Unblinded sample-size reassessment in time-to-event clinical trials

Dominic Magirr[1,*], Thomas Jaki[2], Franz König[1] and Martin Posch[1]

[1]Institut für Medizinische Statistik, Medical University of Vienna
[2]Department of Mathematics and Statistics, Lancaster University
∗current affiliation: Astrazeneca, Cambridge, U.K.

September 2016

Supported by the Austrian Science Fund (FWF):P23167 and EU FP7: "Asterix" 603160

Outline

Lung cancer trial (Schäfer and Müller, 2001)

- Patients randomized to "Radiotherapy + Chemotherapy" (E) or "Chemotherapy" (C)
- Median survival on C $\approx$ 14 months
- Anticipated survival on E $\approx$ 20 months
- Sample size: 255 events ($\alpha = 0.025$, $\beta = 0.2$)
- Exponential model ... this could be achieved with 40 months recruitment and 20 months min follow-up.

40 months into the trial...

(a) patient recruitment was much slower than expected
  - only 136 patients had been randomized

(b) the hazard rate had been over-estimated in the planning
  - only 56 events had been observed

Recommendation of Schäfer and Müller:

"abandon the trial because there [is] no chance of achieving the planned sample size within a reasonable time"

Counterproposal of study group

- Look at the data to see if there is a larger treatment effect than originally anticipated.
- If so, reduce the initially planned sample size (required number of events).
- Larger the observed treatment effect $\rightarrow$ earlier the study ends.

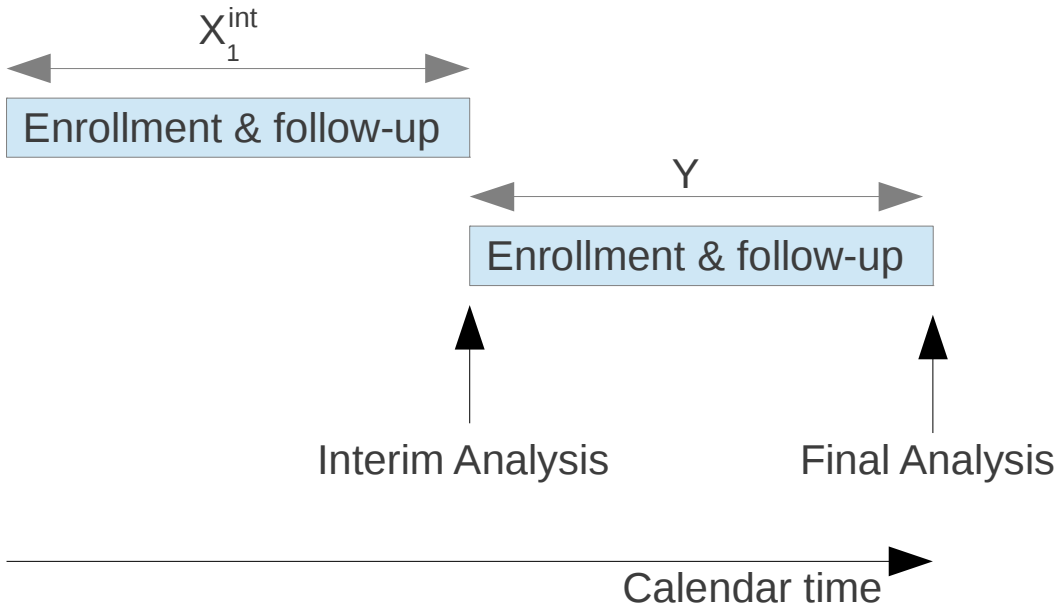
A closely related scenario

Proschan & Hunsburger (1995); Irle & Schäfer (2012)

- Look at data to see if there is a smaller than anticipated treatment effect.
- If so, increase the sample size (required number of events) to give a better chance of achieving a statistically significant result.

Standard analysis will not control the type I error rate...

