

Treatment effect measures when using recurrent event endpoints

Discussion Meeting for Qualification Opinion

10th July, 2018

European Medicines Agency - London, UK

Attendees

Face-to-face:

- Mouna Akacha
- Frank Bretz
- Denis Burkhalter
- Arno Fritsch
- Mireille Muller
- Tobias Mütze
- James Roger
- Patrick Schlömer

By telephone:

- Franco Mendolia
- Heinz Schmidli
- Jiawei Wei

Outline

1. Introduction to qualification opinion request
2. Responses to the list of issues



Introduction to qualification opinion request

Qualification opinion request

We seek an 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.

Motivation

- Recurrent event endpoints are **well suited to characterize disease burden or progression**
 - Established in indications where the rate of terminal events is very low (e.g. multiple sclerosis, asthma, migraine, epilepsy)
 - Less established in indications where the rate of terminal events is high (e.g. chronic heart failure)
- **Key question: how to measure the treatment effect** – especially in the setting where patients may die for disease-related reasons?
 - Death introduces conceptual challenges: patients that die can no longer experience morbidity events while patients in less serious conditions remain on the trial and experience many events
- Depending on the specific clinical trial setting **different treatment effect measures (estimands) can be considered**

Claim and its substantiation

“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.”

- Illustration of use of recurrent event endpoints in clinical practice
 - Discussion for settings with and without terminal events
 - Motivating examples: relapsing-remitting multiple sclerosis (RRMS) and chronic heart failure (CHF)
- Discussion of various estimands based on recurrent event and time-to-first-event endpoints
- Discussion of properties of associated estimators
- Comparison of efficiency for methods targeting recurrent event and time-to-first-event estimands

In-scope & out-of-scope

In-scope

- Confirmatory trials
- Recurrent events that are related to efficacy
- Recurrent events that are clinically meaningful (e.g. characterize disease burden or progression)
- Treatments which impact first and subsequent events

Out-of-scope

- Recurrent events that are related to safety
- Ranking and utility approaches
- Recommendation of specific estimands
- Efficiency comparison for estimands taking into account duration or severity of events

Responses to the list of issues

Suggested order for pending issues

- Question 2.3*
- Additional question 2**
- Question 2.4*
- Additional question 3**
- Question 2.5*
- Question 2.2*
- Question 1.2*
- Additional question 1**

* Received March 9th, 2018

** Received June 14th, 2018

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?

- **Estimand 1** (HHF) compares the hospitalization for heart failure (HHF) rate while alive for test and control treatment (rate ratio) – *effect on hospitalisations*
- **Estimand 2** (HHF+CVD) compares the “bad event” rate while alive for test and control treatment (rate ratio) – *overall effect combining both hospitalisations and cardiovascular death (CVD)*
- Estimand 1 and Estimand 2
 - Appear to be understandable and meaningful for patients and clinicians
 - Have a causal interpretation
 - Are estimable with minimal assumptions

Question 2.3:

continued

Estimand 1

- Estimand 1 (HHF) favors a treatment with a worse effect on CVD
 - That is, a worse effect on CVD (larger HR_{CVD}) leads to better values for the HHF estimand (lower values) and power (higher)
- Such an effect was seen for Estimand 1 in the simulation studies
 - Severely ill patients have many hospitalizations and die earlier than patients in less serious conditions
 - Delaying time to death thus results in observing more hospitalizations than in the case of a neutral effect on CVD
- Estimand 1 appears suitable only in settings where test and control treatment are very unlikely to differ with respect to their effect on CVD

Question 2.3:

continued

Estimand 2

- Estimand 2 (HHF+CVD) does not favor a treatment with worse effect on CVD across a range of realistic scenarios (see original request document* + response to Question 2.2)
- 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
- Estimand 2 provides an overall treatment effect, and could be used in settings where treatment effects on both HHF and CVD are plausible

Question 2.3: ... the question remains open, how a strategy for decision making should be constructed that optimizes positive conclusions regarding a treatment effect at least excluding a detrimental effect on mortality.

- Agree that the scenario of a detrimental treatment effect on mortality is an important scenario to consider
- Considerations to address the effect on mortality are not fundamentally different to those applied for traditional time-to-first event studies
 - Effect on mortality would be evaluated in a component analysis and is not directly related to the choice of the primary efficacy analysis (e.g. time-to-first-event or recurrent event)

Additional question 2: ValHeft is not considered a useful example to discuss the application of recurrent events of worsening of heart failure for decision making. The key result in ValHeft was an increased mortality in patients on background ACE-inhibitor and beta blocker therapy (n = 1610), which was considered a robust result, and a decreased mortality in the other patients (n = 3400). Overall, this led to an apparent neutral effect in mortality in the study. The applicant is asked to comment on how such different results in subgroups in mortality can be detected if studies are designed based mainly on recurrent hospitalisation events and how such heterogeneity is accounted for in the modelling approaches.

- Considerations regarding subgroup effects on mortality are unrelated to the chosen primary estimand, i.e. are essentially the same for a time-to-first-event estimand and a recurrent event estimand
- Ability to investigate subgroup effects on mortality is expected to be comparable in studies based on recurrent HHF estimands as these are not expected to be of considerably smaller size than those based on time-to-first-event estimands (e.g. PARAGON-HF n = 4822).

Additional question 2:

continued

- Due to multiplicity issues and small sample sizes, findings from subgroup analyses are often not reproducible
 - Various modeling approaches to limit the risk of spurious findings have been proposed (e.g. Bornkamp et al., 2017)
- The robustness of the subgroup results in ValHeft is controversial, e.g. was not reproduced in CHARM-Added (White, 2003), see back up slides

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?

- In two unrealistic scenarios Estimand 1 fulfills the criteria for such an alternative estimand. These scenarios are
 1. No treatment effect on CVD and no treatment discontinuation
 2. HHF and CVD are independent and no treatment discontinuation
- If these two scenarios hold then estimation is simple, e.g. LWYY
- If these two scenarios don't hold then an 'independent evaluation of the true effect on the recurrent event independent of the terminal event' is not possible
- In the following, two more realistic scenarios are discussed (for simplicity without treatment discontinuation):
 - HHF and CVD are correlated and test treatment has a **positive** or **negative** effect on CVD, respectively

Question 2.4:

continued

HHF and CVD are correlated and

- test treatment has a **positive** effect on CVD (e.g. **HR=0.8**):
 - As time progresses **less** patients in the test group die early therefore **more** hospitalizations are observed
 - ⇒ Observed treatment effect on HHF at e.g. year 5 is **smaller** than what would be observed in the absence of a treatment effect on CVD which is reflected in Estimand 1 values **> 0.7**
- test treatment has a **negative** effect on CVD (e.g. **HR=1.25**):
 - As time progresses **more** patients in the test group die early therefore **less** hospitalizations are observed
 - ⇒ Observed treatment effect on HHF at e.g. year 5 is **larger** than what would be observed in the absence of a treatment effect on CVD which is reflected in Estimand 1 values **< 0.7**

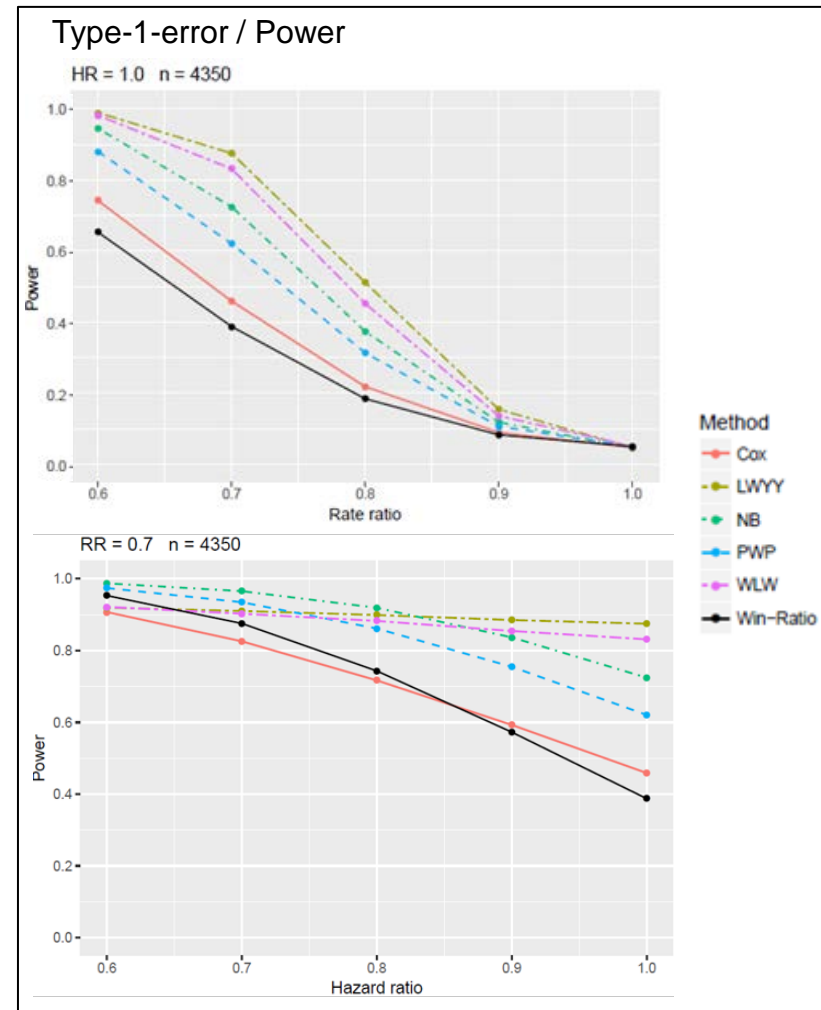
Additional question 3: Please discuss examples of clinical trials, where an analysis of rates of rehospitalisation for worsening heart-failure was helpful for decision making about the efficacy of a drug, or where results on HFH and mortality led to different conclusions. Please discuss this also in the context of an overall assessment of benefit and risks.

- CHAMPION trial
 - Significant result for 6-months rate of HHF as primary endpoint
 - CardioMEMS™ was approved by FDA
 - Effect on time-to-first composite of smaller magnitude and only borderline significant
- CHARM-Preserved trial
 - Borderline result for time-to-first composite event analysis
 - Post-hoc analyses with recurrent HHF showed statistically persuasive evidence of efficacy (p-values $\ll 0.05$)
 - These results suggest a more positive benefit-risk assessment
- In all CHF trials that we are aware of, observed \widehat{HR}_{HHF} for time-to-first HHF were either similar to \widehat{HR}_{CV} or \widehat{HR}_{CV} were closer to 1 than \widehat{HR}_{HHF} , see back up slides

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).

Win ratio (WR) approach:

- Unclear how to describe the WR using the ICH E9(R1) framework
- Depends on censoring distribution, and lacks transitivity
- Controls type I error at nominal level
- Has generally a lower power* than recurrent event methods
- Has higher power than time-to-first event approach when effect on HHF is small and effect on CVD is large



*More scenarios are shown in the back up slides

Question 2.5:

continued

Weighted composites, e.g. Estimand 2:

Defining weights in a manner that is scientifically justified and agreed upon within the clinical community is challenging:

- 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.”

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.

