Two-Stage Design with Sample Size Re-estimation using gsDesign

Keaven M. Anderson and Alison Pedley

Abstract The effect size and nuisance parameters needed to appropriately size a clinical trial are generally not adequately understood before a trial begins. Through pre-planned early stopping rules, a conservatively planned group sequential design can effectively adapt the sample size for a clinical trial for a new treatment that is ineffective, very effective or minimally effective. One potential issue with such designs is that a substantial number of patients can be enrolled after a data cutoff for an interim analysis while data are being entered, cleaned, analyzed and discussed. A strategy of re-estimating the sample size at an interim analysis based on conditional power has been proposed to reduce somewhat this enrollment overrun issue. A purported advantage sometimes claimed is a smaller up-front planned sample size than a conservatively planned group sequential design.
We demonstrate derivation of 2-stage group sequential and conditional power designs using the gsDesign R package and suggest comparing designs with comparable power using expected sample size calculations. While there are cases where conditional power designs may have small advantages, it is quite easy to derive very inefficient conditional power designs. This, along with the fact that a conditional power design may reveal something about the interim treatment effect, will often leave a group sequential design as the design of choice.

The objective of this paper is to demonstrate a tool to derive and evaluate 2-stage designs that adapt sample size at an interim analysis based on conditional power. The rationale for conditional power designs is to 'rescue' trials that appear to have less than the desired power based on an interim analysis. Such designs have been proposed for at least 25 years (Bauer, 1989; Cui et al., 1999; Proschan and Hunsberger, 1995) and refinements have been made to improve design properties while either maintaining the use of conditional power; *e.g.*, (Liu and Chi, 2001; Gao et al., 2008; Mehta and Pocock, 2011) or not (Posch et al., 2003; Lokhnygina and Tsiatis, 2008; Schmitz, 1993). Since it is easy to produce an inefficient conditional power design (Jennison and Turnbull, 2003), it seems valuable to consider the overall power gains and cost of said gains in terms of expected sample size. Using the gsDesign R package we compare power and sample size between a 2-stage conditional power design and a group sequential designs. While we try to provide clarity, we also assume the reader will install the package and examine its extensive help files where some of the code can be further clarified.

Example trial

We will begin with a fixed design sample size for all of the methods considered. We will assume 80% power (Type II error 20%) and 2.5% one-sided Type I error for all cases. Clinical trials with a control and experimental treatment arm will be considered. For a continuous, normally distributed outcome and reasonably large sample sizes, the function `nNormal` that assumes known variance should be adequate. We consider an example from Wang et al. (2012). They compared normally distributed means with a standard deviation of 1 for treatments. Early studies indicated a treatment benefit with a new, experimental therapy to be between .27 and .33. We begin by computing the total required sample size for a new trial using the more optimistic of these assumptions with $\delta = \mu_1 - \mu_2 = .33$.

```
library(gsDesign)
# compute sample size with 1:1 randomization
nNor <- nNormal(delta1=.33, sd=1, alpha=.025, beta=.2)
# round up to an even number
2*ceiling(nNor/2)

## [1] 290
```

For smaller sample sizes a routine based on the t-distribution from base stats may be preferred, yielding a total sample size of 292 in this case: `power.t.test(delta=.33,power=.80)`. Note that `nNormal()` allows unequal randomization and designing for non-inferiority, while these capabilities are not built into `power.t.test()`.

If, in truth, the effect size were the smaller $\delta = .27$ instead of $\delta = .33$, the sample size required to power the trial would be 432 instead of 290. If the effect size were correctly specified as $\delta = .33$, but standard deviation were $\sigma = 1.5$ instead of $\sigma = 1$, the required sample size would be 650. Methods in subsequent sections will attempt to adapt from the original assumptions at an interim analysis to power the trial appropriately.

