

Surviving the Interim: Insights Into Interim Survival Analysis

Venita DePuy, Bowden Analytics

ABSTRACT

Pre-specified interim analyses may be performed to evaluate whether a clinical trial can be halted prematurely for overwhelming efficacy and/or futility. This approach is somewhat more complex when the primary endpoint is a survival analysis. This paper provides an overview of performing the initial calculations and actual interim analyses using SAS® PROC SEQDESIGN and PROC SEQTEST, EAST® software, and PASS® software.

INTRODUCTION

We'll compare sample size calculations, interim analysis specifications, and interim analysis results for the following scenario across SAS, EAST, and PASS software systems:

The overall study intends to enroll subjects into treatments A and B, randomized 1:1. The primary endpoint is the increase in lab value X, which has been suppressed by the disease. Subjects will each be followed for the 6 month treatment period, although we expect the median time until lab value X increases to the target to be 20 days for treatment A (active) and 30 days for treatment B (control). It's a little hard to recruit subjects with this particular indication, so we estimate that it'll take 6 months to enroll 50 subjects (and subsequent subjects will take the same amount of time).

Therefore, we'll be testing $H_0: \theta = 0$ against $H_a: \theta \neq 0$, where $\theta = -\ln(\lambda)$, and λ is the hazard ratio. In other words, if the two treatment arms had the same duration, we would have $\theta = -\ln(1) = 0$, so testing if $\theta = 0$ is the same as testing whether the durations are the same. While this is considered a survival analysis, it can more accurately be referred to as a time-to-event analysis. In other words, we use the same statistical techniques as a true time-to-death (survival) analysis, although we do expect that subjects survive past the end of study participation.

SAMPLE SIZE CALCULATIONS FOR THE OVERALL STUDY (SAS)

Based on the above description, we calculate that the overall study needs 198 subjects (99 per group) using PROC POWER:

```
proc power;
  twosamplesurvival test=logrank alpha=0.05
    groupmedsurvtimes=(20 30)
    npergroup=.
    accrualtime=730
    followuptime=180
    power=.80;
run;
```

I calculated the accrual time of 730 days from the initial estimate of 6 months to enroll 50 subjects; since there are almost 200 subjects being enrolled, 6 months x 4 = 730 days. I admit that I use sort of an iterative process (the accrual time being based on the sample size, which is somewhat based on the accrual time). In this case, the follow-up time and accrual time are so long in comparison to the time to event, that accrual time doesn't affect the calculation. In general, I would recommend starting with a long accrual time for your initial sample size estimate, then shortening it to something that is appropriate for that initial sample size and recalculating the sample size.

We can also familiarize ourselves with the PROC SEQDESIGN output by calculating the same sample size, using NSTAGES=1 to indicate that there will only be a final analysis (NSTAGES=2 indicates a single interim analysis in addition to the final analysis). For clarity, we name the design OverallStudy.

```
proc seqdesign boundaryscale=StdZ;
  OverallStudy: Design nstages=1 alpha=0.05 beta=.2 ;
  samplesize model=TWOSAMPLESURVIVAL
    (medsurvtime=20 nullmedsurvtime=30 ref=hazard
      accrual=uniform acctime=730 foltime=180 );
run;
```

which produces the following output, in addition to a figure showing the recommended rejection/acceptance region (not presented):

The SEQDESIGN Procedure Design: OverallStudy																	
Design Information									Boundary Information (Standardized Z Scale) Null Reference = 0								
Stage	Statistic Distribution								Normal								
	Boundary Scale								Standardized Z								
	Alternative Hypothesis								Two-Sided								
	Alternative Reference								-0.40547								
	Number of Stages								1								
	Alpha								0.05								
	Beta								0.2								
	Power								0.8								
	Max Information (Percent of Fixed Sample)								100								
	Max Information								47.74201								
	Null Ref ASN (Percent of Fixed Sample)								100								
	Alt Ref ASN (Percent of Fixed Sample)								100								
Method Information									Sample Size Summary								
Boundary	Alpha	Beta	Alternative Reference	Drift	Test	Two-Sample Survival											
Upper Alpha	0.02500	0.20000	0.405465	2.801585	Null Hazard Rate	0.023105											
Lower Alpha	0.02500	0.20000	-0.40547	-2.80159	Hazard Rate (Group A)	0.034657											
					Hazard Rate (Group B)	0.023105											
					Hazard Ratio	1.5											
					log(Hazard Ratio)	0.405465											
					Reference Hazards	Alt Ref											
					Accrual	Uniform											
					Accrual Rate	0.261731											
					Accrual Time	730											
					Follow-up Time	180											
					Total Time	910											
					Max Number of Events	190.968											
					Max Sample Size	191.0639											
					Expected Sample Size (Null Ref)	191.0639											
					Expected Sample Size (Alt Ref)	191.0639											
Stage	Numbers of Events (D) and Sample Sizes (N) Two-Sample Log-Rank Test																
	Fractional Time								Ceiling Time								
	D	D(Grp 1)	D(Grp 2)	Time	N	N(Grp 1)	N(Grp 2)	Information	D	D(Grp 1)	D(Grp 2)	Time	N	N(Grp 1)	N(Grp 2)	Information	
1	190.97	95.52	95.44	910.0	191.06	95.53	95.53	47.7420	190.97	95.52	95.45	911	191.06	95.53	95.53	47.7426	

Output 1. Output from PROC SEQDESIGN with NSTAGES=1

Some parameters are relevant only to an interim analysis (i.e. what percent of the fixed sample size design is used for this analysis), so are 100, indicating that 100% of the data collected in the study will be used at this final analysis. The information, I_0 , is calculated as a function of α , β , and θ_1 , where α and β are the desired type I and II error values, respectively, and θ_1 is the expected negative log hazard ratio. In this case:

$$\begin{aligned}\theta_1 &= -\ln(30/20) = -0.40547 \\ I_0 &= (\phi^{-1}(1-\alpha/2) + \phi^{-1}(1-\beta))^2 / \theta_1^2 = (1.96 + 0.842)^2 / (-0.40547)^2 = 47.74201 \\ \theta_1 \sqrt{I_0} &= -2.80159\end{aligned}$$

Since the default is BOUNDARYSCALE=STDZ, the alternative reference and boundary values are displayed on the standardized normal Z scale. As shown in the Boundary Information output table, the hypothesis of $\theta = 0$ is rejected if $|Z| \geq 1.96$.

The Sample Size Summary output table displays the hazard rates (λ_A and λ_B) for the two treatment groups. The hazard rates can be derived from the median times to event ($t_{0.5}$), where $\lambda_x = \ln(2) / t_{0.5}$. We can easily verify that $\lambda_A = \ln(2) / 20 = 0.034657$ and $\lambda_B = \ln(2) / 30 = 0.023105$, which yields the expected log hazard ratio of $\log(\lambda_B / \lambda_A) = -0.40547..$ These do assume that the times to event follow an exponential distribution.

The maximum number of events needed, when subjects are randomized equally to the two treatment groups, is $4 * I_0$. In other words, you need $4 * 47.74201 = 190.968$ subjects (rounded to 191 subjects) who experience an event – in this case, to have lab X increase above the threshold – in order to have 80% power to detect the expected treatment difference, given the other assumptions.

This result is slightly different than the PROC POWER result that requires a total of 198 subjects. The difference may be due to the fact that the PROC SEQDESIGN is specifying that you need to have 191 subjects who experience events (values of lab X increasing above the threshold), as opposed to subjects who never experience the event for the duration of the study, or who drop out of the study prior to experiencing an event. It may also be due to differences in statistical methods between the two procedures (PROC POWER uses the Lakatos normal approximation). If you include accrual time and follow-up time in PROC SEQDESIGN, as above, it will also produce Expected Sample Size calculations. In this case, the 191.0639 (rounds to 192) subjects is still less than the 198 subjects produced in PROC POWER. These types of small differences are not uncommon when comparing results from two different statistical approaches to the same problem. In my personal experience, PROC POWER is more commonly used for power calculations for studies without interim analyses. I would always recommend documenting how power calculations were produced – both what assumptions, approaches and parameters used as well as what software and version – since there are often slight differences between power calculations produced using different software.

We shall perform the interim analysis design with the assumption that 198 subjects will be enrolled in the study. This is because, in my experience, clinical trials are designed to enroll a set number of patients, not to enroll subjects until a given number of subjects experience an event.

OVERVIEW OF INTERIM ANALYSES

Briefly, interim analyses are pre-planned looks at the data, at specific intervals, to see if the data is so overwhelmingly conclusive that it warrants stopping the trial. These should be spelled out in both the protocol and the statistical analysis plan (SAP), or an interim SAP, if that is logistically preferable to including it in an overall study SAP.

There are a variety of different statistical approaches to interim analyses; I have personally seen Pocock, O'Brien-Fleming, and the Lan-DeMets approach to O'Brien-Fleming used most frequently, although I also have seen others used (e.g. Haybittle-Peto). The particular test that best fits a given scenario will depend on the desired operating characteristics for that clinical trial; for example, Pocock has narrower boundaries earlier in the trial, which makes it more likely to be able to stop the study early, but O'Brien-Fleming has a smaller maximum sample size [Jennison & Turnbull 2000]. Those two tests, and others, were originally designed for a fixed number of equally spaced interim analyses; the Lan-DeMets error spending approach allows the interim to be conducted at non-equal intervals (for example, a single analysis after 40% of planned enrollment). I highly recommend Jennison & Turnbull's text for an explanation of the statistical details of the different options.

For this particular example, we are going to use the Lan-DeMets error spending approach that most closely approximates O'Brien-Fleming boundaries. We will conduct it after 85 subjects ($85/198=42.9\%$) have either experienced an event or completed participation in the trial (withdrew prior to experiencing the event, or completed 6 month participation without experiencing it). We will allow stopping for futility, and have the option of either stopping for efficacy or continuing to enroll the rest of the subjects. In other words, if it takes subjects on treatment A sufficiently longer until lab X values hit the threshold than treatment B, we will halt enrollment due to futility. If it takes subjects on treatment A sufficiently less time, we can either halt enrollment for efficacy, or continue to enroll the rest of the subjects. If neither threshold is met at the interim analysis, the study will continue to enroll the rest of the subjects.