Motivated by the discussion at the telephone conference on April 10th we focus on: increased CVD (20%, 40%) and unchanged HHF rates

Type-1-error

- **Estimand 1:** further increased for $HR_{CV} = 1.25$ compared to setting with lower mortality rates
- **Estimand 2:** controlled at nominal level

Endpoint	HR_{CV}	Method	Mort.Rate = 20 %		Mort.Rate = 40 %	
			Estimate	Type I error	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.094	0.001	1.280	0.000
		NB	1.120	0.002	1.298	0.000
		LWYY	1.191	0.000	1.374	0.000
		WLW	1.168	0.000	1.380	0.000
		PWP	1.089	0.000	1.273	0.000
	0.8	Cox	1.045	0.006	1.124	0.001
		NB	1.057	0.007	1.130	0.004
		LWYY	1.087	0.002	1.157	0.001
		WLW	1.078	0.003	1.159	0.001
		PWP	1.043	0.004	1.121	0.001
	1.0	Cox	1.000	0.025	1.003	0.026
		NB	1.000	0.025	1.005	0.028
		LWYY	1.000	0.023	1.003	0.025
		WLW	1.000	0.025	1.003	0.025
		PWP	1.000	0.025	1.003	0.025
	1.25	Cox	0.951	0.090	0.886	0.170
		NB	0.941	0.087	0.884	0.111
		LWYY	0.913	0.162	0.863	0.211
		WLW	0.920	0.151	0.862	0.212
		PWP	0.953	0.115	0.887	0.187
Estimand 2 (HHF+CVD)	1.0	Cox	1.001	0.026	1.001	0.025
		NB	1.001	0.026	1.002	0.024
		LWYY	1.001	0.024	1.002	0.022
		WLW	1.000	0.025	1.002	0.023
		PWP	1.001	0.025	1.001	0.024

Mean treatment effect estimates and one-sided type-1-error rates, $RR_{HHF}=1$, independent treatment discontinuation, $n=4350$

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.

Power considerations*

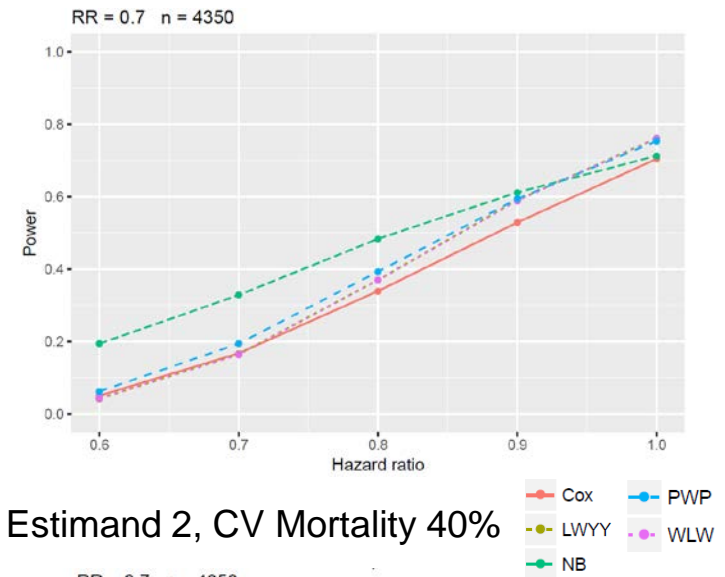
Estimand 1

- Power decrease, especially for $HR_{CV} \ll 1$
- Favors treatment with worse CVD effect

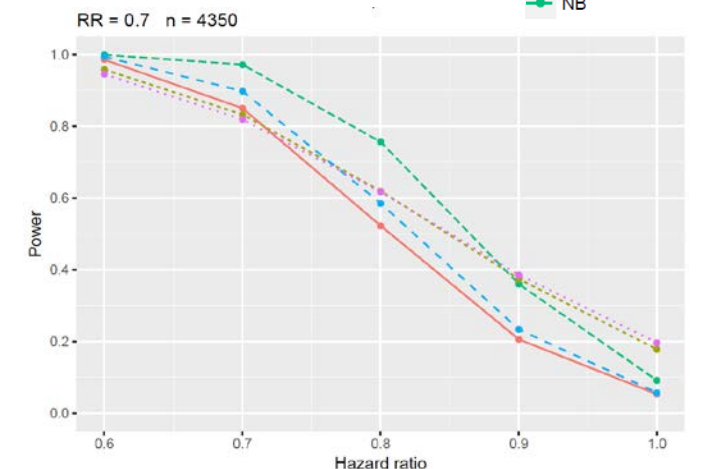
Estimand 2

- Power increase for small HR_{CV} , power decrease for HR_{CV} close to 1
- Favors treatment with better CVD effect
- Less power difference between Cox and recurrent event methods as compared to the case with lower mortality rate

Estimand 1, CV Mortality 40%



Estimand 2, CV Mortality 40%



Question 1.2: For the smallest investigated sample-size all estimates are still less than one. Please comment whether lacking asymptotic normality can be excluded as a reason and bias is truly absent (e.g. by providing results for an even larger sample-size n).

For small sample sizes, a small bias could exist due to lack of asymptotic properties

- Considering the simulation variability (Monte Carlo standard error of approx. 0.003) this bias appears to be negligible
- Further simulations for $n=50$ show that the RR is scattered above and below 1, see back up slides
- Results for larger sample sizes are also scattered above and below 1

	Method	$n = 150$		$n = 200$		$n = 250$	
		RR	Type I error	RR	Type I error	RR	Type I error
Scenario 1: Non-informative (Hypothetical)	Cox	1.001	0.025	0.999	0.025	1.001	0.023
	NB	1	0.026	1	0.026	1.001	0.026
	LWYY	1.001	0.026	1	0.026	1.001	0.025
	WLW	1.001	0.025	1	0.026	1.002	0.025
	PWP	1.001	0.025	1	0.025	1.001	0.026
Scenario 2: Informative (Hypothetical)	Cox	1.001	0.025	0.999	0.025	1	0.024
	NB	1.001	0.026	0.998	0.026	1	0.026
	LWYY	1.002	0.026	0.998	0.026	1	0.025
	WLW	1.002	0.024	0.998	0.025	1	0.025
	PWP	1.001	0.025	0.998	0.025	1	0.025
Scenario 3: Non-informative (Treatment-policy)	Cox	1.001	0.026	0.999	0.024	1.001	0.023
	NB	1	0.025	0.999	0.026	1.001	0.025
	LWYY	1	0.026	0.999	0.026	1.001	0.025
	WLW	1	0.026	0.999	0.026	1.001	0.026
	PWP	1	0.025	1	0.026	1.001	0.024
Scenario 4: Informative (Treatment-policy)	Cox	0.999	0.026	1	0.024	0.999	0.024
	NB	1.001	0.025	1	0.026	1.001	0.025
	LWYY	1.001	0.025	1	0.025	1.001	0.025
	WLW	1.001	0.024	1	0.026	1.001	0.025
	PWP	1.001	0.026	1	0.025	1.001	0.024

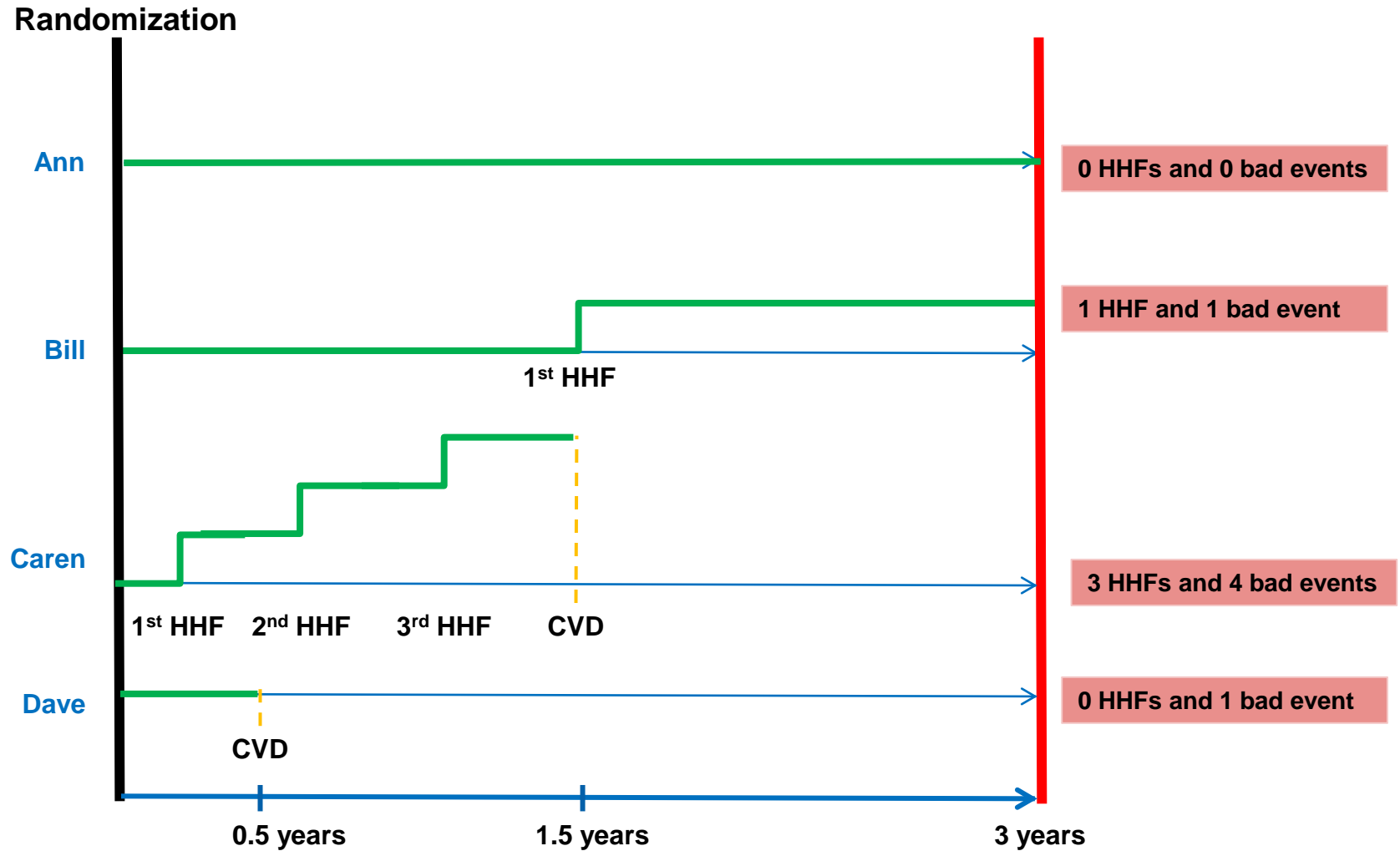
Additional question 1: The use of the frailty model requires further justification because preference would always be given to not add unstructured variability to the model: **a.** Is it impossible to explain the high variability in the frequency of rehospitalisation by means of co-variates? **b.** If there have been attempts to explain this high variability, which models have been investigated? **c.** Please discuss examples, where modelling of the high variability in rehospitalisation-rates has been attempted and in how far this has been successful / not successful.