The efficient estimate for the parameter of interest in this case is $\hat{\delta} = \bar{X}_1 - \bar{X}_2$ where $\bar{X}_j$ is the sample mean after $n/2$ observations for group j, $j = 1,2$. For testing in this case, we would use $Z = \sqrt{n}(\bar{X}_1 - \bar{X}_2)/(2\sigma) = \sqrt{n}\hat{\delta}/(2\sigma)$ which has a standard normal distribution with mean $\sqrt{n}\hat{\delta}/(2\sigma)$. The standardized effect size θ is defined as $\delta/(2\sigma)$ for a two-sample normal test which implies $\hat{\theta} = \hat{\delta}/(2\sigma)$ and

$$Z \sim \text{Normal}(\sqrt{n}\theta, 1).$$

The sample size to achieve power $1 - \beta$ with one-sided Type I error α is

$$\left(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\theta}\right)^2$$

where $\Phi^{-1}()$ is the inverse of the cumulative standard normal distribution function.

Group sequential design

Group sequential design will adapt to different effect sizes if a minimum effect size of interest can be specified and early stopping for efficacy or futility is used to adapt sample size. With a 2-stage group sequential design, we perform a single interim analysis and stop the trial for positive efficacy if the interim test statistic is very positive, stop for futility if the results are discouraging, and otherwise continue the trial to the end. The basic algorithm (Jennison and Turnbull, 2000) inflates a sample size from a fixed design to produce a group sequential design as derived with the code below and shown in Figure 1. Given the cost of powering for the more conservative effect size of .27, assume we wish to start with the more optimistic treatment benefit of .33. The following code plots the expected sample size by underlying treatment effect assuming an enrollment overrun of 75 enrolled beyond those with outcome data available at the interim analysis. The values `k=2` and `timing=.5`, respectively, specify two analyses with the interim analysis including half of the planned observations. The design bounds and sample size are derived using spending functions (Lan and DeMets, 1983). Specifically, we use the power spending function of Kim and DeMets (1987) ($\alpha(t) = \alpha t^{\rho}$). The upper (efficacy) spending function spending function in the argument `sfu`, lower (futility) in `sfl` in order to simply make bounds for alternate designs comparable. The value of the parameter `sfupar` here was chosen to fully specify the upper spending function is to match the commonly used O'Brien and Fleming (1979) bound, while the lower bound (parameter `sflpar`) is intended to provide a 'reasonable' probability of stopping the trial for futility at the interim analysis if underlying treatment effect is less than targeted. Assuming the default one-sided Type I error `alpha=.025` and $\rho = 3.275$ in the power spending function chosen, the nominal Type I error at the interim analysis half-way through the trial ($t = .5$) is $.025(.5)^{3.275} = 0.0026$ which can be seen if you enter the command `gsBoundSummary(gsD)`. You can also see that the probability of crossing the futility bound is $.2(.5)^{1.5} = 0.0707$, where $\beta = .2$ is the Type II error, $\rho = 1.5$ is the power spending function parameter and, again, $t = .5$ to indicate the analysis is done after half of the final planned sample size is available.

```
# extend the fixed design sample size to a group sequential design
gsD <- gsDesign(k=2, timing=.5, beta=.2, alpha=.025, n.fix=nNor, delta1=.33, overrun=75,
                sfu=sfPower,sfl=sfPower,sfupar=3.275,sflpar=1.5)
# plot z-value bounds for the group sequential design
plot(gsD, cex=.8, main="Base group sequential design")
```

If the interim can be conducted promptly after the interim patient set is enrolled, this can result in substantial savings compared to the fixed design sample size of 290; Figure 2 shows the expected sample size with the assumed interim analysis overrun of 75 patients.

```
# plot expected sample size by underlying treatment effect
plot(gsD, plottype="asn",
     main="Expected sample size by underlying treatment effect; overrun = 75",
     xlab=expression(paste("Effect size, ",delta,"; alternate hypothesis is ",
                           delta,"=.33.")))
```