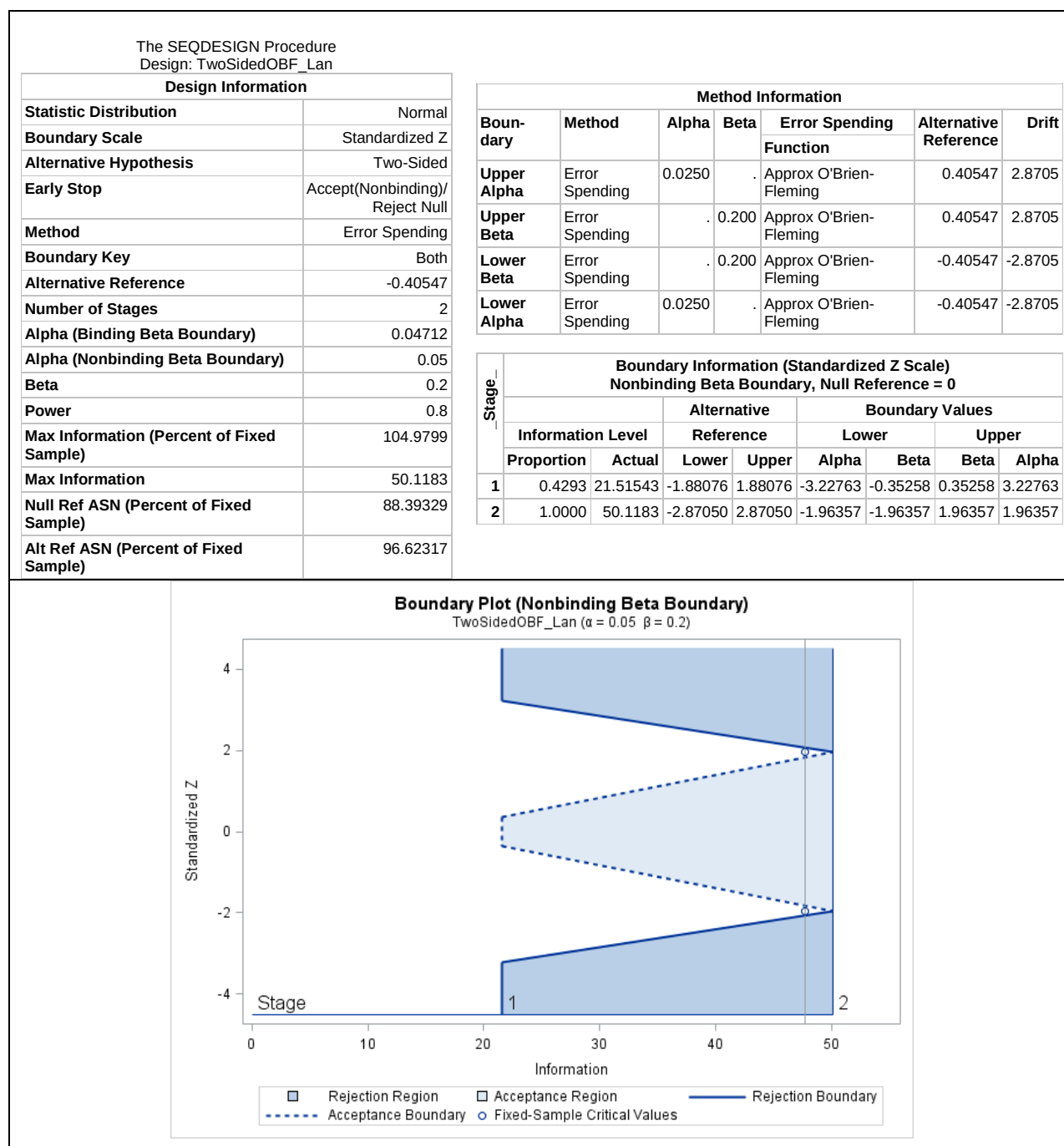
DESIGNING THE INTERIM ANALYSIS IN SAS

We design the interim analysis described above in SAS using PROC SEQDESIGN.

```
proc seqdesign boundaryscale=stdz ALTREF = -0.40547;
  TwoSidedOBF_Lan: design nstages=2
                    info=cum(85 198)
                    alpha=0.05
                    beta=.2
                    method=errfuncOBF
                    stop=both (betaboundary=nonbinding);
ods output Boundary=BoundZ;
run;
```

As above, the alternative reference is $\theta_1 = -\ln(30/20) = -0.40547$. We are choosing to have boundary information displayed in the default (standard normal scale) units, using the BOUNDARYSCALE option. The output will be labeled TwoSidedOBF_Lan to easier identify the output, and NSTAGES=2 indicates that there is one interim and one final analysis. The INFO statement indicates that the first analysis will be after 85 subjects contribute information, and the final after 198 subjects contribute information. This design optimistically assumes that none of the 198 subjects will drop out prior to reaching endpoint, given the long study participation (6 months) as compared to the estimated time to endpoint (20-30 days). The METHOD option indicates that a Lan-DeMets error spending approach that approximates O'Brien-Fleming is used. The STOP option indicates that we can stop for either efficacy or futility,

but that we are not required to stop for efficacy. Finally, the ODS output option is used to generate a dataset with the recommended stopping boundaries. The resulting dataset will be used in the future PROC SEQTEST analyses.



Output 2. Output from PROC SEQDESIGN with NSTAGES=2

Design Information

This section of the output focuses on summarizing the specific parameters entered, and comparing the sample size requirements of this trial to the same trial without any interim analysis (referred to as a fixed-sample design). The Null Ref ASN (Percent of Fixed Sample) is the expected sample size required under the null hypothesis (i.e., there is no difference between treatments), expressed as a percent of the corresponding fixed-sample design; in other words, the percent of the sample size needed to run the trial without an interim analysis. Similarly, the Alt Ref ASN (Percent of Fixed Sample) is the expected sample size, as a percentage of the corresponding fixed-sample design, if the alternative hypothesis is true and there is a significant treatment difference. The Maximum Information (Percent of

Fixed Sample) shows that this design only requires a slight increase in sample size, *if* the trial continues after the interim analysis, as compared to the sample size of the same trial without an interim analysis (i.e., the corresponding fixed-sample design). This is calculated as the ratio of the maximum information ($I_X = 50.1183$) to the amount of information in the fixed-sample design ($I_0 = 47.74201$), converted to a percentage; $50.1183/47.74201 = 1.0498$, which (after adjusting for rounding) is equivalent to the Max Information being 104.9799 percent of that required for the fixed-sample design. The total information available, referred to as $I_{\max}$ in Johnson & Turnbull and as I_X in SAS output, is derived in SAS.

Method Information

The Method Information output table summarizes the pre-specified alternative reference (± 0.40547) and the associated drift parameters (± 2.8705). It is important to note that $-\ln(30/20) = -0.40547$ and $-\ln(20/30) = +0.40547$; the difference between the two is simply, which treatment's hazard rate is larger than the other one. We are effectively evaluating whether the hazard rates are the same, since $-\ln(1) = 0$, so we are interested in looking at whether the hazard rate for A is significantly larger or significantly smaller than that of B.

The derived drift parameter (the standardized alternative reference) is calculated as $\pm \theta_1 \sqrt{I_X} = \pm 0.40547 * \sqrt{50.1183} = \pm 2.87046$.

Boundary Information (Standardized Z Scale)

While the pre-specified alternative reference (-0.40547) remains the same as the earlier output without an interim analysis, the vast majority of the other information has changed. To understand what the output is saying, I find it easiest to start at the Boundary Information output table. `_STAGE_ = 1` indicates the interim analysis, while 2 indicates the final analysis. $85/198 = 42.93\%$ of subjects will have information at the interim analysis; therefore the proportion of the information is 0.4293 for the interim, and 100% (1.0000) for the final analysis. In other words, 42.93% of the total information available (50.1183) = 20.51543 is available at the interim.

The alternative references of ± 1.88076 for the interim and ± 2.87050 for the final analysis are derived by SAS; you can see that the alternative reference for the final analysis is approximately the same as the derived drift parameter.

The boundary values for the interim (± 0.35258 for Beta, ± 3.22763 for Alpha) indicate what values would be required to reject the null hypothesis (no difference between treatments) at the interim analysis. This indicates that if $|Z| < 0.32538$, we would consider accepting $H_0: \theta = 0$ (i.e. stopping for futility, although recall that this boundary is non-binding). Similarly, if $|Z| > 3.22763$, we would reject the null hypothesis and stop the study early for overwhelming efficacy. The boundary values for the final analysis are the same for alpha and beta (± 1.96357), indicating that there is a single cutoff, and we would reject the null hypothesis if $|Z| > 1.96357$. For comparison, we recall that $Z = \pm 1.96$ as the standard cutoff for $\alpha=0.05$, for studies without an interim analysis, so we can see that there is just a slightly more strict criteria for the final analysis as a result of including the interim analysis.

This is also displayed visually in the Boundary Plot shown in Output 2. It is easy to see the narrow range of values at which the null hypothesis would be accepted or rejected at the interim analysis (information level 21.51543), and that the criteria for acceptance or rejection would increase with a larger sample size. I'm not clear as to why the acceptance region is not displayed on the plot; it does show if you reduce censoring to 15% ($Y > 0.85$) and change the average time for treatment A to $\text{SIGMA}=22$.

But What Does That ****Mean****?

At each interim stage, if the standardized Z test statistic is less than or equal to the corresponding lower alpha boundary, H_0 is rejected for efficacy and the study enrollment will be halted; if it's greater or equal than the upper alpha boundary, H_0 is rejected for futility and the study enrollment may be halted. Otherwise, we continue with the study.

Boundaries in Different Units

You can also elect to have the results displayed in different units. If you are performing the interim analysis at the time, you will need to produce the design in either Score scale (for PROC LIFETEST) or MLE scale (for PROC PHREG). My personal opinion is that, if you are designing the interim analysis for use in writing a protocol, SAP, or other document, it is easiest to have it in the Z scale as most statisticians, etc. can mentally compare the results to $Z=1.96$ to get a sense of scale, but don't have the same sense of the scale for score or MLE results. Regardless, I recommend that interim survival analysis design results in a protocol or other document should not specify an exact number (i.e. we will stop the study if $|Z| > 3.22763$) because the exact criteria is dependent on the amount of subjects and the amount of drop-out at the time of the interim analysis, so that number is sure to change, as discussed in following sections. Interim analysis design results are presented here for comparison between scales, and comparison with similar output from EAST and PASS software systems, as well. It is important to note that there exists a unique transformation between any two of these scales, as described in SAS online documentation; if you have the output in one scale, you can uniquely derive the boundary values of statistics for other scales.

Score Statistic Scale

The boundaries on the score statistic scale are produced by specifying BOUNDARYSCALE = SCORE. The score scale results from PROC SEQDESIGN are what will be fed into PROC SEQTEST and used, in conjunction with the interim data, to perform the interim analysis *if* the interim analysis is performed using PROC LIFETEST (whose parameter result is a type of score statistic, per SAS online documentation). More details on this are given in subsequent sections.

