EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

9 March 2018
EMA/CHMP/SAWP/179482/2018
Procedure No.: EMEA/H/SAB/090/1/2018
Product Development Scientific Support Department

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

Summary

A request for a qualification opinion entitled “Clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses” has been issued by a number of renowned statisticians. The request is centred on two examples (relapsing remitting multiple sclerosis (rrMS) and chronic heart failure (CHF)) representing situations where recurrent event analyses may offer opportunities to describe a clinically relevant treatment effect. Whereas in the first example recurrent event analyses for relapses have been used for decision making during drug licensing, experiences are still limited with the use of recurrent re-hospitalisations for worsening heart failure in the latter indication that is distinct in that death is still frequent in heart failure studies, and, as a minimum, mortality should not be adversely affected by medical treatment. From a statistical perspective the intercurrent event “death” obviously censors further observation of the recurrent event endpoint under investigation. In consequence, both, statistical challenges regarding methodological aspects and medical interpretation of outcome will have to be addressed in the end.

Scientific discussion

A wealth of information has been provided and a large number of simulations have been done and discussed. Groundwork regarding recurrent event analysis has been prepared in an extensive report that is under review regarding the aforementioned challenges.

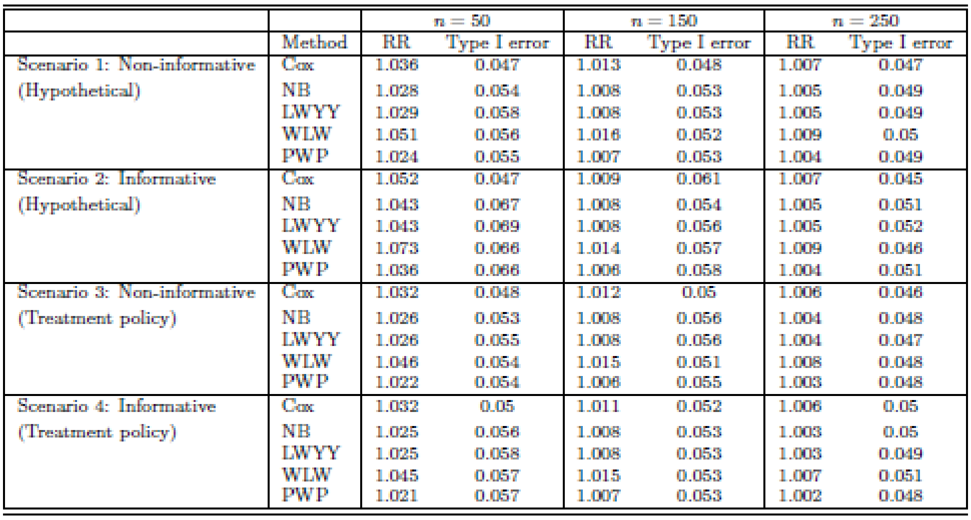
In an initial phase of assessment some open issues have been identified that require clarification and possibly an amendment and even extension of the currently provided simulation exercises.

Regarding scenarios without a terminal event, some questions arise relating to the simulations presented in table 7. Firstly, it is not clear why every estimate of the RR in the table should be over 1.0. A random scattering of values above and below 1.0 would have been expected. One possible reason would be if the averaging across simulation runs had been done on the arithmetic rather than logarithmic scale. If this is not the reason, an explanation for and discussion of the systematic finding would be helpful.

Regarding the apparent loss of type I error control with smaller sample sizes, interpretation of the table would be easier if the simulations were done using 1-sided tests at the 2.5% level, rather than 2-sided 5% tests.