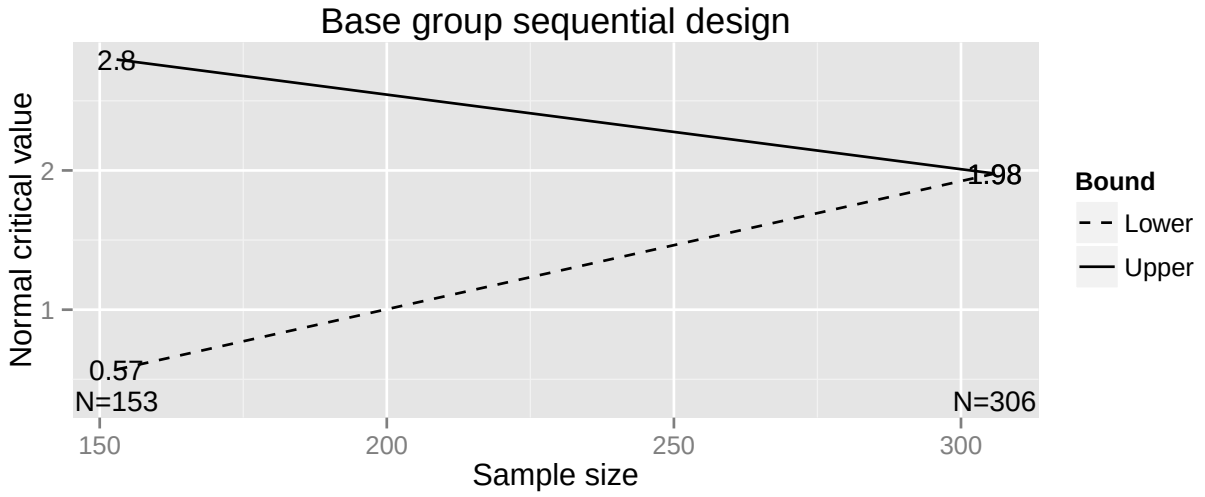

Figure 1: Group sequential design sample size and normal test statistic bounds at interim and final analysis for a standard deviation of 1 and a difference in group means of .33.

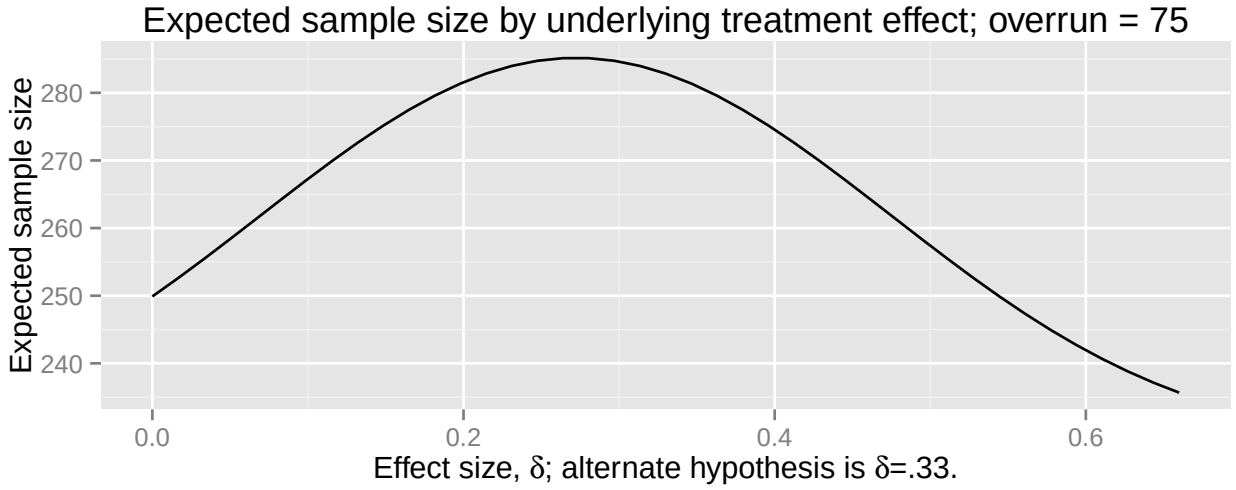

Figure 2: Expected sample size by treatment effect for the planned 2-stage group sequential design powered for a treatment effect of $\delta = .33$ assuming a standard deviation of 1 and enrollment overrun of 75 at the interim analysis.

