

Data and Safety Monitoring Boards: *Statistical Principles*

Ian Marschner

NHMRC Clinical Trials Centre

ACTA Webinar
24 November 2020



THE UNIVERSITY OF
SYDNEY



NHMRC Clinical Trials Centre

Data Monitoring in Trials

- › Identify issues that threaten the:
 - feasibility of completing a trial
 - methodological rigour of the trial
 - quality of the data
- › Review sample size assumptions

*Without knowledge of
treatment allocation*

- › Use interim information to determine if experimental treatment:
 - is harmful
 - is beneficial
 - is unlikely to be appreciably better than the control treatment (standard care)

*With knowledge of
treatment allocation*

Statistical Roles in Data Monitoring

› Blinded Statistician:

- Conduct interim analyses **without knowledge of treatment allocation**
- Report to the Trial Management Committee (TMC)

› Unblinded Statistician:

- Conduct interim analyses **with knowledge of treatment allocation**
- Report to the Independent Safety and Data Monitoring Committee (IDSMC) and remain “firewalled” from the Blinded Statistician and TMC

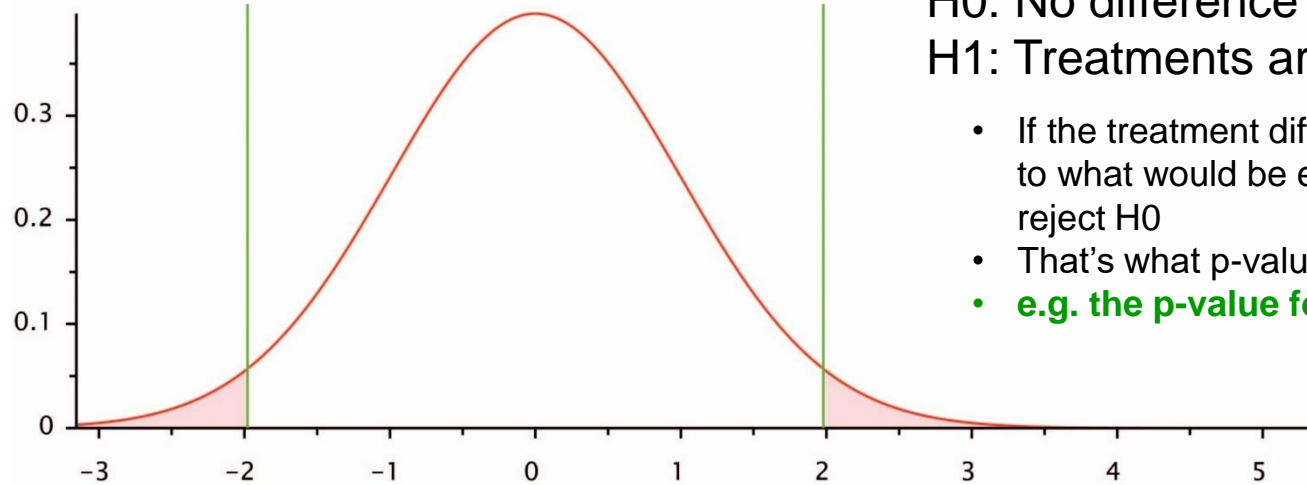
› IDSMC Statistician:

- Member of the IDSMC
- Provides statistical interpretation and recommendations based on unblinded IDSMC analysis reports

Statistical Principles for IDSMCs

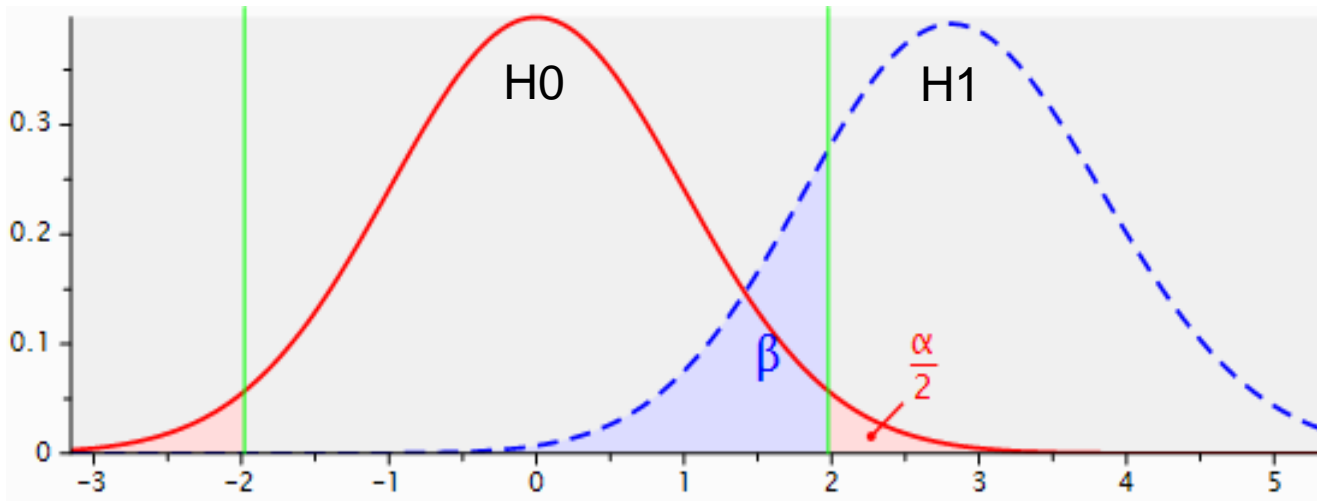
- › A key purpose of interim statistical analysis is to inform a decision about whether to stop the study early because:
 - Intervention is found to be overwhelmingly beneficial compared to the control (**superiority**)
 - Intervention is found to be detrimental / unsafe compared to the control (**inferiority / safety**)
 - Intervention is unlikely to be found to be different from control (**futility**)
- › A key statistical principle for conducting these interim analyses is the need to control the false positive rate, usually called alpha / significance level

Statistical Significance



H0: No difference between treatments
H1: Treatments are different

- If the treatment difference is extreme relative to what would be expected under H0 then reject H0
- That's what p-values measure
- e.g. the p-value for a $|Z|$ of 1.96 is ~5%



α =type I error rate
 β =type II error rate
(1- β)=power

α is the false positive rate

Early Stopping

- › If we conduct an interim analysis and find statistical significance then the IDSMC may want to stop the study for ethical reasons
- › But ... what happens to our false positive rate if we conduct multiple interim analyses (together with the final analysis) and test for statistical significance in the usual way, each time with a false positive rate of 5%?

Multiple Analyses

Multiple significance tests

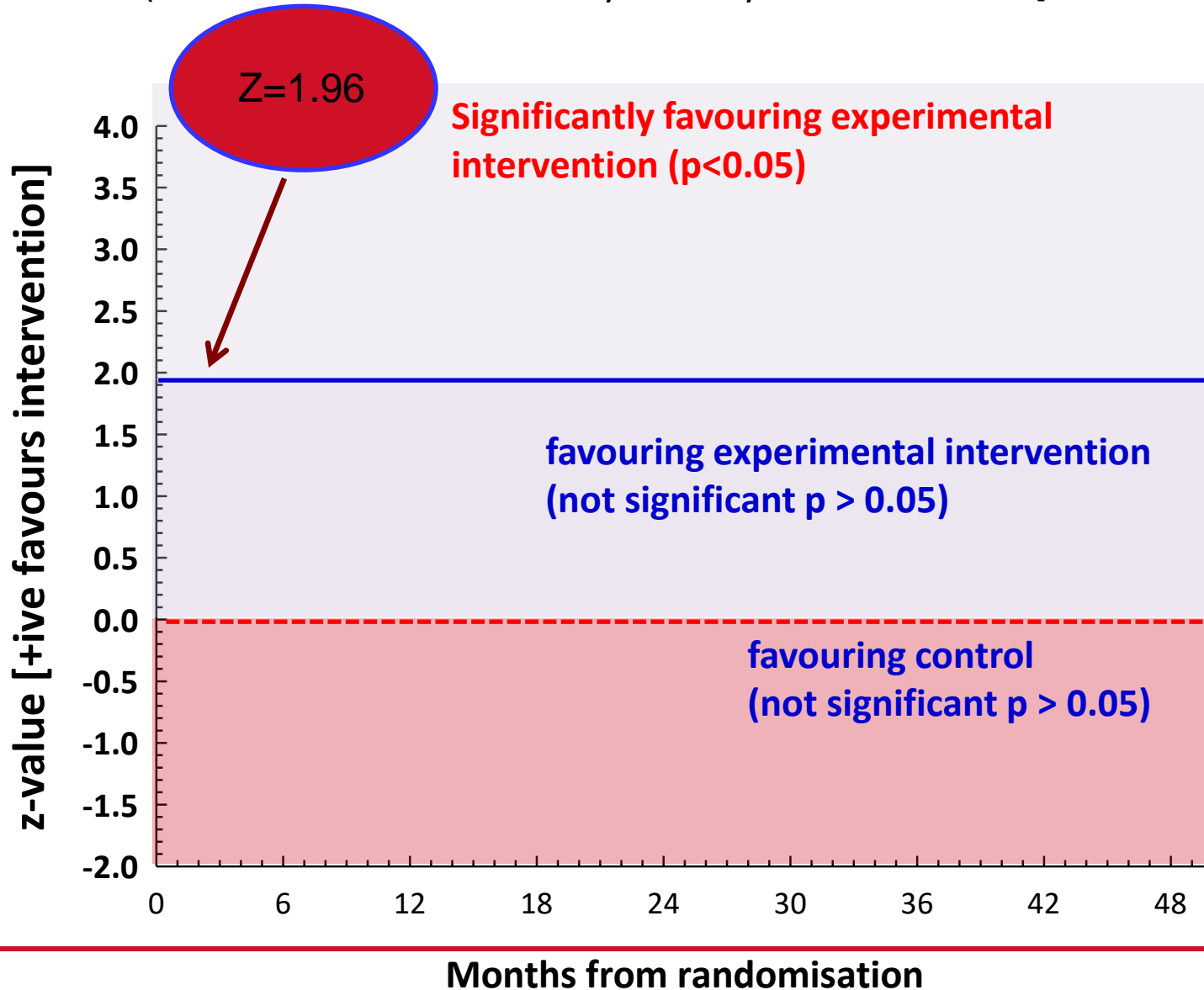
Number of repeated tests at the 5% level	Overall significance level (false positive)
1	0.05
2	0.08
3	0.11
5	0.14
10	0.19
20	0.25

Multiple analyses increase the chance that we will wrongly conclude there is a treatment difference

Source : Armitage, P., et al (1969) Repeated significance tests on accumulating data. *J.R.Statist.Soc.* **132(A)**; 235-244.