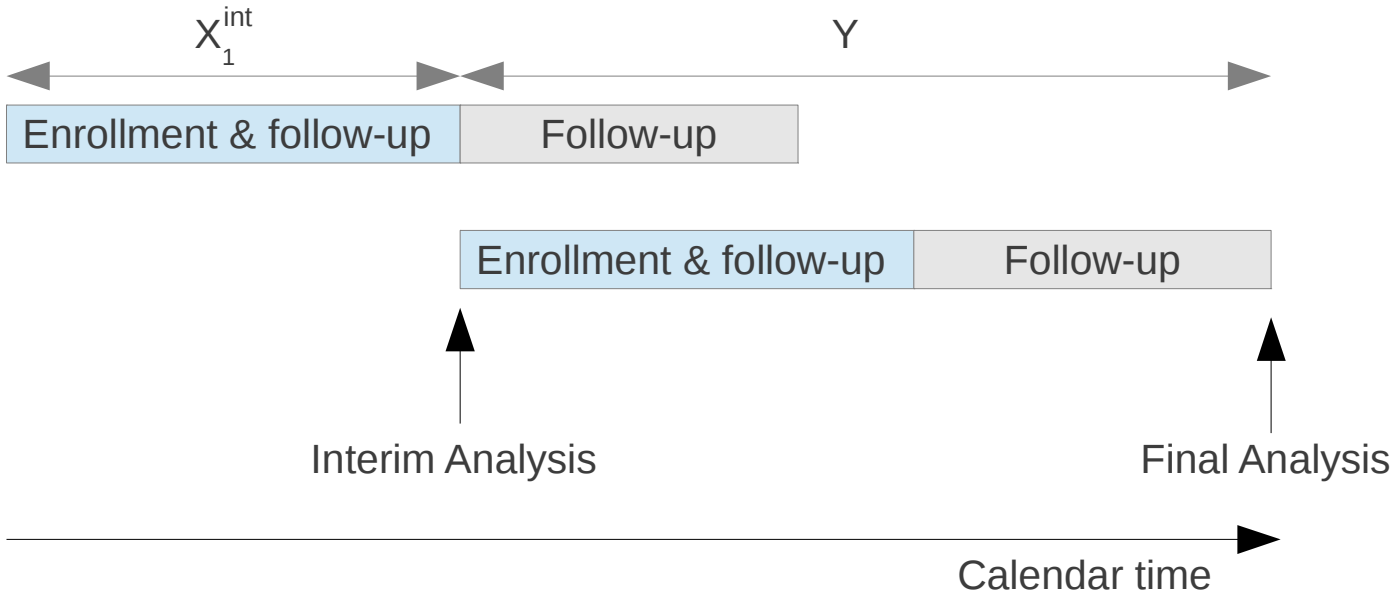
Adaptive design with immediate responses



E.g., under H_0,

$$\frac{1}{\sqrt{2}}\Phi^{-1}\left\{1-p_1(X_1^{\text{int}})\right\}+\frac{1}{\sqrt{2}}\Phi^{-1}\left\{1-p_2(Y)\right\}\sim\mathcal{N}(0,1)$$

Adaptive design with delayed responses



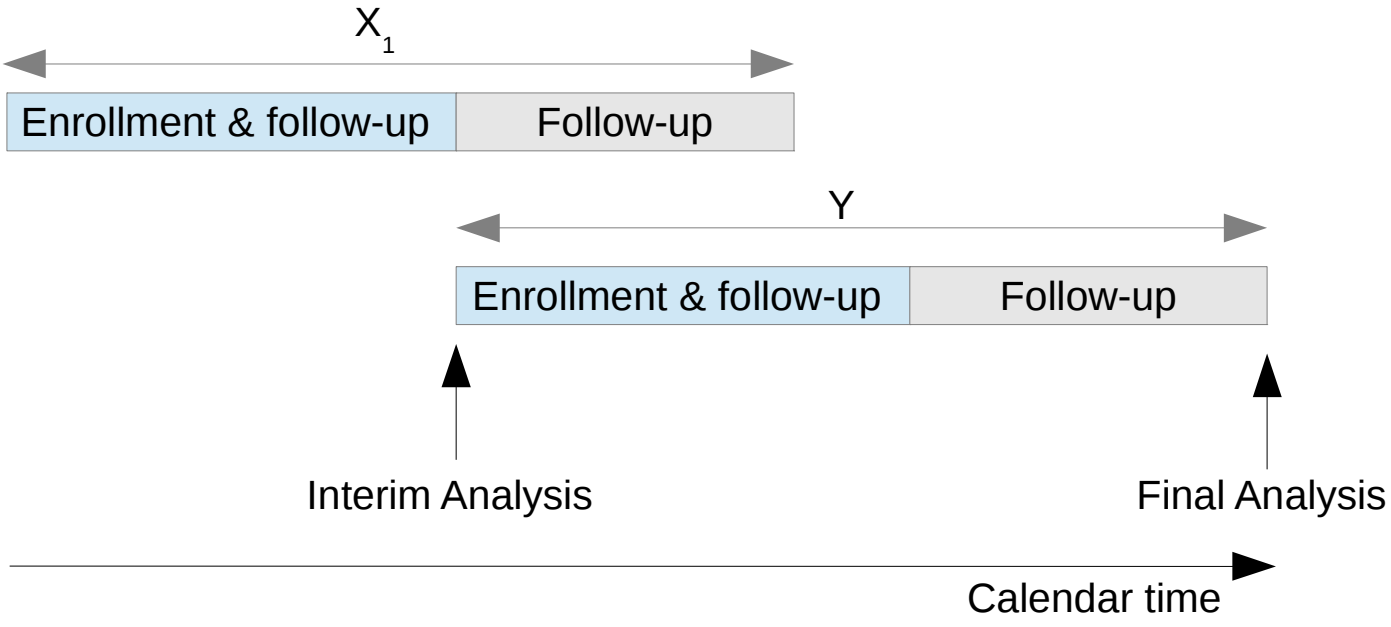
$$\frac{1}{\sqrt{2}}\Phi^{-1}\left\{1-p_1(X_1^{\text{int}})\right\}+\frac{1}{\sqrt{2}}\Phi^{-1}\left\{1-p_2(Y)\right\} \overset{?}{\sim} \mathcal{N}(0,1)$$

When is this valid?