- Clarity is needed on the meaning of ‘unstructured variability’
 - For example, the NB model and an MMRM* model with compound symmetry covariance structure are also frailty models
- In joint frailty models, the frailty term is modelling i) variability in HHFs, ii) variability in survival times and iii) the association between the HHFs and survival times
 - a) Covariates can help reducing the variability in the HHF part, however, it is unrealistic that all covariates that induce heterogeneity are assessed as baseline variables
 - b) We are not aware of any such investigation
 - c) See b) and clarity is needed on the meaning of ‘successful’

Back-up

Question 2.3

Number of HHFs and number of 'bad' events



Estimand 1 and Estimand 2

CHF study with 3 years of follow-up, test treatment

Patient	#HHF	CVD	#bad events	Time of death/study end (years)
Ann	0	0	0	3.0
Bill	1	0	1	3.0
Caren	3	1	4	1.5
Dave	0	1	1	0.5
...				
AVERAGE	1.0		1.5	2.0

Test treatment: $1.0/2.0 = 0.50$ HHF rate while alive
Control treatment: 1.50 HHF rate while alive
Estimand 1 (HHF): $0.50/1.50 = 0.33$ rate ratio

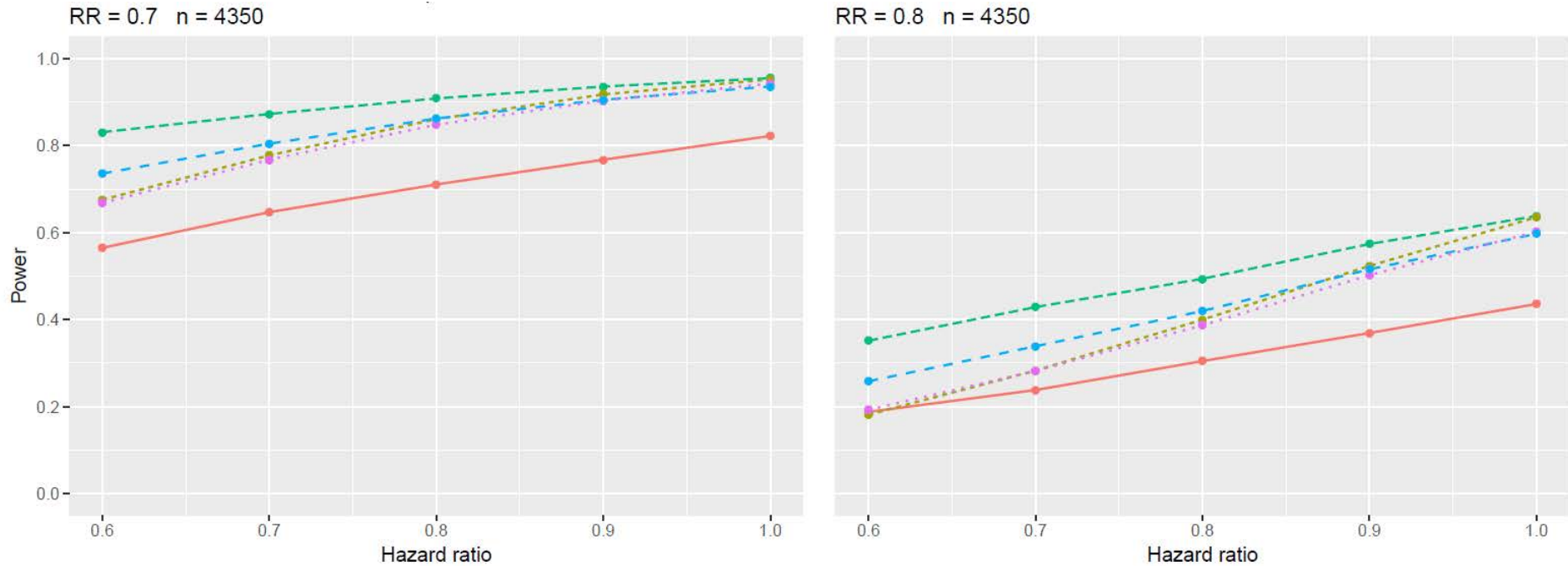
Estimand 1 and Estimand 2

CHF study with 3 years of follow-up, test treatment

Patient	#HHF	CVD	#bad events	Time of death/study end (years)
Ann	0	0	0	3.0
Bill	1	0	1	3.0
Caren	3	1	4	1.5
Dave	0	1	1	0.5
...				
AVERAGE	1.0		1.5	2.0

Test treatment: $1.5/2.0 = 0.75$ bad event rate while alive
Control treatment: 2.00 bad event rate while alive
Estimand 2 (bad events): $0.75/2.00 = 0.38$ rate ratio

Power for Estimand 1 (HHF)

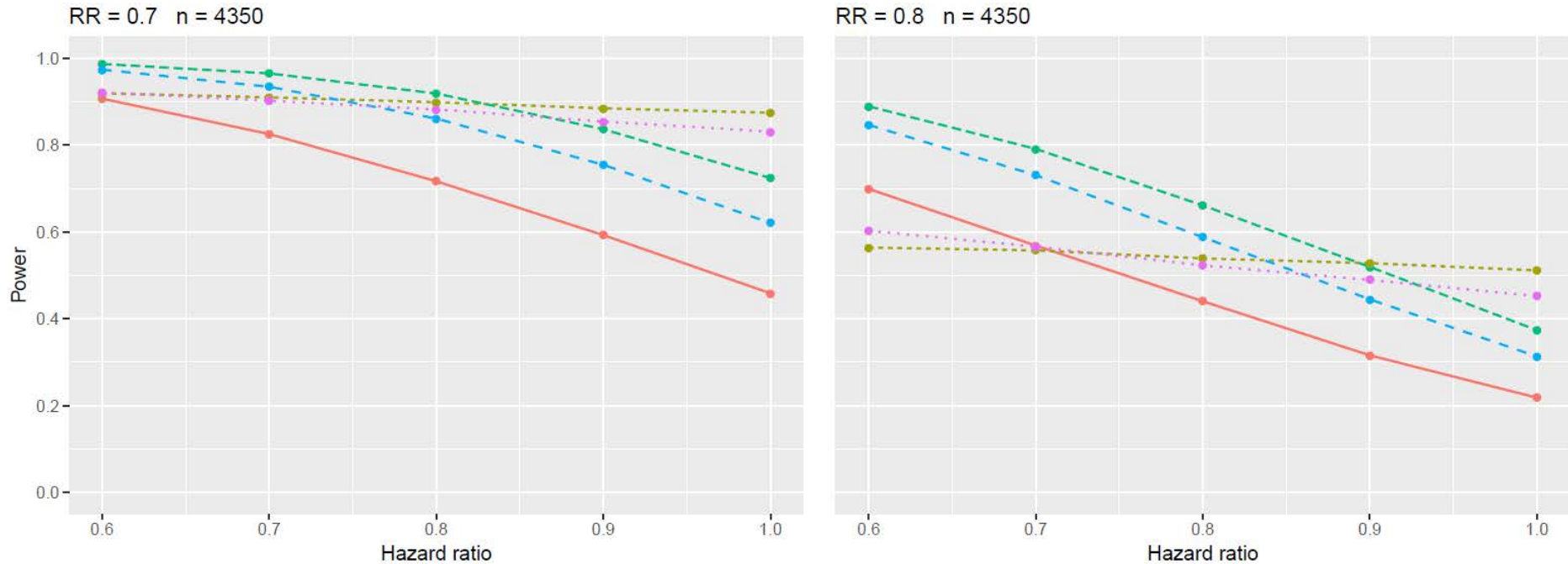


Estimand 1 (HHF) favors a treatment with worse effect on CVD, resulting in higher power as HR_{CV} is increasing

Method

- Cox
- LWYY
- NB
- PWP
- WLW

Power for Estimand 2 (HHF+ CVD)

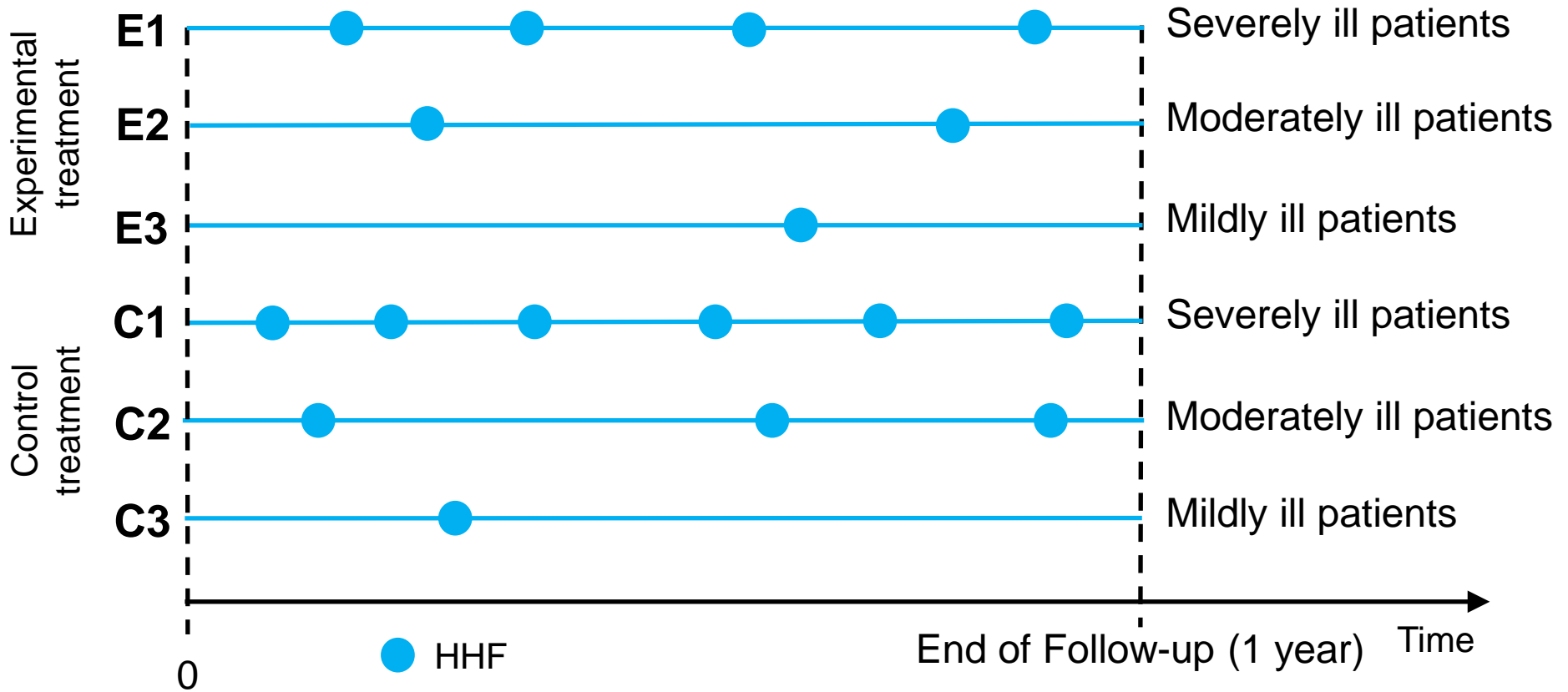


In realistic scenarios, Estimand 2 (HHF+CVD) favors a treatment with positive effect on CVD, resulting in higher power than Estimand 1 (HHF) for small HR_{CV}