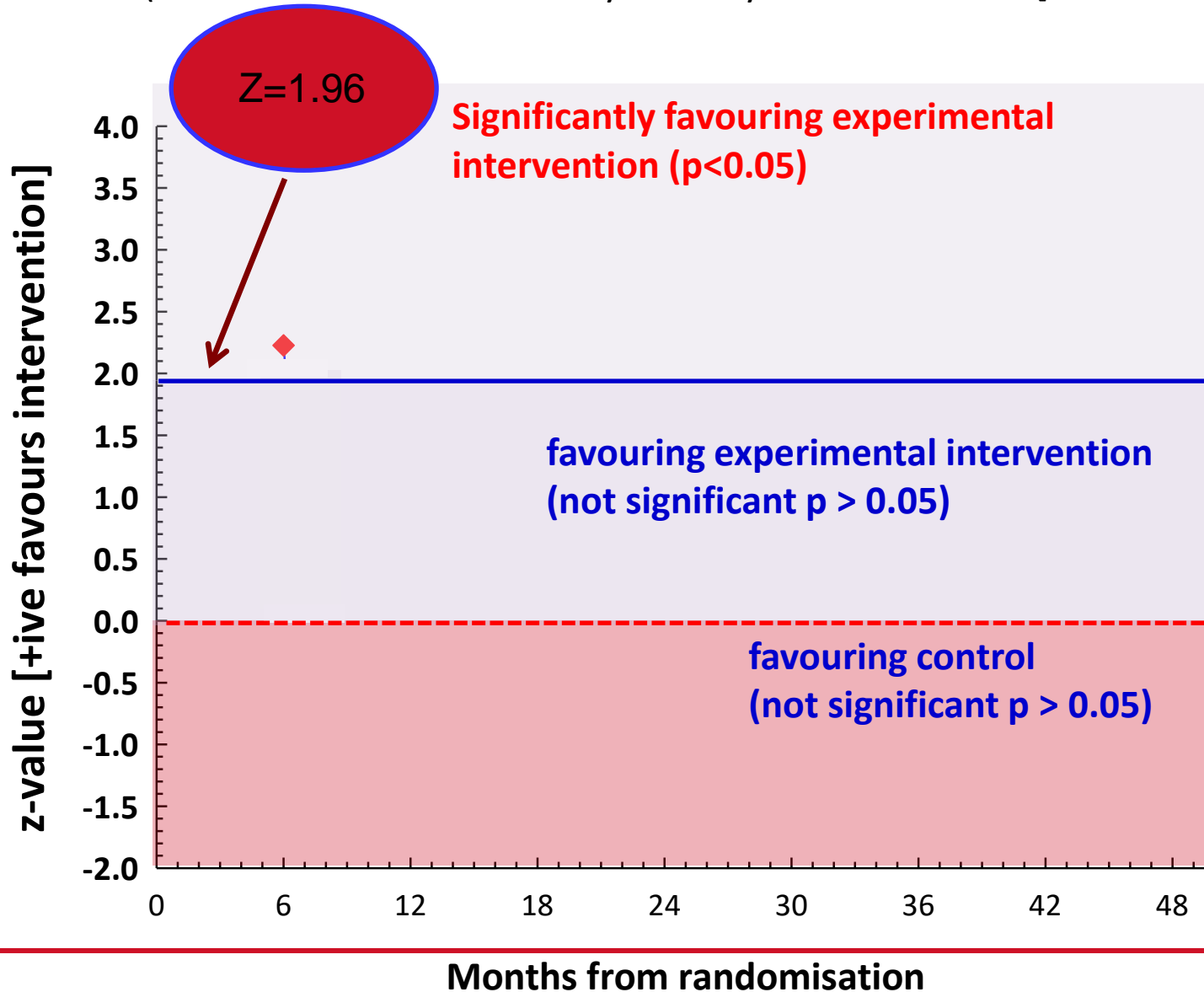
Naïve monitoring

(Simulated RCT with an analysis every month $N=1000$ [H1 is true])



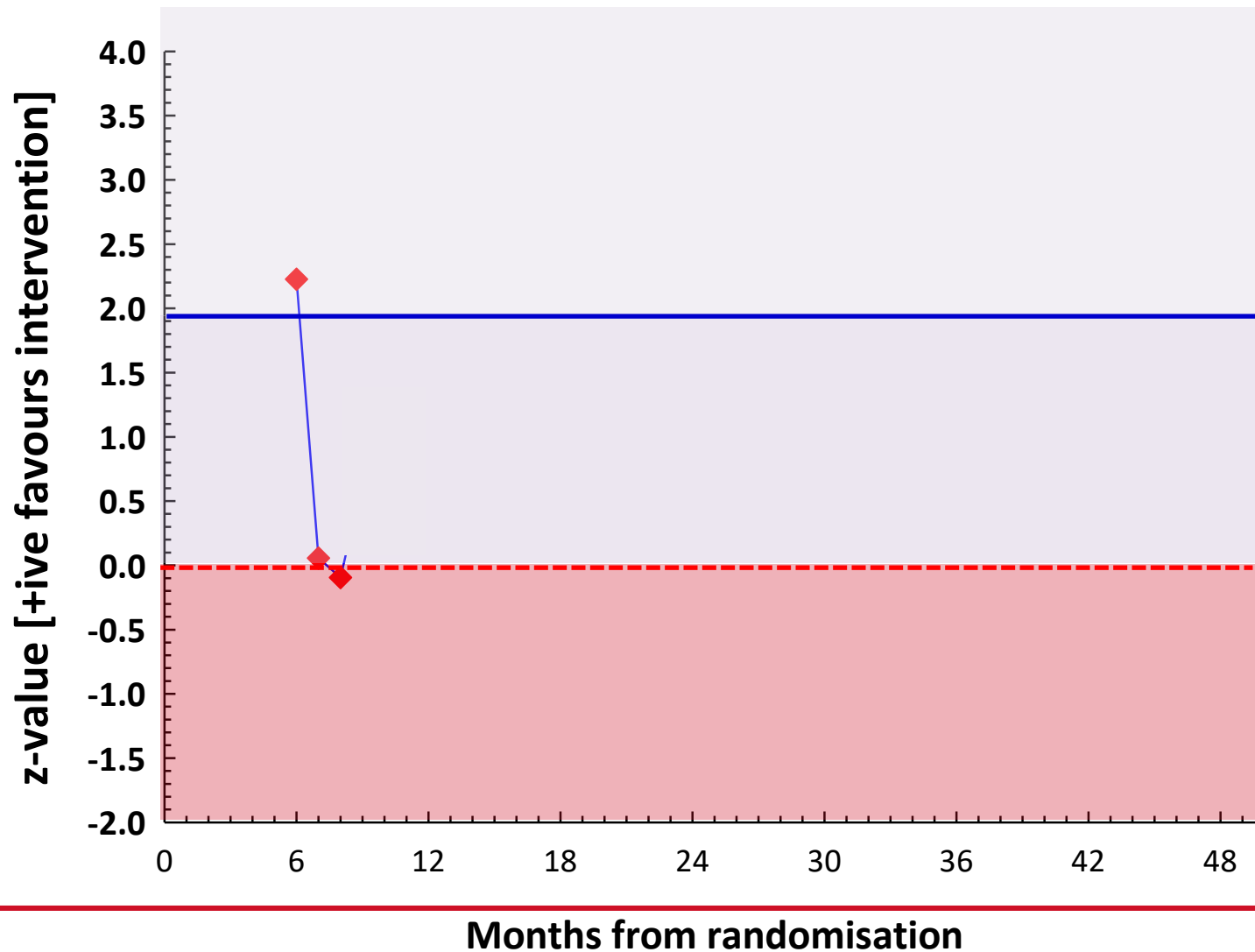
Naïve monitoring

(Simulated RCT with an analysis every month $N=1000$ [H1 is true])



