\boldsymbol{T_m}}{\sum_{i=1}^{m} \boldsymbol{T_i}} \\ &= \frac{\mathbf{1}}{\sum_{i=1}^{m} \boldsymbol{T_i}} \sum_{i=1}^{m} \boldsymbol{T_i} * \boldsymbol{R_i} \\ &= \sum_{i=1}^{m} \boldsymbol{w_i}\, \boldsymbol{R_i}, \qquad \text{where } \boldsymbol{w_i} = {\boldsymbol{T_i}}/{\sum_{i=1}^{m} \boldsymbol{T_i}}.\end{aligned} \qquad \text{(Equation 1)}$$

As implied through the name 'exposure-weighted rate', the individual rate for subject i is weighted by a term that depends on the exposure $\boldsymbol{T_i}$.

Note that the last equation can also be rewritten as follows

$$\frac{\mathbb{E}(\boldsymbol{N})}{\mathbb{E}(\boldsymbol{T})} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{i=1}^{m} \boldsymbol{w_i}'\, \boldsymbol{R_i}, \qquad \text{where } \boldsymbol{w_i}' = {\boldsymbol{T_i}}/{\overline{\boldsymbol{T}}}$$

i.e**.** a patient with exposure $\boldsymbol{T_i}$ equal to the average exposure $\overline{\boldsymbol{T}} = \sum_{i=1}^{m} \boldsymbol{T_i}\,/\boldsymbol{m}$ has weight 1, and a patient with longer/shorter exposure will have a weight larger/smaller than 1.

In contrast, the **equal-weighted rate** is derived by averaging the individual rates while alive, i.e.

$$\begin{aligned}\mathbb{E}\left(\frac{\boldsymbol{N}}{\boldsymbol{T}}\right) &= \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{i=1}^{m} \frac{\boldsymbol{N_i}}{\boldsymbol{T_i}} \\ &= \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{i=1}^{m} \boldsymbol{R_i}.\end{aligned} \qquad \text{(Equation 2)}$$

For the equal-weighted rate, the individual rates while alive all receive the same weight 1.

For toy example A, the average exposure is $(3 + 3 + 1.5 + 0.5)/4 \;=\; 2$, and we obtain

- exposure-weighted rate: $\frac{1}{4}\left(\frac{3}{2} \times 0 + \frac{3}{2} \times \frac{1}{3} + \frac{1.5}{2} \times 2 + \frac{0.5}{2} \times 0\right) = \frac{4/4}{8/4} = \; 0.5$;
- equal-weighted rate: $\frac{1}{4}(0 + \frac{1}{3} + 2 + 0) \approx 0.58$.

Remarks:

- m is the size of the entire population of interest and not the number of patients in the study.
- The two rates will coincide if $\boldsymbol{T_i}$ (time of death or end of study) or the individual rates $\boldsymbol{R_i}$ are the same for all patients. They will differ more the larger the variability in the individual rates while alive is.

- The rates above are defined for one treatment arm. Treatment comparisons can be based on ratios or differences of the treatment-specific rates. In the original request document we have generally focused on ratios.

Clinical interpretability of the rates:

For a given study of a certain length the two different rates can be interpreted as follows:

- **Exposure-weighted rate**: the average number of HHFs patients suffer over the length of the study or until death - whatever comes first - relative to how long patients can expect to live over the course of the study.
- **Equal-weighted rate**: the average number of HHF a patient can expect per study year s/he is alive [regardless of whether the patient will live for long or short].

The interpretation is further illustrated through the following **toy example B**. Assume the whole population of interest consists of only $m = 2$ patients and we run a study for 3 years, i.e. 156 weeks. The patient experiences are as follows:

- Patient 1 experiences $\boldsymbol{N_1 = 2}$ HHFs and dies after two weeks ($\boldsymbol{T_1 = \frac{2}{52}}$ year), i.e. the individual rate while alive for Patient 1 is $\frac{\boldsymbol{2\ HHF}}{\frac{2}{52}\text{year}}$ which corresponds to one event per week alive and an individual annualized rate while alive of $\boldsymbol{R_1 = 52}$.
- Patient 2 experiences $\boldsymbol{N_2 = 4}$ HHFs and dies after 2 years, i.e. $\boldsymbol{T_2 = \frac{104}{52}}$ years. The individual rate while alive for Patient 2 is $\frac{\boldsymbol{4\ HHF}}{\frac{104}{52}\text{years}}$ which corresponds to an individual annualized rate while alive of $\boldsymbol{R_2 = 2}$.

The exposure-weighted annualized rate is

$$\frac{\frac{\boldsymbol{N_1 + N_2}}{\boldsymbol{2}}}{\frac{\boldsymbol{T_1 + T_2}}{\boldsymbol{2}}} = \frac{\frac{\boldsymbol{2+4}}{\boldsymbol{2}}}{\frac{\boldsymbol{2}}{\boldsymbol{52}} + \frac{\frac{\boldsymbol{104}}{\boldsymbol{52}}}{\boldsymbol{2}}} = \frac{\frac{\boldsymbol{6}}{\boldsymbol{2}}}{\frac{\boldsymbol{106}}{\frac{\boldsymbol{52}}{\boldsymbol{2}}}} = \frac{\boldsymbol{3}}{\frac{\boldsymbol{53}}{\boldsymbol{52}}} \approx \boldsymbol{2.94}$$

i.e. a patient can expect to suffer 3 events relative to the expected survival time of 53 weeks (=53/52=1.02 years). Equivalently, the exposure-weighted annualized rate can also be expressed as

$$\boldsymbol{\frac{1}{2}\left(\frac{T_1}{\frac{T_1+T_2}{2}} \times R_1 + \frac{T_2}{\frac{T_1+T_2}{2}} \times R_2\right) = \frac{1}{2}\left(\frac{2}{53} \times 52 + \frac{104}{53} \times 2\right) \approx 2.94}$$

which highlights the different weights that are assigned relative to the length of the follow-up time.