Method

- Cox
- LWYY
- NB
- PWP
- WLW

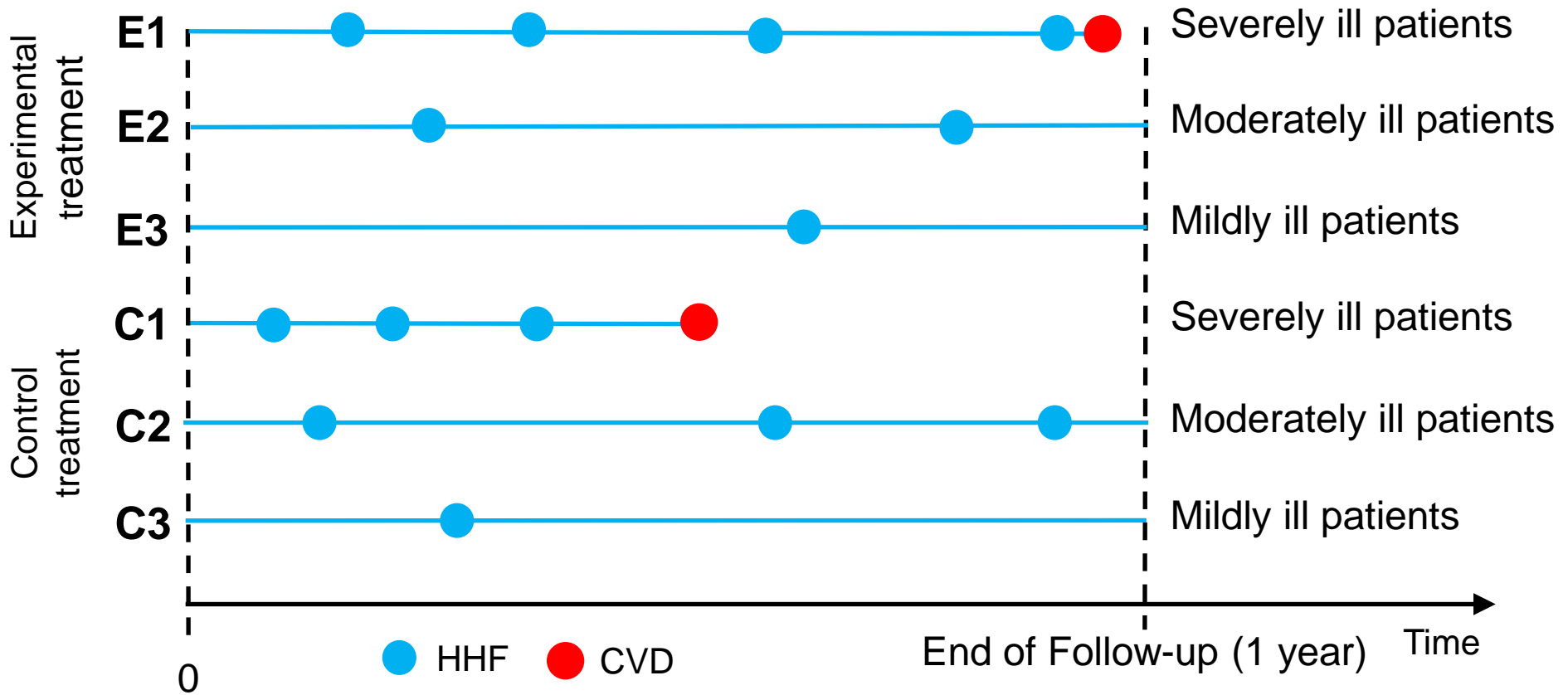
Illustration that Estimand 1 favors a treatment with worse CVD effect - Slide 1 of 2



Without CVD

- Severely ill patients have more HHF
- Experimental treatment delays HHF compared to control

Illustration that Estimand 1 favors a treatment with worse CVD effect - Slide 2 of 2



With CVD

- CVD is more likely to occur in severely ill patients
- Experimental treatment delays CVD compared to control
- Early CVD decreases the number of observed HHF in the control group and makes it look better than in the situation without CVD

Weighting of bad events

- 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
- It may seem that Estimand 2 puts less weight on CVD than a time-to-first-composite-event analysis, since a patient might have several HHFs but can have only one CVD - however:
 - If a patient has a HHF at 9mo, and a CVD at 15mo, Estimand 2 gives equal weight to both events, while the time-to-first-composite-event estimand gives zero weight to the CVD
 - Anker and McMurray (2012): About 30% of events are CVD, regardless of whether time-to-first-composite-event or recurrent-event estimand used
- Other weightings of bad events putting more weight on CVD could also be considered*

Additional question 2

ValHeft subgroup effects on mortality not reproduced

White HD (2003): “The CHARM trials are yet another instance in which prospective testing of a previous hypothesis-generating subgroup finding has shown that the subgroup finding in the previous trials probably occurred by chance.”

- ValHeft results mentioned in the question:
 - 42% ‘increased mortality’ of ACEIs+ β -blockers+valsartan vs ACEIs+ β -blockers+placebo
 - 33% ‘reduced mortality’ in patients not given ACEIs or β -blockers
- CHARM-Added
 - Similar mortality in patients given ACEIs+ β -blockers as in patients not given β -blockers
 - Similar effect of candesartan on CVD and HHF in patients given guideline-recommended doses of ACEIs as in patients not given guideline-recommended doses of ACEIs

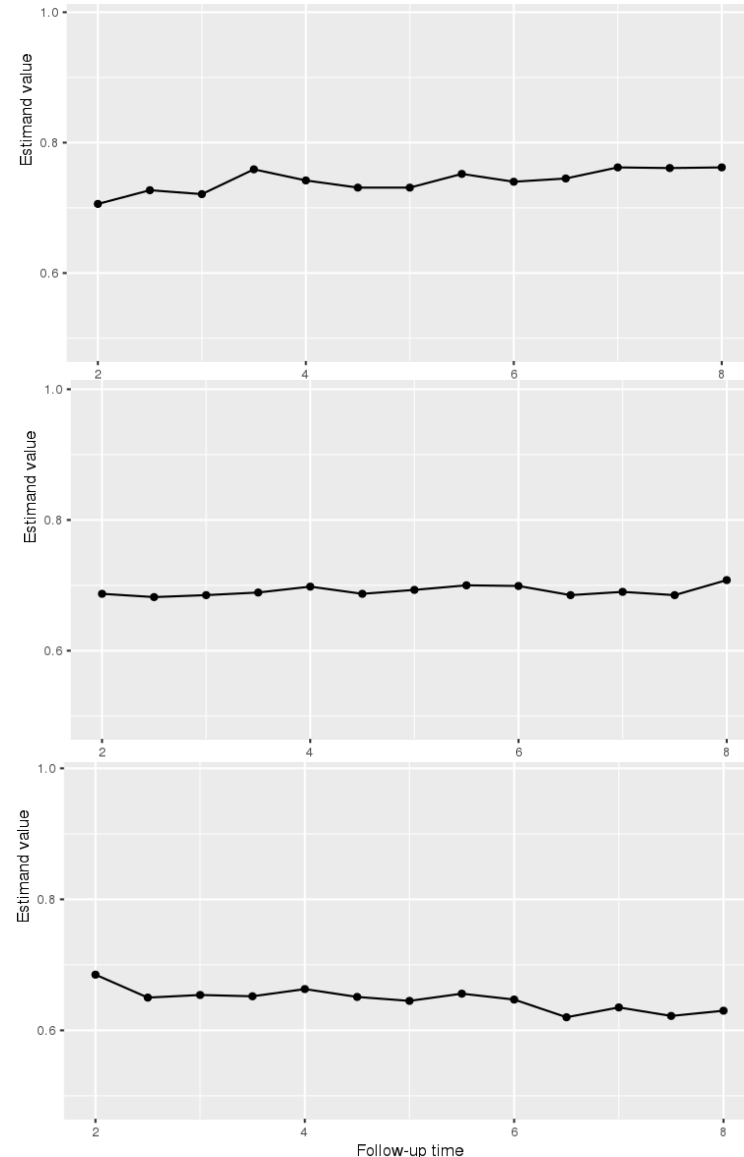
Question 2.4

Estimand 1 (HHF) values for different follow-up times and treatment effects on CVD

HHF and CVD are correlated and test treatment has a **positive effect on CVD (HR=0.8)**

HHF and CVD are correlated and test treatment has **no effect on CVD (HR=1)**

HHF and CVD are correlated and test treatment has a **negative effect on CVD (HR=1.25)**



Additional factors that drive the selection process for Estimand 1

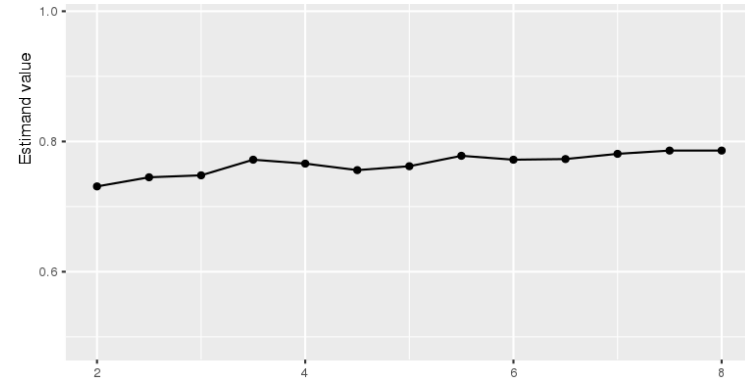
- **Correlation between HHF and CVD**
 - Treatment effect on HHF decreases (Estimand 1 closer to 1) as frailty parameter α (that determines the correlation between HHF and CVD) increases
 - In other words, a larger correlation between HHF and CVD results in a bigger impact of CVD on the HHF treatment effect
- **Treatment discontinuation rate**
 - The treatment effect on HHF decreases (Estimand 1 closer to 1) as discontinuation rate increases, since more patients discontinue treatment and hence no longer benefit from the treatment

How does the selection process impact Estimand 2

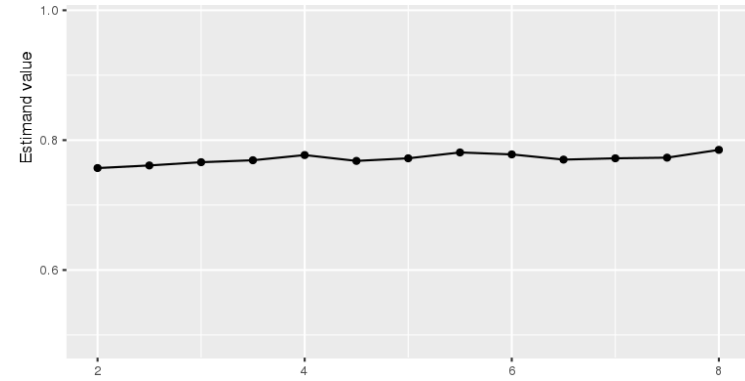
- Estimand 2 will never allow an independent evaluation of the effect on the recurrent event independent of the terminal event, since a composite endpoint (HHF+CVD) is considered
- When HHF and CVD are correlated and test treatment has a negative effect on CVD (e.g. HR=1.25)
 - The treatment effect on HHF+CVD is smaller than that on the HHF only
- When HHF and CVD are correlated, the treatment effect on HHF+CVD **slightly** increases as the treatment effect on CVD increases
- When HHF and CVD are independent, the treatment effect on HHF+CVD increases as the treatment effect on CVD increases

Estimand 2 (HHF+CVD) values for different follow-up times and treatment effects on CVD

