METRUM RESEARCH GROUP

An Issue in Controlling the Type I Error Rate with Adaptive Designs Using Time-to-Event Endpoints

James A. Rogers, Ph.D.

Metrum Research Group LLC

April 28, 2010

3rd Annual FDA/MTLI Medical Device and IVD Statistics Workshop

Recent Planning for Phase 3 Device Trial

- What we know about the **event rate**:
  - Some components of our (composite, time-to-event) endpoint have been studied in similar populations but,
  - Sufficient differences in our endpoint definition & population $\Rightarrow$ event rate only **"known" $\pm 2-3$ percentage points**.
- What we know about the **effect size**:
  - Phase 2 endpoint widely believed to have a predictive and causal relationship with Phase 3 endpoint but,
  - No clear quantitative link in the population of interest $\Rightarrow$ **magnitude of treatment effect highly uncertain**.
  - Fixed design based on minimum clinically meaningful effect could potentially be over-sized.

Group Sequential (GS) Approach versus Unblinded Sample Size Re-estimation

A My initial reaction was: “We can probably meet our right-sizing objectives using a group sequential design”.

Group Sequential (GS) Approach versus Unblinded Sample Size Re-estimation

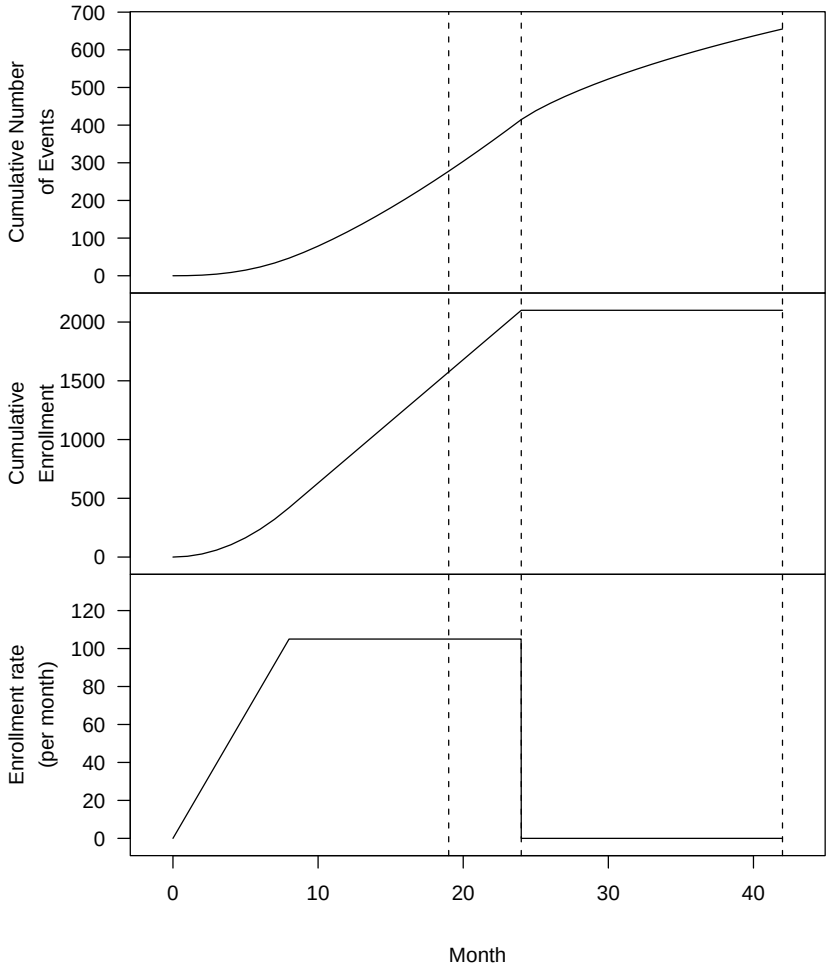
A My initial reaction was: “We can probably meet our right-sizing objectives using a group sequential design”.

B My final recommendation was: “We can meet our right-sizing objectives using a group sequential design”.

Group Sequential (GS) Approach versus Unblinded Sample Size Re-estimation

- A My initial reaction was: “We can probably meet our right-sizing objectives using a group sequential design”.
- B My final recommendation was: “We can meet our right-sizing objectives using a group sequential design”.
- A $\rightarrow$ B Along the path from A to B, we looked at other approaches, including non-group-sequential designs using unblinded sample size re-estimation. Reasons for considering this included:
  - GS approach helps to “right-size” the trial with respect to number of events but,
  - GS not designed to “right-size” the trial with respect to enrollment (enrollment could easily complete prior to crossing of a group sequential boundary).