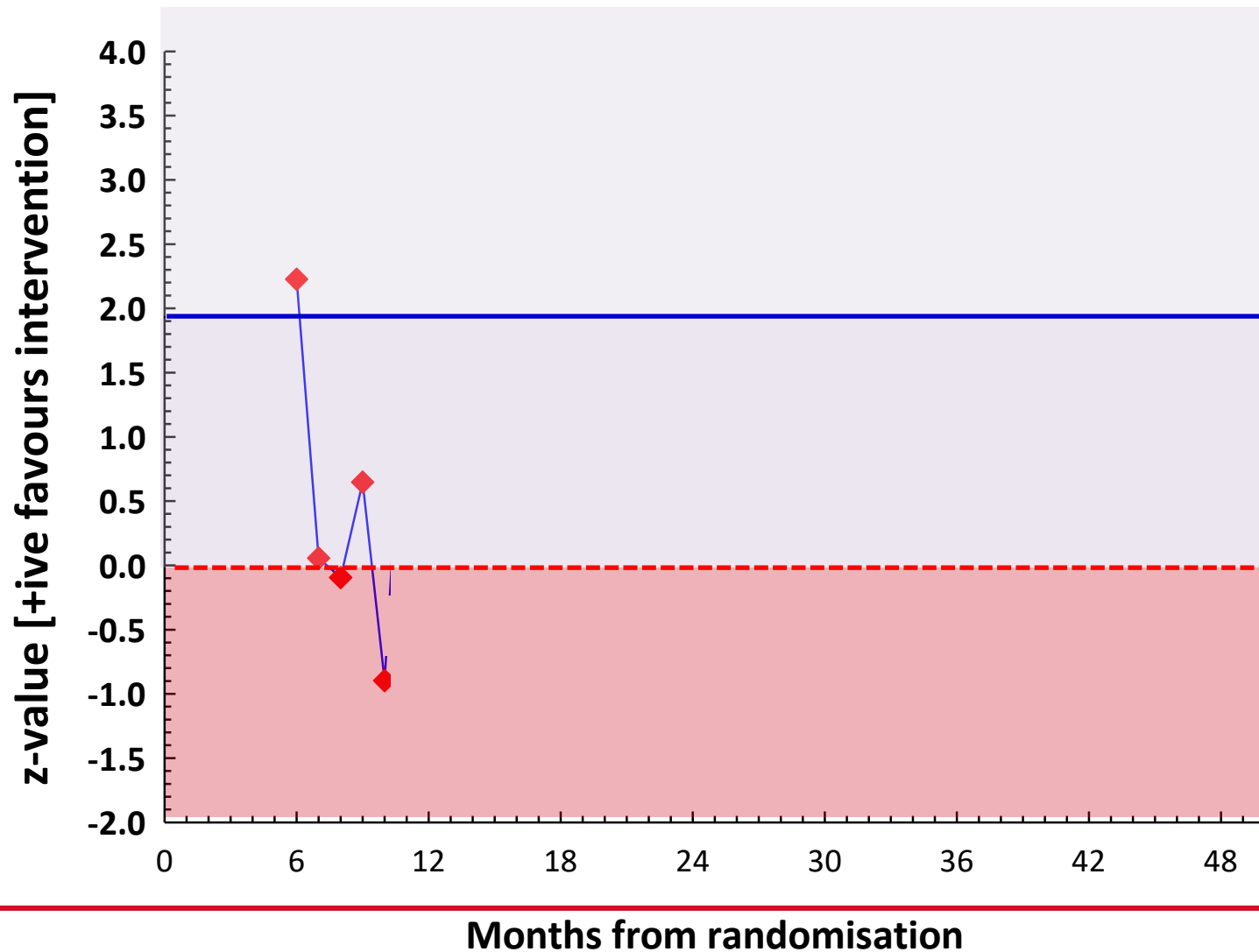
Early variability

(Simulated RCT with an analysis every month N= 1000 [H1 is true])



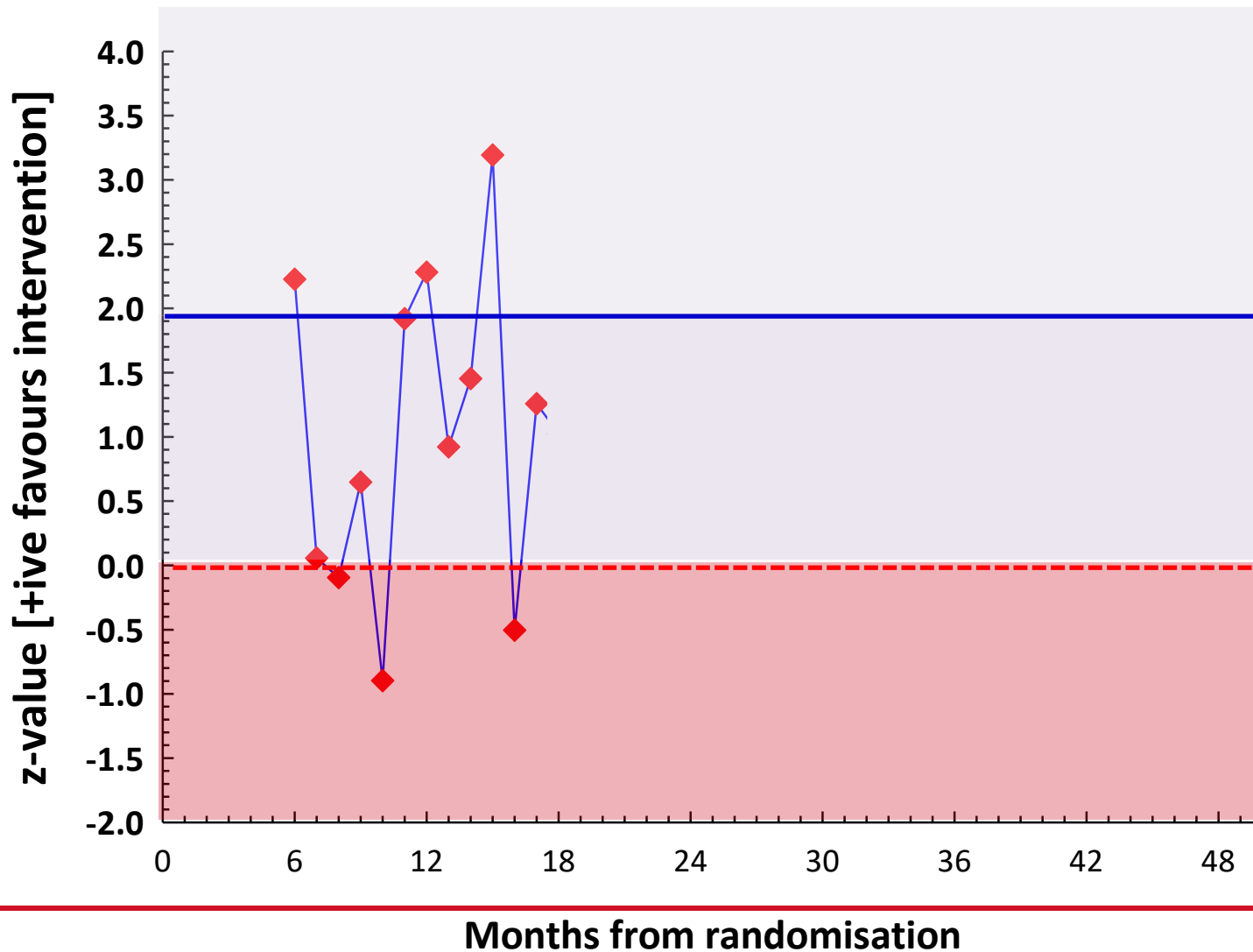
Early variability

(Simulated RCT with an analysis every month N= 1000 [H1 is true])



Early variability

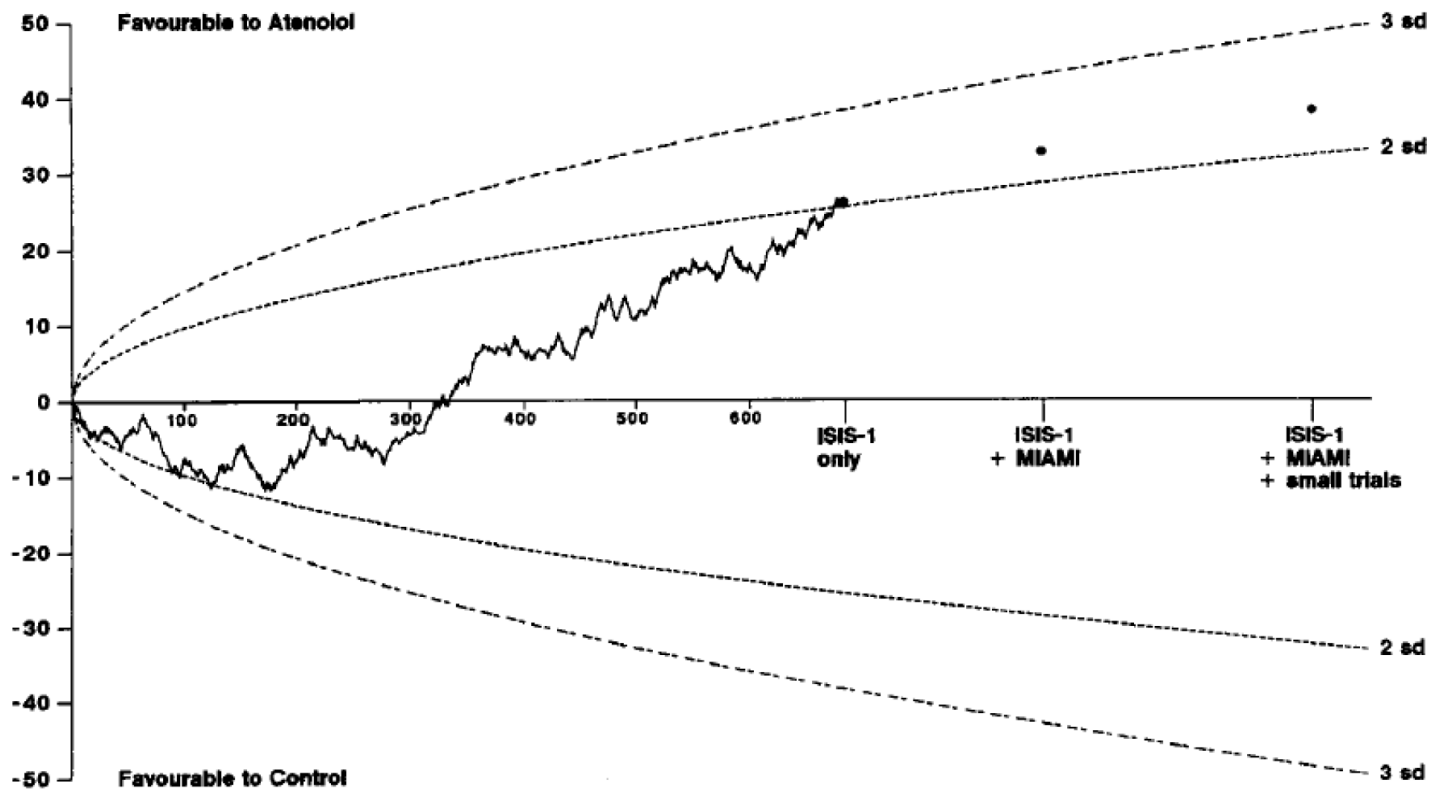
(Simulated RCT with an analysis every month N= 1000 [H1 is true])



Real example

ISIS-1 study: atenolol vs control, coronary death in AMI

Yusuf (2000), *Am Heart J*



ISIS-1: Evolution of treatment difference as information accrued
Y-axis: Observed minus Expected Control deaths in days 0-7
X-axis: Total number of deaths (non-linear)

Issues for the IDSMC to consider

- › Best interests of participants
- › Avoiding unwarranted trial termination
- › Emerging evidence from external sources
- › Statistical guidance to protect against reaching erroneous conclusions

Stopping rules / guidelines

- › Statistical methods for interim analyses have traditionally focused on *stopping rules*
- › Stopping rules are significance tests applied to the accumulating data that provide sufficient early evidence for a treatment difference while controlling the false positive rate
- › If a statistical stopping rule is met, then the study may need to stop or have randomisation suspended
- › Statistical stopping rules should more appropriately be regarded as *guidelines*, rather than *rules*
- › These need to be considered alongside other information, requiring judgement from an IDMSC