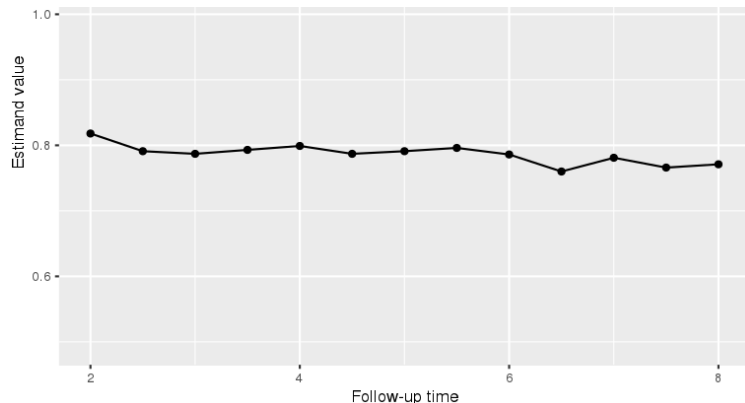
HHF and CVD are correlated and test treatment has a **positive effect on CVD (HR=0.8)**



HHF and CVD are correlated and test treatment has **no effect on CVD (HR=1)**

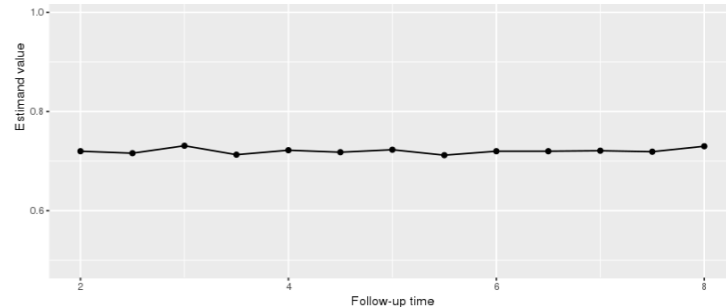


HHF and CVD are correlated and test treatment has a **negative effect on CVD (HR=1.25)**

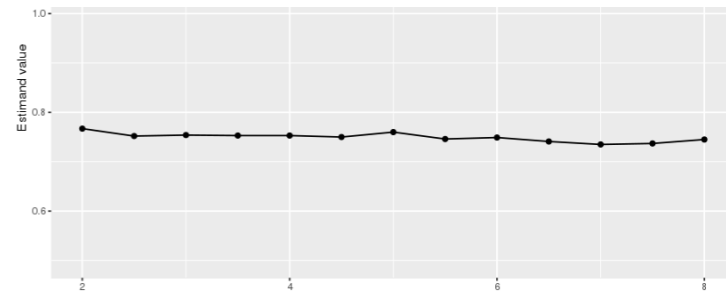


Estimand 2 (HHF+CVD) values for different follow-up times and treatment effects on CVD

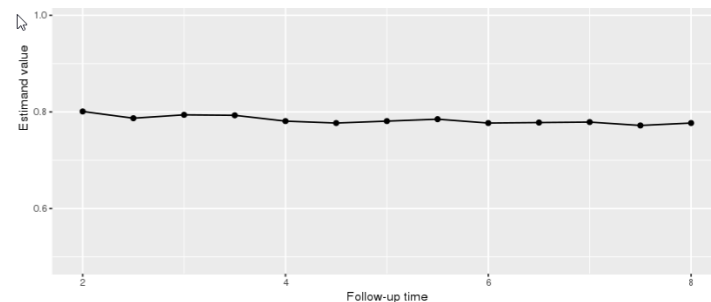
HHF and CVD are independent and test treatment has a **positive effect on CVD (HR=0.8)**



HHF and CVD are independent and test treatment has **no effect on CVD (HR=1)**



HHF and CVD are independent and test treatment has a **negative effect on CVD (HR=1.25)**



Additional question 3

CHAMPION*

- Pulmonary artery pressure measurement system CardioMEMS™
- NYHA Class III heart failure + HHF in last 12 months
- Randomized single-blind study with N=550 (270 treatment, 280 control)

Endpoint	Treatment # Events	Control # Events	HR/RR [95% CI]
HHF rate up to 6 months	84	120	0.72 [0.60-0.85] ¹
HHF rate ²	182	279	0.67 [0.55-0.80] ¹
Time-to-first composite ²	121 (44.8%)	145 (51.8%)	0.77 [0.60-0.98] ³
Mortality ²	50 (18.5%)	64 (22.9%)	0.80 [0.55-1.15] ³

¹ Negative binomial model
² During entire randomized follow-up (mean = 15 months, max = 31.5 months)
³ Cox proportional hazards model

- GUIDE-HF trial (N=3600) initiated to further assess effects on mortality with a composite primary endpoint very similar to Estimand 2

*Food and Drug Administration. CardioMEMS™ HF system summary of safety and effectiveness data. 2014 May 28; http://www.accessdata.fda.gov/cdrh_docs/pdf10/P100045b.pdf [Accessed July 6, 2018]

CHARM-Preserved Trial

- CHARM-Preserved studied candesartan vs placebo
- Patients with HF-pEF (NYHA class II-IV and LVEF>40%)
- N=3023 patients observed over a median follow-up of 36.6 months
- Primary endpoint was time-to-first composite HHF or CVD (Yusuf et al 2003), post-hoc analysis of recurrent events by Rogers et al (2013)

Yusuf, S et al (2003): Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved trial. *Lancet*, 362:777-781

Rogers, JK et al (2014). Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. *European Journal of Heart Failure*, 16:33–40.

CHARM-Preserved Results

Observed treatment effects for different estimands/ approaches

		HR/RR/WinRatio (95%-CI)	p-value
Estimand 1	Cox	0.85 (0.72 – 1.01)	0.072
	LWYY	0.71 (0.57 – 0.88)	0.002
	NB	0.68 (0.54 – 0.85)	<0.001
Estimand 2	Cox	0.86 (0.74 - 1.00)	0.050
	LWYY	0.78 (0.65 - 0.93)	0.006
	NB	0.75 (0.62 - 0.91)	0.003
CVD	Cox	0.99 (0.80 - 1.22)	0.918
WinRatio	Unmatched	1.16 (1.00 – 1.35)	0.062
	Matched	1.19 (1.00 – 1.40)	0.049

Detrimental effects on CVD not observed in past CHF studies

Observed Hazard ratio (HR) for time to CVD and time to first HHF for published heart failure trials

	HR CVD (95%-CI)	HR HHF (95%-CI)
ASTRONAUT	0.92 (0.68 – 1.26)	0.90 (0.72 – 1.12)
CHARM-Added	0.84 (0.72 – 0.98)	0.83 (0.71 – 0.96)
CHARM-Alternative	0.85 (0.71 - 1.02)	0.68 (0.57 – 0.81)
CHARM-Preserved	0.99 (0.80 – 1.22)	0.86 (0.74 – 1.00)
EMPHASIS-HF	0.77 (0.62 – 0.96)	0.61 (0.50 – 0.75)
I-Preserve	1.01 (0.86 – 1.18)	0.95 (0.81 – 1.10)
PARADIGM-HF	0.80 (0.71 – 0.89)	0.79 (0.71 – 0.89)
SHIFT	0.91 (0.80 – 1.03)	0.74 (0.66 – 0.83)
TOPCAT	0.90 (0.73 – 1.12)	0.83 (0.69 – 0.99)
Val-HeFT	1.01 (0.89 – 1.16)	0.73 (0.64 – 0.84)

Recurrent events as post-hoc analysis in CHF trials

Estimand 2 - observed treatment effects RR/HR (with 95% CI)

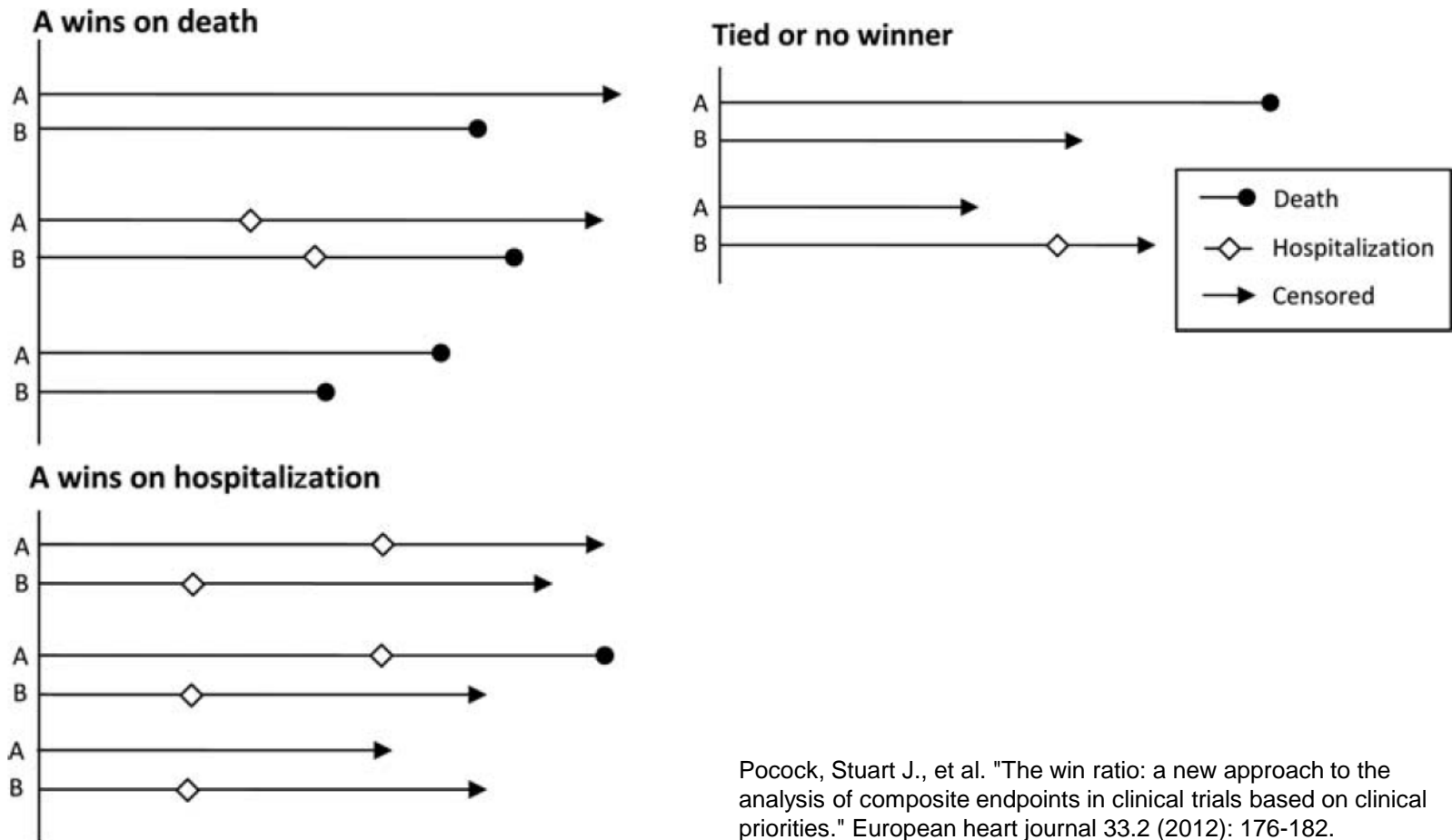
	LWYY	NB	Cox
ValHeFT	0.83 (0.75 - 0.93)	0.84 (0.72 - 0.95)	0.89 (0.81 - 0.98)
CHARM- Preserved	0.78 (0.65 - 0.93)	0.75 (0.62 - 0.91)	0.86 (0.74 - 1.00)
CHARM- Added	/	0.75 (0.62 - 0.91)	0.83 (0.74 - 0.94)
CHARM- Alternative	/	0.65 (0.51 - 0.82)	0.77 (0.67 - 0.89)