In contrast, the equal-weighted annualized rate is

$$\frac{\boldsymbol{R_1 + R_2}}{\boldsymbol{2}} = \frac{\boldsymbol{52 + 2}}{\boldsymbol{2}} = \boldsymbol{27}$$

i.e. a patient can expect to be hospitalized 27 times per calendar year s/he is alive.

In summary, the exposure-weighted and equal-weighted annualized rates lead to very different answers for toy example B, which have to be interpreted with considerable care.

In a third **toy example C**, we consider a population of $m = 3000$ patients and we run a study over 3 years.

- Patient 1 – Patient 2990 experience no HHF and do not die within the study period of 3 years, i.e. $\boldsymbol{N_i = 0}$ HHFs and $\boldsymbol{T_i = 3}$ years for $i \in \{1, \dots, 2990\}$ i.e. the individual annualized rate while alive $\boldsymbol{R_i}$ for Patient 1 to Patient 2990 is $\boldsymbol{0}$.

- Patient 2991 – Patient 3000 experience $\boldsymbol{N_j = 1}$ HHFs and die after one day ($\boldsymbol{T_j = \frac{1}{364}}$ year[1]) for $j \in \{2991, \dots, 3000\}$, i.e. the individual rate while alive for these patients is $\frac{\boldsymbol{1\ HHF}}{\boldsymbol{\frac{1}{364}\ year}}$ which corresponds to one event per day alive and an individual annualized rate while alive of $\boldsymbol{R_j = 364}$.

The exposure-weighted annualized rate is

$$\frac{\frac{\boldsymbol{N_1 + \cdots + N_{3000}}}{\boldsymbol{3000}}}{\frac{\boldsymbol{T_1 + \cdots + T_{3000}}}{\boldsymbol{3000}}} = \frac{\frac{\boldsymbol{2990 \times 0 + 10 \times 1}}{\boldsymbol{3000}}}{\frac{\boldsymbol{2990 \times \frac{156}{52} + 10 \times \frac{1/7}{52}}}{\boldsymbol{3000}}} \approx \frac{\boldsymbol{0.003}}{\boldsymbol{2.99}} \approx \boldsymbol{0.001}$$

i.e. a patient can expect to suffer 0.003 events relative to the expected survival time of 2.99 years.

In contrast, the equal-weighted annualized rate is

$$\frac{\boldsymbol{R_1 + \cdots + R_{3000}}}{\boldsymbol{3000}} = \frac{\boldsymbol{2990 \times 0 + 10 \times 364}}{\boldsymbol{3000}} \approx \boldsymbol{1.21}$$

i.e. a patient can expect to be hospitalized 1.21 times per calendar year s/he is alive.

Arguably, the toy examples B and C are based on somewhat extreme scenarios. However, they illustrate that the distribution of the individual rates while alive (i.e. $\boldsymbol{R_i}$) can be highly skewed leading to considerable differences between the exposure-weighted and equal-weighted rates. This skewness is caused by a few patients that suffer HHFs early and die early. While the weights for the patients are equal, a few such patients will have a big influence on the equal-weighted rate. The proportion of such patients will generally be larger when considering the number of bad events, i.e. the composite for HHFs and CVD, as every patient that dies early will have a relatively large individual rate while alive (i.e. $\boldsymbol{R_i}$).

It is therefore questionable that an average of the individual rates (i.e. the equal-weighted rate) forms the basis for a meaningful and interpretable summary measure and ultimately estimand. There are at least two additional reasons in support of this statement:

[1] We consider a year to have 52 weeks and therefore 7×52=364 days.

- For the considered application of heart failure trials most patients (about 60-80% in previous trials) do not experience any bad event. The individual rates for all these patients are 0, regardless of their length of follow-up. Thus, important information on follow-up times is disregarded.
- The fact that patients with relatively short follow-up times can have a relatively large influence suggests that the equal-weighted rate can also be highly sensitive to time-changing rates. For example, if the event rate is high early in the trial, this early high rate would have a large influence on the estimand value through these patients.

As illustrated through the toy examples, the exposure-weighted rate appears to have a meaningful and transparent interpretation. It is not highly influenced through patients with short follow-up times and it captures relevant information on follow-up times for all patients. None of these points can be made in favour of the equal-weighted rate. These considerations provide the reason for our preference of exposure-weighted rates over equal-weighted rates, and why we used the exposure-weighted rate in the original request document.

b) Please repeat the simulations looking at the performance characteristics of recurrent event analysis methods using estimates of estimand 1 and 2 calculated on a per patient basis, and discuss whether this reduces the issue of the estimated treatment effect on HHR being impacted by the treatment effect on CVD.

The simulations presented in the original request document focused on various estimators that are well-established in the literature. One of the aims of the simulation study was to investigate whether these estimators target the estimands of interest across a range of plausible scenarios.