Boundary Information (Score Scale) Nonbinding Beta Boundary, Null Reference = 0								
Stage			Alternative		Boundary Values			
	Information Level		Reference		Lower		Upper	
	Proportion	Actual	Lower	Upper	Alpha	Beta	Beta	Alpha
1	0.4293	21.51543	-8.72386	8.72386	-14.97129	-1.63544	1.63544	14.97129
2	1.0000	50.1183	-20.32147	20.32147	-13.90099	-13.90099	13.90099	13.90099

Maximum Likelihood Estimator Scale

The boundaries on the score statistic scale are produced by specifying BOUNDARYSCALE = MLE. These is actually the scale that you will use for many interim analyses – those using PROCs GENMOD, LOGISTIC, MEANS, PHREG, and REG, per the SAS online documentation. So, if your interim survival analysis is performed using PROC PHREG, you will need to produce your results in this scale. Note that for survival, the MLE presented is the log hazard ratio.

Boundary Information (MLE Scale) Nonbinding Beta Boundary, Null Reference = 0								
Stage			Alternative		Boundary Values			
	Information Level		Reference		Lower		Upper	
	Proportion	Actual	Lower	Upper	Alpha	Beta	Beta	Alpha
1	0.4293	21.51543	-0.40547	0.40547	-0.69584	-0.07601	0.07601	0.69584
2	1.0000	50.1183	-0.40547	0.40547	-0.27736	-0.27736	0.27736	0.27736

P-value Scale

The results can be provided on the p-value scale, using BOUNDARYSCALE = PVALUE. If we recall that we are used to seeing $|Z| = 1.96$ associated with a $\alpha=0.05$, where we split $\alpha/2 = 0.025$ to either end of the normal distribution and end up with p values of 0.025 and $1-0.025 = 0.975$, it is easy to see that the slightly adjusted $|Z| = 1.96357$ from the previous output equates to p values of 0.02479 and 0.97521 for the final analysis. We can simply use the Excel function =NORM.S.INV(xx) to convert the p-values to Z-values and verify that the results produced here are the same as those on the Z scale. That said, you should be careful to interpret the values correctly; results of 0.99938 indicate a very large difference in a specific direction, and do not indicate no difference.

Boundary Information (p-Value Scale) Nonbinding Beta Boundary, Null Reference = 0								
Stage			Alternative		Boundary Values			
	Information Level		Reference		Lower		Upper	
	Proportion	Actual	Lower	Upper	Alpha	Beta	Beta	Alpha
1	0.4293	21.51543	-1.88076	1.88076	0.0006241	0.36220	0.63780	0.99938
2	1.0000	50.1183	-2.87050	2.87050	0.02479	0.02479	0.97521	0.97521

PERFORMING THE SAMPLE SIZE AND INITIAL INTERIM ANALYSIS DESIGN IN EAST

These sample size and initial interim analysis calculations were replicated in EAST, using EAST 6. Full details of how this was performed are included in Appendix 1. EAST uses a point-and-click interface, so screen shots and tips for actually using it are presented in the appendix in lieu of lines of code. There are a number of options to personalize the output (axis labeling, etc.). I found both the EAST 6 User Manual and the EAST 6 Tutorial (particularly Chapter 3.2: Group Sequential Design for a Survival Superiority Trial) to be extremely helpful. There were quite a number of different options available (piecewise dropout rates, numerous boundary type options, etc.)

EAST does offer a variety of boundary families (including a few that SAS doesn't), as shown in the appendix.

As with PROC SEQDESIGN, EAST determined that 191 events were needed for the given study parameters, and that a sample of 192 subjects would be needed to achieve 191 events. For simplicity, I chose to force it to 198 subjects in order to have an apples-to-apples comparison to the results produced by SAS.

EAST can provide results in the Z score, hazard ratio, p-value, and other scales. The information proportion (0.4293) is the same as that produced by SAS, and the interim analysis boundary values are very close but not exactly the same as those produced by SAS, as shown in Output 3. Further details, and results in different scales, are shown in Appendix 1.

Spacing of Looks ☐ Equal ☒ Unequal Boundary Scale: Z Scale									
Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	3.219	-3.219	0.024	0.168	-0.168
2	1	☒	☒	0.050	1.964	-1.964	0.196	1.964	-1.964

Output 3. Output from EAST Interim Analysis Design

PERFORMING THE SAMPLE SIZE AND INITIAL INTERIM ANALYSIS DESIGN IN PASS

These sample size and initial interim analysis calculations were replicated in PASS, using PASS 2008[®]. Full details of how this was performed are included in Appendix 2. Like EAST, PASS uses a point-and-click interface, so screen shots and tips for actually using it are included in lieu of lines of code. There are a number of options to personalize the output (axis labeling, etc.). While PASS does provide a users' manual, it provided a single example for performing interim survival analyses. It did clarify that, while the types of interim analyses can be selected from the Spending Function drop-down menu, the O'Brien-Fleming and Pocock selections are actually the Lan-DeMets spending functions that approximate the O'Brien-Fleming and Pocock boundaries. Three other spending functions of the Lan-DeMets type are also available. It does not have an option to select that one boundary be non-binding, so this example was performed assuming that both boundaries were binding.

I attempted to perform the same sample size calculation as performed in SAS and EAST, which resulted in an "Out of memory" pop-up window each time I attempted it with these parameter values. I couldn't figure out what caused that.

I continued with the initial interim analysis design using $n=198$, in order to compare results between the software packages. PASS does provide an easy-to-read report for its output (included in Appendix 2). PASS did specify in the sample size calculation that T1 should be the median survival time for the control group, and T2 for the treatment group. While the treatment group order was not specified in the interim analysis stage, I continued that treatment assignment. That resulted in a hazard ratio of $20/30 = 0.66667$, although the output report printed it as 0.6000.

When we compare the output to that from SAS, we see that it is on the standardized Z-scale, like the default SAS output. The upper and lower boundaries correspond exactly to the upper and lower Alpha boundaries in the SAS output. The drift parameter is quite different (3.52838 vs. 2.8705). The incremental alpha and total alpha, and incremental power and total power are not explicitly presented in the SAS output on the standardized Z-scale. However, if we look at the SAS output in p-value scale, we see that the $\alpha/2$ boundary values of 0.0006241 and 0.02479, when doubled, are equivalent to the 0.0012482 and 0.049580 listed as the nominal alpha.

Details when Spending = O'Brien-Fleming, N = 198, d = 191, S1 = 0.0098, S2 = 0.0625									
Look	Time	Info	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total Power
1	85.0000	0	-3.22763	3.22763	0.001248	0.001248	0.001248	0.179880	0.179880
2	198.0000	1	-1.96357	1.96357	0.049580	0.048751	0.050000	0.761393	0.941273
Drift	3.52838								

Output 4. Abbreviated Output for Initial Interim Analysis from PASS

CREATING THE DATA FOR THE INTERIM ANALYSIS

In my personal opinion, interim analyses for survival (or other time-to-event analyses) are much harder to explain simply than interim analyses for proportions, for example, when you can figure out ahead of time what percentage of successes you need at a given time point in order to have an interim analysis that results in stopping the trial. Time-to-event analyses do not depend just on the number of subjects achieving the endpoint, but on the pattern of times

when they did, *and* on the pattern of censoring. In other words, it is much more difficult to guess the likelihood of not continuing the trial, without actually performing the analysis.

So, we create a sample dataset on which to perform the interim analyses. The 85 subjects are randomly assigned to TRT = 1 or 2; subjects in each treatment group have times following an exponential distribution with mean times of 20 days (TRT = 1, group A) or 30 days (TRT = 2, group B). A randomly chosen 20% of subjects are censored, with censoring time exponentially distributed with mean of 14 days. Any subject still on the study is censored at the end of the 180 day follow-up. Times are rounded up to whole days, on the assumption that the lab values would not hit the threshold until at least day 1 (assuming the study treatment occurred at day 0). For the purpose of this illustration, we assume that the interim analysis is performed after the first 85 patients have completed participation in the study (either finished the 6 month follow up and have information on the number of days until the lab X value recovered, or finished the follow up period without the lab value ever reaching the threshold, or withdrew from the study after lab X values recovered, or withdrew from the study without the lab X values ever having recovered).

```
data samp;
  call streaminit(123459);
  do id = 1 to 85 ;
    trtran=rand("Uniform");

    if trtran > .5 then do;
      TRT=1;
      sigma = 20;
    end;
    else do;
      TRT=2;
      sigma = 30;
    end;

    x = sigma*rand("Exponential");
    y=rand("Uniform");
    censtime = 14*rand("Exponential");
    days = ceil(x);

    *censoring;
    if y > .8 then do;
      CENSOR=1;
      mTIME=ceil(censtime);
    end;
    else do;
      CENSOR = 0;
      mtime=days;
    end;
    if mtime > 180 then do;
      censor=1;
      mtime=180;
    end;
    output;
  end;
run;
```

This results in 38 subjects in group A and 47 in group B at the interim analysis, with 5 and 8 subjects censored, respectively. The mean time to event in uncensored subjects is 22.4 and 28.5 days, respectively. This data was then output to a .csv file for use in the other software systems.

PERFORMING THE INTERIM ANALYSIS IN SAS

The first step in the interim analysis, once the data is available, is to perform the regular analysis; in this case, the survival analysis. The following code indicates that time is in the mTIME variable, subjects with CENSOR=1 are censored, the difference between treatment groups TRT is tested, and the parameters needed for further calculations are output into dataset PARMS.

```
proc lifetest data=samp;
  time mTIME*CENSOR(1);
  test TRT;
  ods output logunichisq=PARMS;
run;
```

The contents of the PARMS dataset are shown below.

Variable	Statistic	StdErr	ChiSq	ProbChiSq
trt	4.6995	4.0794	1.3271	0.2493

The next step is to reformat the PARMS dataset so that PROC SEQTEST can use it. It is important to note that, just as PROC SEQDESIGN can produce results on the Z, p, score, and MLE scales, PROC SEQTEST can evaluate results on those same scales; you just need to make sure that both SEQDESIGN and SEQTEST are done on the same scale.

HELP – HOW DO I TELL WHAT SCALE TO USE?

You need to use the scale that corresponds with the statistic being output from your analysis, *and* make sure that your PROC SEQDESIGN was also performed in that scale (or re-performed on that scale, if it was performed in a different scale earlier during the design phase).