- ✓ Interim decision strategy based solely on (primary endpoint) treatment effect estimate.
- × Interim decisions are based on partial information from patients who are yet to provide full primary endpoint response

  e.g. second-stage sample size is chosen on basis of progression-free survival when primary endpoint is overall survival.

Potential solution



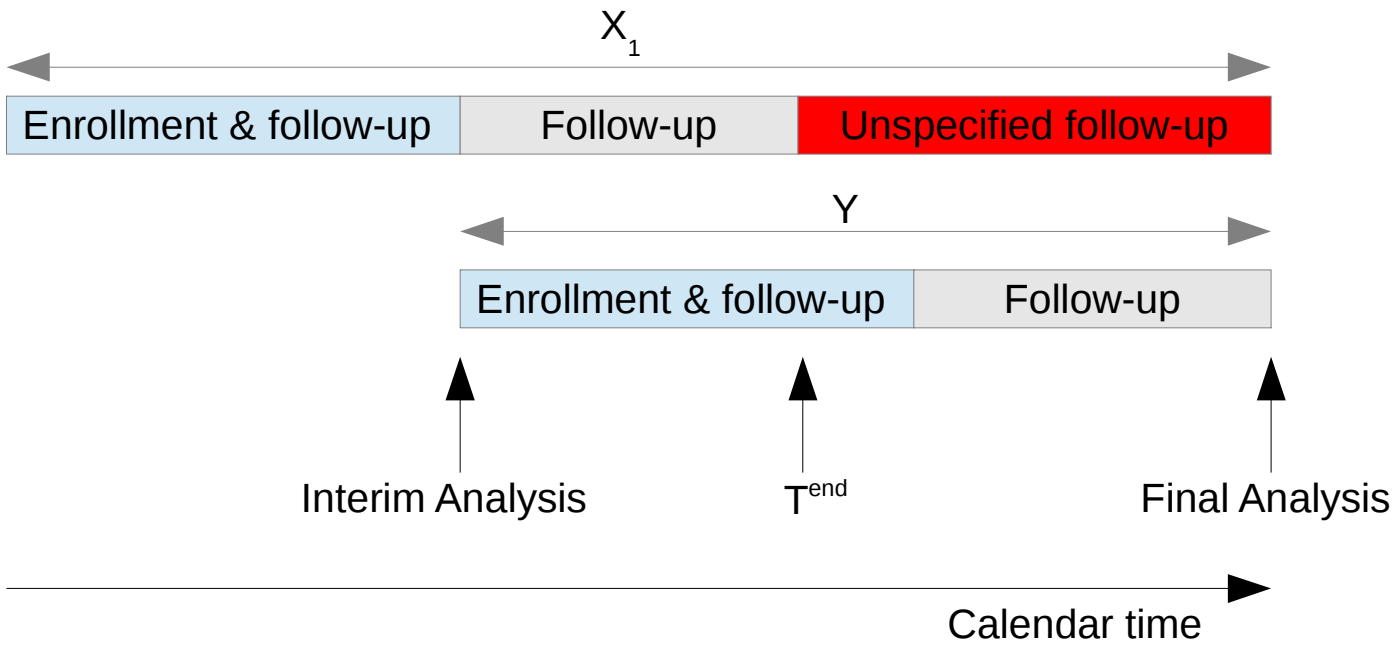
$$\frac{1}{\sqrt{2}}\Phi^{-1}\{1-p_1(X_1)\}+\frac{1}{\sqrt{2}}\Phi^{-1}\{1-p_2(Y)\}\sim\mathcal{N}(0,1)$$

e.g., Liu & Pledger (2005) – Gaussian responses

Schmidli, Bretz & Racine-Poon (2007) – Binary responses

Extra problem with time-to-event endpoint?

Jenkins, Stone & Jennison (2011); Irle & Schäfer (2012)



- Must pre-specify end of follow-up of first-stage patients, T^{end}, in definition of p_1.
- Otherwise, $p_1(X_1) \overset{\times}{\sim} U[0,1]$ under H_0, and type I error may be inflated.

Some survival times are ignored

- Final test decision only depends on a subset of the recorded survival times; part of the observed data is ignored.
- Particularly damaging if long-term survival is of most concern (it is the survival times of earliest recruited patients that is ignored).
- Therefore, we (Magirr et al., 2016) investigated the effect of naïvely incorporating this illegitimate data into the final test statistic...