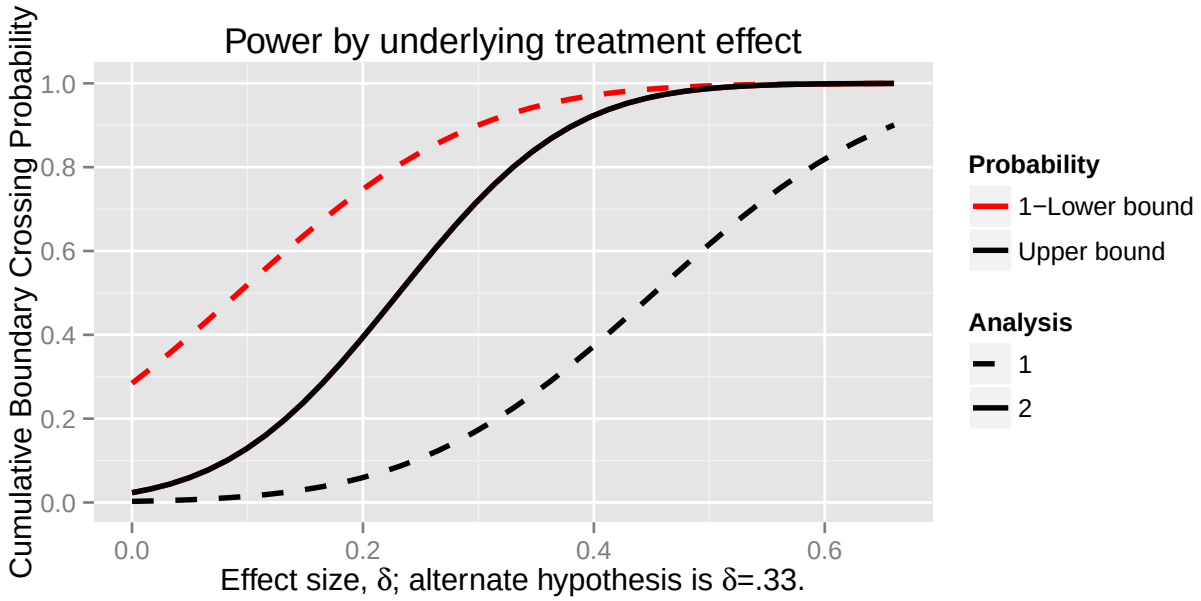

Figure 3: Power by treatment effect for the planned 2-stage group sequential design powered for a treatment effect of $\delta = .33$ assuming a standard deviation of 1.

We also show how to print the expected sample sizes given $\delta = 0$, .27 and .33, respectively.

```
# print power for delta=0, .27, .33 recalling that standardized effect size
# theta=delta/2; the gsProbability function can add values of theta into a
# gsDesign object
gsProbability(d=gsD,theta=c(0,.27,.33)/2)$en

## [1] 249.8941 285.1678 282.8383
```

We can also plot power as shown in Figure 3. The solid black line shows that there is a sharp drop in overall study power (positive result at interim or final analysis) from 80% with $\delta = .33$ to 63% if the less optimistic $\delta = .27$ holds. The dashed black line shows the probability of crossing the efficacy bound at the interim analysis, while the dashed red line shows one minus the probability of crossing the futility bound at the interim analysis. Thus, the lines divide the probability of all possible boundary crossing possibilities for each underlying effect size.

```
# plot expected sample size by underlying treatment effect
plot(gsD, plottype="power",
     main="Power by underlying treatment effect",
     xlab=expression(paste("Effect size, ",delta,"; alternate hypothesis is ",
                           delta,"=.33.")))
```

A group sequential design fully powered for $\delta = .27$ would be one way of ensuring adequate power for this smaller effect size. Another way of ensuring adequate power for the smaller effect size would be to design a group sequential trial with an interim analysis after 50% of the trial has been analyzed and adapt the sample size up if interim results are in the continuation region of the group sequential design and 'sufficiently encouraging.' This is done with the thought of improving the power of the design without incurring the additional associated expense if such an increase is not needed. The discussion that follows shows some suggested adaptation methods from the literature as well as a method to evaluate the impact of such strategies on the expected sample size (cost) required and study power.