To our knowledge neither estimators nor statistical tests for estimands based on equal-weighted rates are discussed in the scientific literature. In a simulation we investigate whether the equal-weighted rate based estimand is targeted by any of the established statistical methods which were investigated in the original request document. Results for a selection of scenarios are shown in Table A and reveal that none of the established approaches targets the equal-weighted rate based estimand. In contrast, the results provide reassurance that LWYY, see Appendix A.2.3.1 in the original request document, targets the exposure-weighted rate based estimand for all considered scenarios. Note that the true estimand values were derived based on approximate calculations for a population of size $m = 100.000$. Focusing on the true estimand values, we notice the large values for the equal-weighted rate based composite estimand of HHF and CVD which are caused by patients that die relatively early. These conclusions also apply to all additional scenarios considered in the original request document.[2]

While no established estimator is readily available for the equal-weighted rate based estimand one could in principle use a plug-in estimator, i.e. non-parametric estimates for $\boldsymbol{R_i}$ are plugged into the formula that is used for the estimand derivation, see Equation (2). Associated confidence intervals and hypothesis tests could then be based on bootstrap approaches and permutation tests, respectively. To get a glimpse into the operating characteristics of such an estimator we performed an additional simulation where we investigate the variability of the plug-in estimators for both the exposure-weighted and equal-weighted rate based estimands.

[2] Tables summarizing the findings for the additional scenarios are omitted but can be shared in case of interest.

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.

We focus on the same scenarios as presented for Table A and a sample size of 4350. For the exposure-weighted rate based estimand the plug-in estimator is derived by plugging non-parametric estimates for $\mathbb{E}(N)$ and $\mathbb{E}(T)$ into Equation (1). The resulting rate ratio estimates are presented in Table B. In order to appreciate the variability of the plug-in estimators we report the standard deviation (SD) based on the 1000 clinical trial simulations as well as the minimum (min) and maximum (max) estimates.

Table B: Terminal event case: Treatment effect estimates (i.e. rate ratios) based on the plug-in estimators for the exposure-weighted and equal-weighted rate based estimands. Results are based on 1000 simulations, sample size $N = 4350$, $RR_{HHF} = 0.7$, $HR_{CV} = 0.8; 1.0; 1.25$.

		Exposure-weighted based estimand*				Equal-weighted based estimand			
	HR_{CV}	Mean	SD	Min	Max	Mean	SD	Min	Max
Scenario 1: Non-informative HHF	0.80	0.769	0.067	0.571	0.979	0.729	0.114	0.442	2.484
	1.00	0.721	0.062	0.560	0.928	0.724	0.102	0.347	1.353
	1.25	0.670	0.057	0.522	0.886	0.713	0.102	0.399	1.301
Scenario 2: Informative HHF	0.80	0.769	0.067	0.570	1.043	0.737	0.103	0.220	1.235
	1.00	0.726	0.066	0.532	0.978	0.727	0.107	0.281	1.235
	1.25	0.678	0.059	0.488	0.883	0.720	0.124	0.312	2.589
Scenario 3: Non-informative HHF+CVD	0.80	0.794	0.059	0.623	0.981	1.104	2.666	0.001	66.606
	1.00	0.797	0.058	0.637	1.002	1.999	21.026	0.039	663.342
	1.25	0.799	0.055	0.628	0.984	1.930	10.360	0.014	319.777
Scenario 4: Informative HHF+CVD	0.80	0.794	0.058	0.599	1.004	1.295	4.654	0.003	109.483
	1.00	0.800	0.060	0.622	1.012	1.825	15.447	0.022	475.896
	1.25	0.806	0.057	0.627	0.981	3.969	75.013	0.043	2371.761

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

The following observations can be made:

- When focusing on HHF only, the SD for the equal-weighted rate based plug-in estimator is increased by a factor of 1.5 to 2.1 compared to the SD for the exposure-weighted rate based plug-in estimator. This implies that the sample size needed to show the same effect size based on the equal-weighted rate based estimand also increases by a factor of 2.25 to 4.41 compared to that needed based on an exposure-weighted rate based estimand.

- The increase in SD is more dramatic in case of the composite of bad events, i.e. HHF+CVD.

- When focusing on HHF only, the estimated rate ratios based on the equal-weighted rate based estimators are relatively stable across different treatment effects on CVD. In contrast, the estimated rate ratios based on the exposure-weighted rate based estimators are more sensitive to changes in the treatment effect on CVD. More specifically, the treatment effect on the rate ratio gets larger with worse effects on CVD, see also the discussion during the Meeting on July 10th, 2018.

- For the composite of HHF+ CVD, the plug-in estimators for the equal-weighted rate based estimand are highly influenced by the skewed distribution of the individual rates while alive. In our simulations we observed estimated rate ratios as large as 2372 bad events per calendar year alive.

The performed simulations highlight that

- none of the established approaches targets the equal-weighted rate based estimand;

- the plug-in estimator for the equal-weighted rate based estimand has a substantially larger variability than that for the exposure-weighted counterpart.

Taking these findings together with the lack of a transparent and meaningful interpretation of the equal-weighted rate as expressed in the response to Question 1a) further supports our preference for the exposure-weighted rate based estimand.

Question 2:

Please provide sample size calculations to show how the sample size and power vary when recurrent event analysis is used as primary compared to time to event analysis in differing scenarios, some of which should be based on past clinical trials.

Response:

Our response will focus on the comparison of sample size and power between time-to-first-event analysis and recurrent event analysis for the chronic heart failure scenario. Time-to-first-event analysis has been commonly used as primary analysis method for the composite event of heart failure hospitalization (HHF) and cardiovascular death (CVD). Thus for the comparison of time-to-first-event analysis and recurrent event analysis, our response will focus on Estimand 2, i.e. the exposure-weighted rate based estimand for the composite of a recurrent HHF and CVD. We are not performing a comparison of power and sample size for methods estimating Estimand 1, i.e. exposure-weighted rate based estimand for HHF, because Estimand 1 is considered only appropriate in cases with negligible death rates or where it can be reasonably assumed that there is no treatment effect on the terminal event.