Question 2.5

Defining a winner when comparing two subjects

- Prioritizing mortality over heart failure hospitalizations
- Subject A wins over subject B with respect to mortality if subject A has a longer time from randomization to CVD than subject B
 - Straightforward if both or neither have a CVD
 - If only one subject has a CVD, the other must be followed for longer in order to know definitely who had CVD first
- Compare time from randomization to first HF hospitalization in the same manner if it cannot be determined who had a CVD first

Defining a winner - Example

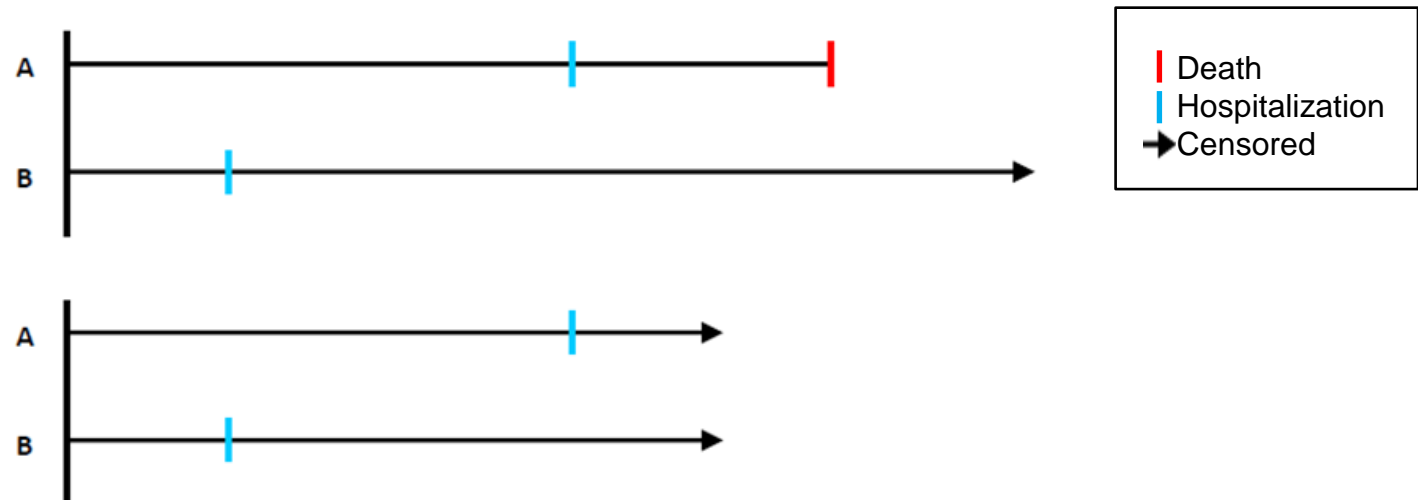


Pocock, Stuart J., et al. "The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities." *European heart journal* 33.2 (2012): 176-182.

The win ratio approach according to Pocock et al. (2012)

- To calculate the win ratio, pairs including a patient of the treatment group and the control group, respectively, are formed
- Subjects within each pair are compared as described before to determine winner/loser/tie
- Number of winners in treatment group: N_W
- Number of losers in the treatment group: N_L
- Win ratio: $WR = \frac{N_W}{N_L}$
- Win ratio either based on comparison of matched pairs or unmatched pairs
 - Matched pairs are formed based on individual patient risk
 - Unmatched pairs approach considers each possible pair of subjects between two groups
- Here: Focus on unmatched pairs approach

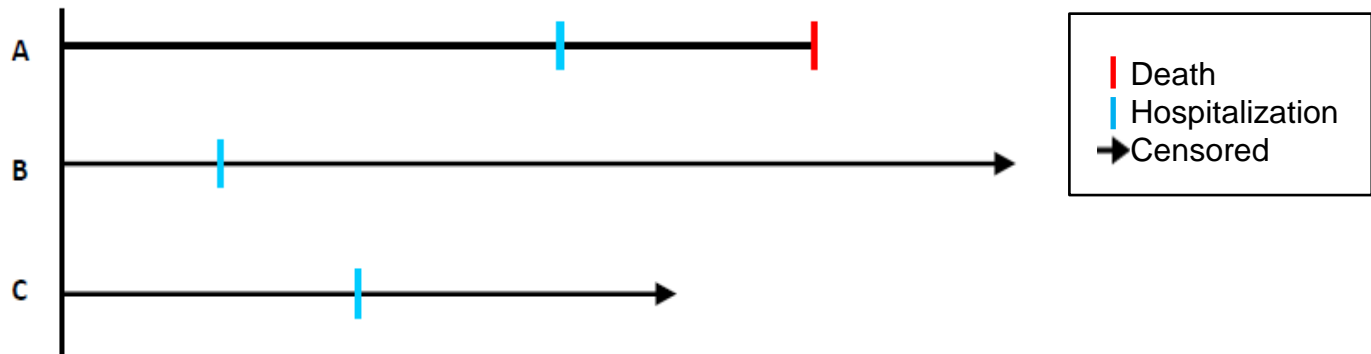
Win ratio depends on follow-up and censoring time



- Top: B wins over A, because of longer survival
- Bottom: If follow-up time shortened, A wins over B, because of longer time until first HHF.

Win ratio is not transitive

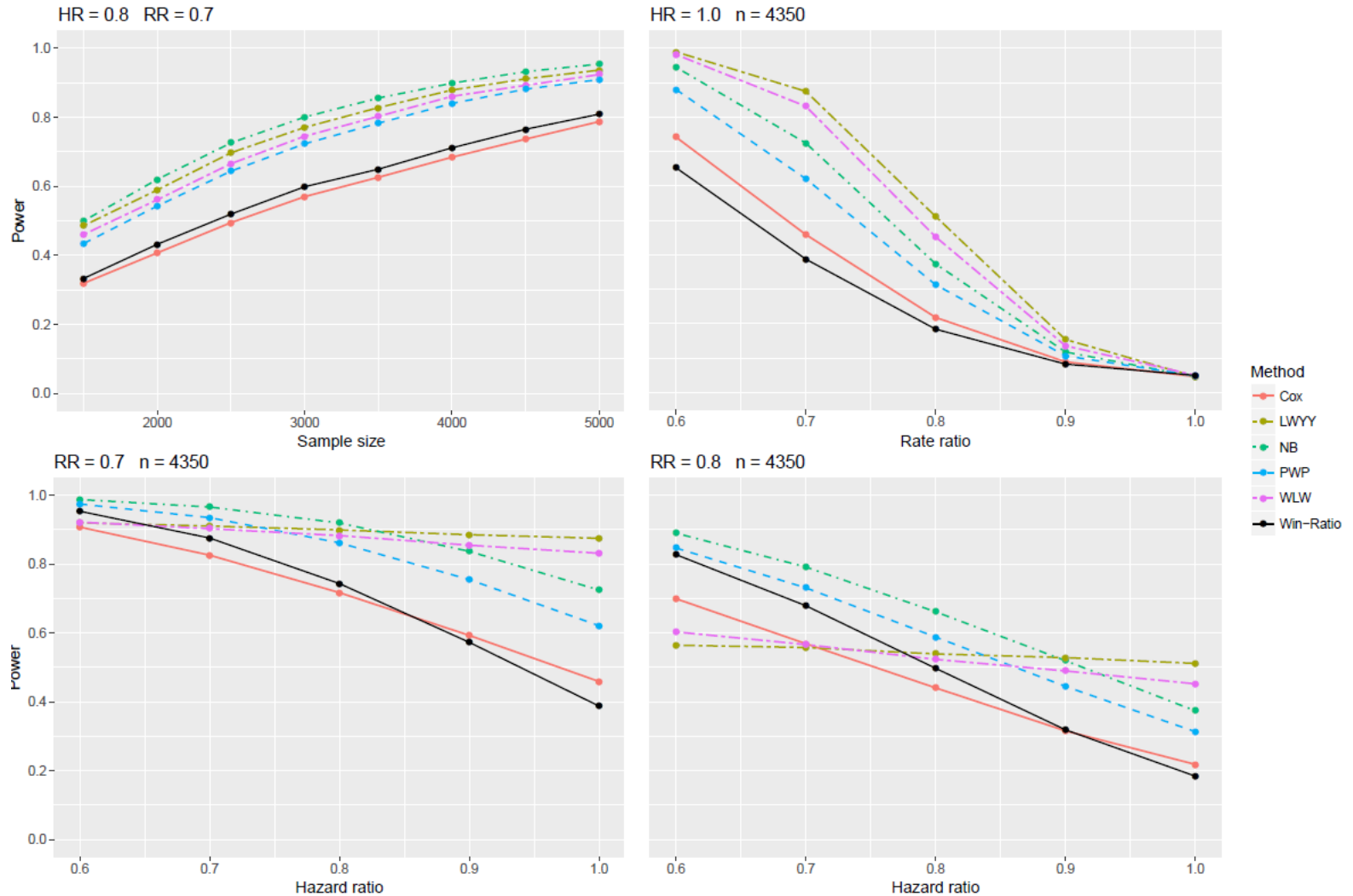
- Transitivity (example): If $a < b$ and $b < c$, then $a < c$



- B wins over A, because of longer survival
- C wins over B, because of longer time to first HHF
- Should follow: C wins over A
 - But this is not the case: A wins over C, because of longer time to first HHF

Results of simulation study

Comparison with Estimand 2 (HHF+CVD)



Results of simulation study

Comparison with Estimand 2 (HHF+CVD)

- Win ratio approach has for practically relevant scenarios in general a lower power than the recurrent event methods
- Win ratio approach has higher power than Cox (time-to-first) when effect on HHF is small and effect on CVD large
- Practically relevant scenarios:
 - In practice, it seems unlikely that the effect on CVD would be larger than the one on HHF. At least the effects observed in previous trials have either similar magnitude or the treatment effect on HHF has been larger

Weighted composites

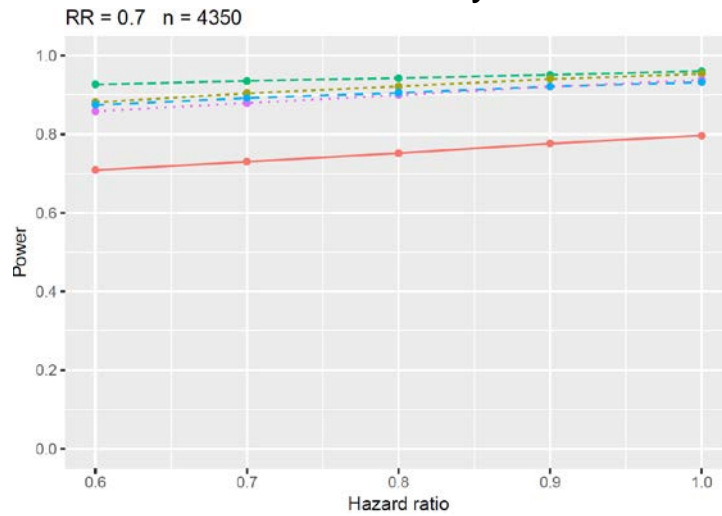
- Idea: the individual components of the endpoint are assigned weights which reflect the clinical importance of the component
- Defining weights in a manner that is scientifically justified and agreed upon within the clinical community is challenging from a clinical perspective
 - Anker et al. (2016): “[Statistical methods to weight outcomes] are limited by lack of consensus on the relative weighting of events and inconsistency across studies.”
- Estimand 2 constitutes a weighted endpoint in the sense that a cardiovascular hospitalization is weighted the same as a cardiovascular death