Predicted Enrollment and Information Accrual



Opportunity to make reasonably informed decision prior to full enrollment?

Conditional Error Approach to Unblinded Sample Size Re-estimation for Time-to-Event Endpoint

STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:3741–3751 (DOI: 10.1002/sim.1136)

Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections‡

Helmut Schäfer*,† and Hans-Helge Müller

Institute of Medical Biometry and Epidemiology, Philipps-University Marburg, Bunsenstraße 3, D-35037 Marburg, Germany

SUMMARY

A method is presented which allows us to adapt the sample size as well as the number and time points of interim analyses to the treatment difference observed at an interim look during the course of a clinical trial with censored survival time as the endpoint. The method allows the inclusion of data inspections during the course of the trial and redesign of the trial on the basis of the observed treatment difference without affecting the type I error risk. Formulae for recalculating the required number of events and the number of further patients to be randomized as a function of the observed hazard rates and the detectable hazard ratio are given.

Notation for Evolution of Logrank Statistic Over Time

$j = 1, \ldots, J(t)$: indices of ordered event times (relative to randomization times), as evaluated at absolute time t

$O_j(t)$: indicator of whether j^{th} event (as evaluated at absolute time t) occured in the control group

$N_j^C(t),\ N_j^T(t)$: numbers at risk for control and treatment arms just prior to event j (as evaluated at absolute time t)

$E_j(t) = N_j^C(t)/(N_j^C(t) + N_j^T(t))$: probability that event j occurs in control group, under the null hypothesis

$S(t) = \sum_{j:1,\ldots,J(t)} (O_j(t) - E_j(t))$: logrank score at time t

$V(t)$: estimated variance of logrank score at time t

Schäfer and Müller Result in Context of Simple Design with Only One Potential Adaptation

Consider a design that is initially specified to continue until time t^F. Asymptotically, we know:

$$P_{H_0}\left(Z(t^F) > z_{1-\alpha}\right) = \alpha, \text{ where } Z(t^F) = S(t^F)/V(t^F)$$

Suppose that at interim time t^I we observe $Z(t^I) = z^I$, and we calculate:

$$P_{H_0}\left(Z(t^F) > z_{1-\alpha} \;\middle|\; Z(t^I) = z^I\right) = \alpha^*$$

Result: Any design modification that preserves this conditional Type I error rate at level α^* will preserve the unconditional (overall) Type I error rate at level α.

Asymptotic Joint Distribution of Increments to Logrank Statistic

Let:

$$Z_1 = Z(t^I) \;\; ; \;\; Z_2 = \frac{S(t^F) - S(t^I)}{\sqrt{V(t^F) - V(t^I)}}$$

Then, under the null hypothesis, Z_1 and Z_2 have independent Standard Normal distributions. Therefore:

$$\begin{aligned} & P_{H_0}\left(Z_2 > z_{1-\alpha^*} \mid Z_1 \equiv Z(t^I) = z^I\right) \\ = \quad & P_{H_0}\left(Z_2 > z_{1-\alpha^*}\right) = \alpha^* \end{aligned}$$

Consequently, following the interim analysis we may modify the design and replace the originally specified primary analysis with the test that rejects if $Z_2 > z_{1-\alpha^*}$.

An Important Caveat

STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:1333–1335 (DOI: 10.1002/sim.1759)

LETTER TO THE EDITOR

Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections

by H. Schäfer and H.-H. Müller, *Statistics in Medicine* 2001; **20**:3741–3751

From: *P. Bauer and M. Posch*
Department of Medical Statistics
University of Vienna

The Caveat of Bauer and Posch

Let D represent information, beyond that contained in $Z(t^I)$, that is available to the interim decision maker. Then the relevant conditional error probability at the interim is

$$P_{H_0}\left(Z_2 > z_{1-\alpha^*} \;\middle|\; Z(t^I) = z^I \text{ and } D = d\right)$$

and this is not necessarily equal to α^*!

If the true conditional error probability is less than α^*, then the final analysis test that rejects if $Z_2 > z_{1-\alpha^*}$ will lead to an inflated Type I error rate overall.

In particular, Bauer and Posch note that knowledge of covariate values at the interim could lead to such Type I error inflation.