We will address this question in two parts. Firstly, we compare the sample size and the power of recurrent event analyses with a time–to-first-event analysis based on the base case scenarios with non-informative treatment discontinuation, which we described in Section 5.2.1 of the original request document. Secondly, power and sample size of the different methods are compared based on data from the Val-HeFT study (Cohn et al., 2001) since the Val-HeFT study was also considered in Section 4.2 of the original request document.

Sample size comparison for the base case scenario

For the first part, the sample size required to achieve 80 % and 90 % power, respectively, for a time-to-first-event analysis (Cox regression) and two recurrent event methods (negative binomial (NB) model, see Appendix A.2.2.4 in original request document, and LWYY model) are discussed based on the base case scenario. Different treatment effects on recurrent HHF (RR_{HHF}) and on CVD (HR_{CV}) are considered, assuming that the treatment effect on CVD was neutral or positive (i.e. $HR_{CV} \leq 1$), and also equal or smaller (i.e. closer to 1) than the treatment effect on heart failure hospitalizations. These are reasonable planning assumptions based on the effects observed in previous heart failure trials.[3] We do not consider scenarios involving detrimental effects on CVD, i.e. $HR_{CV} > 1$, since a trial would not be conducted if there was good reason to assume a detrimental CVD effect. Furthermore, the results based on non-informative and informative treatment discontinuations are very similar. We thus only present the former results.

The required sample size is computed based on the simulation model that was also used for the original request document, see Section E.2. The power is simulated on a grid of sample size values (N=2000 to N=6000 in steps of 50 and N=6100 to N=8000 in steps of 100). In Table C we present the minimum sample size required to achieve the target power level. For each scenario, 10.000 simulation runs are performed. In case the required sample size exceeds 8000, the observed power for N=8000 is provided. Note that in sample size planning for time-to-first-event analyses, patient specific frailties are usually not considered. However, these

[3] See also the presentation slides for the additional question 3 presented at the discussion meeting on July 10th, 2018.

have been included here, as they allow controlling the ratio of all events to first events as well as creating a linkage between HHF process and CVD process – two important features relevant for recurrent event sample size planning. Without frailties, the ratio of all events to first events would not be in the order observed in previous trials and the CVD process would be independent of the HHF process, which does not seem reasonable. Including these patient specific frailties also for the time-to-first-event analysis allows for a fair comparison between time-to-first and recurrent event methods, even though typical time-to-first-event sample size calculations would assume no patient-specific frailties but smaller treatment effects to be detected.

The simulation results are presented in Table C. As expected, the required sample size to achieve the target power is reduced when using recurrent event methods as compared to a time-to-first-event analysis. However, sample sizes for recurrent events would still require several thousand patients.

Table C Required sample size to achieve target power for recurrent event methods and time-to-first composite event analysis based on the base case scenario with non-informative treatment discontinuation.

Target Power	RR_{HHF}	HR_{CV}	Cox		LWYY		NB	
			N	Power*	N	Power*	N	Power*
80 %	0.7	0.70	4100	0.802	3150	0.813	2400	0.804
		0.80	5300	0.803	3250	0.801	3050	0.807
		0.90	7100	0.809	3400	0.805	3900	0.806
		1.00	> 8000	0.719**	3550	0.803	5150	0.805
	0.75	0.75	6300	0.804	4850	0.800	3700	0.809
		0.80	7300	0.806	4900	0.808	4150	0.806
		0.90	> 8000	0.709**	5050	0.804	5600	0.807
		1.00	> 8000	0.539**	5250	0.800	7700	0.802
90 %	0.7	0.70	5500	0.903	4200	0.901	3200	0.901
		0.80	7100	0.901	4350	0.900	4100	0.901
		0.90	> 8000	0.854**	4500	0.901	5200	0.902
		1.00	> 8000	0.719**	4700	0.902	7000	0.902
	0.75	0.75	> 8000	0.889**	6500	0.905	4900	0.904
		0.80	> 8000	0.839**	6700	0.903	5550	0.902
		0.90	> 8000	0.709**	6800	0.901	7400	0.901
		1.00	> 8000	0.539**	7100	0.902	> 8000	0.813**

*Observed power for the given sample size.

**Observed power for N=8000.

With respect to the two recurrent events methods, we can see that while the sample size required for LWYY is roughly stable across different HR_{CV} values, the sample size required for the NB model changes with changing treatment effect on CVD. When the treatment effect on CVD is large (i.e. $HR_{CV} \leq 0.8$), the sample size required for the NB model is smaller than that required for the LWYY model to achieve the same power, whereas when the treatment effect on CVD is small (i.e. $HR_{CV} = 0.9 \; or \; 1.0$), the required sample size for the LWYY model is smaller. These observations are in line with those made previously for the mean estimated treatment effect and power, see e.g. Table 9 and Figure 10 of the original request document.

Furthermore, when the effect on CVD is large (i.e. $HR_{CV} \leq 0.8$), the sample size to achieve 80% power using a time-to-first-event analysis roughly results in 90 % power or more when using a recurrent event method. For example, for $RR_{HHF} = 0.7$ and $HR_{CV} = 0.7$ we require $N = 4100$ patients to achieve 80 % power using the Cox model, while only 3150 and 2400 patients are required for the LWYY and the NB model, respectively. The 4100 patients needed for the Cox model, however, would give close to 90 % power using the LWYY model

(n=4200 patients required), and more than 90 % using the NB model (n=3200 patients required for 90% power). Therefore, in terms of study planning, one could consider to plan a study with 80 % power for a time-to-first-event analysis but use a recurrent event analysis for a primary recurrent event estimand, while keeping the time-to-first-event analysis as secondary time-to-first-event estimand. This would ensure that enough power is available for the traditional time-to-first-event analysis and allows gaining more insights into mortality effects. When the effect on CVD is assumed to be small (i.e. $HR_{CV} = 0.9 \; or \; 1.0$), planning a study with recurrent event estimands remains feasible, while a time-to-first-event estimand would result in studies with a sample size larger than 7000. However, a treatment with a large effect on recurrent HHF and a small effect on CV death could still be of interest for patients and thus could be investigated via a recurrent event estimand.