Two-stage sample size re-estimation

The notation for the above group sequential design is n_1 observations at the interim analysis, an additional n_2 for a total of $N_2 = n_1 + n_2$ at the final analysis, and for analyses $i = 1, 2$, futility bounds a_i and efficacy bounds

b_i. We assume tests statistics Z_1 and Z_2 that approximately follow a bivariate normal distribution with means $\sqrt{n_1}\theta$ and $\sqrt{N_2}\theta, respectively, variances of 1, i=1,2 and covariance n_1/N_2$. Let $w_i = \sqrt{n_i/N_2}$ for $i = 1,2$ and assume $Y_2 \sim \text{Normal}(\sqrt{n_2}\theta, 1)$. Letting

$$Z_2 = w_1 Z_1 + w_2 Y_2$$

produces the same multivariate normal distribution for Z_1 and Z_2. For our specific example examining the difference in means of two normal samples with a common variance and sample sizes, we have Z_1 based on the first $n_1/2$ observations for each group and Y_2 based on the next $n_2/2$ observations for each group.

We consider a two-stage adaptive design here that is a generalization building on the group sequential design just outlined. The approach taken assumes a second stage sample size $n_2(Z_1)$, and a conditionally independent normally distributed random variable

$$Y_2(Z_1)|(Z_1 = c) \sim \text{Normal}\left(\sqrt{n_2(c)}\theta, 1\right).$$

We denote the total sample size as

$$N_2(Z_1) = n_1 + n_2(Z_1).$$

Note that the functions $N_2()$, $n_2()$ and the random variable $Y_2(Z_1)$ are all generalizations from their respective simplest forms N_2, n_2 and Y_2 from the group sequential design. We also assume that the trial is declared positive if $Y_2(Z_1) \geq c(Z_1, n_2(Z_1))$ for functions $c(,)$ and $n_2()$. For the group sequential case, since the trial is declared positive if $Z_2 \geq b_2$, we have

$$c(Z_1, n_2(Z_1)) = \frac{b_2 - \sqrt{w_1}Z_1}{\sqrt{w_2}}. \tag{1}$$

We assume the trial stops at stage 1 for futility with sample size n_1 included in the analysis if $Z_1 < a_1$ or for efficacy if $Z_1 \geq b_1$. Letting $\phi()$ denote the standard normal density function, the probability of a positive trial is

$$\begin{aligned}\alpha(\theta) &= P_\theta\{Z_1 \geq b_1\} + P_\theta\{Y_2(Z_1) \geq c(Z_1, n_2(Z_1))\} \\ &= 1 - \Phi(b_1 - \sqrt{n_1}\theta) + \int_{a_1}^{b_1} \phi(z_1 - \sqrt{n_1}\theta)\left(1 - \Phi\left(c(z_1, n_2(z_1)) - \sqrt{n_2(z_1)}\theta\right)\right) dz_1.\end{aligned} \tag{2}$$

Equation (2) is general in that it can apply to a 2-stage group sequential design, various conditional power designs (Cui et al., 1999; Proschan and Hunsberger, 1995; Lehmacher and Wassmer, 1999; Liu and Chi, 2001; Chen et al., 2004; Mehta and Pocock, 2011), or other sample size adaptation methods such as Posch et al. (2003). The optimal two-stage design of Lokhnygina and Tsiatis (2008) also takes this form, although computing $n_2(z_1)$ requires dynamic programming.

Next we focus on the choice of $c(z_1, n_2(z_1))$ where z_1 is assumed to be an observed value of Z_1. There are two options we will consider here, both of which use a weighted combination of Z_1 and $Y_2(Z_1)$ taking the a generalization of the form in equation (1):

$$c(z_1, n_2(z_1)) = \frac{b_2 - z_1 w_1(z_1)}{w_2(z_1)}. \tag{3}$$

- Using a combination test weighting Z_1 and $Y_2(Z_1)$ as planned in the underlying group sequential design (Cui et al., 1999) guarantees control of Type I error by down-weighting observations composing $Y_2(Z_1)$ if the sample size is increased. Noting that we are using the group sequential design n_2, not the more general $n_2(Z_1)$, the weights for $i = 1,2$ are:

$$w_i(z_i) = \sqrt{n_i/(n_1 + n_2)}. \tag{4}$$

  Defining $Z_2(Z_1) = w_1 Z_1 + w_2 Y_2(Z_1)$ we note that the pair $(Z_1, Z_2(Z_1))$ has the same distribution as the pair (Z_1, Z_2) from the group sequential design and, thus, the testing has not changed from the original group sequential test.