Question 2.2

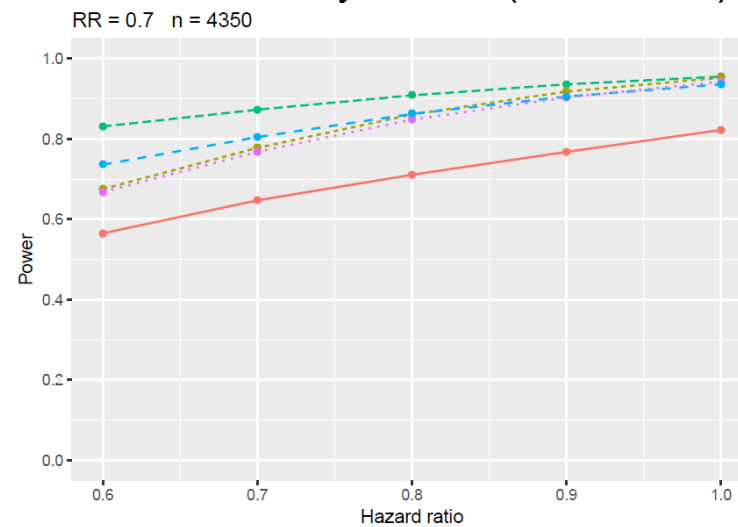
Increased mortality – Power

Estimand 1, $RR_{HHF}=0.7$

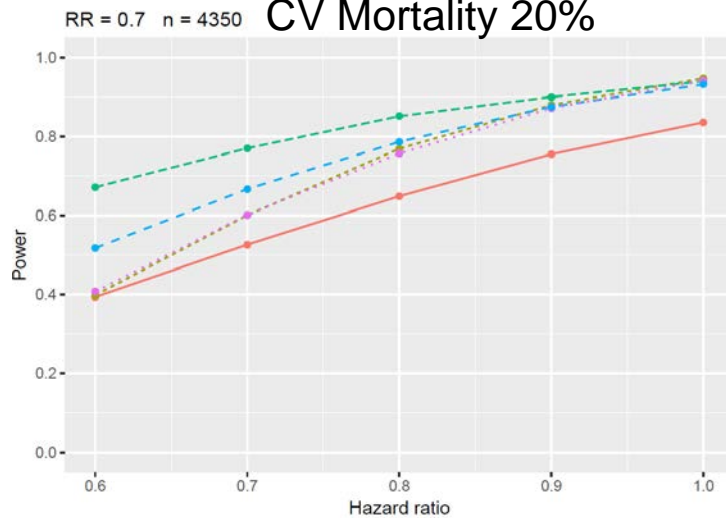
CV Mortality 5%



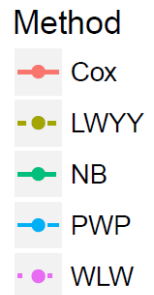
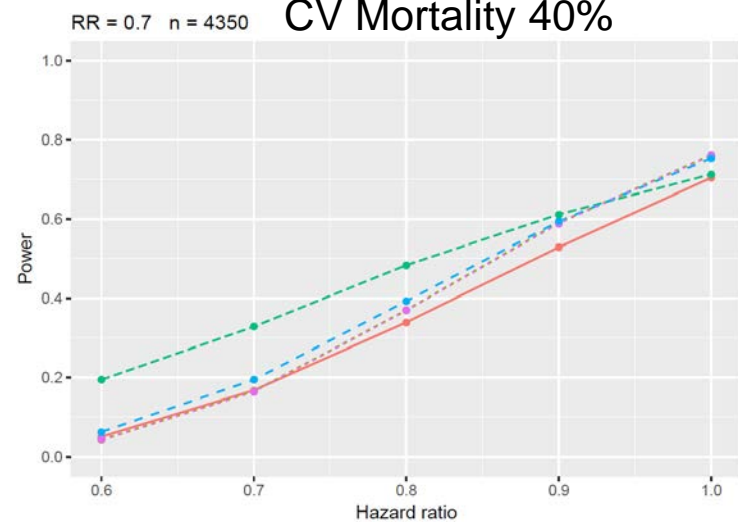
CV Mortality 12.5% (Base case)



CV Mortality 20%



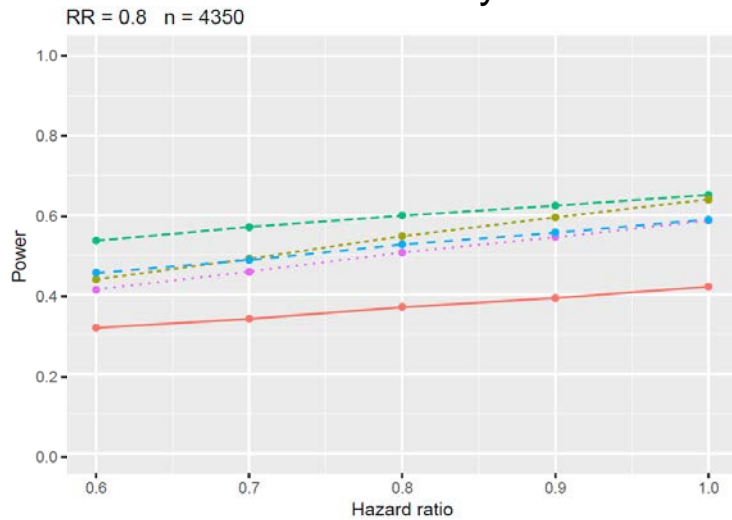
CV Mortality 40%



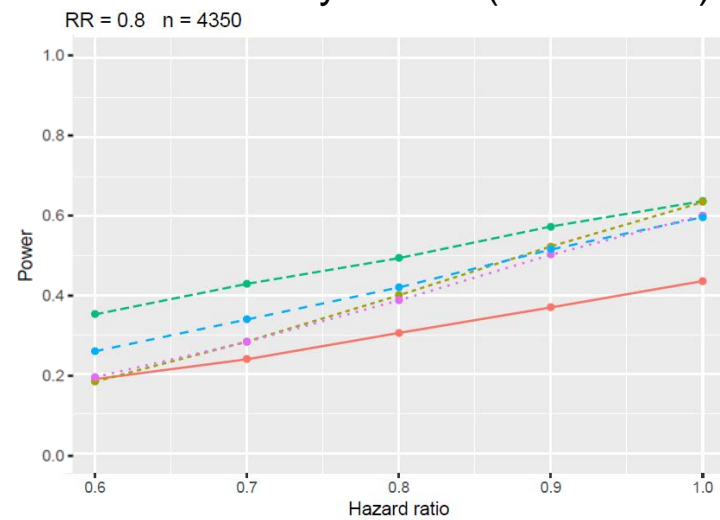
Increased mortality – Power

Estimand 1, $RR_{HHF}=0.8$

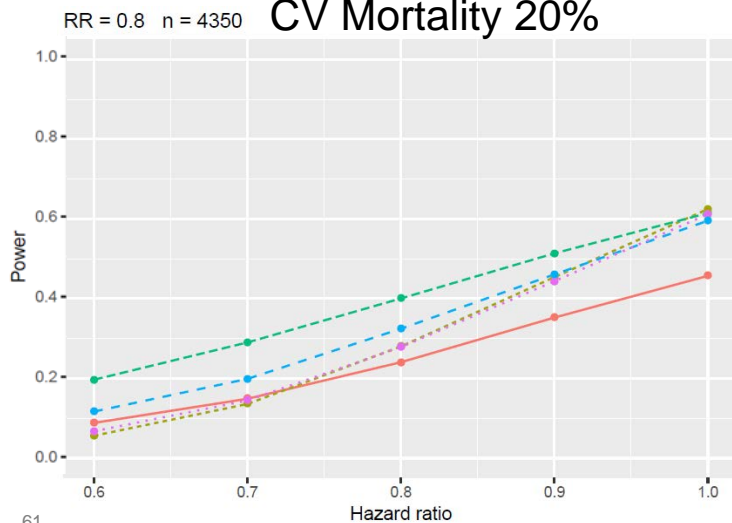
CV Mortality 5%



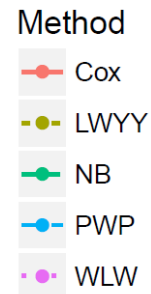
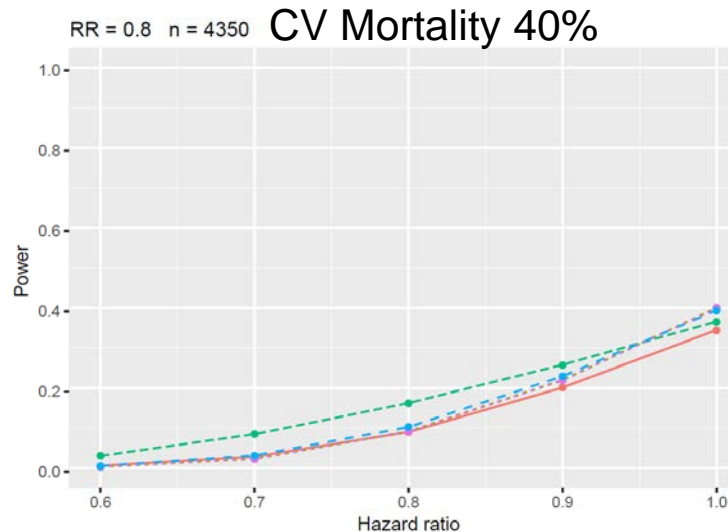
CV Mortality 12.5% (Base case)