Finally, the decision to perform a recurrent event analysis instead of a time-to-first-event analysis should of course not be guided only by the resulting lower sample size needed. Instead, the decision should be based on discussions around the fact that a recurrent event estimand might characterize the disease under investigation better than a time-to-first-event estimand.

Sample size comparison based on Val-HeFT study

In the second part of the response, we compare the recurrent event with time-to-first-event analysis methods concerning sample size and power based on data from the Val-HeFT study. This was a parallel group, placebo-controlled, double blind clinical trial with 5010 patients suffering from CHF of New York Heart Association (NYHA) class II, III or IV that were randomly assigned to receive test treatment valsartan or placebo in a 1:1 ratio. The trial was designed with two primary endpoints: time to all-cause mortality and time to a combined endpoint of mortality and morbidity, defined as the incidence of cardiac arrest with resuscitation, HHF or receipt of intravenous inotropic or vasodilator therapy for at least four hours. The trial results showed that overall mortality was similar in the two groups. The risk of the combined endpoint, however, was 13.2% lower with test treatment than with placebo (HR = 0.87; 97.5% confidence interval, 0.77 to 0.97; p-value: 0.009), predominantly because of a lower number of patients hospitalized for HF: 455 (18.2%) on placebo and 346 (13.8%) on test treatment (p-value < 0.001). The comparison of both endpoints between test treatment and placebo was performed using a log-rank test.

The comparison of recurrent event with time-to-first-event analysis methods is performed by resampling data from the Val-HeFT study. The resampling scheme is illustrated in Figure 2.

In detail, a random bootstrap sample of size m is drawn from the Val-HeFT data ($n = 5010$). In our investigations, the sample size m is varied between $m = 2000$ and $m = 10500$ in steps of size 100. The data for the treatment group of the random sample is drawn from the treatment group of the Val-HeFT data and the data for the control group of the random sample is drawn from the control group of the Val-HeFT data. The drawn sample is then analyzed. The analysis methods of interest are a time-to-first composite event analysis (Cox regression) and two recurrent event methods (negative binomial model and LWYY model). For each method, no covariate other than treatment is considered. In the negative binomial model, the logarithm of the time to study termination is considered as an offset. This process of randomly drawing a sample and analyzing it is repeated $b = 10000$ times. Based on the $b = 10000$ analyzed random samples of size m from the Val-HeFT data, the power for the analysis methods of interest for a given sample size m is then calculated.

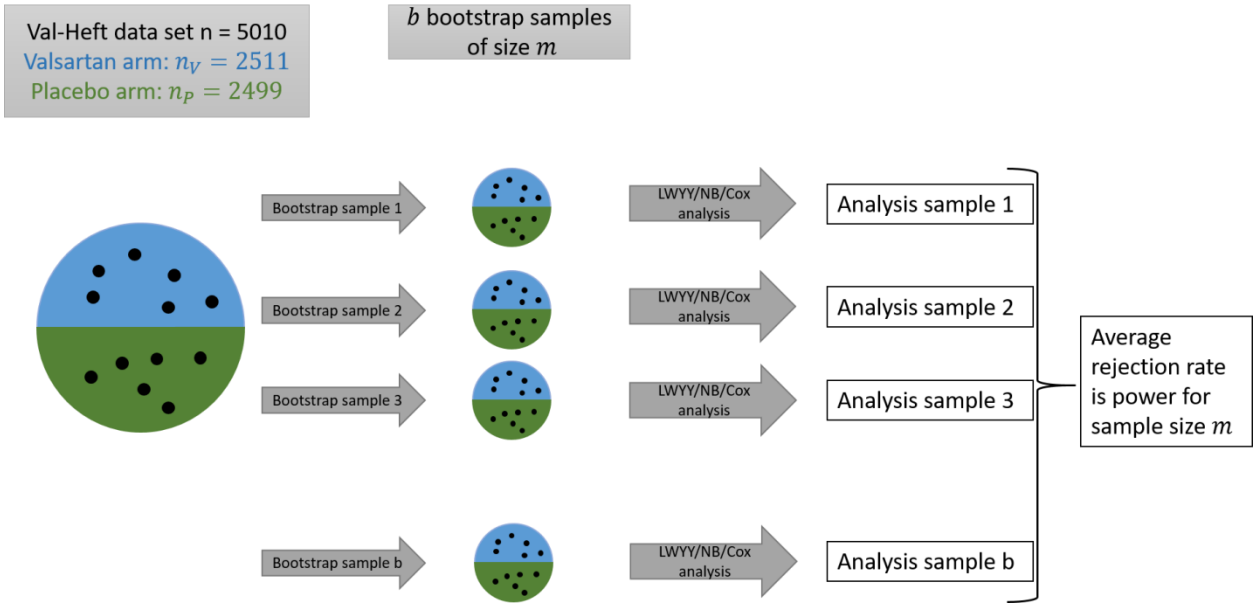

Figure 2 Schematic illustration of bootstrapping of Val-HeFT data.