My Initial Interpretation of the Preceding Literature

From conversation between me and myself, ca. Oct. 2009:

"We can structure the adaptation rule any way we want, as long as:

- the final analysis uses the logrank test that rejects if $Z_2 > z_{1-\alpha^*}$, and
- we stay blind to covariate values at the interim analysis. "

(I later realized this wasn't necessarily the whole story)

So Then I Thought . . .

Use interim decision criteria based on a parametric Bayesian analysis?

- We have good reason to believe that a Weibull model will be adequate.
- We have sufficient basis for a reasonably informative prior for the Weibull shape parameter.
- A decision that leverages this information must be better than a decision just based on the interim logrank statistic.

Recent Precedent for a Similar Approach

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1445–1463
Published online 5 March 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3559

Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology

Werner Brannath[1], Emmanuel Zuber[2,*,†], Michael Branson[2], Frank Bretz[2,3], Paul Gallo[4], Martin Posch[1] and Amy Racine-Poon[2]

[1]*Medical University of Vienna, Vienna, Austria*
[2]*Novartis Pharma AG, Basel, Switzerland*
[3]*Department of Biometry, Medical University of Hannover, 30623 Hannover, Germany*
[4]*Novartis Pharmaceuticals, East Hanover, U.S.A.*

SUMMARY

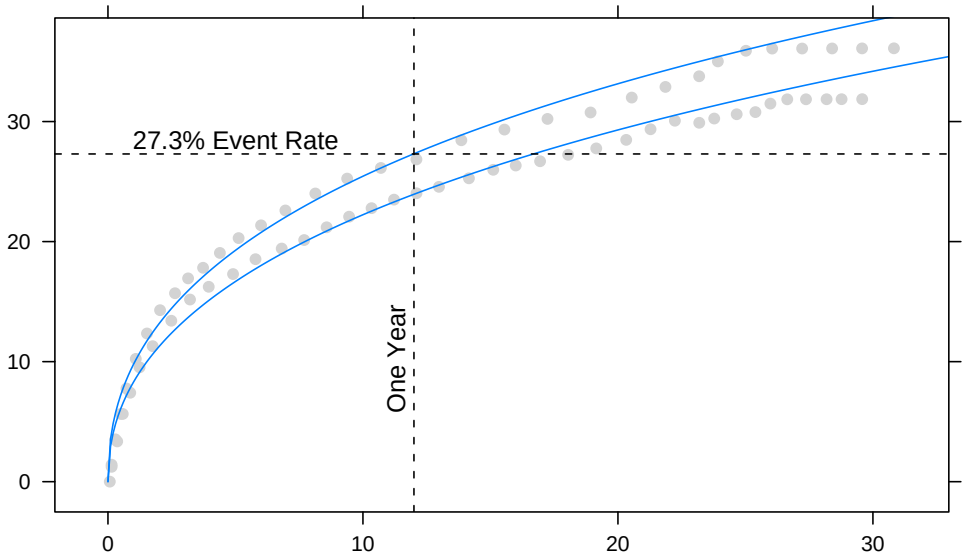
The ability to select a sensitive patient population may be crucial for the development of a targeted therapy. Identifying such a population with an acceptable level of confidence may lead to an inflation in development time and cost. We present an approach that allows to decrease these costs and to increase the reliability of the population selection. It is based on an actual adaptive phase II/III design and uses Bayesian decision tools to select the population of interest at an interim analysis. The primary endpoint is assumed to be the time to some event like e.g. progression. It is shown that the use of appropriately stratified logrank tests in the adaptive test procedure guarantees overall type I error control also when using information on patients that are censored at the adaptive interim analysis. The use of Bayesian decision tools for the population selection decision making is discussed. Simulations are presented to illustrate the operating characteristics of the study design relative to a more traditional development approach. Estimation of treatment effects is considered as well. Copyright 2009 John Wiley & Sons, Ltd.

Adaptations in that case were for subset selection, but the methodology was fundamentally similar:

- Bayesian interim analysis
- Frequentist methods to control Type I error (they used a combination p-value, which is closely related to the conditional error approach).

Plausible Placebo Event Incidence for the Phase 3 Endpoint I Was Working With

Weibull model fit to digitized data from published K-M curves:



Note: high initial hazard followed by lower hazard (typical of survival patterns following acute events)

A Potential Problem?