Per Examples 78.1 – 78.8 in the SAS online documentation, output from PROCs GENMOD, LOGISTIC, PHREG, and REG produce MLE statistics, but the log-rank statistic produced by PROC LIFETEST is a score statistic. Therefore, for our particular example, we will use the output from PROC SEQDESIGN where BOUNDARYSCALE=SCORE; but it is important to note that, if you are doing your survival analysis in PROC PHREG, you would need to do it on the MLE scale.

PERFORMING THE ACTUAL INTERIM ANALYSIS

In this case, I am renaming the dataset with a “S” suffix to remind myself that it’s in the score scale. `_STAGE_ = 1` is because this is the first (and only) interim analysis; if you had 2 interim analyses, you would use `_STAGE_ = 2` for the second interim. *Note that I have also (re)run the PROC SEQDESIGN using BOUNDARYSCALE=SCORE prior to performing the PROC SEQTEST.*

```
data PARMSs (keep = VARIABLE _SCALE_ _STAGE_ STDERR ESTIMATE);
  set PARMS (rename=(Statistic=Estimate));
  if VARIABLE='TRT'; *make sure that capitalization matches;
  _SCALE_='Score';
  _STAGE_=1;
run;
```

After reformatting the parameter dataset, we perform the PROC SEQTEST:

```
proc seqtest boundary=BoundS
  parms(Testvar=trt) = parmsS
  infoadj=prop
  boundaryscale=Score;
ods output test=testS;
run;
```

Results are shown in Output 5 on the following page.

Overall, the important piece to know is that the result of the interim analysis is that insufficient information was present to be able to stop the trial, so the trial continues. The figure shows where the result was (the dot) and how extreme it would have to have been in order for the trial to halt enrollment for futility or efficacy (the acceptance and rejection regions, respectively).

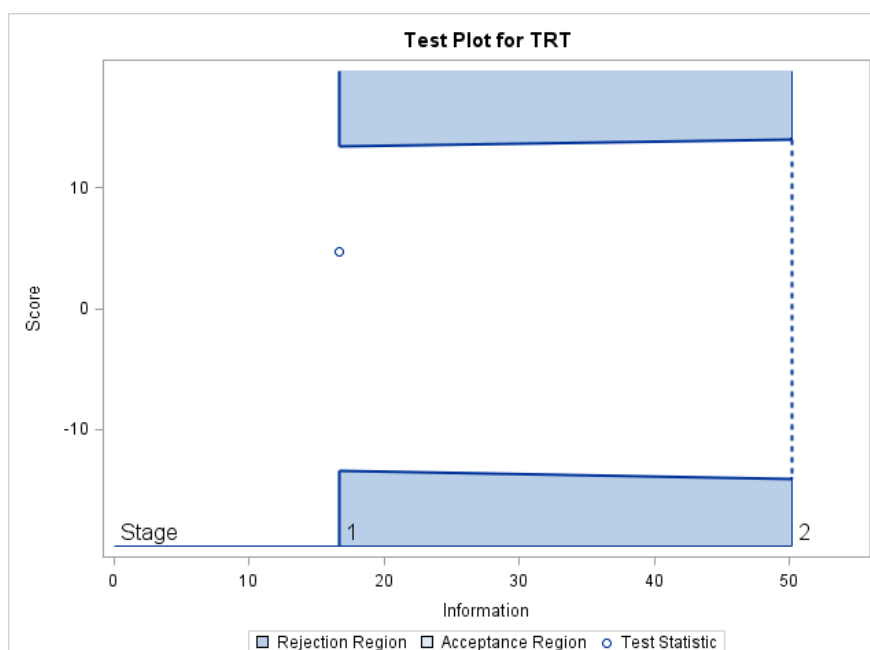
It is also important to note that, with approximately 20% of subjects (n=13) dropping out in our sample, it now estimates that only 33.2% of information from the overall study is available, instead of the 42.9% of information anticipated. The less information available, the stricter the criteria to stop the study. Another option might be to specify that the interim analysis is performed after 85 subjects have had lab X values recover, while on study – this example can be examined by changing it to n=103 subjects in the interim analysis, using the above code. (That results in 41.2% of information, an estimate of 3.72, and a recommendation of continuing). We will continue with the

scenario that the interim analysis will be performed after 85 subjects have either completed participation or met the endpoint, as that scenario is likely to occur in clinical trials.

Overall: The cutoffs used in the actual interim analysis, for a survival analysis, do not appear to ever be exactly the same as the cutoffs calculated prior to the start of the study. This is most likely due to the presence of censoring, which seems almost unavoidable in clinical trials. Therefore, one should consider this when documenting the planned interim analysis in the protocol or other a priori documents.

Design Information	
BOUNDARY Data Set	WORK.BOUNDS
Data Set	WORK.PARMSS
Statistic Distribution	Normal
Boundary Scale	Score
Alternative Hypothesis	Two-Sided
Early Stop	Accept/Reject Null
Number of Stages	2
Alpha	0.04712
Beta	0.18877
Power	0.81123
Max Information (Percent of Fixed Sample)	100.1983
Max Information	50.118299
Null Ref ASN (Percent of Fixed Sample)	100.1337
Alt Ref ASN (Percent of Fixed Sample)	96.86224

Test Information (Score Scale) Null Reference = 0										
Stage	Information Level		Alternative Reference		Boundary Values				Test TRT	
	Proportion	Actual	Lower	Upper	Lower		Upper		Estimate	Action
					Alpha	Beta	Beta	Alpha		
1	0.3320	16.6411	-6.74747	6.74747	-13.46356	.	.	13.46356	4.69946	Continue
2	1.0000	50.1183	-20.32147	20.32147	-14.08309	-14.08309	14.08309	14.08309	.	.



Output 5. Output from PROC SEQTEST, based on PROC LIFETEST

RESULTS IN DIFFERENT UNITS, IN SAS

It is important to match the scale of the parameters produced by the analysis (for example, score statistic from PROC LIFETEST, MLE from PROC PHREG) to the scale of the results in PROC SEQDESIGN that are fed into PROC SEQTEST. Otherwise, if you take results from PROC SEQDESIGN on the Z scale, and merge that with results from PROC LIFETEST and incorrectly label those parameter results as being on the Z scale, you will get erroneous results.

It may be preferable to produce PROC SEQDESIGN results on the Z scale, or p-value scale, for the analysis design phase to be described in the protocol, and later repeat the PROC SEQDESIGN in the appropriate scale for performing the interim analysis.

SO... HOW WOULD I KNOW IF THE INTERIM ANALYSIS RESULTED IN STOPPING?

In the sample code to generate the dataset, change SIGMA to 10 for TRT=1, and rerun both PROC LIFETEST and PROC SEQTEST; you'll see the output result in an estimate of 13.90491 and ACTION = Reject Null. The graph will also show that the dot is in the rejection region. In this scenario, you would stop the study at the interim analysis for overwhelming efficacy (i.e. treatment A takes so much less time for lab X values to increase to the threshold).

In the same code, if you change SIGMA to 29.5 for TRT=1 and change the censoring criteria to Y > 1 (so no censoring) and rerun everything, you'll see the output result in an estimate of -1.45914 and ACTION = Accept Null. The graph will show that the dot is in the acceptance region. In this scenario – *if* you chose to halt the study for futility – you would stop the study because the results of the two treatment arms were so close that you accepted the null hypothesis of no difference.

THAT'S NICE AND ALL, BUT I REALLY WANTED TO USE PROC PHREG, NOT LIFETEST....

That's straightforward; you just need to make sure you're doing your SEQDESIGN & SEQTEST in terms of MLE instead of score statistics. First, run your PROC SEQDESIGN with the BOUNDARYSCALE=MLE option. Then, the following code will perform the survival analysis, using the placebo group (TRT=2) as the reference group:

```
proc phreg data=samp;
  class trt (ref='2');
  model mtime*censor(1) = trt;
  ods output parameterestimates=parm;
run;
```

which results in the following parameter estimate:

Parameter	DF	Estimate	StdErr	ChiSq	ProbChiSq	HazardRatio	Label
TRT	1	0.27532	0.23973	1.3190	0.2508	1.317	TRT 1

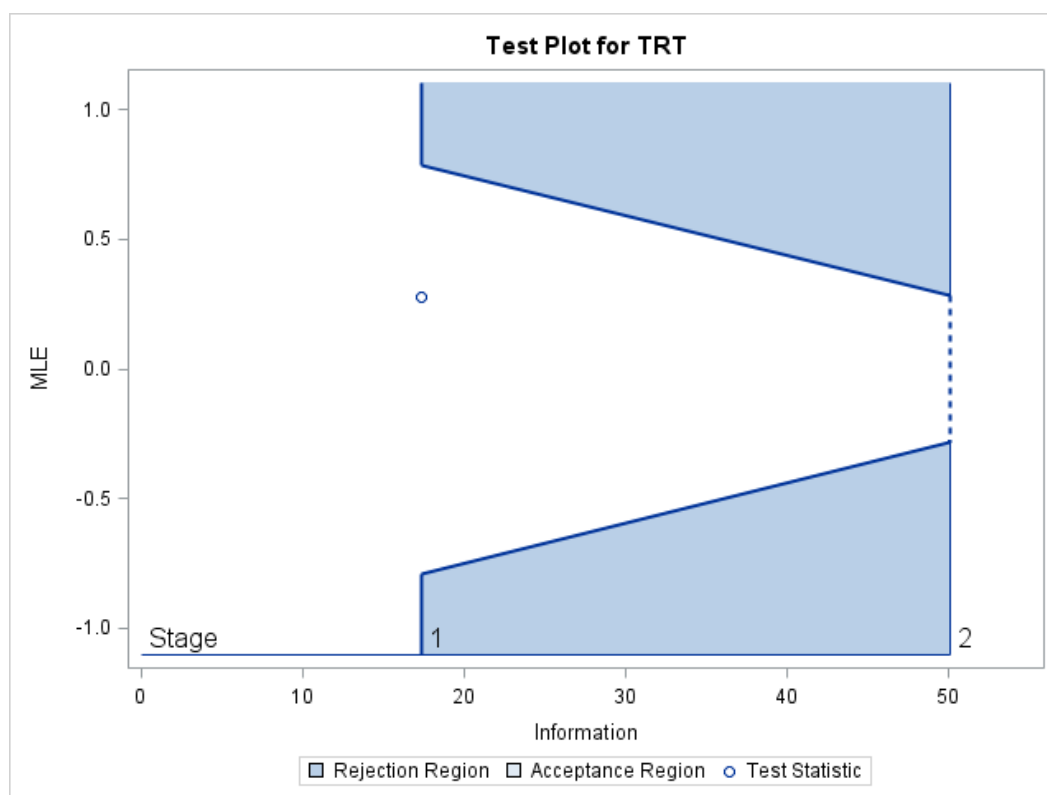
The reformatting of the parameter estimate file is similar to that for the earlier procedure:

```
data PARMSm (keep = PARAMETER _SCALE_ _STAGE_ STDERR ESTIMATE);
  set PARM ;
  if PARAMETER='TRT'; *make sure that capitalization matches;
  _SCALE_='MLE';
  _STAGE_=1;
run;
```

After that, and changing to BOUNDARYSCALE=MLE in PROC SEQTEST, and being sure that the dataset name changes (PARMSm for MLE parameter estimate) are incorporated, we obtain the output shown in Output 6, which is very close to that obtained using PROC LIFETEST.

