EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

18 June 2019
EMA/CHMP/SAWP/291384/2019
Committee for Medicinal Products for Human Use (CHMP)

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

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

Comments should be provided using this template. The completed comments form should be sent to Qualification@ema.europa.eu

Keywords	Recurrent Events, Estimand, Chronic Heart Failure, Mortality

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

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

Based on the coordinators' reports the CHMP gave the following answers:

Question

Does the CHMP agree that the results described in this request support the claim that treatment effect measures can be defined based on recurrent event endpoints that are clinically interpretable and allow for efficient statistical analyses?

CHMP answer

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

Recurrent events refer to the repeated occurrence of the same type of event over time for the same patient. They are related to disease burden and may indicate disease progression in some instances. Recurrent event endpoints are well established in indications where the rate of terminal events (e.g. death) is very low and reduction in mortality is not a primary goal of treatment. Examples include relapses in multiple sclerosis (CHMP, 2015), exacerbations in pulmonary diseases (e.g. chronic obstructive pulmonary disease (CHMP, 2012a) and asthma (CHMP, 2010a)), headache attacks in migraine (CHMP, 2007, 2016a), hypoglycemia episodes in diabetes mellitus (CHMP, 2012b), and seizures in epileptic disorders (CHMP, 2010b, 2016b). In these chronic diseases, time-to-first-event endpoints that focus on the treatment effect on the first event are clinically less meaningful and hence rarely used. Experience with recurrent event endpoints is more limited in indications where the rate of terminal events is high and the clinical meaningfulness is an issue of discussion if the impact of a therapeutic intervention on mortality is of key importance. Chronic heart failure treatment is an indication to exemplify the need for a thorough discussion, both, from a clinical, as well as from a methodological perspective.

1. Clinical background: Recurrent event analyses in chronic heart failure

The Applicant emphasized the example of chronic heart failure (CHF). In the European regulatory framework the primary analysis in pivotal trials in this disease usually is based on a time-to-first-event endpoint, i.e. mortality alone or as a component of a composite endpoint in combination with endpoint(s) related to worsening of heart failure as time to first heart failure hospitalization (HFH). (Guideline on clinical investigation of medicinal products for the treatment of chronic heart failure, CPMP/EWP/235/95, Rev.2, 20, July 2017). Assessment of mortality in confirmatory trials should include both all-cause mortality and cardiovascular mortality. The guideline summarizes on the issue of recurrent HFH as follows: ”reoccurring hospitalisations for heart failure (HFH) are relatively common in patients with CHF and despite their significance they are rarely used as an endpoint in clinical trials compared to time to first HF hospitalisation”. It is further stated that “the main therapeutic goals in the treatment of CHF are to reduce cardiovascular mortality and to prevent deterioration of the clinical status and hospitalisations; these goals should represent the primary aim of new agents developed for the treatment of CHF […] endpoints accounting for recurrent HFH events may under certain conditions better characterise the prognosis of patients with CHF. Recurrent events are also important as they represent a large burden to patients. The inclusion of recurrent events as co-primary endpoint may be considered, but this setting needs further justification, adjudication of the events and a clear methodological strategy”.

In this aspect the ability to appropriately estimate the effect of treatment on recurrent hospitalization is of importance.

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

Mortality

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

Although mortality rates in CHF have decreased over the last decades, all-cause mortality remains high. In the European ESC-HF pilot study, covering a period between October 2009 to May 2010, 12-month all-cause mortality rates for hospitalized (acute heart failure) and stable/ambulatory HF patients were 17% and 7%, respectively, with 12-month hospitalization rates of 44% and 32%, respectively. Similar numbers were observed in the PARADIGM HF trial (Murray et al., N Engl J Med 2014; 371: 993 and EPAR EMEA/H/C/004062/0000, run in 2009 through 2012, stopped 2014) that may serve as an example for mortality rates in present clinical heart failure studies. 17.0% and 19.8% of the patients died in the LCZ696 and the Enalapril group, respectively, after a median follow-up of 27 months. The rate per 100 patient years (95% CI) was: all-cause mortality 7.6 (7.1, 8.2) vs. 9.0 (8.3; 9.7), CV death: 6.0 (5.5; 6.5) vs. 7.5 (7.0; 8.1) and first HFH 6.2 (5.7; 6.7) vs. 7.8 (7.2; 8.4), respectively. The significant treatment effect was observed for CV death and all cause-mortality, first HFH and for the primary endpoint, the composite of CV mortality and first HFH. The statistically significant result was to a large degree based on efficacy in earlier stages of the disease (NYHA I – II). The study is an example for a reasonably sized study (8442 patients) able to provide the data needed for assessment of effects on mortality and hospitalization for patients as included in this study.