CV Mortality 20%



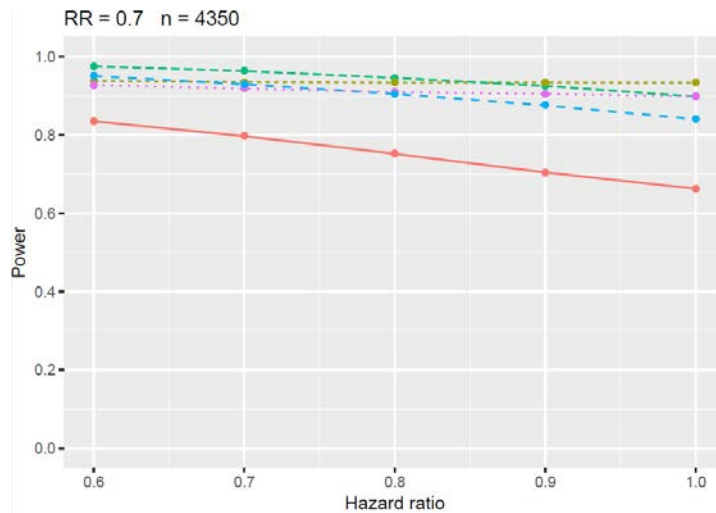
CV Mortality 40%



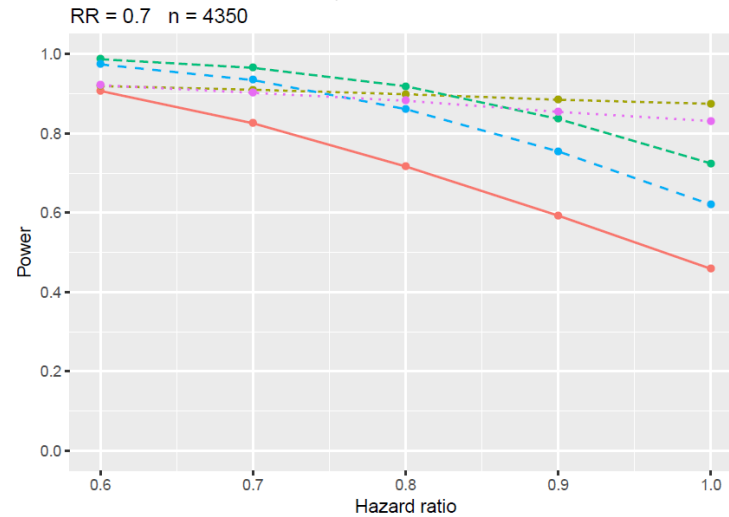
Increased mortality – Power

Estimand 2, $RR_{HHF}=0.7$

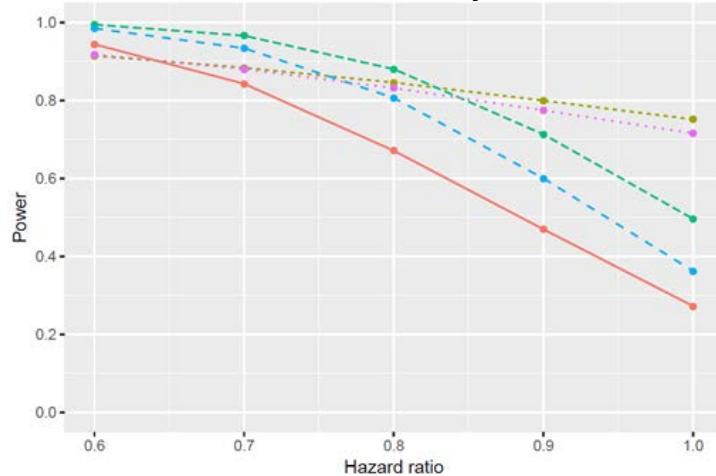
CV Mortality 5%



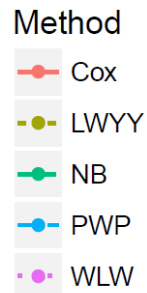
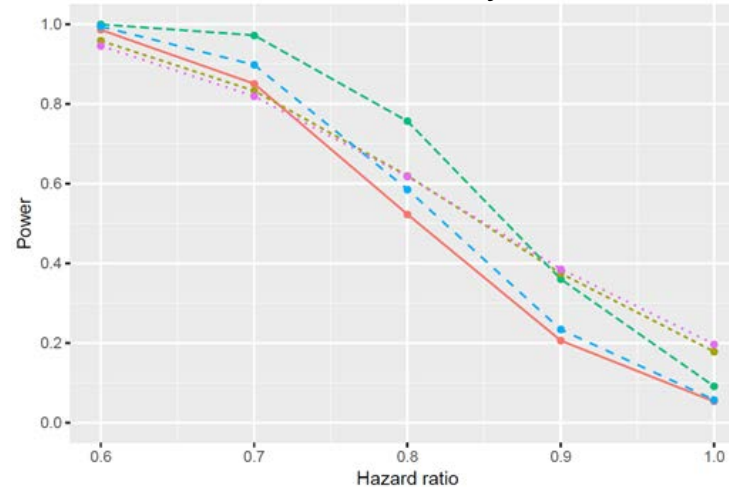
CV Mortality 12.5% (Base case)



CV Mortality 20%



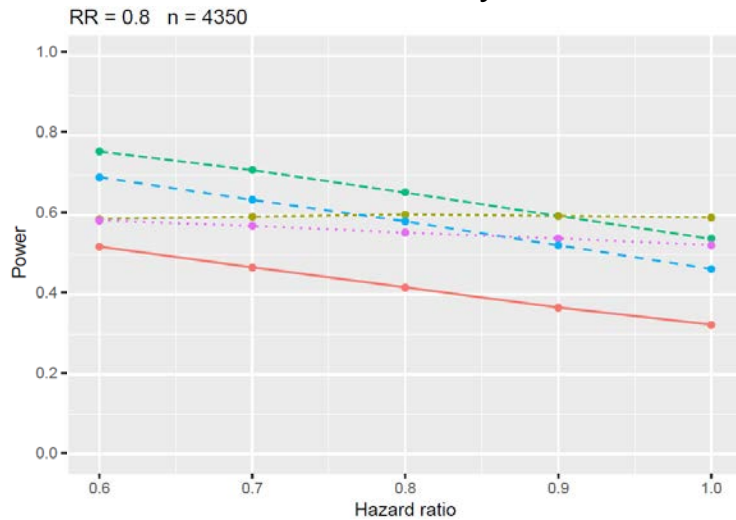
CV Mortality 40%



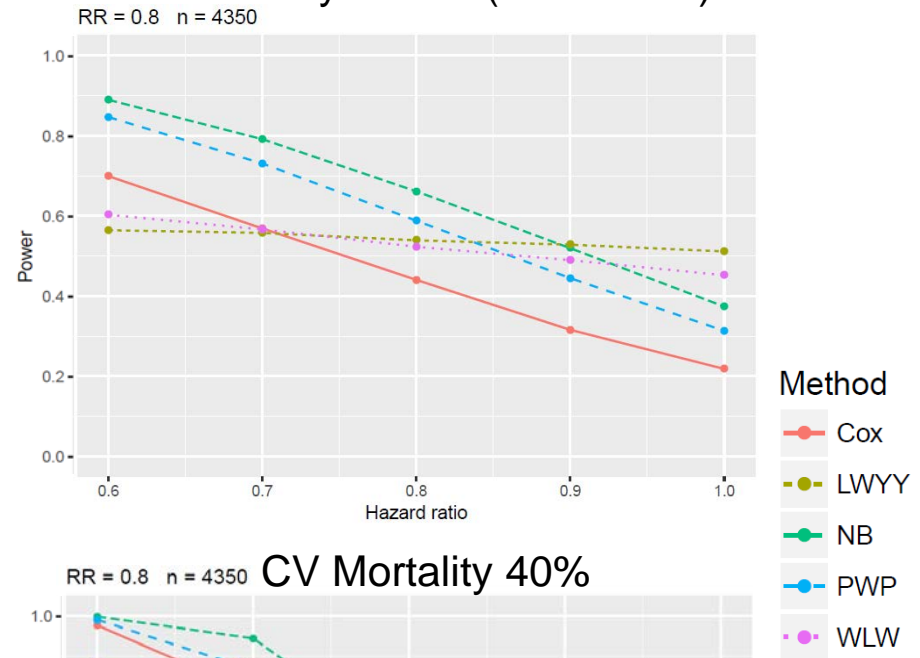
Increased mortality – Power

Estimand 2, $RR_{HHF}=0.8$

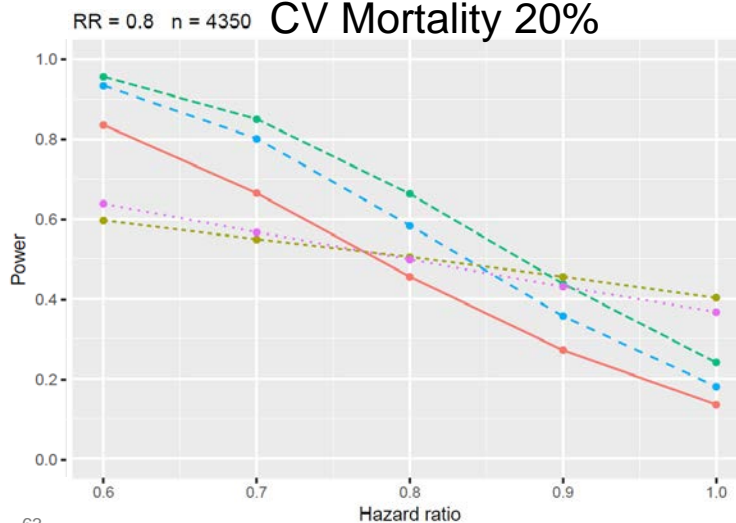
CV Mortality 5%



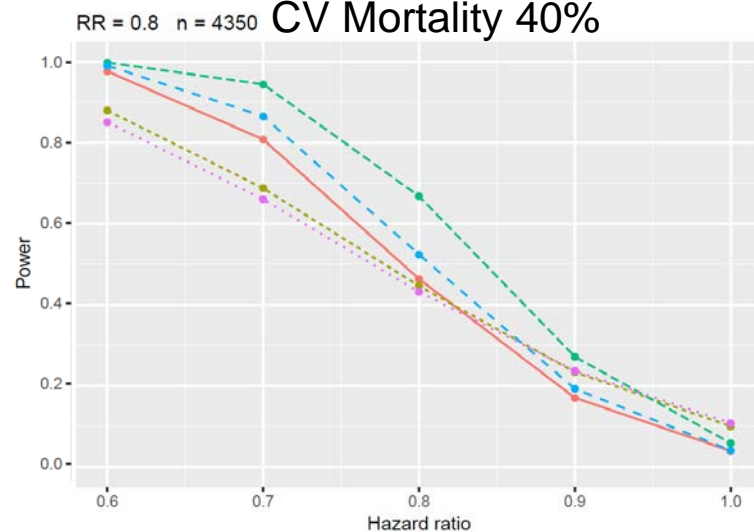
CV Mortality 12.5% (Base case)



CV Mortality 20%



CV Mortality 40%



Question 1.2

Results for smaller sample size with a different random number seed

Table 7A.1: 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 $\theta = 0.25$.

	Method	$n = 50$		$n = 75$		$n = 125$	
		RR	Type I error	RR	Type I error	RR	Type I error
Scenario 1: Non-informative (Hypothetical)	Cox	1.003	0.023	1	0.025	1.002	0.026
	NB	1.001	0.028	0.999	0.027	1.001	0.026
	LWYY	0.999	0.027	1.001	0.026	0.999	0.026
	WLW	0.998	0.027	1	0.026	0.997	0.028
	PWP	0.999	0.025	1.001	0.023	0.999	0.027
Scenario 2: Informative (Hypothetical)	Cox	1.005	0.023	1	0.026	1.001	0.023
	NB	1.003	0.028	0.999	0.027	1.002	0.027
	LWYY	0.997	0.03	1.001	0.027	0.998	0.027
	WLW	0.996	0.028	1	0.026	0.997	0.028
	PWP	0.997	0.028	1	0.026	0.998	0.027
Scenario 3: Non-informative (Treatment-policy)	Cox	1.003	0.024	1	0.024	1.002	0.025
	NB	1.001	0.027	0.998	0.029	1.001	0.026
	LWYY	0.999	0.026	1.001	0.027	0.999	0.026
	WLW	0.998	0.027	1.001	0.027	0.998	0.026
	PWP	1	0.026	1.001	0.025	0.999	0.027
Scenario 4: Informative (Treatment-policy)	Cox	1.003	0.025	0.999	0.027	1.002	0.026
	NB	1.001	0.029	0.999	0.027	1.001	0.025
	LWYY	0.999	0.029	1.001	0.026	0.998	0.027
	WLW	0.998	0.028	1.001	0.024	0.997	0.026
	PWP	0.999	0.03	1.001	0.025	0.999	0.026



Thank you