Design Information	
BOUNDARY Data Set	WORK.BOUNDM
Data Set	WORK.PARMSM
Statistic Distribution	Normal
Boundary Scale	MLE
Alternative Hypothesis	Two-Sided
Early Stop	Accept/Reject Null
Number of Stages	2
Alpha	0.04712
Beta	0.18876
Power	0.81124
Max Information (Percent of Fixed Sample)	100.1962
Max Information	50.118299
Null Ref ASN (Percent of Fixed Sample)	100.1302
Alt Ref ASN (Percent of Fixed Sample)	96.58627

Test Information (Score Scale) Null Reference = 0										
Stage	Information Level		Alternative Reference		Boundary Values				Test TRT	
	Proportion	Actual	Lower	Upper	Alpha	Beta	Beta	Alpha	Estimate	Action
1	0.3472	17.4007	-0.40547	0.40547	-0.78819	.	.	0.78819	0.27532	Continue
2	1.0000	50.1183	-0.40547	0.40547	-0.28100	-0.28100	0.28100	0.28100	.	



Output 6. Output from PROC SEQTEST, based on PROC PHREG

Although the information proportion calculated is slightly different than the one based on PROC LIFETEST, and the scale of the results and shape of the graph are noticeably different due to being on the score statistic scale versus the MLE scale, we can see that the result (continue with the trial) is the same for both approaches.

PERFORMING THE INTERIM ANALYSIS IN PASS

PASS appears to be designed to perform the initial calculations, but not to perform the actual interim analysis. NCSS statistical software, produced by the same software company as PASS, appears to perform the interim analyses. However, I do not have access to that software so am unable to compare the results of that software package to the results obtained from SAS.

PERFORMING THE INTERIM ANALYSIS IN EAST

After importing the .csv file into EAST, it was fairly simple to perform the survival analysis. There is only one option (unlike being able to choose between PROCs LIFETEST and PHREG in SAS), and I just had to specify the trial type (superiority), variable names, and variable values for treatment vs. control and censoring vs. having an event. The full output is given in Appendix 2, and limited output given below in Output 7.

Summary of Observed Data:

Treatment ID	No. of Subjects	Events		Censored	
		Count	%	Count	%
2	47	39	82.979%	8	17.021%
1	38	33	86.842%	5	13.158%
Total	85	72	84.706%	13	15.294%

Parameter Estimates:

Hazard Ratio (HR)	95% Confidence Interval(2-Sided)	
	Lower Limit	Upper Limit
1.317	0.799	2.171

Test of Hypothesis:

Log Rank Score	Std. Error	Standardized Test Statistic	(1-Sided)		(2-Sided)
			Tail	p-value	p-value
4.699	4.023	1.168	G.E.	0.121	0.243

Estimated Hazard Rates:

Control (λ_c)	0.033
Treatment ($\lambda_c * HR$)	0.043

Output 7. Output from Logrank Analysis in EAST

As before, these results are fairly close to those produced by SAS. The log rank score of 4.699 is similar to the score statistic produced by PROC LIFETEST (4.6995), and the hazard ratio of 1.317 matches that produced by PROC PHREG. There is a setting that allows you to change the number of decimal places displayed for each statistic.

After the logrank analysis is performed, the interim analysis can be performed as described in Appendix 2. The number of cumulative events (72), estimate of δ , and standard error of δ need to be entered. While the number of cumulative events is straightforward from the default output, and we can easily calculate δ as the natural log of the hazard ratio, I did not see a simple way to obtain the standard error directly from the EAST output, although it is straightforward to do algebraically (as described in the appendix). After entering that information, and having it calculate the test statistic of 1.148, it produced the following output. If the result was that the study should be halted, a pop-up box would appear to state that. Therefore, EAST also recommended that the trial continue at this interim analysis. EAST also provides some additional information on the repeat p-value, predictive power, etc..

Look #	Information Fraction	Cumulative Events	Test Statistic	Est. of δ	Std. Error of Est. of δ	Efficacy		Futility		95% RCI for HR		Repeat ... p-value	CP	Predicti... Power
						Upper	Lower	Upper	Lower	Upper	Lower			
1	0.367	72	1.148	0.275	0.24	3.518	-3.518	0.072	-0.072	3.061	0.567	0.706	0.465	0.481
2														

CONCLUSION

Interim survival analyses can be challenging, both to design and to perform.

It may be easier to design them in terms of a commonly understood scale like the Z scale, but after dropout and censoring, it is very unlikely that the amount of information available at the time of the analysis is exactly the same as it was in the original design. Therefore, I would recommend specifying the software, version, assumptions, and so forth used in the interim analysis design in the study protocol, SAP, or other documentation, but stating that the analysis will be performed after “approximately xx subjects have either had an event, completed participation in the study, or withdrew from the study” or “after approximately xx subjects have had an event”, or something similar, will provide more flexibility and allow you to use all of the available information at the time of the analysis.

When performing the interim analysis, the first step is to perform the survival analysis – whether PROC LIFETEST, PROC PHREG, or the logrank analysis in EAST. Those results are then incorporated into the actual interim analysis. In SAS, it is imperative to be sure that results from PROC SEQDESIGN, in the correct units (score scale for PROC LIFETEST, MLE scale for PROC PHREG) are used in the PROC SEQTEST. EAST is somewhat simpler in that regard, in that it does not require you to specify which units are being used.

Results from the two software systems seem similar overall.

REFERENCES

- Cytel, Inc. 2016. *EAST 6. Statistical software for the design, simulation, and monitoring of clinical trials*. Cambridge, MA.
- Cytel, Inc. 2014. *EAST version 6.3.1 Tutorial*. Cambridge, MA: Cytel, Inc.
- Cytel, Inc. 2014. *EAST version 6.3.1 User Manual*. Cambridge, MA: Cytel, Inc.
- Hintze, J. 2008. *PASS 2008*. Kaysville, Utah: NCSS, LLC.
- Jennison, C. and B.W. Turnbull. 2000. *Group Sequential Methods: Applications to Clinical Trials*. Boca Raton, FL: CRC Press LLC
- SAS Institute, Inc. 2013. *SAS/STAT Users Guide*. Cary, North Carolina: SAS Institute, Inc. (In particular, examples 78.6, 83.14, 87.14, and 87.15)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Venita Bowden, PhD
Owner, Bowden Analytics
Bowden.analytics@gmail.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

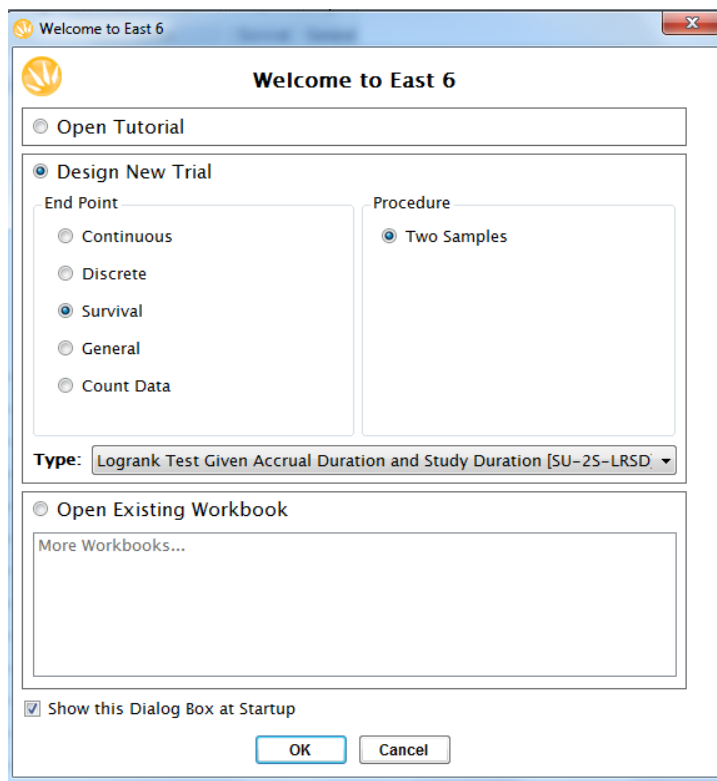
Other brand and product names are trademarks of their respective companies.

APPENDIX 1: SAMPLE SIZE AND INTERIM ANALYSIS REPLICATED IN EAST

This is not intended to be a comprehensive reference; both the EAST Users Manual and EAST Tutorial have a great deal more information in them. An interim survival analysis is included in the EAST Tutorial.

SAMPLE SIZE CALCULATIONS

- 1) The initial sample size calculation is requested by selecting Design New Trial, End Point = Survival, Procedure = Two Samples, Type = Logrank Test Given Accrual Duration and Study Duration from the start-up dialog box.



- 2) For the initial sample size calculation, we set Number of Looks = 1 (i.e., only the final analysis), Design Type = Superiority (the other option is Non-Inferiority), Test Type = 2-sided, alpha = 0.05, Power=0.8, allocation ratio between treatment arms = 1, and variance of log hazard ratio = null (the default). The hazard rate is constant throughout (# of Hazard Pieces = 1), and we selected the input method of Median Survival Times entered the median survival time in the control group (30 d) and median survival time in the treatment group (20 d), and it automatically calculated the hazard ratio and ratio of medians.

Design: Survival Endpoint: Two-Sample Test - Parallel Design - Logrank Given Accrual Duration and

Design Type: **Superiority** Number of Looks: **1**

Design Parameters **Accrual/Dropout Info**

Test Type: **2-Sided** # of Hazard Pieces: **1** Input Method: **Median Survival Times**

Type I Error (α): **0.05**

Power: **0.8**

Sample Size (n): **Computed**

No. of Events: **Computed**

Allocation Ratio: **1**
(n_t/n_c)

☐ Hazard Ratio (Optional) **Alternative**

☐ Hazard Ratio (λ_t/λ_c) **1.5**

☐ Ratio of Medians (m_t/m_c) **0.667**

Period #	Med. Surv. Time (Control)	Med. Surv. Time (Treatment: Alt.)
1	30.000	20.000

Variance of Log Hazard Ratio

☒ Null ☐ Alternative

Then, we chose the Accrual/Dropout Info tab to enter the accrual duration (730 d) and study duration (730 d + 180 d duration of participation = 910 d), then hit 'Compute' in the lower right hand corner (not shown).