Statistical Guide for Early Rejection of H_0

- i) After $k = 1, \dots, K - 1$ looks if $|z_k| \geq c_k$ stop and reject H_0
otherwise continue to the next look
- ii) After look K if $|z_K| \geq c_K$ stop and reject H_0
otherwise stop and accept H_0

› Critical values, c_k , chosen to ensure

Overall Type I
error rate

$$\Pr\left(|z_k| \geq c_k \text{ for some } k=1, \dots, K \mid \theta = 0\right) = \alpha$$

Group Sequential Methods

- › Various rules
 - Pocock
 - Haybittle-Peto
 - O'Brien Fleming

- › K analyses (equally spaced)
 - e.g. 20%, 40%, 60%, 80%, 100% (K=5)

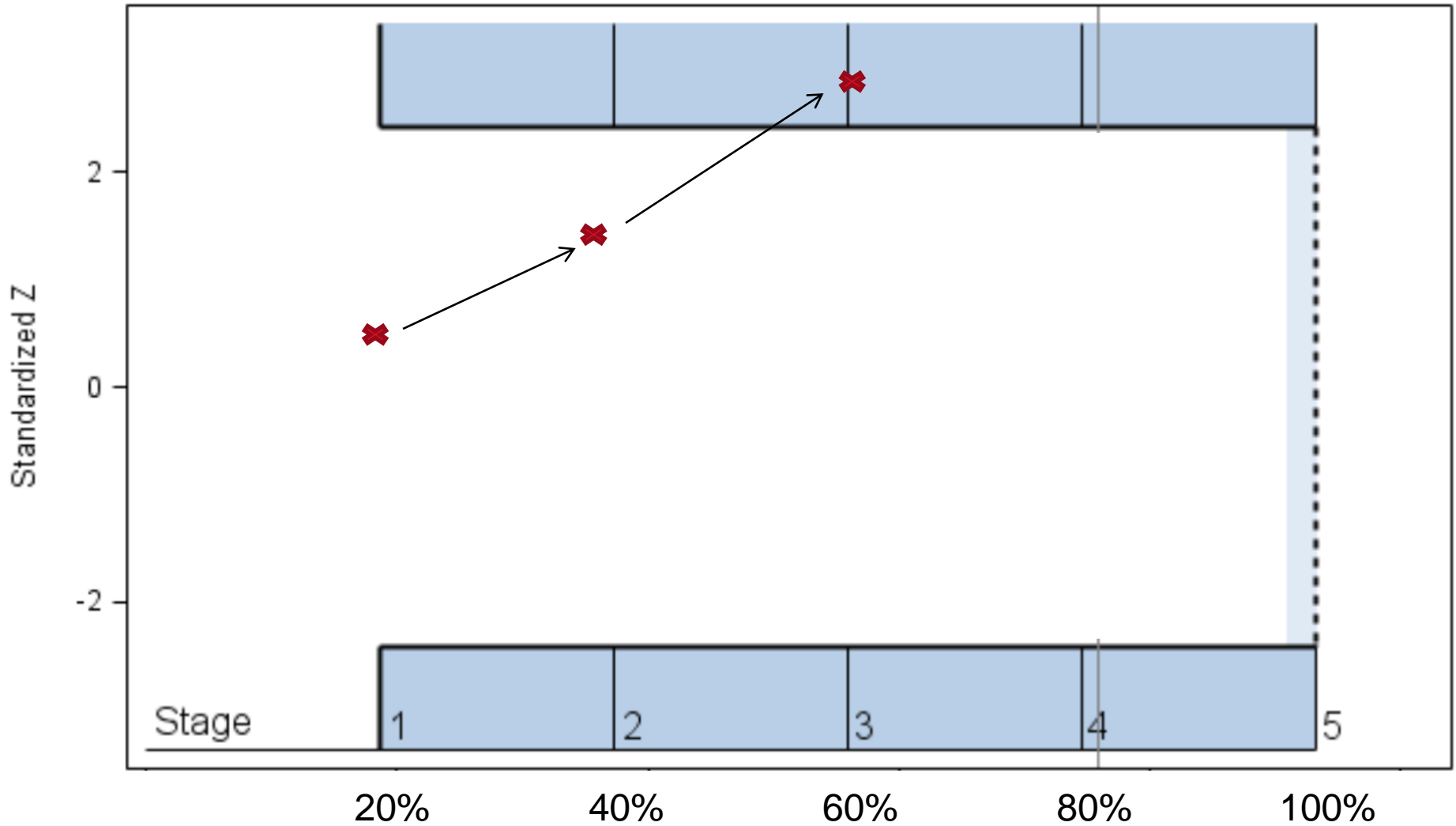
- › Sidedness of test & Overall Type I error rate
 - e.g. 2 sided test at 5% significance level

Information fraction

- › Interim analyses occur after a fraction of information has been collected
- › E.g. 5 equally spaced analyses would have interim analyses after 20%, 40%, 60% and 80% of the information has been collected
- › The term “information” refers to the amount of statistical information
- › Statistical information is measured differently for different endpoints
- › Continuous (normal) endpoints: sample size
- › Time-to-event: number of events
- › More generally: standard error (or variance) of the treatment effect estimate

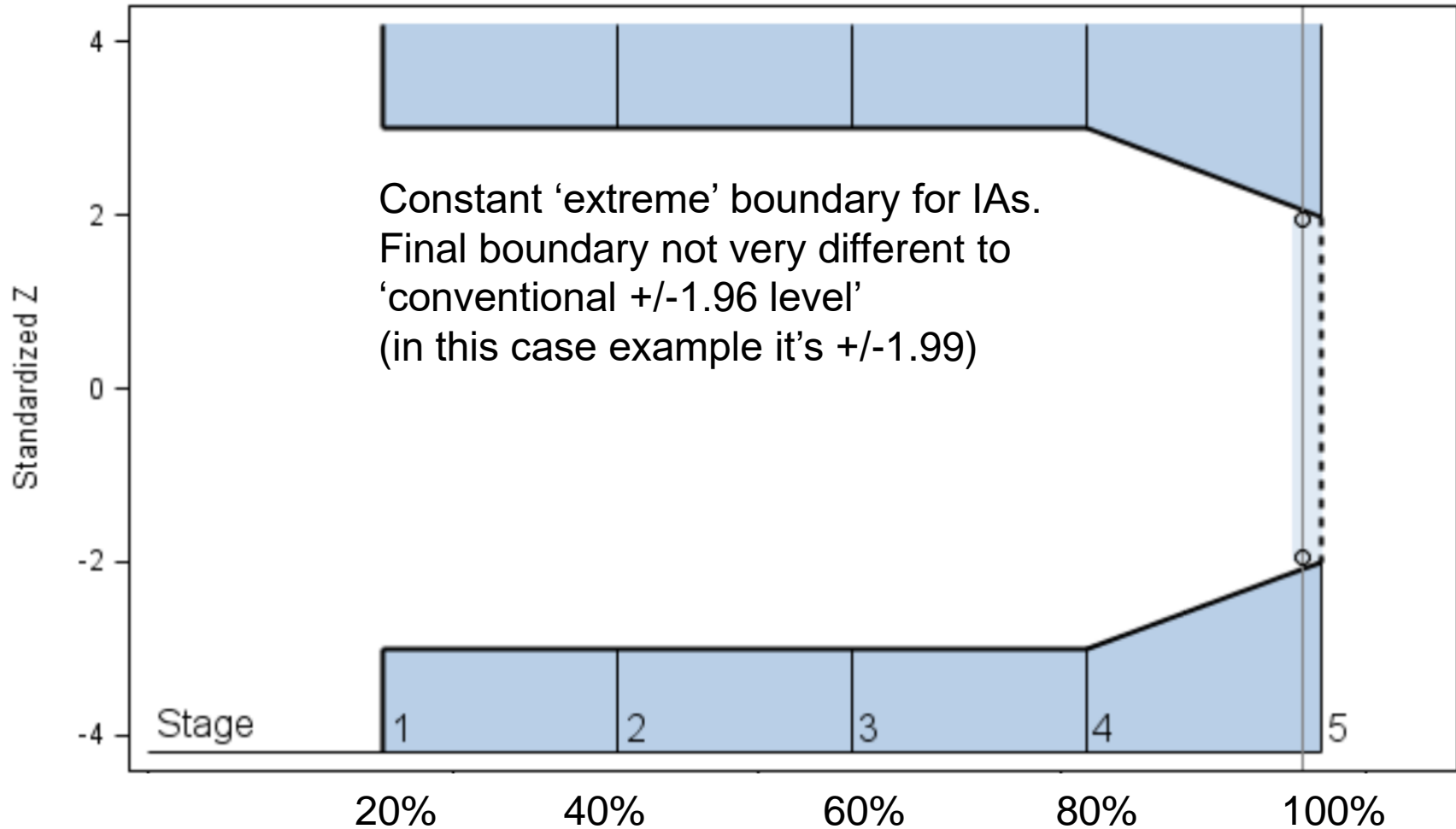
Pocock Rule for 5 Looks

$$P(|Z_1|, |Z_2|, |Z_3|, |Z_4| \text{ or } |Z_5| > \underline{2.413} \mid H_0 \text{ true}) = 0.05$$