Adaptive log-rank test

"Correct" adaptive test statistic

$$Z^{\text{CORRECT}} = w_1 L_1(T^{\text{end}}) + w_2 \Phi^{-1}(1 - p_2)$$

"Naïve" adaptive test statistic

$$Z^{\text{NAIVE}} = w_1 L_1(T^*) + w_2 \Phi^{-1}(1 - p_2)$$

- $L_1(t)$ is the log-rank statistic based on Stage 1 patients, followed up until calendar time t.
- w_i are explicitly (Jenkins et al.) or implicitly (Irle & Schäfer) fixed weights with $w_1^2 + w_2^2 = 1$.
- T^{end} is the (implicitly) fixed end of first-stage follow up.
- T^* is the time of final analysis (dependent on interim decisions).

Worst-case assumption

- The null distribution of Z^{CORRECT} is $\mathcal{N}(0,1)$.
- The null distribution of Z^{NAIVE} is completely unknown.
- However, we can look at the stochastic process

$$Z(t) = w_1 L_1(t) + w_2 \Phi^{-1}(1 - p_2), \qquad t \in [T^{\text{end}}, T^{\text{max}}].$$

Worst-case: the interim data (PFS, early endpoints, etc) can be used to predict exactly when $L_1(t)$ reaches its maximum.

Upper bound on type I error

An upper bound can be found assuming second-stage design is engineered such that T^* coincides with $\arg\max L_1(t)$:

$$\max \alpha = P_{H_0}\left\{ \max_{t \geq T^{\text{end}}} w_1 L_1(t) + w_2 \Phi^{-1}(1-p_2) > 1.96 \right\}$$

$$= \cdots$$

$$\approx \int_0^1 P_{H_0}\left[\max_{u=u_1}^{1} B(u) > \sqrt{u}\frac{\left\{1.96 + w_2\Phi^{-1}(x)\right\}}{w_1} \right] \mathrm{d}x,$$

with
$u_1 = \left\{\#\text{ stage 1 events at } T^{\text{end}}\right\} / \left\{\#\text{ stage 1 events at } T^{\max}\right\}$