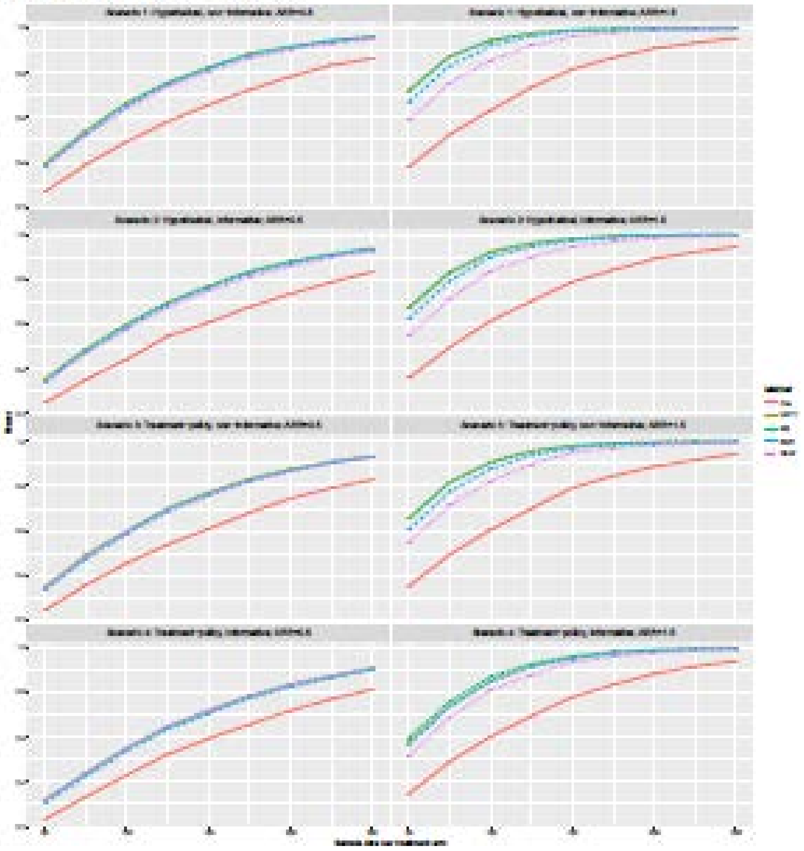
It would also be valuable to include the log-rank test in the table (although no estimate of the RR would be available) as this is often the method used for the initial significance test in time-to-first-event analyses. This would also be valuable in the simulations of power, such as were presented in Figure 7.

Table 7: Settings without terminal event: Mean treatment effect estimates and type I error rate under four scenarios based on 10'000 clinical trial simulations, $RR = 1$, $\theta = 0.25$, $\lambda_0 = 0.5$.



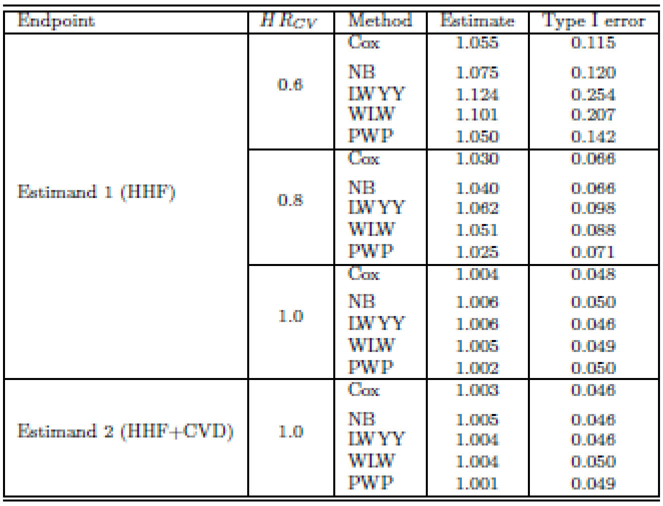
		$n = 50$		$n = 150$		$n = 250$	
	Method	RR	Type I error	RR	Type I error	RR	Type I error
Scenario 1: Non-informative (Hypothetical)	Cox	1.036	0.047	1.013	0.048	1.007	0.047
	NB	1.028	0.054	1.008	0.053	1.005	0.049
	LWYY	1.029	0.058	1.008	0.053	1.005	0.049
	WLW	1.051	0.056	1.016	0.052	1.009	0.05
	PWP	1.024	0.055	1.007	0.053	1.004	0.049
Scenario 2: Informative (Hypothetical)	Cox	1.052	0.047	1.009	0.061	1.007	0.045
	NB	1.043	0.067	1.008	0.054	1.005	0.051
	LWYY	1.043	0.069	1.008	0.056	1.005	0.052
	WLW	1.073	0.066	1.014	0.057	1.009	0.046
	PWP	1.036	0.066	1.006	0.058	1.004	0.051
Scenario 3: Non-informative (Treatment policy)	Cox	1.032	0.048	1.012	0.05	1.006	0.046
	NB	1.026	0.053	1.008	0.056	1.004	0.048
	LWYY	1.026	0.055	1.008	0.056	1.004	0.047
	WLW	1.046	0.054	1.015	0.051	1.008	0.048
	PWP	1.022	0.054	1.006	0.055	1.003	0.048
Scenario 4: Informative (Treatment policy)	Cox	1.032	0.05	1.011	0.052	1.006	0.05
	NB	1.025	0.056	1.008	0.053	1.003	0.05
	LWYY	1.025	0.058	1.008	0.053	1.003	0.049
	WLW	1.045	0.057	1.015	0.053	1.007	0.051
	PWP	1.021	0.057	1.007	0.053	1.002	0.048

Figure 7: Setting without terminal event: Statistical power at varied sample size under four scenarios based on 10'000 clinical trial simulations, $RR = 0.65$, $\theta = 0.25$, $\lambda_0 = 0.5, 1.5$.