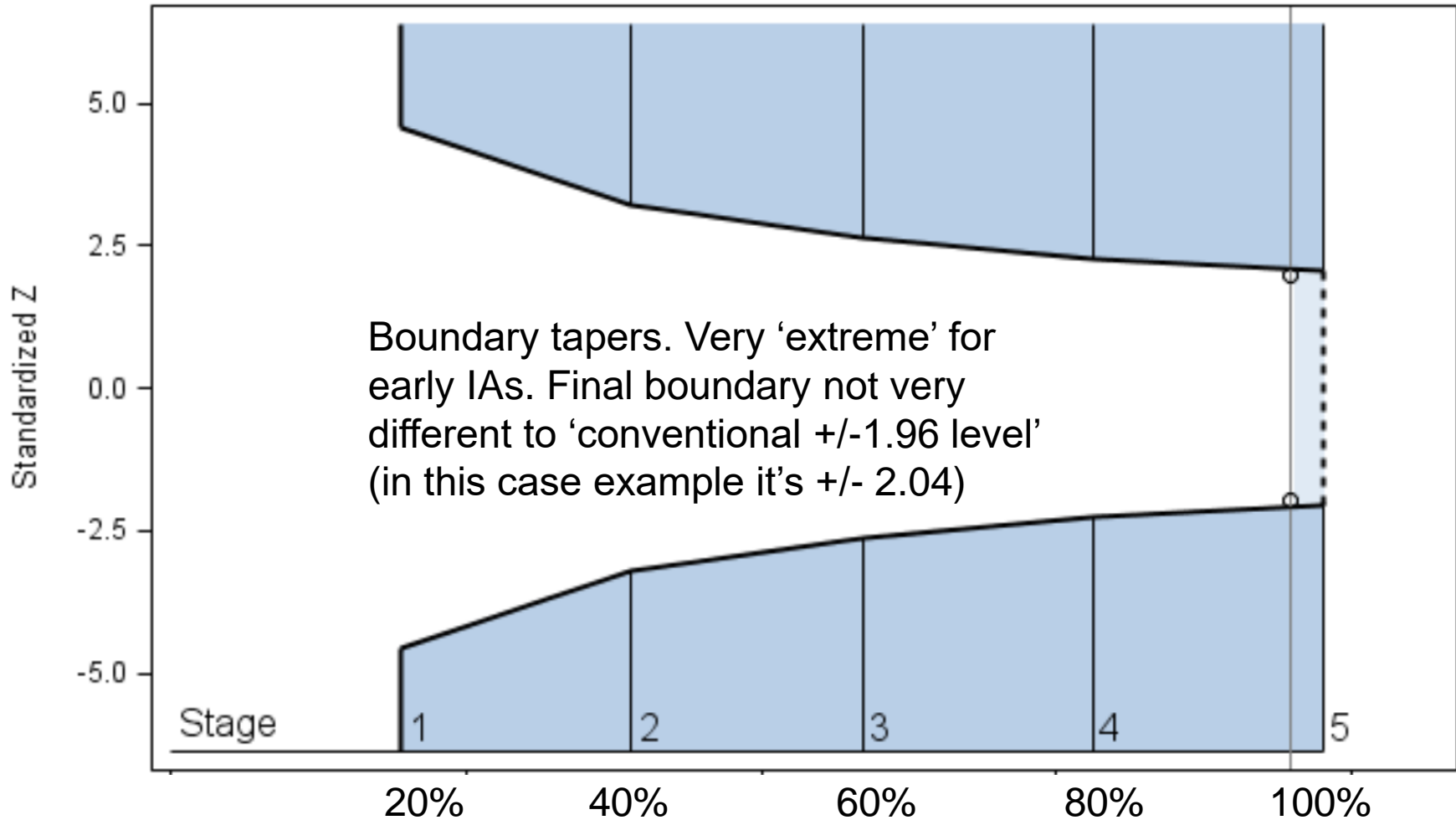
Haybittle-Peto Rule for 5 Looks

$$P(|Z_1|, |Z_2|, |Z_3|, |Z_4| \text{ or } |Z_5| > \text{boundary} | H_0 \text{ true}) = 0.05$$



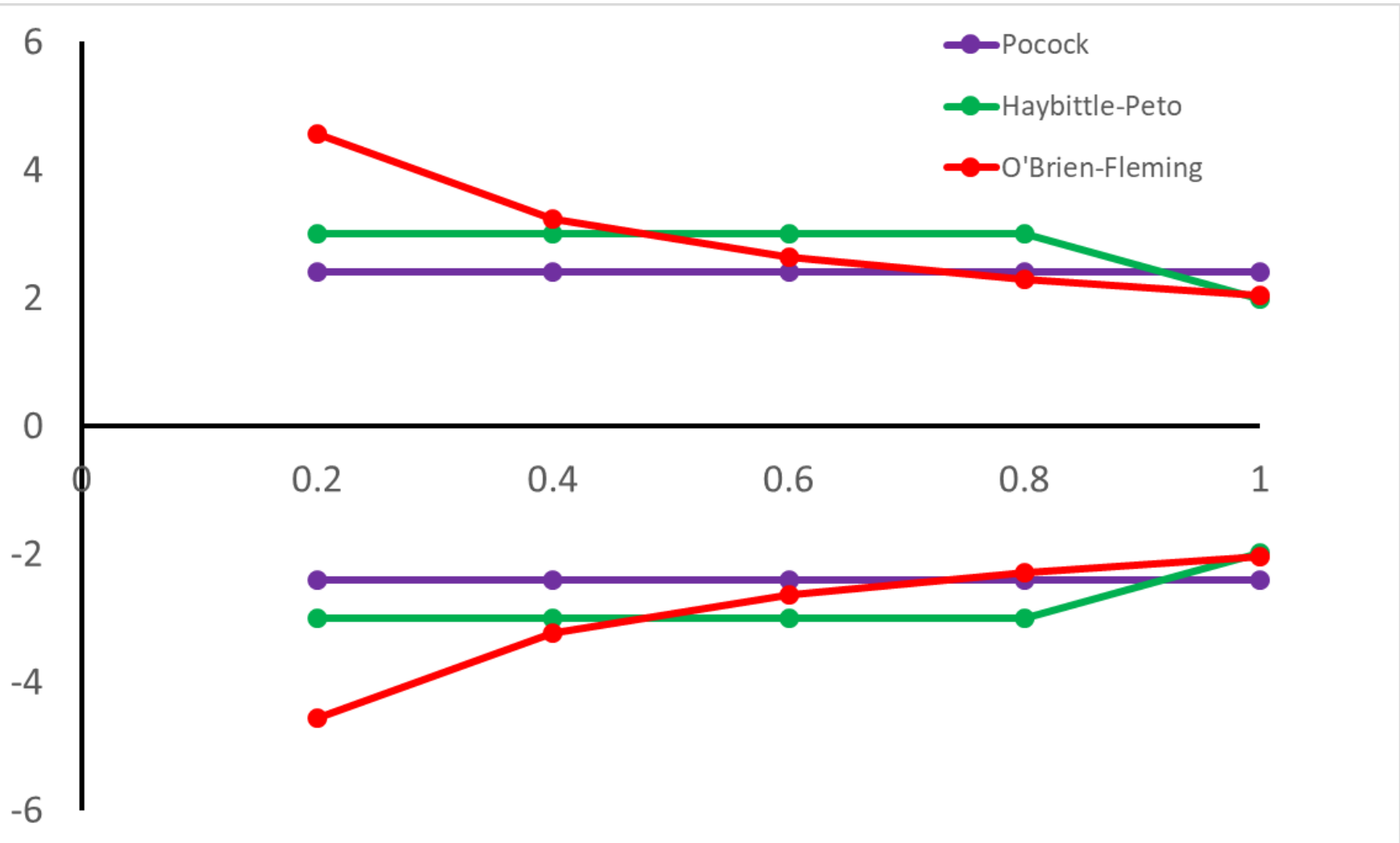
O'Brien-Fleming Rule for 5 Looks

$$P(|Z_1|, |Z_2|, |Z_3|, |Z_4| \text{ or } |Z_5| > \text{boundary} | H_0 \text{ true}) = 0.05$$