Let:

T_i	be the (possibly unobserved) event time for subject i
$C_i(t)$	be the administrative censoring time for subject i at absolute time t

For subjects still in the risk set at time t^I, $T_i > C_i(t^I)$. This knowledge is leveraged in a parametric interim analysis and, for non-exponential event times:

$$P\left((T_i - C_i(t^I)) \leq x \;\middle|\; T_i > C_i(t^I)\right) \neq P(T_i \leq x)$$

So, could interim knowledge of $C_1(t^I), \ldots, C_J(t^I)$ introduce a risk of inflating the Type I error?

What I Still Don't Know For Sure

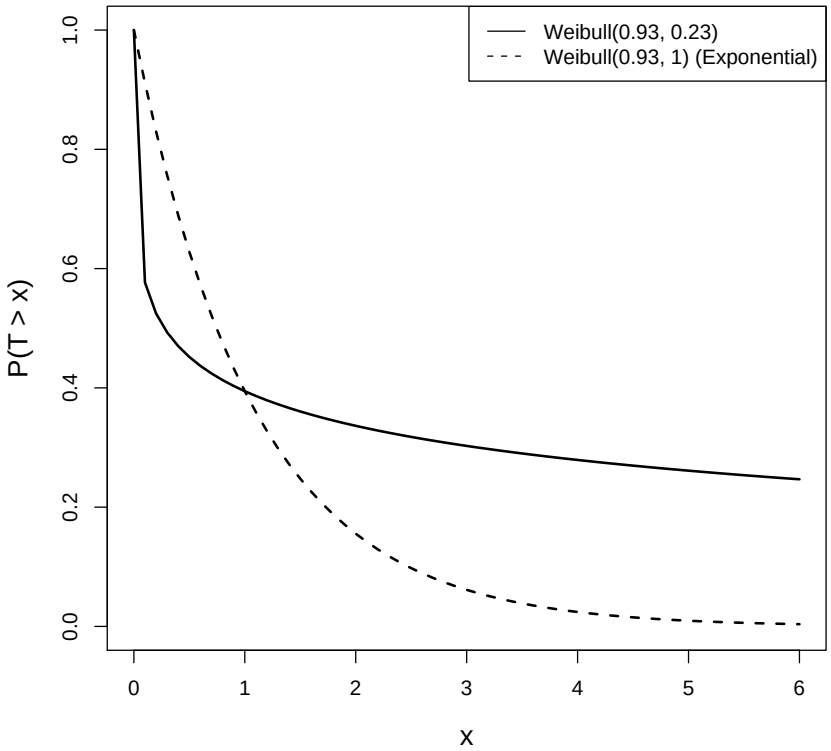
Is it the case that the increment to the logrank statistic Z_2 satisfies:

$$P\left(Z_2 > z_{1-\alpha^*} \;\middle|\; Z_1,\ \underline{C_1(t^I), \ldots, C_J(t^I)}\right) \quad = \quad P(Z_2 > z_{1-\alpha^*}) \ ?$$

- Intuitively, I would not expect the equality to hold, since Z_2 is a function of event times that can be partially predicted by $C_1(t^I), \ldots, C_J(t^I)$.
- Theoretically, I can't seem to prove one way or the other if it holds (maybe others can).
- In the cases I have simulated, the equality does in fact seem to hold.

Trying to Find an Example Where Z_2 Can Be Partially Predicted Using Interim Censoring Times