For the settings with a terminal event similar issues arise regarding type I error, as shown in table 11, with all estimates above 1.0 (even in the global null situation) and the difficulty of interpreting 2-side 5% tests as compared to 1-sided 2.5% tests. To match with table 7 presentation of results with varying sample size would be useful. Also a row could be provided for HRCV = 1.25 to match other tables. HRCV could also be varied for Estimand 2, despite this meaning that the figures would no longer strictly represent type I error for that estimand.

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, $RR_{HHF} = 1$ 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

In the scenario with a terminal event two estimands were considered. Firstly, the ratio of the number of recurrent events (in this case hospitalisations), and secondly the ratio of events, where the terminal even (death) was also counted as an event.

Both these estimands seem to exhibit concerning properties. Neither truly estimate the effect on the recurrent event independent of the terminal event, or present a coherent combination of the two for an overall evaluation of the two factors together (adding one additional recurrent event to represent a terminal even seems arbitrary).

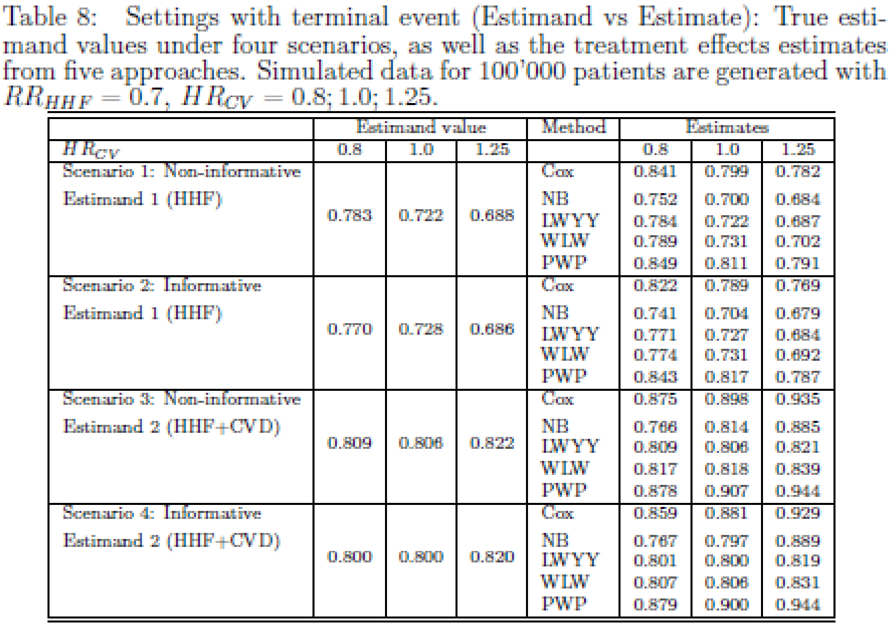


Table 8: Settings with terminal event (Estimand vs Estimate): True estimand values under four scenarios, as well as the treatment effects estimates from five approaches. Simulated data for 100'000 patients are generated with $RR_{HHF} = 0.7$, $HR_{CV} = 0.8; 1.0; 1.25$.

	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 risk ratio for hospitalization is 0.7, but depending on the rates of terminal events the estimand value alters, and with estimand 1 gets more impressive if the treatment has an adverse effect on the terminal events. Similarly, treatments which are reducing the rate of terminal events are penalised. This does not occur with estimand 2 in these examples, but that is partly a function of follow-up time and the rates of each type of event, and it seems likely that it would happen for other durations of study or different choices of event rate parameters. If the intention here is to use the estimation methods to estimate the effect on hospitalisations independent of the effect on the terminal event, then an estimand that gives 0.7 regardless of the terminal even effect would seem to be desirable. If this is not the intention and a combination of terminal and recurrent events is the intention, then a more sophisticated joint modelling approach than just adding in the terminal event as an additional event seems required. Thoughts turn to rank-based approaches where patients are ordered based on their outcome on both variables.

While the methods, particularly LWYY seem to estimate the estimands well, the estimands themselves are currently questioned. Rank based methods such as the win-ratio, while maybe lacking power, at least do not have that property of these estimands, though they do lead to weighting issues regarding the importance of the terminal event.

List of issues to be addressed only in writing

Based on the coordinators' reports the Scientific Advice Working Party (SAWP) determined that the Applicant should discuss the following points, before advice can be provided:

Issues to be addressed in writing by 6 April 2018

Based on the coordinators' reports the Scientific Advice Working Party (SAWP) determined that the Applicant should discuss the following points, before advice or a qualification opinion can be provided:

For the simulations of scenarios with no terminal event:

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.
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.

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.

For the situation where there is a terminal event:

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 HRCV=1.25, add the log-rank test to the table, vary HRCV for estimand 2 and provide results for varying sample size.
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 HRCV in these situations.
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?
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?
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).
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.