Design: Survival Endpoint: Two-Sample Test - Parallel Design - Logrank Given Accrual Duration and Study Duration

Design Type: **Superiority** Number of Looks: **1**

Design Parameters **Accrual/Dropout Info**

Subjects are followed: **Until End of Study**

Accrual Info

Accrual Duration: **730** Study Duration: **910**

of Accrual Periods: **1**

Period #	By Time	Cum. % Accrued
1	730.000	100.000

Piecewise Dropout Information

of Pieces: **0** Input Method: **Hazard Rates**

Period #	Starting at Time	Hazard Rate (Control)	Hazard Rate (Treatment)
----------	------------------	-----------------------	-------------------------

- 3) The resulting output indicated that 191 events were expected, and calculated a sample size of 192 would be needed to achieve 191 events.

	Des 1
Mnemonic	SU-2S-LRSD
Test Parameters	
Design Type	Superiority
No. of Looks	1
Test Type	2-Sided
Specified α	0.05
Power	0.8
Model Parameters	
Allocation Ratio (nt/nc)	1
Hazard Ratio (Alt.)	1.5
Var (Log HR)	Null
Accrual & Dropout Parameters	
Subjects are Followed	Until End of Study
No. of Accrual Periods	1
No. of Dropout Pieces	0
Sample Size	
Maximum	192
Expected Under H0	192
Expected Under H1	192
Events	
Maximum	191
Expected Under H0	191
Expected Under H1	191
Study Duration	
Maximum	910
Expected Under H0	835.258
Expected Under H1	814.914
Accrual Duration	
Maximum	730
Expected Under H0	730
Expected Under H1	730

INTERIM ANALYSIS SPECIFICATION

We continued interim analysis calculations on the assumption that 198 subjects would enroll, in order to have an apples-to-apples comparison with SAS and PASS output.

- 4) The design entry screen is virtually identical to the earlier one, with the main differences being that Number of Looks is set to 2, the sample size is forced to 198, and the power is left to be computed (since it cannot also be specified in this case). The Accrual/Dropout Info tab remains unchanged (not shown). As you can see, adding the interim analysis creates the Boundary Info tab as well.

Design: Survival Endpoint: Two-Sample Test - Parallel Design - Logrank Given Accrual Duration and

Design Type: Number of Looks:

Design Parameters Boundary Info Accrual/Dropout Info

Test Type: # of Hazard Pieces: Input Method:

Type I Error (α):

Power: ☐

Sample Size (n): ☐

No. of Events: ☐

Allocation Ratio: (n_t/n_c)

☐ Hazard Ratio (Optional)

☐ Hazard Ratio (λ_t/λ_c) Alternative

☐ Ratio of Medians (m_t/m_c) Alternative

Period #	Med. Surv. Time (Control)	Med. Surv. Time (Treatment: Alt.)
1	30.000	20.000

Variance of Log Hazard Ratio

☒ Null ☐ Alternative

- 5) The Boundary Info tab allows us to select what type of boundary family (Spending Functions, Haybittle-Peto, or Wang-Tsiatis), and the type of spending function (Lan-DeMets, Gamma family, Rho family, or Interpolated), and specify the particular type of Lan-DeMets with the Parameter (OF = O'Brien-Fleming, PK = Pocock). We are populating both the Efficacy and Futility sides (to only choose Efficacy, you would select boundary family of 'None', and vice versa), and selecting a non-binding futility boundary. In our example, the spacing of looks is unequal, so we select that radio button and then edit the Information Fraction for the first look to be $85/198 = 0.4293$. We leave the Boundary Scale in the default Z-scale units; other options are: p-value, score, HR, and $\ln(HR)$.

After hitting the Compute button (lower right of screen; not shown), the results are populated at the bottom of the window.

Design: Survival Endpoint: Two-Sample Test - Parallel Design - Logrank Given Accrual Duration

Design Type: Superiority Number of Looks: 2

Design Parameters **Boundary Info** **Accrual/Dropout Info**

Efficacy

Boundary Family: Spending Functions

Spending Function: Lan-DeMets

Parameter: OF

Type I Error (α): 0.05

Futility

Boundary Family: Spending Functions

Spending Function: Lan-DeMets

Parameter: OF

Type II Error (β): 0.196

☒ Non-Binding
☐ Binding

Spacing of Looks: ☐ Equal ☒ Unequal

Boundary Scale: Z Scale  

Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	3.219	-3.219	0.024	0.168	-0.168
2	1	☒	☒	0.050	1.964	-1.964	0.196	1.964	-1.964

- 6) Output can easily be changed to different boundary scales by selecting another option (such as the p-value scale or score scale) from the drop-down menu; those output are included below. EAST also provides a nice listing of composite output, from all the variations you've looked at, at the bottom of the page (not shown here).

Spacing of Looks ☐ Equal ☒ Unequal Boundary Scale: **Score Scale**  

Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	14.837	-14.837	0.024	0.774	-0.774
2	1	☒	☒	0.050	13.781	-13.781	0.196	13.781	-13.781

Spacing of Looks ☐ Equal ☒ Unequal Boundary Scale: **HR Scale**  

Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	2.010	0.497	0.024	1.037	0.964
2	1	☒	☒	0.050	1.323	0.756	0.196	1.323	0.756

Spacing of Looks ☐ Equal ☒ Unequal Boundary Scale: **ln(HR) Scale**  

Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	0.698	-0.698	0.024	0.036	-0.036
2	1	☒	☒	0.050	0.280	-0.280	0.196	0.280	-0.280

Spacing of Looks ☐ Equal ☒ Unequal Boundary Scale: **p-value Scale**  

Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	0.001	0.001	0.023	0.872	
2	1	☒	☒	0.050	0.025	0.025	0.194	0.050	

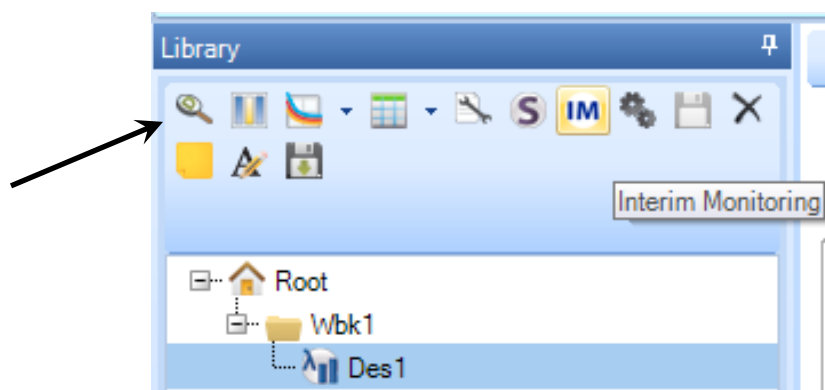
Spacing of Looks ☐ Equal ☒ Unequal Boundary Scale: **cp_delta1 Scale**  

Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	0.001	0.001	0.023	0.872	0.001
2	1	☒	☒	0.050	0.025	0.025	0.194	0.050	0.025

Spacing of Looks ☐ Equal ☒ Unequal Boundary Scale: **cp_deltahat Scale**  

Look #	Info. Fraction	Stop for Efficacy	Stop for Futility	Cum. α Spent	Efficacy Boundary		Cum. β Spent	Futility Boundary	
					Upper	Lower		Upper	Lower
1	0.4293	☒	☒	0.001	0.001	0.001	0.023	0.872	0.001
2	1	☒	☒	0.050	0.025	0.025	0.194	0.050	0.025

- 7) The first design created in EAST is named, by default, Des1. As well as it being displayed at the bottom of the output window, you can also see it displayed on the left hand side, in the Library window.



That will display quite a bit more details about the interim analysis, as shown below:

Design: Survival Endpoint: Two-Sample Test - Parallel Design - Logrank Given Accrual Duration and Study Duration

Test Parameters	
Design ID	Des1
Design Type	Superiority
Number of Looks	2
Test Type	2-Sided
Specified α	0.05
Attained α	0.049
Power	0.802
Model Parameters	
HR = λ_1/λ_0	
Under H0	1
Under H1	1.5
Ratio of Medians	0.667
Var (Log HR)	Null
Allocation Ratio (n_1/n_0)	1
Boundary Parameters	
Spacing of Looks	Unequal
Efficacy Boundary	LD (OF)
Futility Boundary	LD (OF) (NB)
Accrual/Dropout Parameters	
Accrual Duration	730
Max Study Duration	910
Dropout	No

Fixed Follow-up Design: All subjects are followed for maximum 180 time units, or until drop out or failure.

Sample Size Information

	Control Arm	Treatment Arm	Total
Sample Size (n)			
Maximum	99	99	198
Expected H1	93.363	93.363	186.726
Expected H0	92.311	92.311	184.623
Events (s)			
Maximum	97	99	196
Expected H1	96.278	98.329	183.849
Expected H0	97.204	97.204	181.268
Maximum Information (I): 49			

Accrual and Study Duration

	Accrual Duration	Study Duration
Maximum	730	842.072
Expected H1	688.435	788.348
Expected H0	680.679	837.003

Stopping Boundaries: Look by Look

Look #	Info. Fraction (s/s_max)	Events (s)	Cumulative α Spent	Cumulative β Spent	Boundaries			
					Efficacy Z		Futility Z	
					Upper	Lower	Upper	Lower
1	0.429	84	0.001	0.024	3.231	-3.231	0.164	-0.164
2	1	196	0.05	0.198	1.964	-1.964	1.964	-1.964

Events, Sample Size, Pipeline and Analysis Times: Look by Look (Under H0)

Look #	Info. Fraction (s/s_max)	Sample Size (n)	Events (s)	Pipeline (n-s)	Analysis Time	Boundary Crossing Probability (Incremental)		
						Efficacy		Futility
						Upper	Lower	
1	0.429	97	84	13	355.037	6.175E-4	6.175E-4	0.13
2	1	198	196	2	910	0.024	0.024	0.821

⊖ Events, Sample Size, Pipeline and Analysis Times: Look by Look (Under H1)

Look #	Info. Fraction (s/s_max)	Sample Size (n)	Events (s)	Pipeline (n-s)	Analysis Time	Boundary Crossing Probability (Incremental)		
						Efficacy		Futility
						Upper	Lower	
1	0.429	95	84	11	346.864	0.085	1.802E-7	0.024
2	1	198	196	2	842.072	0.717	7.543E-7	0.175

⊖ Survival Information : Median Survival Times

Median Survival Times		Hazard Ratio
Control (mc)	Treatment (mt1)	Alt. (mc/mt1)
30	20	1.5

⊖ Accrual Information

Period #	Starting at Time	Cum. % Accrued
1	730	100

PERFORMING THE INTERIM ANALYSIS

- 8) As with SAS, the first step is to perform the actual analysis. In this case, I exported the dataset created in SAS into a .csv file called interim_analysis.csv, then used the Import button on the top left corner of the window to import the dataset into EAST.