- Using a combination that corresponds to testing at the second analysis with a sufficient statistic, equally weighting observations before and after the interim analysis. This requires some restrictions in order to ensure Type I error is controlled (Chen et al., 2004; Gao et al., 2008). Here we have

$$w_1(z_1) = \sqrt{n_1/(n_1 + n_2(z_1))} \tag{5}$$

$$w_2(z_1) = \sqrt{n_2(z_1)/(n_1 + n_2(z_1))}. \tag{6}$$

Now we need to specify $n_2(z_1)$. For the methods presented here, this will be based on conditional power. Let $\beta^\star$ represent a target conditional Type II error for the trial conditioning on the interim test statistic $Z_1 = z_1$; normally $\beta^\star$ equals the Type II error β planned for the underlying group sequential design. This conditional power is computed under the assumption that $\theta = \theta(z_1)$. Some authors propose an efficient estimator such as $\theta^*(z_1) = \hat{\theta}_1 \approx \sqrt{n} z_1$ (Cui et al., 1999; Proschan and Hunsberger, 1995; Chen et al., 2004; Gao et al., 2008; Mehta and Pocock, 2011). However, Liu and Chi (2001) suggest using the originally specified value for the alternate hypothesis for which the group sequential trial was powered, $\theta(z_1) = \theta_1$. In any case, letting $\Phi()$ represent the standard normal cumulative distribution function, we wish to select $n_2(z_1)$ to satisfy

$$\begin{aligned} \beta^\star &= P\{Y_2(z_1) < c(z_1, n_2(z_1)) | \theta = \theta(z_1)\} \\ &= \Phi\left(c(z_1, n_2(z_1)) - \sqrt{n_2(z_1)}\theta(z_1) \right). \end{aligned} \tag{7}$$

To satisfy this, from equations (3) and (7) we must have

$$n_2(z_1) = \left(\frac{c(z_1, n_2(z_1)) - \Phi^{-1}(1-\beta^\star)}{\theta(z_1)} \right)^2 \tag{8}$$

Note that in the case of equation (4) that equation (8) can be used to directly compute $n_2()$ to achieve the desired conditional power since $c(z_1, n_2(z_1))$ does not depend on the function $n_2()$. When $c(z_1, n_2(z1))$ depends on $n_2(z_1)$ in equations (5)-(6), fixed point iteration can be simply applied as follows:

1. Initialize $n_2(z_1)$ using equations (3) and (4).
2. Compute $c(z_1, n_2(z_1))$ based on and the computed n_2-values and desired $c()$-function.
3. Compute $n_2(z_1)$ based on equation (8).
4. Repeat steps 2 and 3 until n_2 converges to a solution.

This normally can be completed adequately in a small, fixed number of iterations.

Code for deriving conditional power at a group sequential design interim analysis

Before deriving adaptive designs based on conditional power, we showing the conditional power for the group sequential design under consideration (without sample size adaptation) assuming the trial continues past the interim analysis. This is computed using the function `condPower()`. The default assumes the true treatment effect for the second stage of the trial is the estimated treatment effect based on data from the first part of the trial, $\hat{\theta}_1 \approx \sqrt{n} z_1$. The following code generates Figure 4 which demonstrates the conditional power for the group sequential design assuming either 1) the interim treatment effect $\hat{\theta}_1$ or 2) the treatment effect θ_1 under the alternate hypothesis. Note that the conditional power based on the observed effect size equals the design power of 80% when $z_1 \approx 1.8$ and drops off sharply for smaller values of z_1.

```
# select z-statistic values at interim to evaluate
z <- seq(0,4,.0125)
# 2nd stage sample size is difference between total sample size
# at interim and final
n2 <- gsD$n.I[2] - gsD$n.I[1]
# compute conditional power based on observed treatment effect
cphat <- condPower(z1=z,n2=n2,x=gsD)
# compute conditional power based on H1 treatment effect
cp1 <- condPower(z1=z,n2=n2,x=gsD, theta=gsD$delta)
# combine these into a data frame and plot
cp <- rbind(data.frame(z1=z,CP=cphat,grp=1),
            data.frame(z1=z,CP=cp1,grp=2))
cp$grp <- factor(cp$grp)
library(scales)
ggplot(data=cp,aes(x=z1,y=CP,lty=grp))+geom_line()+
  ylab("Conditional Power") + xlab(expression(Z[1]))+
```