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

In general, a medicinal product can be approved based on a beneficial effect on hospitalization rates, even if studies fail to show a mortality benefit. As a prerequisite the data have to provide sufficient reassurance that mortality is not increased to a relevant degree in the overall population and in subgroups. The key example is digoxin. In a placebo controlled study including 6800 patients digoxin

had no effect on all-cause mortality (RR 0.99; 95 % CI 0.91 to 1.07, The Digitalis Investigation Group (DIG), N Engl J Med 1997; 336: 525-533), but significantly improved first HFH rate (26.8 % vs. 34.7 %; RR 0.72; 95 % CI 0.66 to 0.79; P<0.001). The trial was large enough to exclude an increase in all-cause mortality by more than 7% which may be sufficient for a well-established drug. However, careful analysis of the mortality is crucial in such a case since an overall neutral effect on mortality despite of a HFH benefit may well be the result of divergent effects on mortality in subgroups. This has been discussed for the DIG trial. In a post-hoc subgroup analysis in male patients all-cause mortality was decreased at lower digoxin levels, neutral at intermediate digoxin levels and increased in patients with higher digoxin plasma levels. Similarly, in the Val-HEFT study, comparing valsartan with placebo, a beneficial effect was observed on first event HFH (RR 0.87; 97.5 % CI, 0.77 to 0.97; p=0.009) whereas the effect on all-cause mortality was neutral (deaths during the entire trial: RR 1.02 (0.88 – 1.18)). In Val-HEFT, the neutral effect on mortality was the net result of a significantly increased mortality in patients receiving in addition ACE inhibitors and beta blockers, and a significantly decreased mortality in the other patients.

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

Recurrent HFH events

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

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

The study showed some peculiarities when assessing recurrent HFH events. Hospitalization rates were not linearly distributed over time, they clustered at early post-discharge and pre-fatal time. The clinical meaningfulness of recurrent pre-fatal HFH events beyond a statistical booster of mortality remains to be clarified. Furthermore, HFH rate depended on the underlying disease. In ischemic heart failure, where the hospitalization rate was higher, a clear differentiation between heart failure related and ischemia related hospitalization may not be feasible in every case. Recurrent event analyses are currently not accepted in the regulatory context in cardiovascular trials aiming at the prevention of MACE related to ischemic diseases.

Whereas it has been considered that recurrent HFH events may better characterize the prognosis of patients under certain conditions (CPMP/EWP/235/95, Rev.2, 20, July 2017) it cannot be assumed a priori for a new therapeutic agent that HFH is predictive for mortality. HFH or signs and symptoms of heart failure did not exactly mirror the effect of a treatment on mortality in the above mentioned two studies with milrinone and xamoterol. Also the DIG study is an example of discrepant results for both parameters. Furthermore, models predicting mortality in patients with heart failure were reported to have a higher discriminative ability than those designed to predict hospitalization (Rahimi K et al., JACC heart failure 2014; 2: 440 ff; Outwerkerk W JACC heart failure 2014; 2; 429). Among the possible reasons is that hospitalization is more dependent on health care supply indicating that HFH and mortality are not interchangeable parameters for outcome.

In summary, the main therapeutic goals in the treatment of CHF are to reduce cardiovascular mortality and to prevent deterioration of the clinical status and hospitalizations; these goals should represent the primary aim of new agents developed for the treatment of CHF. Recurrent events may represent a large burden to patients and endpoints accounting for recurrent HFH events may under certain conditions better characterise the prognosis of patients with CHF (c.f. CPMP/EWP/235/95, Rev.2, 20, July 2017). Among the challenges when clinically interpreting recurrent event HFH are disease specific differences, clustering of events and factors like health care supply that may have an impact on the event rate. Studies may become smaller when sample sizes are calculated based on recurrent HFH. This has a relevant impact on data available for mortality assessment. Moreover, using a composite of first event HFHs and mortality promotes inclusion of patients at a relevant risk of dying in order to get a sufficient number of endpoint events whereas planning a study based on recurrent HFH as a component of a primary endpoint may stipulate inclusion of patients at lower risk witch may further decrease the robustness of information on mortality. The impact of a new therapeutic agent on mortality, either as a measure of efficacy or at least in order to provide robust reassurance that there is no detrimental effect, is key information expected from a pivotal trial in chronic heart failure. Such data is needed not only for the overall population but also for relevant subgroups. Examples exist, where it was possible to achieve this information with a reasonably sized clinical program based on the requirements as outlined in CPMP/EWP/235/95, Rev.2. Considering requirements to rule out an excess of mortality, the number of patients needed in a study using recurrent HFH events as a component of a primary endpoint may in the end not be lower than in a study designed according to the current guideline.