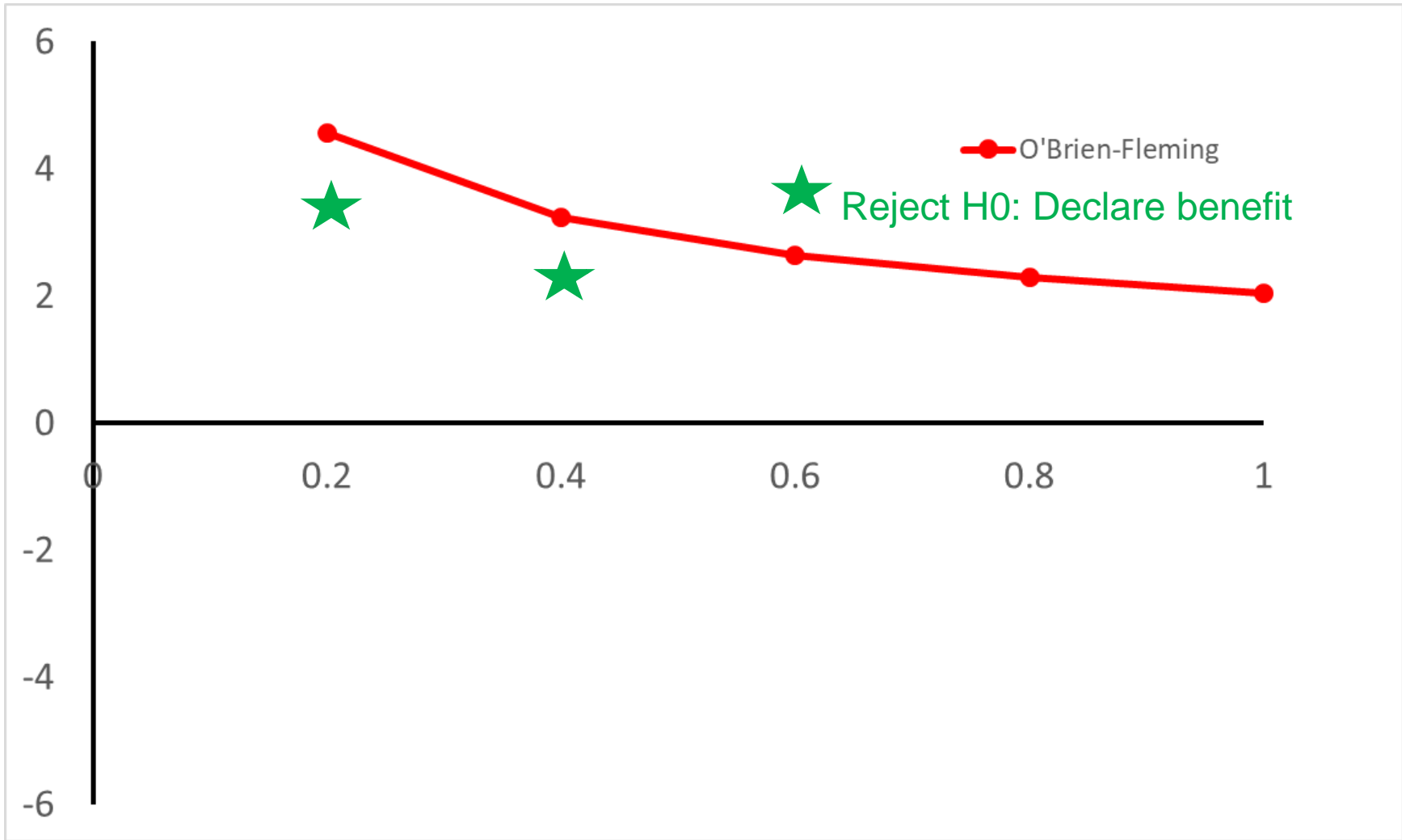


Stopping Boundaries

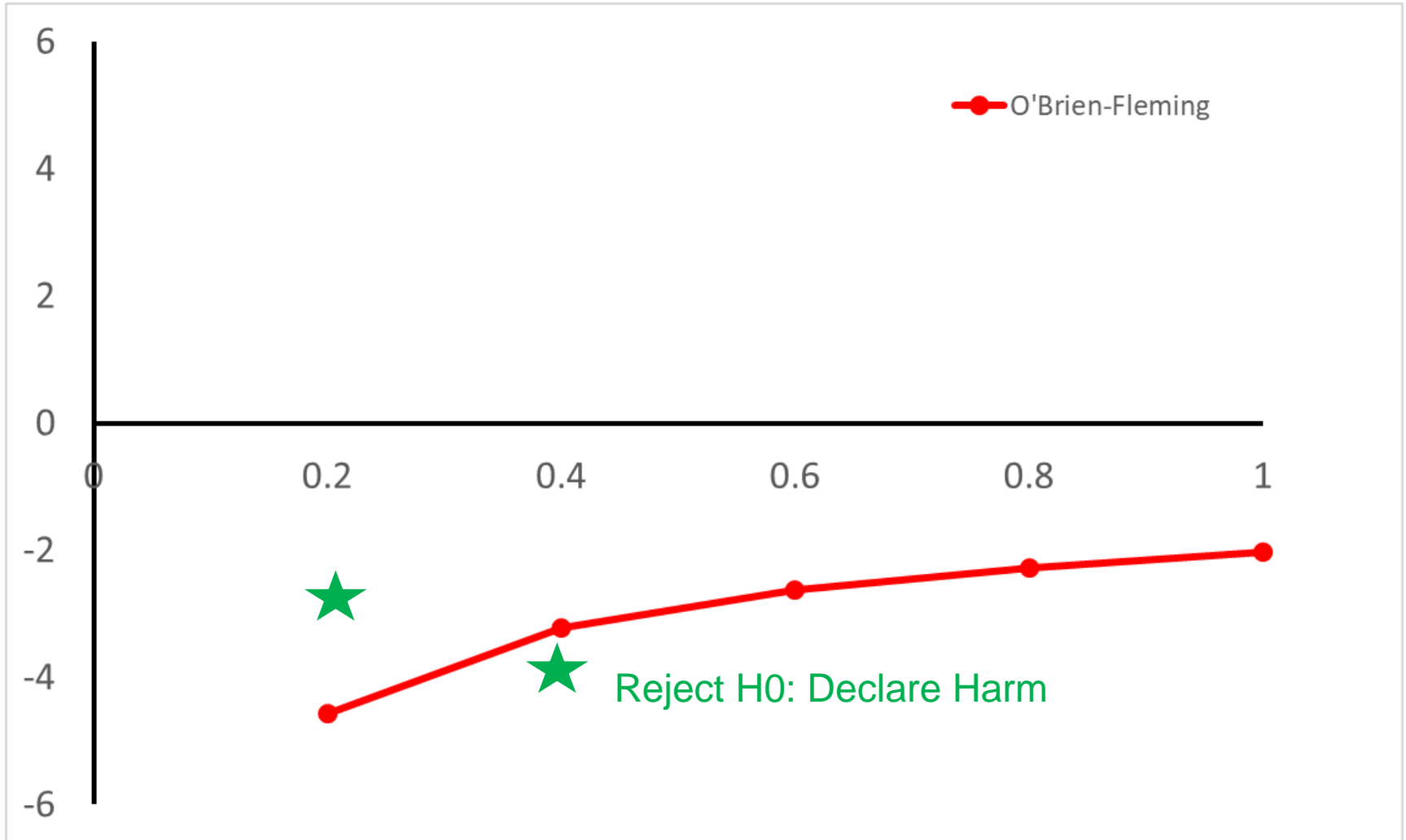
Five interim analyses using an overall 5% significance level



Early Stopping for Benefit



Early Stopping for Harm

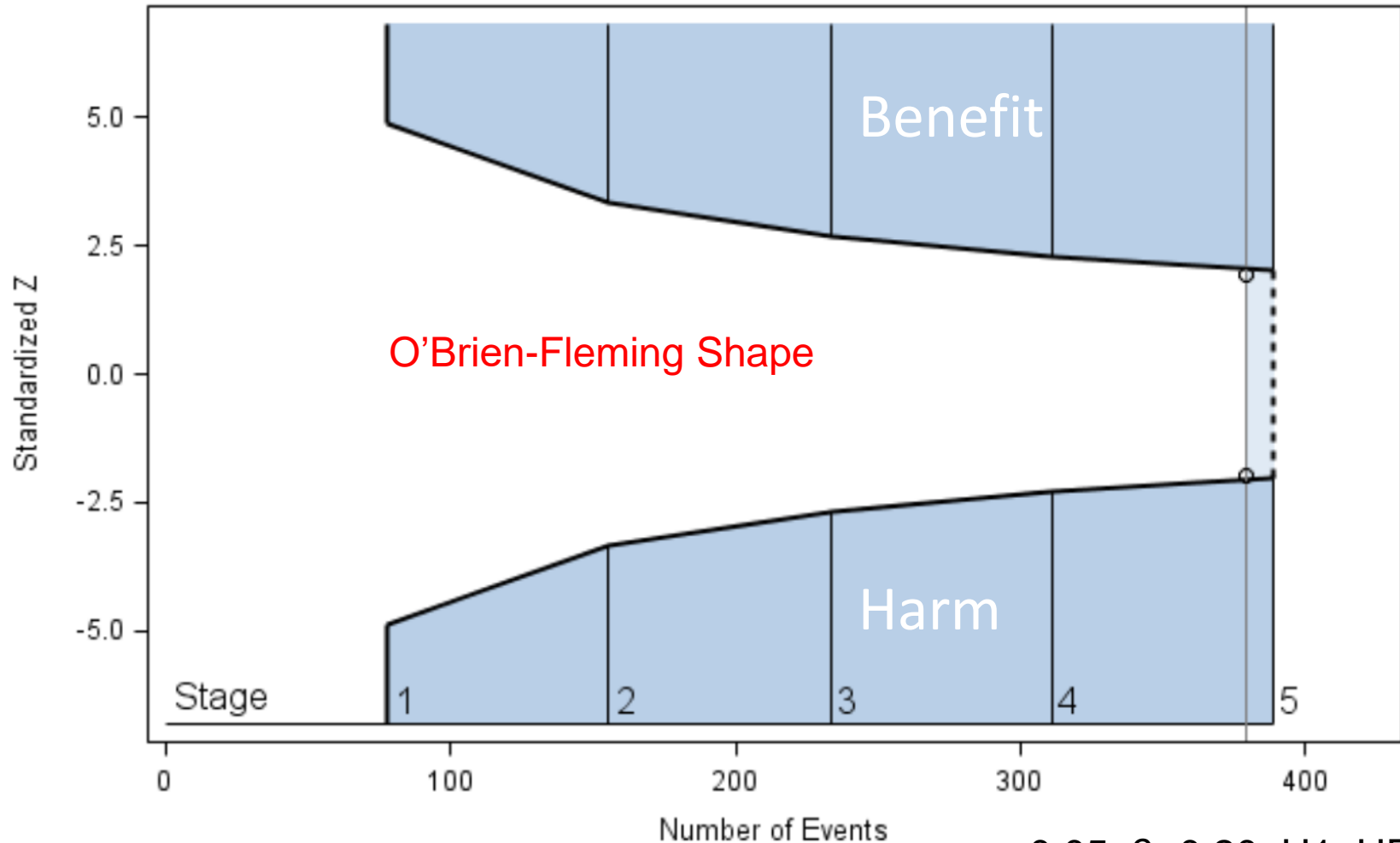


Alpha-Spending

- › Group sequential boundary methods need to be pre-specified (number of analyses, boundary type, timing)
- › **Alpha-spending:** a more flexible method of distributing alpha across the different interim analysis. Unplanned interim analyses can be performed.
- › The function $\alpha(t)$ specifies how the overall 5% false positive rate will be “spent” over time
- › The form of $\alpha(t)$ is chosen so as to provide the desired boundary shape (OB-F, H-B, Pocock, etc.)
- › Importantly, time t = information fraction