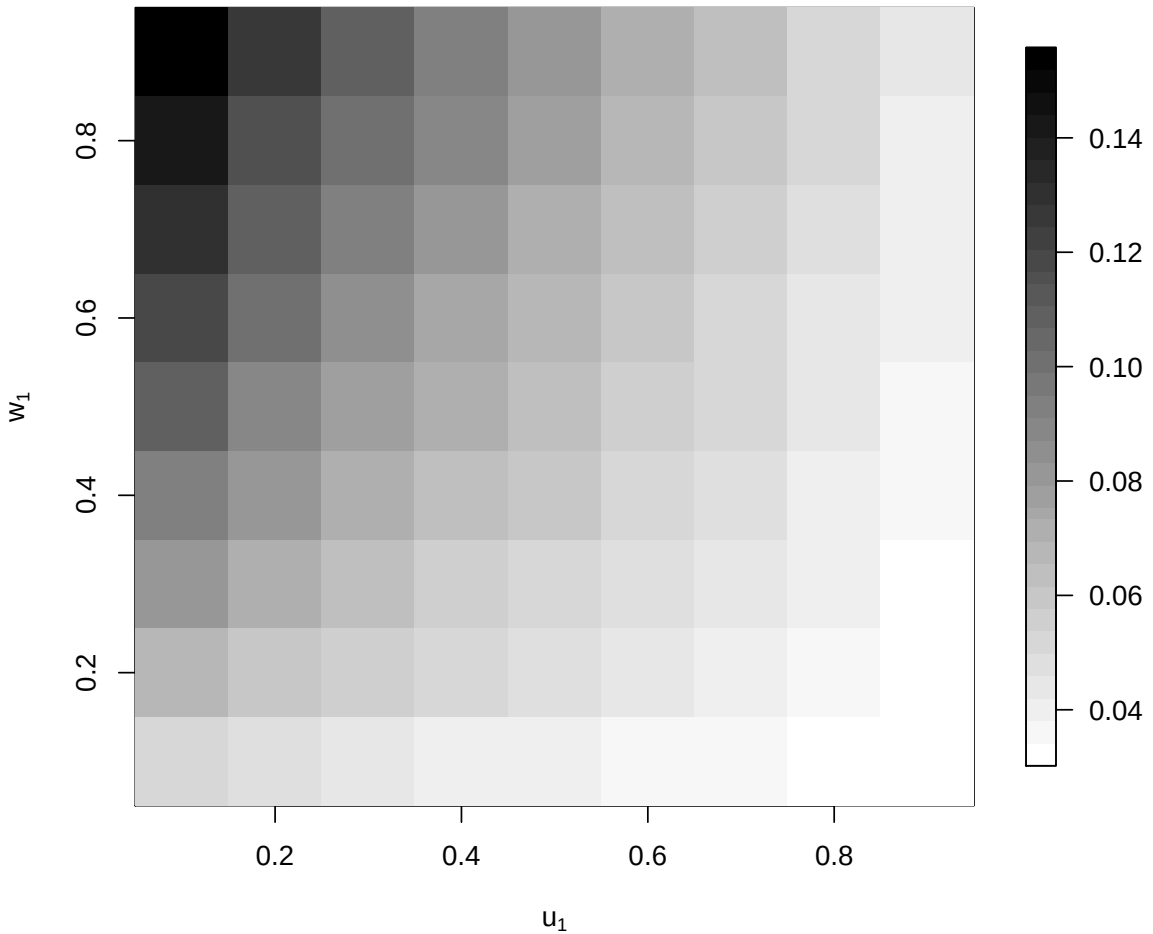
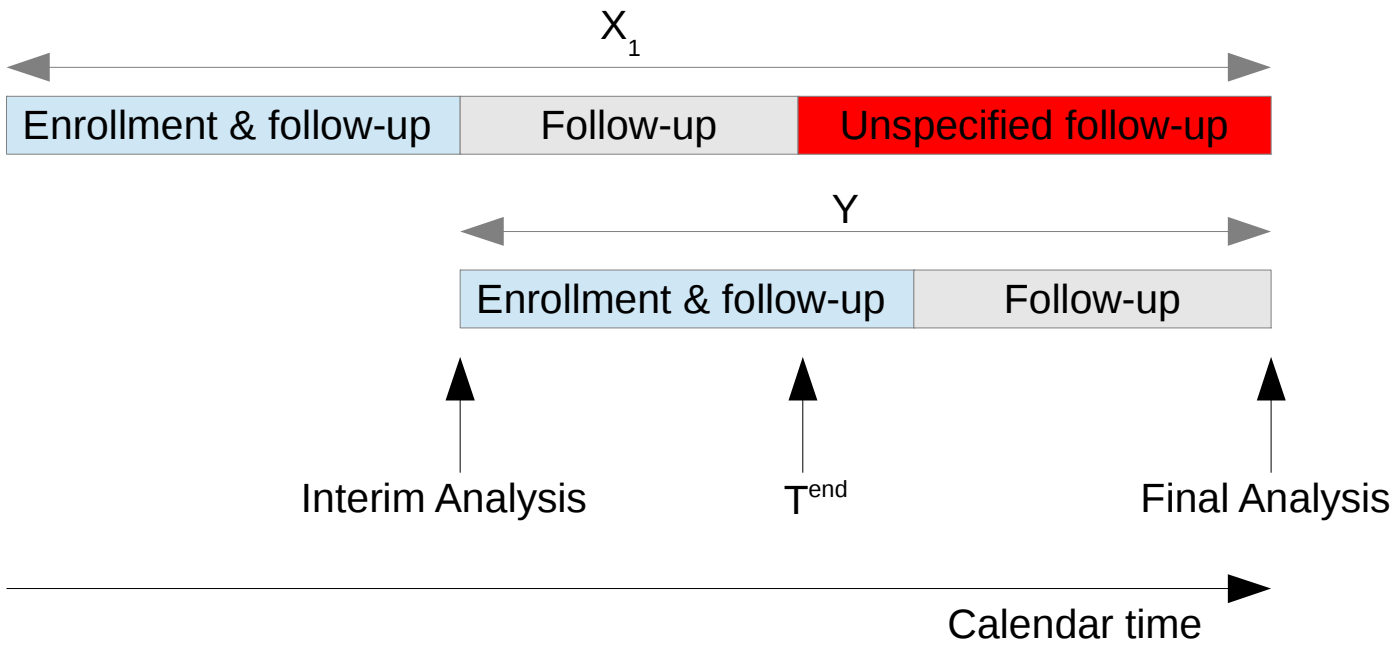