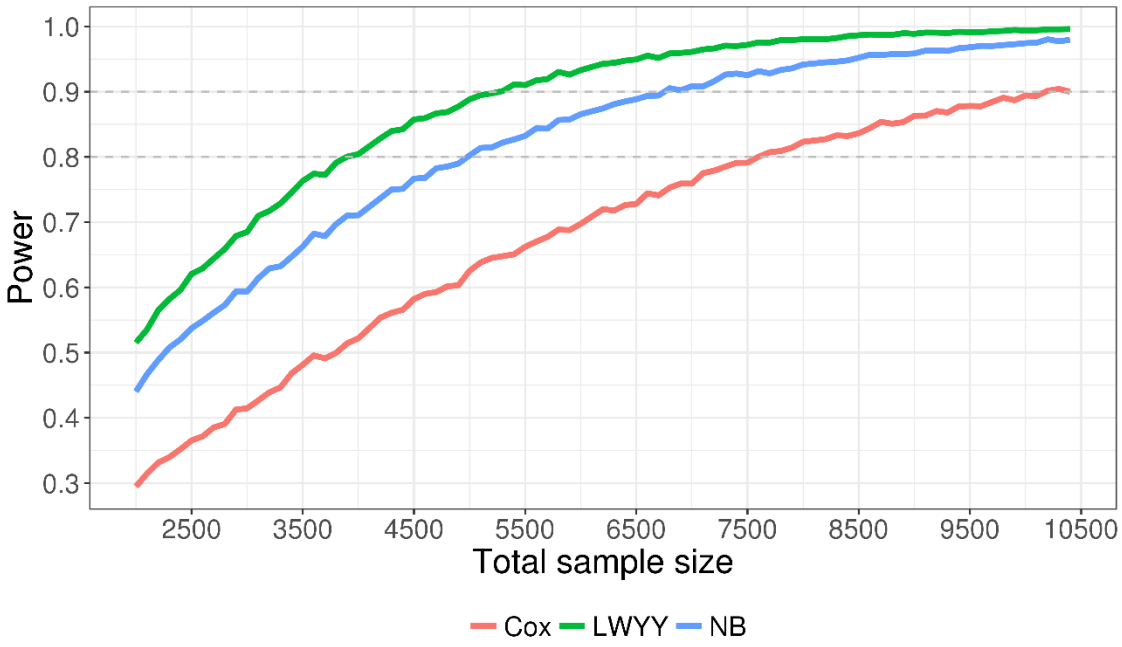

Figure 3 Power of recurrent event analyses (negative binomial, LWYY) and time-to-first composite event analysis based on resampled data from the Val-HeFT study.

Figure 3 shows that based on the Val-HeFT data, the recurrent event methods have a higher power than a time-to-first composite event analysis based on the Cox regression. Moreover, the LWYY method requires the

smallest sample size among all considered methods. The NB method requires a larger sample size than the LWYY method, but a smaller sample size than the Cox regression. As Table D highlights, the sample size required for the time-to-first-event method to achieve a power of 80% or 90% is clearly larger than the sample size required for the recurrent events methods. The sample sizes shown in Table D where determined based on a grid search with step size 100.

Table D Required sample size to achieve target power for recurrent event methods and time-to-first composite event analysis based on resampled data from the Val-HeFT study.

Method	**Total sample size required to achieve target power**	
	Power=80%	Power=90%
Cox	7600	10200
LWYY	3900	5300
NB	5000	6800

The differences in power and sample size between the recurrent event methods and time to event method based on the Val-HeFT data are in alignment with the results shown in the first part of our response to Question 2. In detail, the time-to-first-event method needs a much higher sample size than the recurrent event methods and the NB method requires a higher sample size than the LWYY method for scenarios in which there is a small or neutral effect on CVD, which was the case in the Val-HeFT study.

Question 3:

Please discuss the performance characteristics of recurrent event methods in scenarios where the HHR changes over time, i.e. the HHR generally increases over time. Do (or do not) completed clinical trials in different therapeutic areas indicate changing event rate over time in survivors?

Response:

We split our response into three parts. In the first part, we discuss disease aspects that would result in changing event rates. In the second part of our response, we address the first part of the question about the performance characteristics of recurrent event methods when the event rate changes over time. This is done based on both theoretical considerations as well as simulation results presented in the original request document. Finally, we respond to the second part of the question by discussing whether completed studies in multiple sclerosis and heart failure indicate a changing event rate over time.

Disease aspects resulting in changing event rates over time

When discussing changing event rates over time, it is crucial to differentiate between changes of the event rate on the population level and on the patient level. The population-level event rate is the number of events, which are expected to occur in the population within one time unit. The patient-level event rate is the number of events per unit time which a patient is expected to experience. Below we describe two examples of patient level changes over time:

1. The patient-level event rate might change over time because of the underlying disease process or the definition of the inclusion and exclusion criteria. For instance, if patients are included into a clinical trial in a vulnerable phase (e.g. shortly after a worsening of clinical conditions) the event rate could be high at baseline and decrease thereafter.

2. The event rate could change every time an event occurs because of disease progression. For example, every time a patient is hospitalized for worsening heart failure it might be reasonable to assume that after the hospitalization, the risk of having another hospitalization increases due to the permanent deterioration of the patient's health.

These two patient-level changes in the event rate also directly translate into event rate changes on the population level. In addition to changes on a patient level, changes in the population-level event rate can be due to a selection process in the population. For instance, due to the positive correlation between CVD and HHF, the survivors tend to have lower HHF rates.