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

2. Methodological issues

2.1 Calculation of HHF rate for a treatment-group:

Before going into an in-depth discussion on estimands and corresponding estimates a simplified example is presented to illustrate and discuss two different effect measures: The exposure-weighted and the patient-weighted event rate.

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

Exposure-weighted rate

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

It can be thought of as the total number of events observed in that group divided by the total follow-up time. In the example this gives 4/10.5, i.e. 0.38 events per year.

It could also be thought of as the average number of HHF events per patient, divided by the average follow-up – so in the example 1/2.625, or 0.38 events per year.

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

Patient-weighted rate

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

Comparison

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

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

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

equal weight, while the exposure rated approach would generally give less weight to patients with higher HHF rates.

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

2.2. Calculation of the treatment effect on HHF rate

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

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

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

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

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

Example:

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

Treatment

Patient	HHF	Follow-up (years)	HHF per year
Ann	0	3.0	0
Bill	1	3.0	0.33
Caren	3	3.0	1
Dave	0	3.0	0

Total	4	12	1.33
Average per patient	1	3.0	

Control

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

Annualised HHF rates:

Exposure weighted: Treatment 0.333 per year, Control 0.476 per year; ratio 0.7

Patient weighted: Treatment 0.333 per year, Control 0.667 per year; ratio 0.5

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

Intercurrent events, particularly if terminal / absorbing or impacting differentially (i.e. to a different degree on treated and control patients) on duration of observation by other mechanisms, cause obvious problems with the independent interpretation of treatment effect estimates for differences in recurrent events.

2.3. Applicant proposal

The Applicant's proposals are based on the exposure-weighted rate approach. The reason for this preference is related to the drawbacks of the patient-weighted approach of the high influence of patients who die early leading to high variability and a skewed distribution of results. In addition, they state that none of the established estimators and statistical tests for recurrent events data in the literature target the patient-weighted estimate. However, high variability per se indicates lower confidence for decision making and may be an argument on its own that simply more information is needed to provide robust conclusions (i.e. regarding relevant subgroups of different risks and secondary endpoints).

Four different methods for recurrent event analysis were looked at and compared with Cox regression – which looks at time to first event. NB refers to negative binomial regression, which targets an estimand based on the number of recurrent events. When there is complete follow-up NB provides an

estimate of the RR which is the ratio of the average event numbers in the two groups. LWYY is the Anderson-Gill method, which gives the same point estimate as negative binomial regression. The other two methods, Wei (WLW) and Prentice (PWP) do not have such a clear interpretation. None of these directly offers an opportunity to model terminal intercurrent events.

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

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

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

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

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

Table 7A: Mean treatment effects estimates (geometric mean) and Type I error rates (1-sided tests, nominal significance level $\alpha=0.025$) under four scenarios, with treatment effect size 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

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

Table 5: Settings without terminal event (Estimand vs Estimate): Numerical values of hypothetical estimand and treatment policy estimand under four scenarios. The ratio of the target of estimation (Estimate) for each of the five analysis methods over the corresponding estimand value (Estimand) is also shown. 'Estimand' values are calculated analytically, 'Estimate' values are calculated based on a simulated data set with 100'000 patients with $RR = 0.65$, $\theta = 0.25$, and $\lambda_0 = 0.5, 1.5$. Estimate/Estimand values larger (smaller) than 1 correspond to overestimation (underestimation).

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

Table 6: Settings without terminal event: Mean treatment effect estimates under four scenarios based on 10'000 clinical trial simulations, $RR = 0.65$, $\theta = 0.25$, $\lambda_0 = 0.5, 1.5$.