After walking through the various straightforward questions about file type and delimiting values, it completed importing the dataset into a file called interim_analysis.cydx, as shown below:

East 6 - [Wbk1:interim_a]

	id	TRTran	TRT	sigma	x	y	censtime
1	1	0.337339476	2	30	13.1957086	0.510978833	31.8151582
2	2	0.578239761	1	20	17.7224313	0.150870798	3.48967092
3	3	0.679886389	1	20	5.75225011	0.0925449037	30.2454659
4	4	0.485883104	2	30	11.5201148	0.310745946	7.69897008
5	5	0.512625251	1	20	4.94650159	0.0562038985	3.83283157
6	6	0.0625122695	2	30	16.8073877	0.189678994	2.68826227

- 9) To perform the analysis, I clicked on the Two Samples button in the Events section of the toolbar, and selected the Logrank design as shown above. This resulted in a pop-up window, where I specified the various attributes of the model:

Data Set: interim_analysis.cyx

Main Advanced

Trial Type: Superiority Response Variable: mTIME Frequency Variable:

Population ID: TRT

Control: 2

Treatment: 1

Censor Indicator: CENSOR

Censored: 1

Complete: 0

Which resulted in the following output:

Analysis: Time to Event Response: Logrank Test

Let $\delta = \ln(\lambda_t / \lambda_c)$

$H_0 : \delta = 0$ Vs. $H_1 : \delta \neq 0$ for 2-Sided test

Either $H_1 : \delta > 0$ Or $H_1 : \delta < 0$ for 1-Sided test

Data File: interim_analysis.cyx

Trial Type: Superiority

Population ID: TRT(Treatment=1, Control=2)

Response Variable: mTIME

Censor: CENSOR(Censor Value=1, Complete=0)

Confidence Level: 0.95

Output

Summary of Observed Data:

Treatment ID	No. of Subjects	Events		Censored	
		Count	%	Count	%
2	47	39	82.979%	8	17.021%
1	38	33	86.842%	5	13.158%
Total	85	72	84.706%	13	15.294%

Parameter Estimates:

Hazard Ratio (HR)	95% Confidence Interval(2-Sided)	
	Lower Limit	Upper Limit
1.317	0.799	2.171

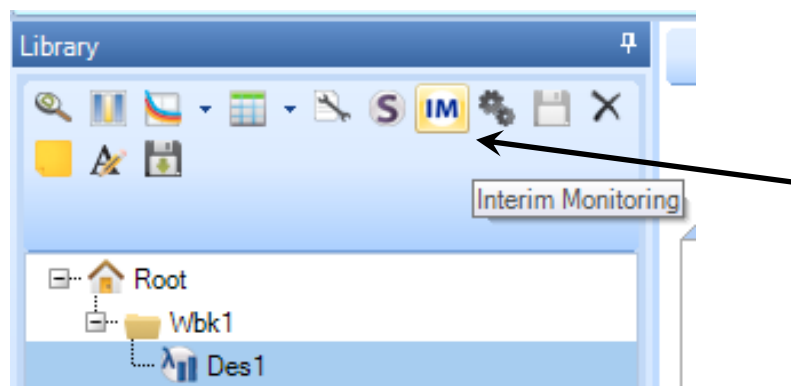
Test of Hypothesis:

Log Rank Score	Std. Error	Standardized Test Statistic	(1-Sided)		(2-Sided)
			Tail	p-value	p-value
4.699	4.023	1.168	G.E.	0.121	0.243

Estimated Hazard Rates:

Control (λ_c)	0.033
Treatment ($\lambda_c^* \text{ HR}$)	0.043

- 10) To perform the interim analysis, you click in the “IM” (interim monitoring) button in the Library window, as shown below.



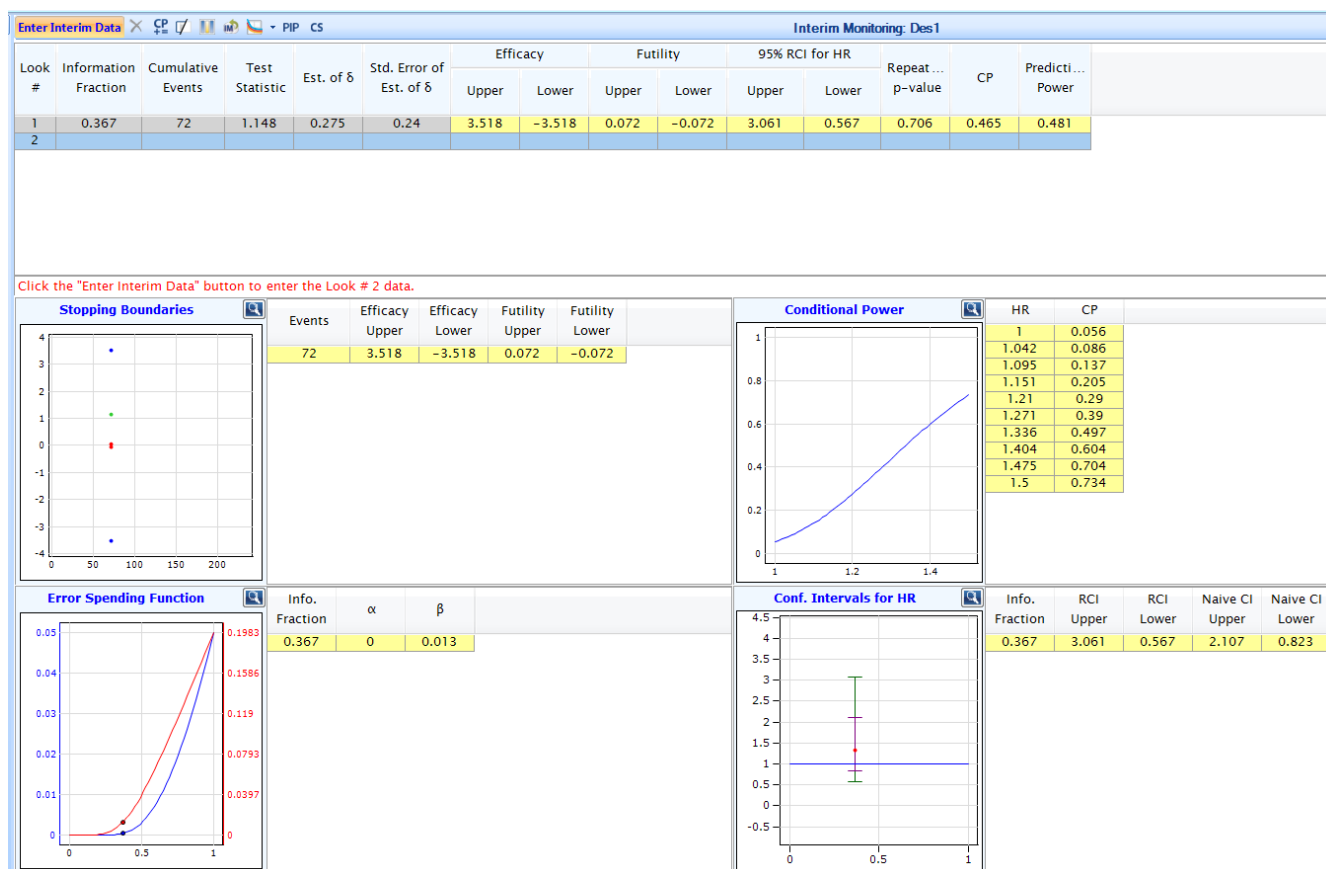
After the next window opens up, hit Enter Interim Data in the top left corner. This results in a pop-up window to enter the pertinent details, as shown below. The number of cumulative events (72) is taken directly from the log-rank analysis output.

We can enter the log hazard ratio, $\ln(\delta)$, of $\ln(1.317) = 0.275$, based on the analysis output. The standard error of the estimate can be calculated based on the confidence interval; the natural log of the lower bound of the confidence interval, minus δ , all divided by -1.96 , results in an estimate of 0.255 for the standard error of δ . Since SAS's PROC PHREG produces these estimates, and to a greater number of decimal points, I entered SAS's estimates of 0.27532 and 0.23973 below.

Hitting the Recalc button will cause it to populate the Test Statistic with 1.148; then hit OK.

 A screenshot of the 'Test Statistic Calculator' dialog box. The title bar says 'Test Statistic Calculator'. Inside, there's a section 'Editing Look #1' with a checkbox 'Set Current Look as Last' which is unchecked. Below this are four input fields: 'Cumulative Events:' with the value '72', 'Input for Survival end point' (a label), 'Estimate of δ :' with the value '0.27532', and 'Standard Error of Estimate of δ :' with the value '0.23973'. Below these is an 'Output' section with a label 'Test Statistic:' and a value of '1.148'. At the bottom are three buttons: 'Recalc', 'OK', and 'Cancel'.

11) The interim monitoring results are displayed below.



APPENDIX 2: SAMPLE SIZE AND INTERIM ANALYSIS SPECIFICATIONS REPLICATED IN PASS

- 1) The initial sample size calculation is requested by selecting Survival Analysis and Reliability / Logrank Tests / Logrank Tests (Lakatos) using Median Survival Time. Power (0.80), alpha (0.05), median survival time in the control group ($T_1 = 30$), median survival time in the treatment group ($T_2 = 20$), accrual time (730), accrual pattern (equal), total time (730 accrual + 180 duration of participation = 910), and test type (two-sided) were entered. PASS produced an "Out of Memory Error" notation.