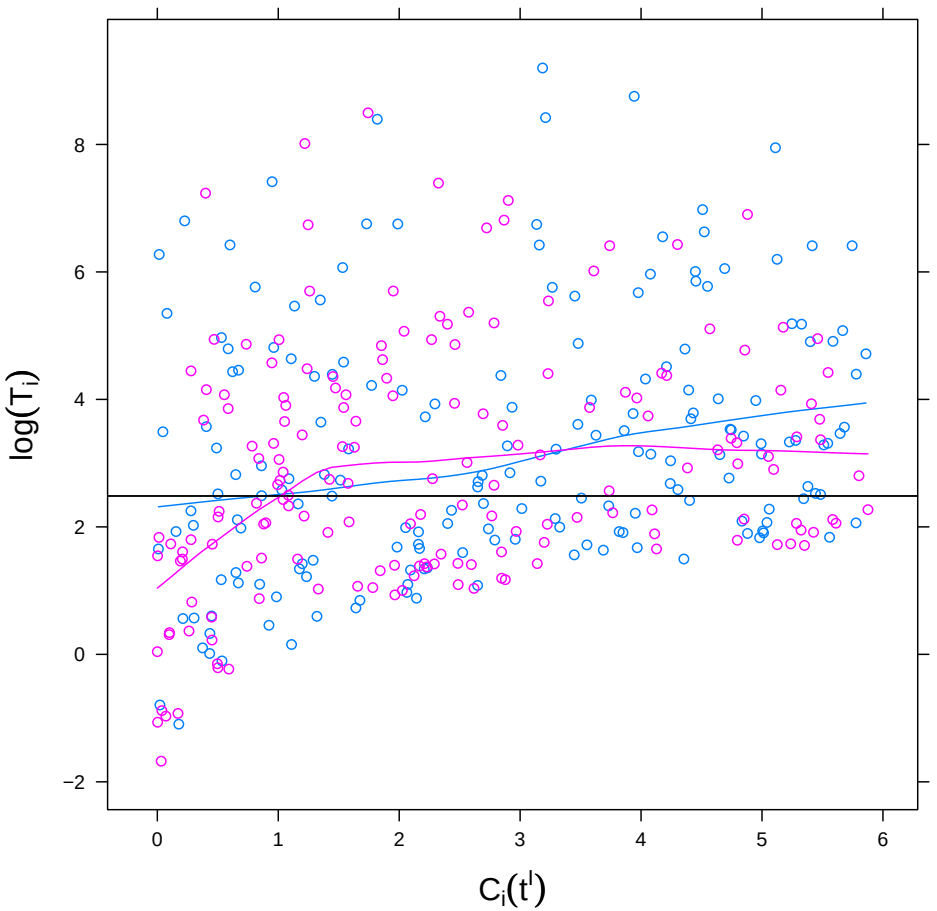
$$P(\text{Weib}(\alpha, \gamma) \leq x) = 1 - \exp(-\alpha(x^{\gamma}))$$



In an attempt to find a situation where Type I error could be inflated, I chose to simulate from a highly non-exponential distribution, so that administrative censoring times would be as informative as possible. Weibull$(0.93, 0.23)$ seemed to fit the bill, without being totally unrealistic.

Randomization Saves the Day (Probably)

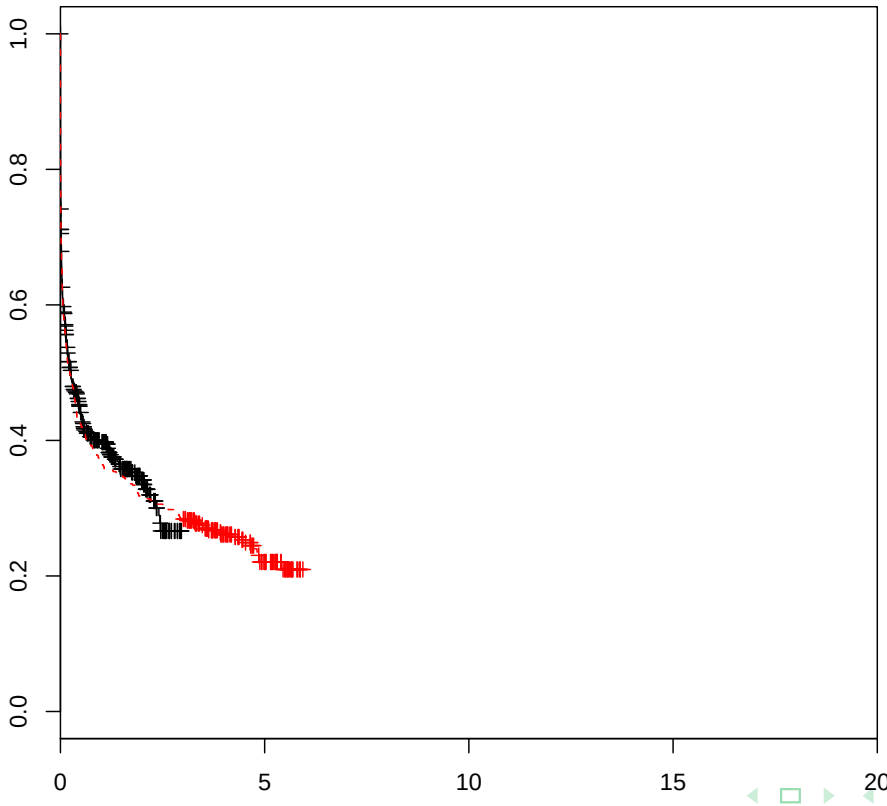
Predicting event times for those still in the risk set:



- Administrative censoring times *can* be used to (partially) predict event times,
- But treatment arms should be balanced with respect to administrative censoring times due to randomization,
- Therefore there is no obvious way to (even partially) predict future *differences* between the treatment arms.