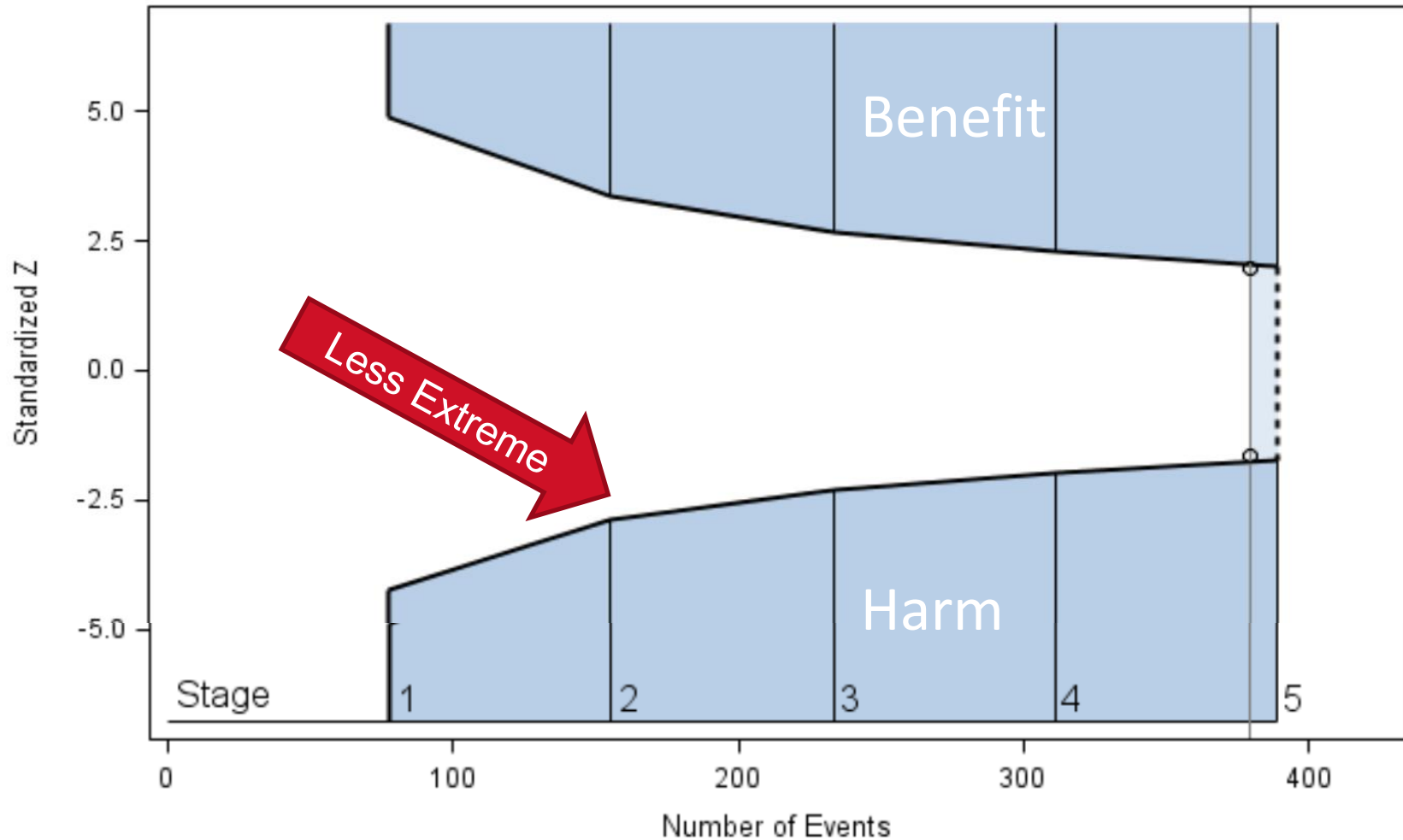
Symmetric Boundaries – Error Spending

Analyses at: 20%, 40%, 60%, 80%, 100%



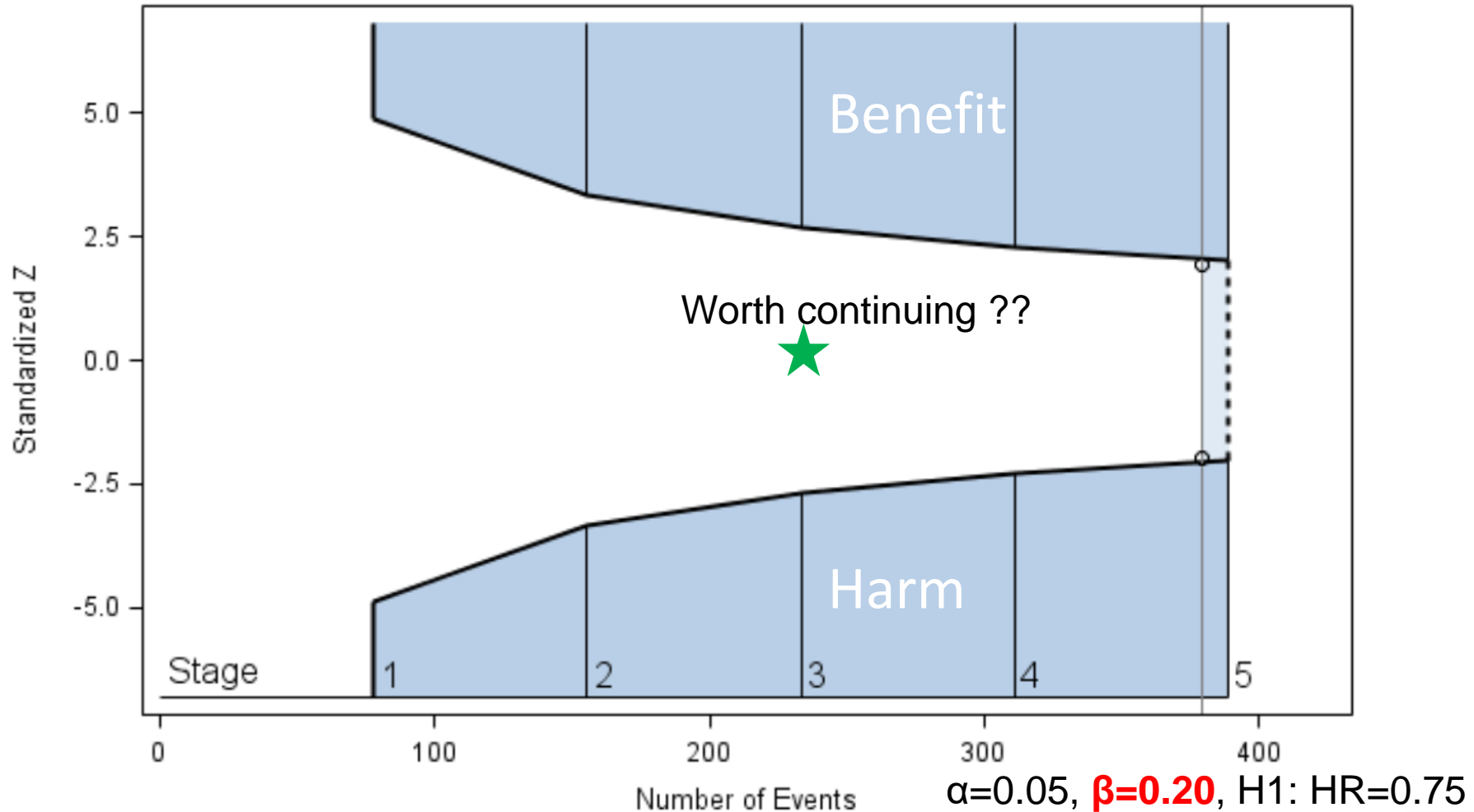
$\alpha=0.05, \beta=0.20, H1: HR=0.75$

Asymmetric boundaries

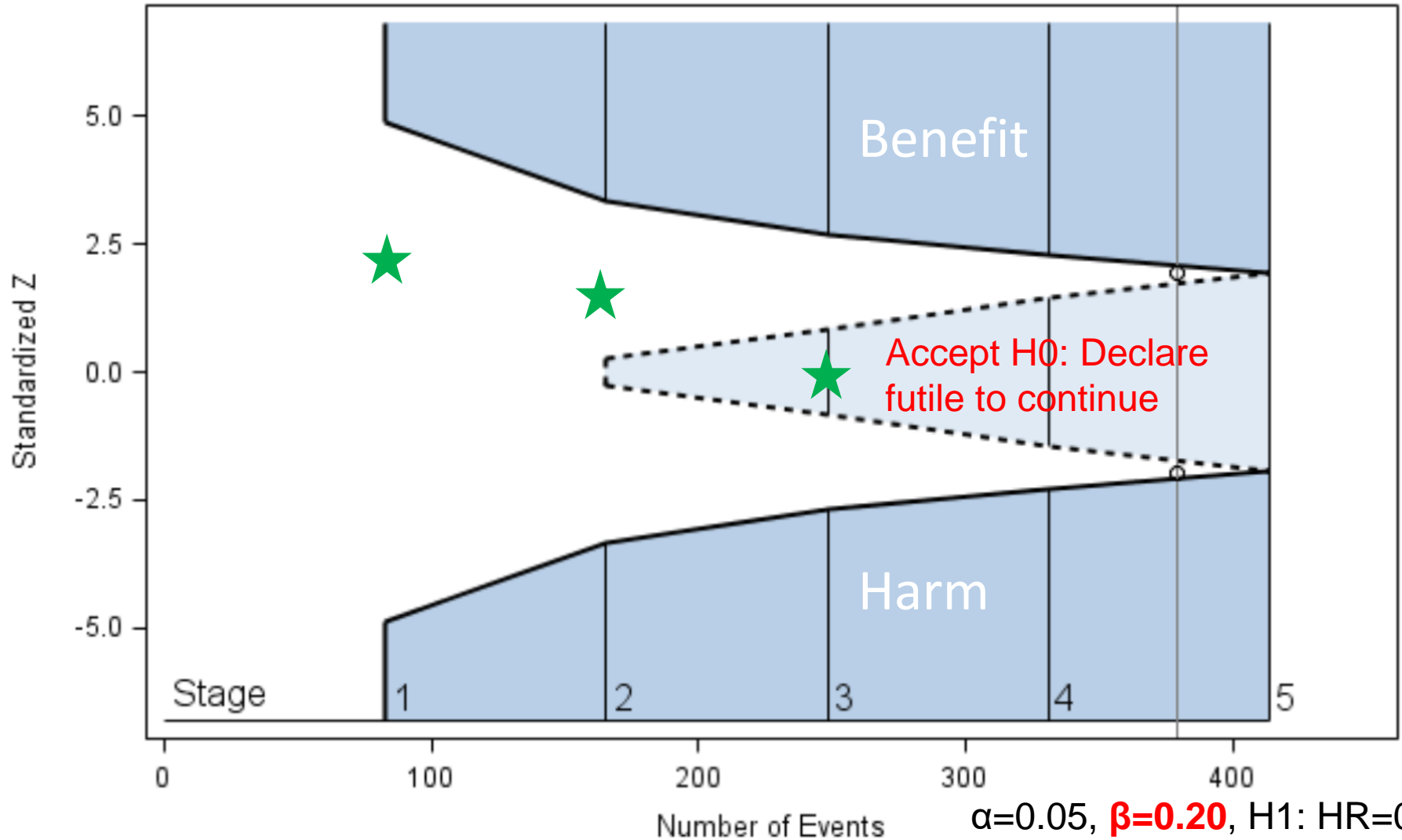


Stopping for Futility

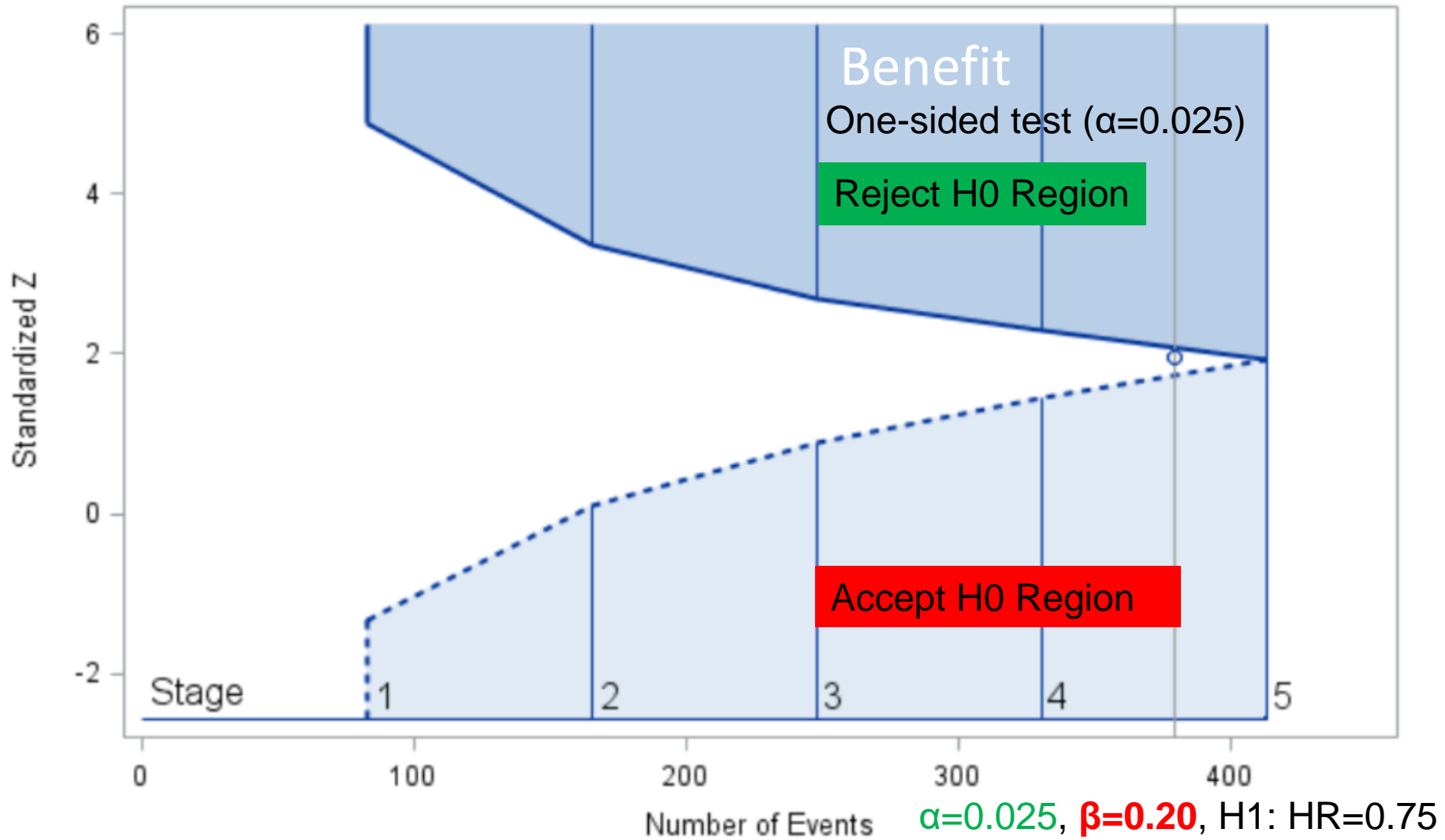
- › What if interim results are inconsistent with (i.e. convincingly lower than) the effect size the trial was powered to detect?



Stopping for Benefit, Harm, or Futility



Stopping for Benefit or Futility



Bayesian Monitoring

- › A Bayesian approach begins with a prior distribution $\pi(\theta)$ for the treatment effect θ
- › This leads to a prior probability for treatment benefit

$$P_0 = \Pr(\theta > 0)$$

- › At interim analysis k , with current data D_k , combine the prior distribution and the current data to produce a current posterior distribution $\pi(\theta | D_k)$ and a posterior probability

$$P_k = \Pr(\theta > 0 | D_k)$$

- › Stop the study at analysis k if P_k exceeds a pre-specified boundary p_k

$$\text{e.g. } p_1 = \dots = p_{K-1} = 0.99$$

- › Interpretation: stop the study early if the probability that the experimental treatment is beneficial exceeds 99%

Bayesian vs Frequentist

- › When we assume a prior distribution for θ with a Bayesian stopping rule

$$P_k > p_k$$

this is equivalent to using a frequentist stopping rule

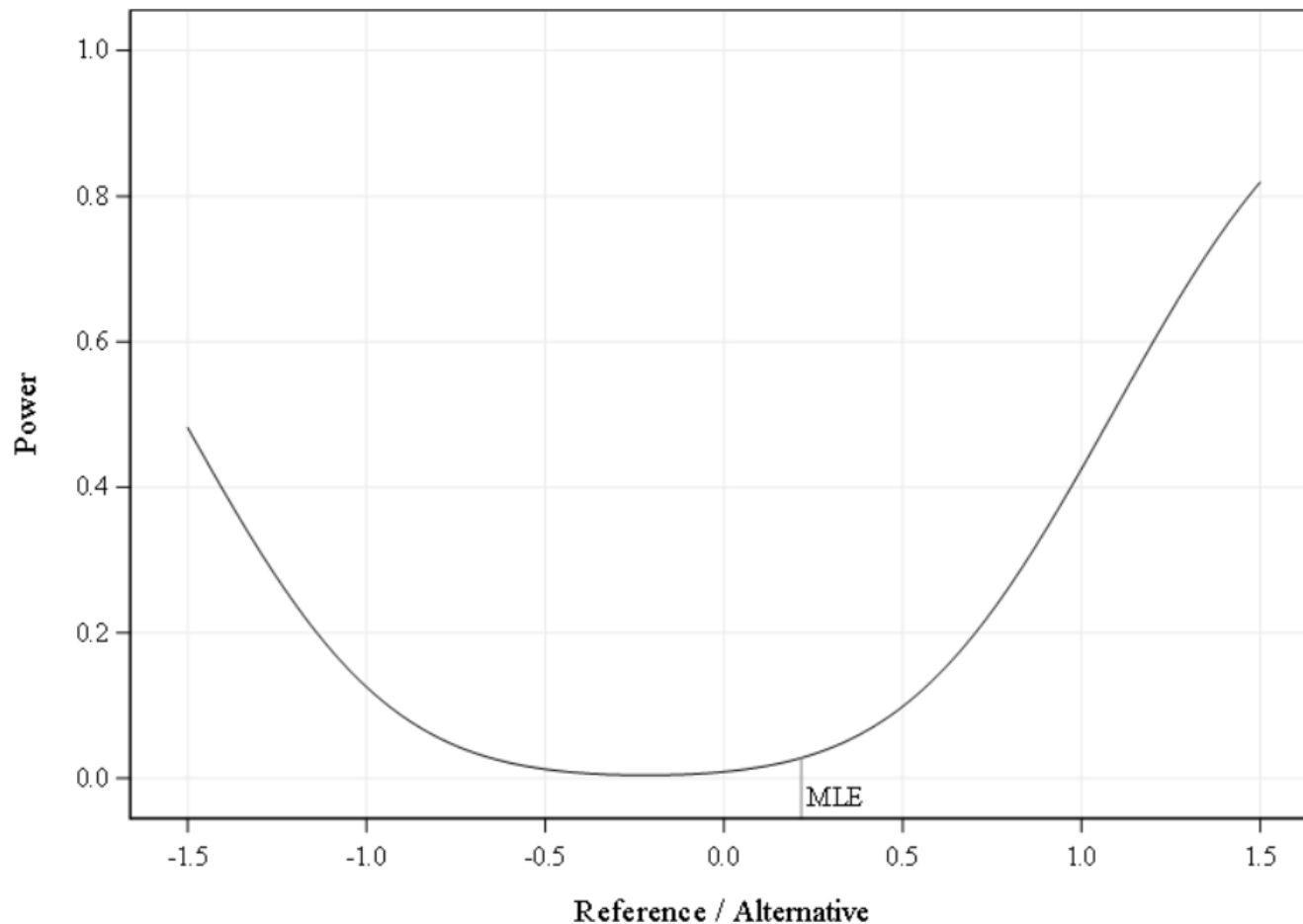
$$Z_k > c_k$$

for appropriate choice of c_k (for normal endpoints or large samples)

- › Pocock monitoring: non-informative prior with $p_1 = \dots = p_K$
- › O'Brien-Fleming monitoring: very informative (pessimistic) prior