PASS: Logrank Tests (Lakatos) [Median Survival Time]

File Run Means Proportions Correlation Regression Survival ROC Variances DOE Tools Win

RUN NEW OPEN SAVE PASS MAP OUT MACRO DIFFS 2PROP Z-S T-TEST T-TEST 1WAY ANOVA RM ANOVA CONTRL MC-S HR LINK LINEAR REG LOGIST REG

Symbols/Back Options Abbreviations Template

Data Reports Axes/Legend Plot Text Plot Type

Solve For

Find (Solve For):
N (Total Sample Size)

Error Rates

Power (1-Beta):
0.80

Alpha (Significance Level):
0.05

Sample Size

N (Total Sample Size):
500

Proportion in Control Group:
0.5

Effect Size

T1 (Median Survival Time - Control):
30

Treatment Group Parameter:
T2 (Med. Survival Time - Treatment)
20

T2 (Median Survival Time - Treatment):
20

HR (Hazard Ratio = T_1/T_2):
0.5

Parameter Conversion

Duration

Accrual Time (Integers Only):
730

Accrual Pattern:
Equal

Total Time (Integers Only):
910

Proportion Lost or Switching Groups

Controls Treatments

Lost:
0.00 0.00

Switch to Treatments: Switch to Controls:
0.0 0.0

Note: Proportions are for a single time period.

Test

Alternative:
Two-Sided

Spreadsheet

Spreadsheet

- 2) For comparability, we continue using the $n=198$ sample size calculation obtained from SAS. The interim analysis calculation is performed by selecting Survival Analysis and Reliability / Logrank Tests / Group-Sequential Logrank Tests. Sample size ($n=198$), alpha (0.05), alternative hypothesis type (two-sided), survival time assumption (exponential), number of looks (2; interim and final), boundary truncation (none) are straightforward to enter.

PASS determines the timing of the interim analysis using the Times, Max Time, and Informations fields. In this case, Times is equivalent to the number of subjects whose survival time is known (or, presumably, censored) at the time of analysis, so we enter the number of subjects (85 and 198), separated by a space, and similarly enter the fractions of information available at those time points (0.4292 and 1.00). The Max Time is the “time” at which 100% of the information is available, i.e. 198 subjects.

PASS: Group-Sequential Logrank Tests

File Run Means Proportions Correlation Regression Survival ROC Variances DOE Tools

RUN NEW OPEN SAVE PASS MAP OUT MACRO DIFFS Z-TEST S-S T-TEST L-RAY ANOVA RM ANOVA CMTL MO-3 HR L-RANK LINEAR REG

Symbols/Back End Plot Axes Options Template

Data Reports Axes/Legend Plot Text Plot Type

Solve For
Find (Solve For):
Power and Beta

Error Rates
Power (1-Beta):
.80
Alpha (Significance Level):
0.05

Sample Size
N (Total Sample Size):
198

Test
Alternative Hypothesis:
Two-Sided
Survival Time Assumption:
Exponential Survival

Effect Size
S1 (Proportion Surviving 1):
9.84313320230369E-03
S2 (Proportion Surviving 2):
6.25000000000001E-02
Survival Parameter Conversion Tool

Look Details
Number of Looks: Boundary Truncation:
2 None
Spending Function: Max Time:
O'Brien-Fleming 198
Times:
85 198
Informations:
0.4292 1.00
Upper Boundaries (Spending = User):
0.25 0.75
Lower Boundaries (Spending = User):

Default values (not yet changed)

- 3) The Effect Size calculations require a little additional data entry, since we know the median times to event (30 & 20 days) and the hazard ratio ($30/20 = 1.5$) but not the effect sizes. We first click on 'Survival Parameter Conversion Tool' to bring up a separate pop-up window. We enter the median survival times ($T1=30$, $T2=20$ since the earlier screen indicated that $T1$ was for the control and $T2$ for treatment) and the other numbers automatically populate. We then change $T0$ to reflect the duration of study participation (180 days), so that the proportions surviving after that point in each treatment group ($S1$ and $S2$) are correct. I then copy/pasted the values for $S1$ and $S2$ back into the original window (note that they did not automatically populate after using this conversion tool window).

Survival Parameter Conversion Tool

Survival Parameters | Proportions (Years to Months)

Use this window to convert among various survival distribution parameters (assumes an exponential survival distribution). The appropriate values may then be copied and pasted.

Survival Parameters

Time ($T0$): 180

Hazard Ratio ($h1/h2$): 0.6666666666666666

Mortality Ratio ($M1/M2$): 0.986301369863014

Group 1

Hazard Rate ($h1$): 2.31049060186648E-02

Proportion Surviving Past $T0$ ($S1$): 1.56250000000001E-02

Mortality until $T0$ ($M1$): 0.984375

Median Survival Time ($T1$): 30.0000000000001

Changing a value will cause all appropriate values to be updated.

Hazard Ratio ($h2/h1$): 1.5

Mortality Ratio ($M2/M1$): 1.01388888888889

Group 2

Hazard Rate ($h2$): 3.46573590279973E-02

Proportion Surviving Past $T0$ ($S2$): 1.95312499999999E-03

Mortality until $T0$ ($M2$): 0.998046875

Median Survival Time ($T2$): 20

1) Enter these

2) Enter this

3) Copy these calculated values to other window

- 4) Then press the Run icon in the upper left corner, to run the calculations

PASS: Group-Sequential Logrank Tests

File Run Means Proportions Correlation Regression Survival ROC Variances DOE Tools Win

RUN NEW OPEN SAVE PASS MAP OUT MACRO DIFFS 2 PROP 2 S T-TEST 1-WAY ANOVA RM ANOVA CNTRL MC-S HR LRNK LINERR REG LOGIST REG

Symbols/Back End Plot Axes Options Template

Data Reports Axes/Legend Plot Text Plot Type

Solve For

Find (Solve For):
Power and Beta

Error Rates

Power (1-Beta):
.80

Alpha (Significance Level):
0.05

Sample Size

N (Total Sample Size):
198

Test

Alternative Hypothesis:
Two-Sided

Survival Time Assumption:
Exponential Survival

Effect Size

S1 (Proportion Surviving 1):
1.5625000000000001E-02

S2 (Proportion Surviving 2):
1.9531249999999999E-03

Survival Parameter Conversion Tool

Look Details

Number of Looks: Boundary Truncation:
2 None

Spending Function: Max Time:
O'Brien-Fleming 198

Times:
85 198

Informations:
0.4292 1.00

Upper Boundaries (Spending = User):
0.25 0.75

Lower Boundaries (Spending = User):

- 5) PASS spits out an easily read report with the results, including references and a few graphs.

The PASS documentation clarifies that the Nominal Alpha is the alpha that would be used for these boundaries if they were used for a single, standalone test; in other words, this is the significance level that should be used in conjunction with a standard statistical package that does not adjust for multiple looks at the data. The Inc (incremental) Alpha is the amount of alpha that is spent by the interim test; it is close to the Nominal Alpha, but slightly different due to the correction that must be made for multiple tests. The Total Alpha is the amount of alpha used up to and including the current test. The PASS documentation does not further specify how the last two differ. The Incremental Power is described as the amounts that are added to the total power at each interim test, which are the "exit probabilities" (the probability that significance is found and the trial is stopped, given the alternative hypothesis). The Total Power is the cumulative power values, i.e. the probability of the trial being stopped at or before the corresponding time.

Group Sequential - Survival

Page/Date/Time 1 9/1/2016 10:33:49 PM

Numeric Results for Two-Sided Logrank Test (Assuming Exponential Survival)

Power	Total Sample Size (N)	Total Required Events	Alpha	Beta	Proportion Surv. (S1)	Proportion Surv. (S2)	Hazard Ratio
0.941273	198	191	0.050000	0.058727	0.0098	0.0625	0.6000

References

Chow, S.C.; Shao, J.; Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.
 Lan, K.K.G. and DeMets, D.L. 1983. 'Discrete sequential boundaries for clinical trials.' Biometrika, 70, pages 659-663.
 O'Brien, P.C. and Fleming, T.R. 1979. 'A multiple testing procedure for clinical trials.' Biometrics, 35, pages 549-556.
 Pocock, S.J. 1977. 'Group sequential methods in the design and analysis of clinical trials.' Biometrika, 64, pages 191-199.
 Reboussin, D.M., DeMets, D.L., Kim, K, and Lan, K.K.G. 1992. 'Programs for computing group sequential boundaries using the Lan-DeMets Method.' Technical Report 60, Department of Biostatistics, University of Wisconsin-Madison.

Report Definitions

Power is the probability of rejecting a false null hypothesis. Power should be close to one.
 N is the total number of items sampled. There are N/2 in each group.
 Events is the number of events that must occur in each group.
 Alpha is the probability of rejecting a true null hypothesis in at least one of the sequential tests.
 Beta is the probability of accepting a false null hypothesis at the conclusion of all tests.
 S1 is the proportion surviving in group 1.
 S2 is the proportion surviving in group 2.
 HR is the hazard ratio. It is calculated using $\text{Log}(S2)/\text{Log}(S1)$.

Summary Statements

A total sample size of 198 (split equally between the two groups), or 191 events, achieves 94% power to detect a hazard rate of 0.6000 when the proportions surviving in each group are 0.0098 and 0.0625 at a significance level (alpha) of 0.050000 using a two-sided log rank test. These results assume that 2 sequential tests are made using the O'Brien-Fleming spending function to determine the test boundaries and that the survival times are exponential.

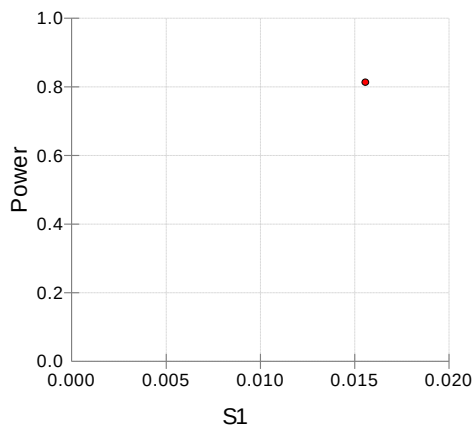
Details when Spending = O'Brien-Fleming, N = 198, d = 191, S1 = 0.0098, S2 = 0.0625

Look	Time	Info	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total Power
1	85.0000	0	-3.22763	3.22763	0.001248	0.001248	0.001248	0.179880	0.179880
2	198.0000	1	-1.96357	1.96357	0.049580	0.048751	0.050000	0.761393	0.941273
Drift	3.52838								

Group Sequential - Survival

Page/Date/Time 2 9/1/2016 10:33:50 PM

Chart Section

Power vs S1 with S2=0.0020 Alpha=0.05 N=198 Surv
Test

Although the title truncates
when copying it from the
output window, the full title
reads: Power vs S1 with
S2=0.0020 Alpha=0.05
N=198 Surv Test

O'Brien-Fleming Boundaries with Alpha = 0.05