Performance characteristics of recurrent event methods when event rates change over time

In what follows, we discuss the performance of two recurrent events methods, i.e. the LWYY model and the negative binomial (NB) model, in settings where the patient-level event rate changes over time. The LWYY method assumes that all subjects share the same unspecified baseline rate function. Therefore, as long as the proportional rates assumption is fulfilled, scenarios with changing patient-level event rate do not violate the assumptions of the LWYY model. In contrast to the LWYY model, the standard NB model makes the

assumption of a constant patient-level rate. However, when all patients have the same follow-up time, the number of events at the end of the trial still follows a NB distribution even if the patient-level event rate changes over time. In this case, the rate estimated by the NB model is equal to the cumulative event rate.

The above mentioned considerations are supported by the different simulations we provided in our original request document. For the setting without terminal events we investigated a non-homogeneous Poisson process (cf. example 1 of patient-level changes mentioned above) with patient-specific frailties and a decreasing intensity function.[4] The results did not differ systematically from those presented for the base case simulations, i.e. both the LWYY model and the NB model preserved the type I error rate for practically relevant sample sizes.[5] With regard to power, only in situations when the annualized relapse rate in the first year was substantially higher than in the second year (ratio of 1.63 in our simulations) the power was about 10-15% lower than in the constant event rate case.[6] For the setting with terminal events, we investigated an increasing autoregressive event rate (cf. example 2 of patient-level changes mentioned above) as variation of the base case.[7] Under the global null hypothesis, i.e. $HR_{CV} = RR_{HHF} = 1$, the LWYY model preserved the type I error rate, while there was an increased type I error rate for the NB model. With regard to power, the relative performance of all methods was similar to that observed for the base case, but the power of all methods was slightly increased.[8] Note that all scenarios for the terminal event case also include a population-level change due to linkage of CVD and HHF frailties.

Indications about changing event rates over time in completed clinical trials

The second part of the question focuses on event rate changes observed in completed clinical trials in different therapeutic areas. For multiple sclerosis we would like to refer to the systematic review by Nicholas et al. (2012). In a meta-analysis including the information from 13 clinical trials in relapsing multiple sclerosis, Nicholas et al. (2012) showed that the annualized relapse rate is decreasing over time. As the above mentioned selection process is negligible in multiple sclerosis trials, i.e. the rate of disease-related death is very low, this finding indicates a decreasing event rate on the patient level. This was also the motivation for including a non-homogeneous Poisson process as one variation of the base case for the simulations in the setting without terminal events.

For heart failure, we are not aware of any similar investigation of the patient-level event rate. However, the mean cumulative functions (MCFs) presented for various heart failure studies (see Figure 4) indicate that in most of the studies there seem to be only minor changes in the event rate over time on the population level, an exception being the CHARM-Alternative study. It should be noted here that a linear MCF would indicate a constant population-level event rate. For the CHARM-Alternative study the MCF seems to indicate an event rate which decreases over time, which might be due to the selection effect described above, i.e. patients with a high rate of HHF die earlier so that the tail of the MCF is primarily driven by the patients with a low rate of HHF. However, this might also be due to a decreasing event rate on the patient level. For the other studies the deviations from linearity are only relatively small, mostly indicating a slightly decreasing population-level event rate. The only study which has some indication of a slightly increasing population-level event rate is

[4] See Appendix D.1 in the original request document.
[5] See Tables 49-52 and Tables 57-60 in Akacha et al., 2017.
[6] See for example Table 1 and Table 41 in Akacha et al., 2017.
[7] See Chapter E.5.2 in the original request document.
[8] See for example Figure 9 and Figure 21 in the original request document.

ValHeft. As there could be different sources for a population-level change that even act in different directions, the interpretation of the MCFs in regards to patient-level changes is inconclusive.