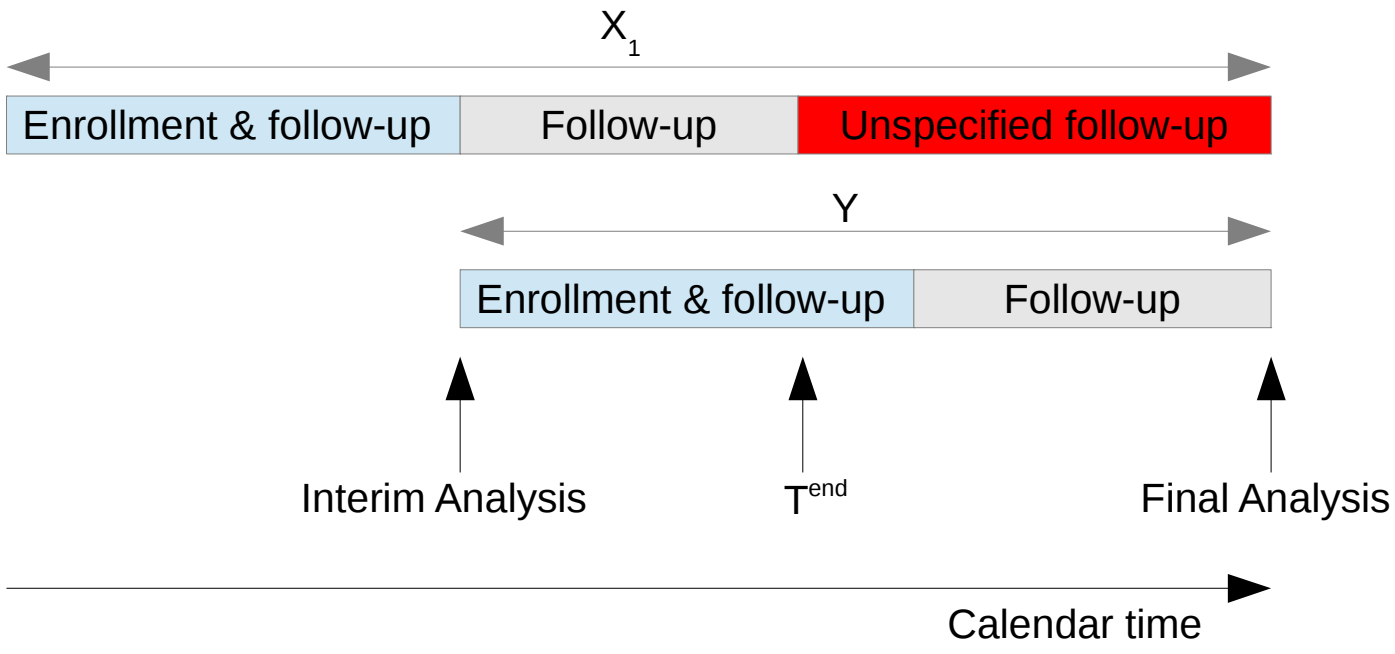
Figure : Worst case type I error for various choices of weights and information fractions.

Choice of w_1 (Jenkins et al.)



- $T^{\text{end}} = \min\{t : \#\text{ stage 1 events } = d_1\}$
- w_1 is fixed in advance.
- E.g., $w_1^2 = d_1/(d_1 + \tilde{d}_2)$, where $\tilde{d}_2$ is the *anticipated* number of stage 2 events at time T^{end}.

Choice of w_1 (Irle & Schäfer)



- $T^{\text{end}} = \min\{t : \text{ total } \# \text{ events } = d\}$ and

$$w_1^2 = \frac{\#\text{ stage 1 events at time } T^{\text{end}}}{\text{total } \#\text{ events at time } T^{\text{end}}}.$$

Choice of w_1

The advantage of the Irle & Schäfer choice of weights is that if the trial concludes as planned after observing d events, then the adaptive test statistic is the same as the standard logrank test statistic (efficient).

Additionally, the timing of the interim analysis need not be pre-specified.

The disadvantage is that it is not possible to change the recruitment rate following the interim analysis.

A comment on u_1

As well as the weight w_1, the extent of the maximum type I error rate also depends on u_1, which is

$$u_1 = \left\{ \# \text{ stage 1 events at } T^{\text{end}} \right\} / \left\{ \# \text{ stage 1 events at } T^{\text{max}} \right\}$$
$$\approx \left\{ \# \text{ stage 1 events at } T^{\text{end}} \right\} / \left\{ \# \text{ patients recruited by interim} \right\}.$$

So, roughly speaking:

Faster recruitment $\rightarrow$ lower u_1 $\rightarrow$ higher max α.

