6th April 2018

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

Reply to
'Treatment effect measures when using recurrent event endpoints – Qualification Opinion List of Issues regarding provided simulation exercises'

Introduction

This preliminary response document addresses the questions raised by the Scientific Advice Working Party (SAWP) on March 21st 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"* and is been submitted in preparation of the clarification TC to be held on 10 April 2018.

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

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

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

Section 1 – 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.

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≥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.

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.

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.

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.

Reply: We believe that this question has already been answered for Question 1.1, therefore we kindly refer you to our reply on Question 1.1.

Section 2 – For the simulations of scenarios with 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.

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

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.

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%?

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?

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).

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?

Reply: It is not entirely clear what is meant by the 'true effect' in the 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.

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).

Reply: We split our answer 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?

Question 2.6:

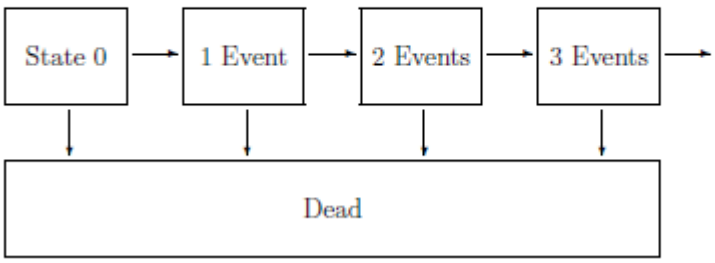
Discuss the utility of multi-stage models to simulate and estimate both, the effect of treatment on mortality and, the effect on HFH. These estimates should be investigated in simulations regarding their statistical properties, interpretability, and yardsticks to their utilization.

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?

References

Andersen et al. (1993). Statistical Models Based on Counting Processes. Springer.

Anker SD and McMurray JJV (2012). Time to move on from 'time-to-first': should all events be included in the analysis of clinical trials? European Heart Journal, 33:2764-2765.

Anker, SD, Schroeder S, Atar D, Bax JJ, Ceconi C, Cowie MR et al. (2016). Traditional and new composite endpoints in heart failure clinical trials: facilitating comprehensive efficacy assessments and improving trial efficiency. European journal of heart failure, 18(5), 482-489.

Buyse M (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Statistics in Medicine, 29:3245–57.

Hernán MA, Robins JM (2018). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book

ICH (2017). Draft ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. ICH.

Lin et al. (2000). Semiparametric regression for the mean and rate functions of recurrent events. J. R. Statist. Soc. B 62, 711-730.

Luo X, Qiu J, Bai S, Tian H (2017). Weighted win loss approach for analyzing prioritized outcomes. Statistics in Medicine, 36:2452-2465.

Mao L, Lin DY (2016). Semiparametric regression for the weighted composite endpoint of recurrent and terminal events. Biostatistics, 17:390-403.

Oakes D (2016). On the win-ratio statistic in clinical trials with multiple types of event. Biometrika, 103:742-745.

Pocock SJ, Ariti CA, Collier TJ, Wang D (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European Heart Journal, 33:176-82.

Rauch G, Jahn-Eimermacher A, Brannath W, Kieser M (2014). Opportunities and challenges of combined effect measures based on prioritized outcomes. Statistics in Medicine, 33:1104-1120.

Rogers JK, Pocock SJ, McMurray JJV, Granger CB, Michelson EL, Östergren J, Pfeffer MA, Solomon S, Swedberg K and Yusuf S (2014). Analysing recurrent hospitalisations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. European Journal Heart Failure, 16:33-40.