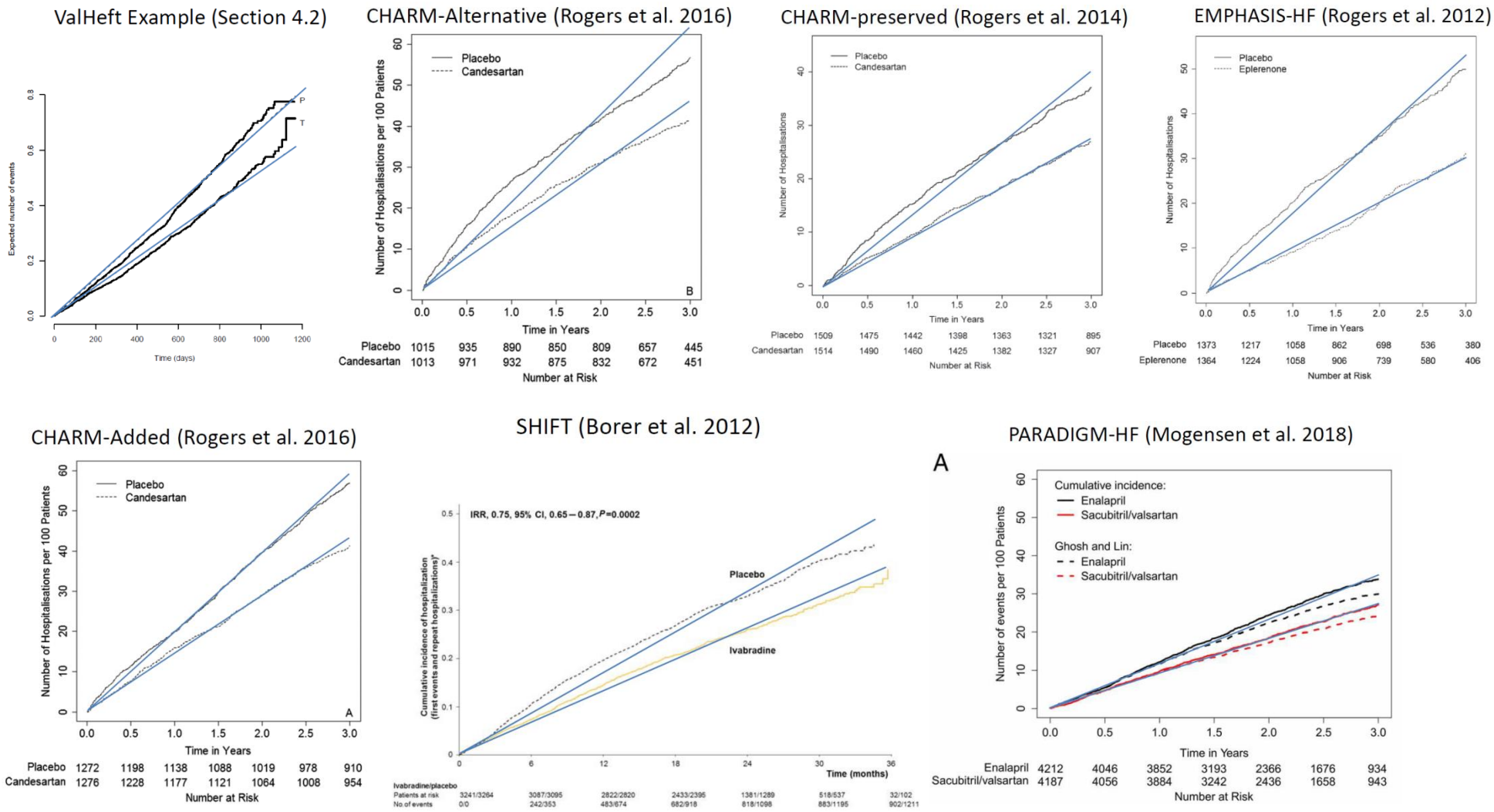

Figure 4: Estimated mean cumulative functions for hospitalizations for heart failure from various publications of heart failure studies. The blue straight lines have been manually added as a visual aid to roughly judge deviations from linearity.

Question 4:

For estimand 2 please clarify how events are counted when a patient is hospitalised and then dies while hospitalised. Would this be counted as one event or two? Please further discuss the clinical interpretation of results on estimand 2.

Response:

In clinical trials, it seems appropriate for Estimand 2 (exposure-weighted bad event rate) to count only one bad event in case death occurs shortly after hospitalization (e.g. less than 24h after admission), and otherwise two bad events. This specification would also be consistent with Hicks et al. (2014), which specify: "Hospitalization is defined as an admission to an inpatient unit or a visit to an emergency department that results in at least a 24 hour stay (or a change in calendar date if the hospital admission or discharge times are not available)." Hence if a patient dies less than 24h after hospital admission, this would not count as a hospitalization, but as a death without hospitalization (i.e. one bad event). If a patient dies at least 24 hours after admission, this could legitimately be considered an additional disease-related bad event. This is also the approach taken in the PARAGON-HF trial (Solomon et al., 2017).

The alternative of always only counting one bad event for death during hospitalization has also been used (Rogers et al., 2014). Here, death may be seen as the worst possible hospitalization.

In the simulation studies presented in the original request document, we used the simplified setting where hospitalization is an instantaneous event without duration, and hence death after hospitalization was always counted as two events.

We think that the interpretation of Estimand 2 (exposure-weighted bad event rate) would not fundamentally change, regardless of which specific definition for bad event counts would be used.

References

Akacha et al. (2017). Simulations for efficacy comparisons of time-to-event with recurrent event analyses. Technical Report. Available at https://www.biostat.uni-hannover.de/fileadmin/institut/pdf/complete.pdf.

Borer et al. (2012). Effect of ivabradine on recurrent hospitalization for worsening heart failure in patients with chronic systolic heart failure: the SHIFT Study. *Eur Heart J 33*, 2813-2820.

Cohn, J. N., Tognoni, G. (2001). A randomized trial of the angiotensin-receptor blocker valsartan in chronic heart failure. New England Journal of Medicine, 345(23), 1667-1675.

Hicks et al. (2014) Standardized Definitions for Cardiovascular and Stroke Endpoint Events in Clinical Trials. Draft Definitions for CDISC August 20, 2014. https://www.cdisc.org/system/files/all/standard/Draft%20Definitions%20for%20CDISC%20August%2020%2C%202014.pdf (accessed August 7, 2018)

Mogensen et al. (2018). Effect of sacubitril/valsartan on recurrent events in the Prospective comparison of ARNI with ACEI to Determine Impact on Global Mortality and morbidity in Heart Failure trial (PARADIGM-HF). *Eur J Heart Fail 20(4)*. 760-768.

Nicholas et al. (2012). Time-patterns of annualized relapse rates in randomized placebo controlled clinical trials in relapsing multiple sclerosis: a systematic review and meta-analysis. *Multiple Sclerosis Journal 18 (9)*, 1290-1296.

Rogers et al. (2012). Eplerenone in Patients With Systolic Heart Failure and Mild Symptoms: Analysis of Repeat Hospitalizations. *Circulation 126(19)*, 2317-23.

Rogers et al. (2014). Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. *Eur J Heart Fail 16(5)*, 33-40.

Rogers et al. (2016). Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model. *Statistics in Medicine 35(13)*, 2195-205.

Solomon SD, Rizkala AR, Gong J, Wang W, Anand IS, Ge J, Lam CSP, Maggioni AP, Martinez F, PackerM, et al. (2017). Angiotensin Receptor Neprilysin Inhibition in Heart Failure With Preserved Ejection Fraction: Rationale and Design of the PARAGON-HF Trial. JACC:Heart Failure 5:471–482.