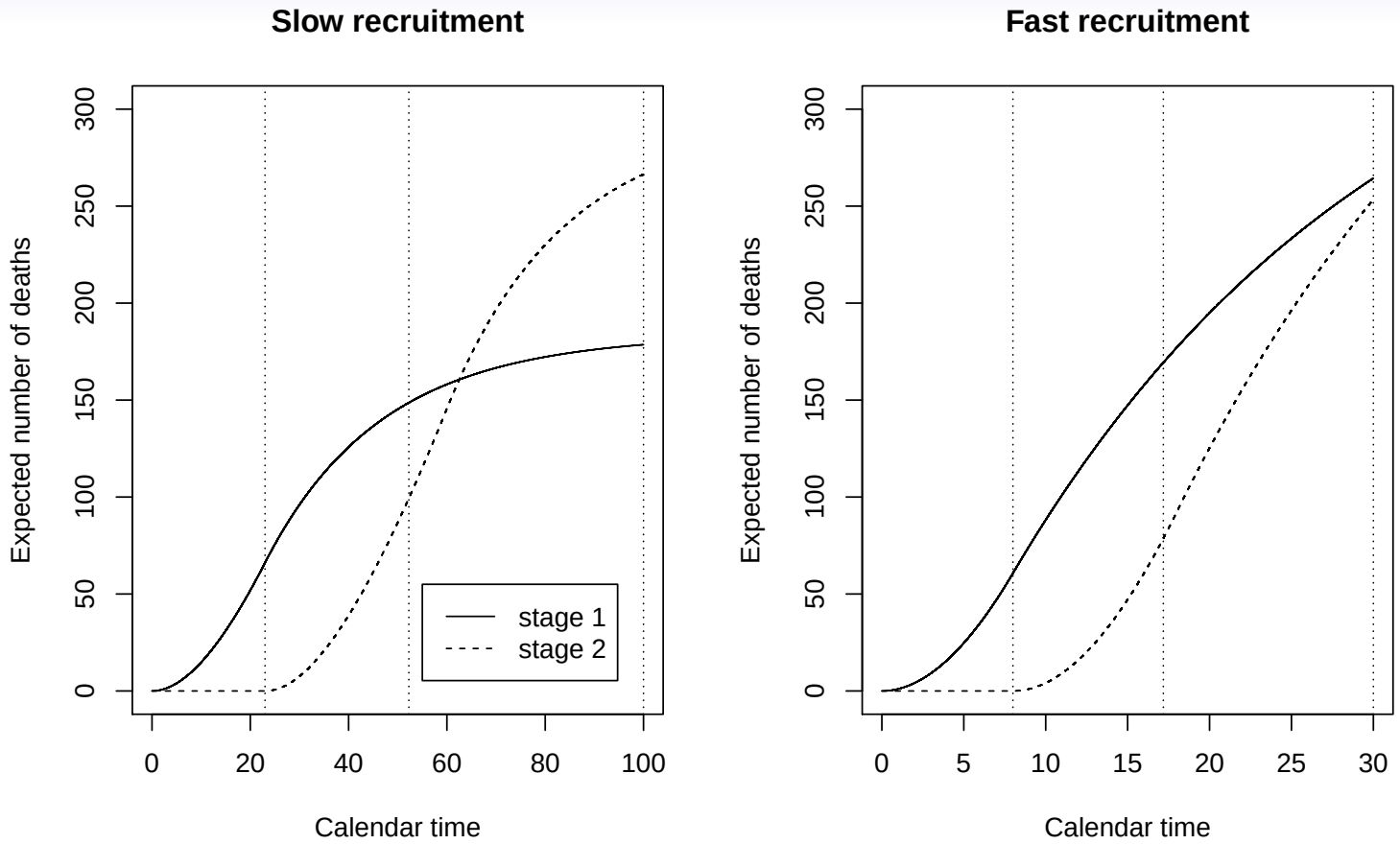


Figure : Expected total number of events as a function of time based on exponential survival with hazard rates $\lambda_C = 0.05$ and $\lambda_E = 0.035$. Slow recruitment: 8 patients per month for a maximum of 60 months; $\max \alpha = 0.035$. Fast recruitment: 50 patients per month for a maximum of 18 months; $\max \alpha = 0.045$. Vertical lines are at T^{int}, T^{end} and $T^{\max}$.

Guaranteed level-α test

Simply increase the cut-off value k^* such that $P_{H_0}\{\max_{t \geq T_1} Z(t) \geq k^*\} = \alpha$.

Table : Cutoff values for corrected level-0.025 test.

		u_1								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	0.1	2.29	2.25	2.21	2.19	2.16	2.13	2.11	2.08	2.04
	0.2	2.41	2.35	2.31	2.27	2.23	2.20	2.16	2.12	2.07
	0.3	2.50	2.43	2.38	2.34	2.30	2.25	2.21	2.16	2.10
	0.4	2.58	2.50	2.44	2.39	2.34	2.30	2.25	2.19	2.12
w_1	0.5	2.64	2.56	2.49	2.44	2.38	2.33	2.27	2.21	2.14
	0.6	2.70	2.60	2.53	2.47	2.42	2.36	2.30	2.23	2.15
	0.7	2.74	2.64	2.57	2.51	2.45	2.39	2.33	2.26	2.17
	0.8	2.79	2.68	2.60	2.54	2.48	2.41	2.35	2.28	2.18
	0.9	2.83	2.72	2.64	2.57	2.50	2.43	2.37	2.29	2.19

Power of the guaranteed level-α test

When we use Z^{NAIVE} in place of Z^{CORRECT} our statistic in increased by

$$Z(T^*) - Z(T^{\text{end}}) = w_1 \left\{ L_1(T^*) - L_1(T^{\text{end}}) \right\}$$

and the α-level cut-off value is increased by

$$k^* - \Phi^{-1}(1-\alpha).$$

The relative power of the guaranteed level-α test (compared to the "correct" adaptive test) depends on which of these differences is larger...