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

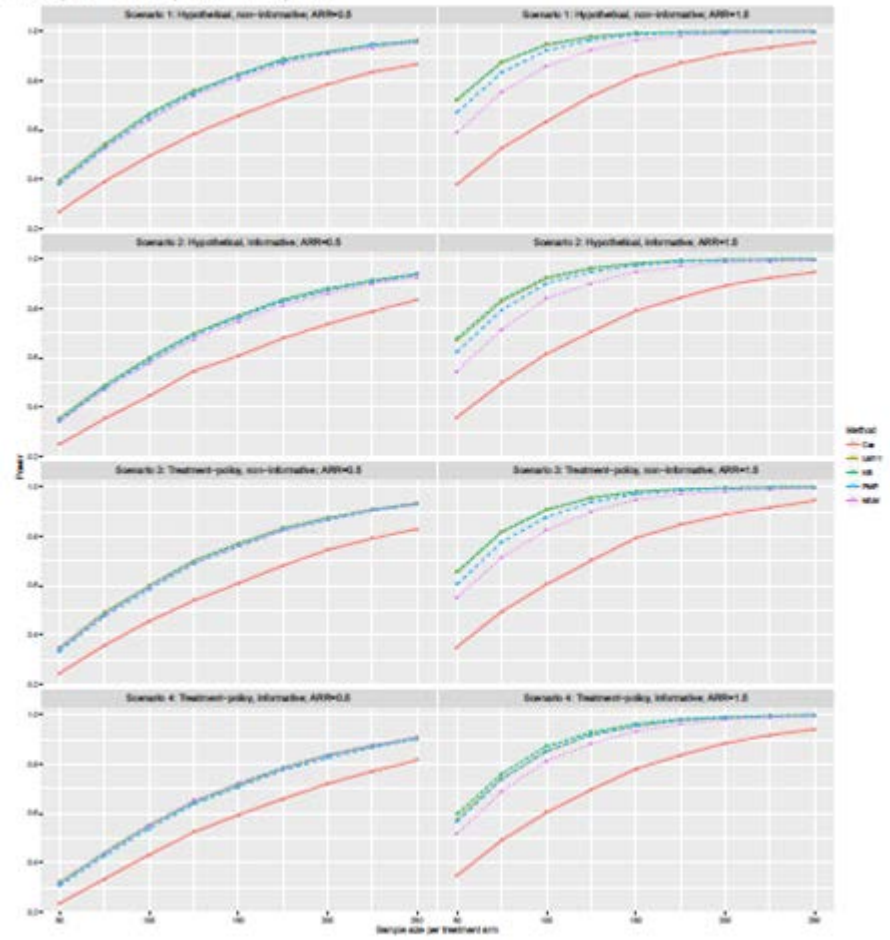
Informative discontinuation means that an effective treatment would keep the patients with a higher event rate on treatment longer allowing them to contribute more events, which explains the conservative estimation in scenario 2. Otherwise there is no suggestion of bias for NB or LWYY. WLW seems to be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

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

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

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

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



Overall, aside from an issue with type I error control for small sample sizes, which should be investigated further, it can be agreed that methods such as negative binomial regression are more efficient than time to first events approaches in a situation where the rate of terminal events is negligibly low. The provided simulations demonstrate increased power, and the estimates of the RR reflect the true treatment effect, except for being conservative for the effect in scenario 2 where the rate of withdrawal from treatment is positively correlated with the rate of recurrent events. Obviously, this correlation may have a different impact on the control of the type-1-error, if non-inferiority (or equivalence) is supposed to be demonstrated.

2.3.2. Scenarios with a terminal event

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

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

Table 11: Settings with terminal event: Mean treatment effect estimates and type I error rates for Estimands 1 and 2 with non-informative treatment discontinuation based on 10'000 clinical trial simulations, $RR_{HHF} = 1$ 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

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

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

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

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

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

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

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

Equal weighted (per patient) estimand

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

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

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

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

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

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

However, there are clear limitations with the patient-weighted approach. The Applicant notes that none of the investigated analysis methods targets the estimand. They also express concern over the likely increased variability of such an estimate, which would necessitate large sample sizes, and potentially lose the efficiency hoped to be gained by using a recurrent events analysis, and its skewed distribution, these issues mainly caused by the weight given to patients who have short follow-up. CHMP considers that this is evidence of population heterogeneity which needs to be understood for decision making about efficacy of the drug under consideration. Patients with short follow-up likely are informative regarding the terminal event and should not be down-weighted with the aim to reduce variability.

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

3. Conclusion- qualification opinion statement

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

Clinical considerations regarding meaningfulness and the loss of information on mortality if studies become smaller when designed based on recurrent events are summarized in section 1 of this document. Methodological considerations in the scenario where there are terminal events are summarized here: the targeted effect on the recurrent event in the exposure-weighted approach alters dependent on the effect on the terminal event, meaning the effects are not clinically interpretable in the way CHMP would ideally require for an individual patient. The effect also alters with other design properties such as the duration of follow-up. There is also a loss of type I error for the individual assessment of the treatment effect on the recurrent event in situations, where the global null-hypothesis is not true and the treatment effect regarding mortality is not neutral.

The CHMP could envisage as an option to provide a basis for decision making an analysis which delivers separate estimates which appropriately summarise the expected effect of the treatment on the annual event rate for the recurrent event while alive, and the effect on the terminal event. These estimates should be unbiased from a statistical perspective.

Use of an approach for the recurrent event analysis where patients are given equal weight in the analysis regardless of the duration of follow-up may have the potential to achieve this objective. There are limitations with this approach, in that it would likely lead to high variability which could reduce the efficiency advantages the use of recurrent event approaches hopes to obtain, but, as elaborated above, this may simply indicate that more information is needed for proper decision making. There are also currently no established methods in the literature which target this estimand. However, based on the information provided this seems to be a possibly fruitful path to investigate and the CHMP would encourage research into devising efficient methods of estimation that target such an estimand.