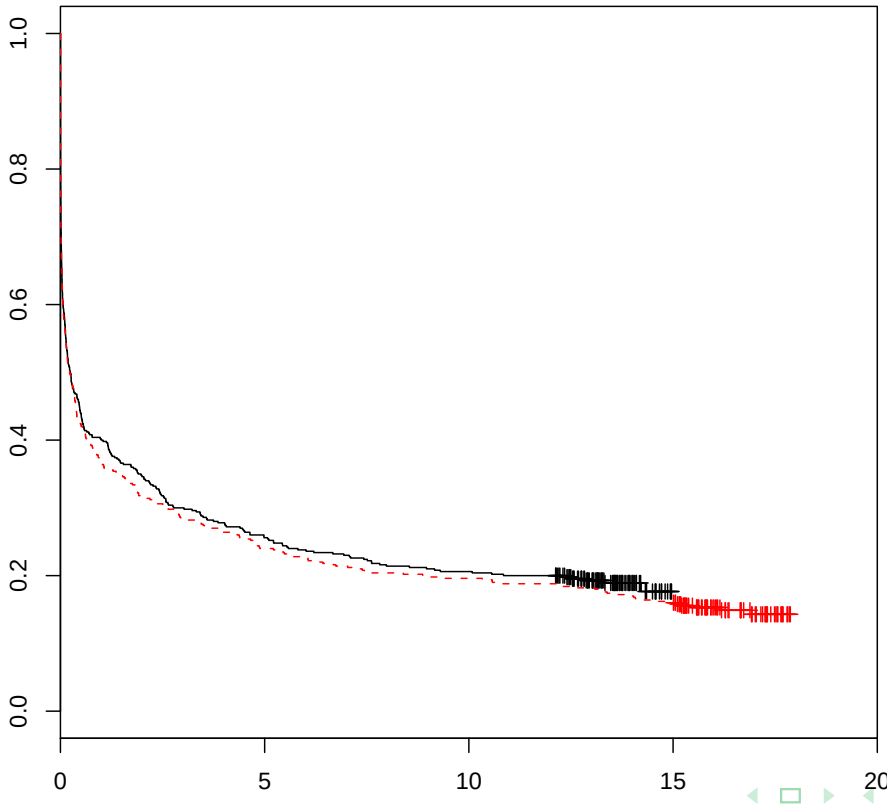
What About When Randomization Happens to be Temporally Unbalanced?

"Easy" to predict basic evolution of K-M curves . . .



What About When Randomization Happens to be Temporally Unbalanced?

. . . but still hard to predict differences between curves



Summary of Results

After running sizeable (B = 10,000) simulations under a variety of enrollment and event hazard scenarios:

- I could not find an example where imbalance in administrative censoring times (as represented by e.g. mean or median difference in enrollment times) was in any way predictive of Z_2.
- (Consequently), I could not find an example where use of a parametric analysis at the interim could lead to an inflated Type I error.
- On the other hand, I can't prove that it's impossible.

Conclusions

- When event times are non-exponential, administrative censoring times are partially predictive of future event times for those still in the risk set at the time of the interim.
- In theory this implies that the conventional computation of conditional Type I error does not apply, and use of that conventional computation could lead to inflation of the Type I error rate overall.
- In theory, related methodologies based on combination *p*-values would be similarly invalidated.
- In practice (as assessed by somewhat limited simulations), there seems to be essentially no real opportunity to inflate the Type I error rate.

Recommendations

- If using conditional error rate approach (or combination *p*-value approach):
  - Validate control of Type I error by simulation, including simulation of non-exponential event distributions (within the range of what is plausible).
  - There may be room to augment existing theory, so that it helps us a little more in this situation (?)
- If using a purely simulation-based adjustment to the nominal alpha level (as in many "Bayesian Adaptive Designs"), investigate sensitivity to non-exponential event distributions.

References

Schäfer, H. and Müller, H.H.

Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections.
Stat Med **20** (2001):3741–51.

Bauer, P. and Posch, M.

Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections, by H. Schäfer and H.-H. Müller, Statistics in Medicine 2001; 20: 3741-3751.
Stat Med **23** (2004):1333–4; author reply 1334–5.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M. and Racine-Poon, A.

Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology.
Stat Med **28** (2009):1445–63.