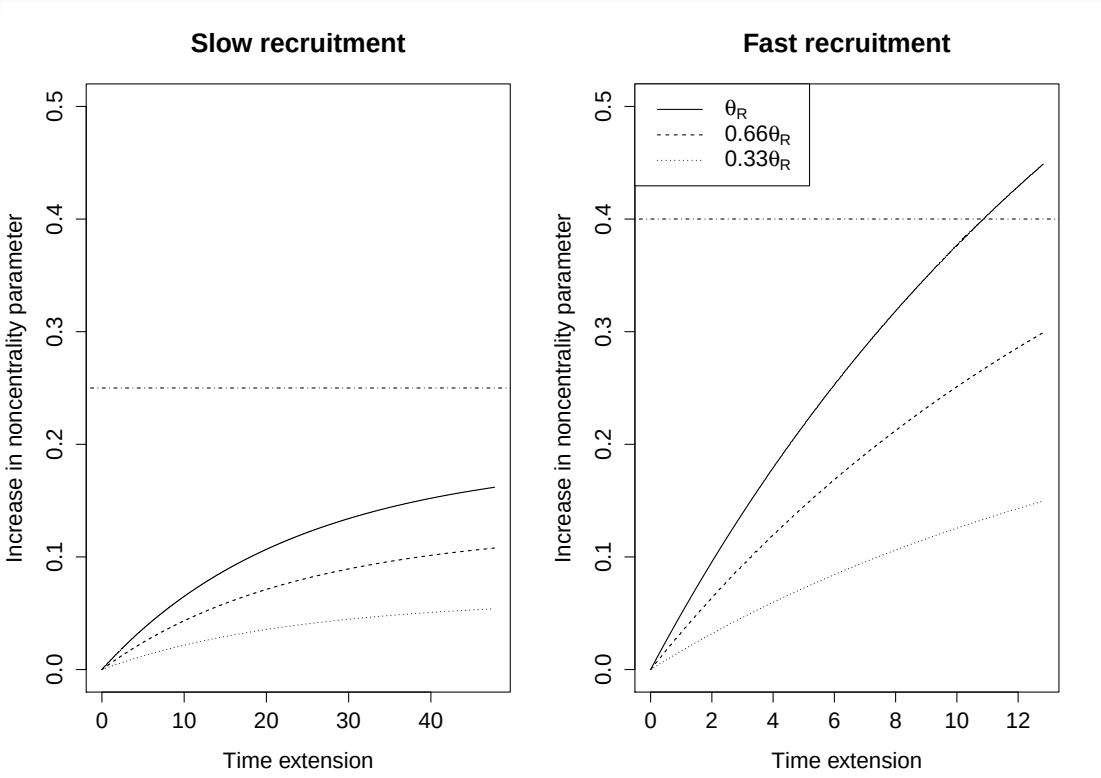


Figure : Difference between the noncentrality parameters of the adaptive test statistics $Z(T^*)$ and $Z(T^{\text{end}})$ as a function of the time extension $T^* - T^{\text{end}} \in \left[0, T^{\max} - T^{\text{end}}\right]$. Horizontal lines are drawn at $k^* - \Phi^{-1}(0.975)$, where k^* denotes the cut-off value of the alternative level-α test.

Unblinded SSR: methods trade-off

	Type I control	Informed interim decisions	All survival times in test	Relative power
"Independent increments"	✓	×	✓	✓
"Correct" adaptive	✓	✓	×	✓
"Naïve" adaptive	×	✓	✓	✓
"Naïve" + k^*	✓	✓	✓	×

References

1. Magirr, D., Jaki, T., König, F., and Posch, M. (2016) Sample Size Reassessment and Hypothesis Testing in Adaptive Survival Trials *PLoS ONE* 11(2).
2. Schäfer, H., & Müller, H. H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in medicine*, 20(24), 3741-3751.
3. Irle, S., & Schäfer, H. (2012). Interim Design Modifications in Time-to-Event Studies. *Journal of the American Statistical Association*, 107(497), 341-348.
4. Liu, Q., & Pledger, G. W. (2005). Phase 2 and 3 combination designs to accelerate drug development. *Journal of the American Statistical Association*, 100(470).

References

5. Schmidli, H., Bretz, F., & Racine-Poon, A. (2007). Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Statistics in medicine*, 26(27), 4925-4938.
6. Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*, 10(4), 347-356.
7. Proschan, M.A. & Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51(4), 1315-1324.
8. Mehta, C., Schäfer, H., Daniel, H., & Irle, S. (2014). Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine*. 33(26), 4515-4531.

dominic.magirr@astrazeneca.com