- › Bayesian and frequentist monitoring may be effectively equivalent but with different interpretations
- › See e.g. Stallard et al. *BMC Med Res Meth* 2020; 20:4

Conditional Power

- › Conditional Power: probability of rejecting H_0 at the end of the trial, given the interim data
- › Stop trial for futility if sufficiently small e.g. 20%



Bayesian predictive probability

- › In order to use conditional power, we must assume a particular value for the treatment effect (e.g. the current MLE)
- › The Bayesian predictive probability is analogous to conditional power but obviates the need to assume a specific treatment effect by averaging over the current posterior distribution for the treatment effect

- › Conditional power:

$$CP(\theta) = \Pr(\text{reject } H_0 \text{ at the end of the study} \mid \text{interim data}; \theta)$$

- › Bayesian predictive probability:

$$BPP = E_{\theta} [\Pr(\text{reject } H_0 \text{ at the end of the study} \mid \text{interim data}; \theta)]$$

(where the expectation is over the current posterior for θ)

- › $CP(\theta)$ and BPP are used in a similar way to stop the study for futility

Conclusion

- › IDSMC monitors data for benefits and hazards, looking for sufficiently persuasive evidence to stop the trial
- › Statistics provides guidelines rather than rules that should be considered along side other information, which demands an element of judgement from an IDSMC
- › This includes external evidence from other trials and an assessment as to whether interim results would be persuasive enough for clinicians to change practice
- › Statistics cannot provide absolute answers but can substantially reduce the chance of an incorrect decision