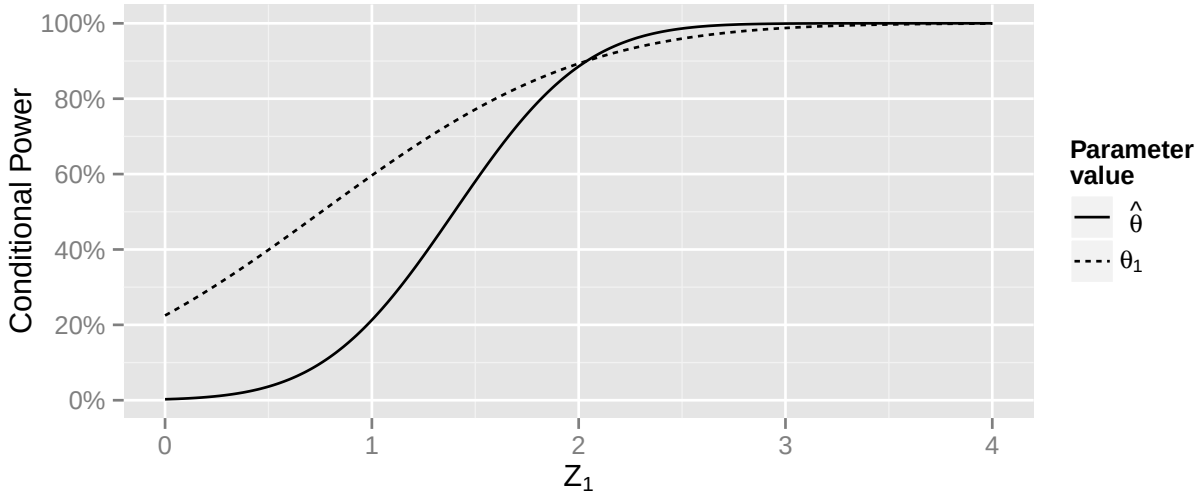

Figure 4: Conditional power based on interim test statistic and assumed parameter value.

```
  scale_y_continuous(labels=percent, breaks=(0:5)/5) +
  scale_linetype_discrete(name="Parameter\nvalue",
                          breaks=c(1,2),
                          labels=c(expression(hat(theta)),expression(theta[1])))
```

While the group sequential design will perform as designed, the presumed 'problem' demonstrated above is that the conditional power for the trial given the interim treatment effect and test statistic is often lower than the originally planned power.

Code for deriving 2-stage conditional power designs

Recall that the above group sequential design was well-powered for the optimistic treatment effect $\delta = .33$, but underpowered for $\delta = .27$. We will attempt to improve power for the lesser treatment effect while maintaining some of the lower sample size benefit of the under-powered group sequential design. We consider three conditional power variations of the group sequential design shown that are based on boosting conditional power:

- "Observed treatment effect" strategy: increase sample size to match conditional power to the originally planned power based assuming the interim test statistic and observed treatment effect.
- "Planned treatment effect" strategy: the same strategy, but assume the originally targeted treatment effect when computing conditional power.
- "Single adapted N" strategy: adapt between only two final sample sizes to improve conditional power.

The first alterative adapts sample size based on conditional power assuming the observed interim treatment effect and the planned final group sequential design sample size, in this case 154, 306. In notation, we use the conditional power $c(z_1, n_2)$ from equation (3) where weights are based on $w_i = n_i/(n_1 + n_2)$, $i = 1,2$. We also demonstrate `Power.ssrCP()` to show exact power for $\delta = .27$ is 68.7%; in the appendix we use `Power.ssrCP()` to create plots of power and expected sample size and in order to compare designs.

```
# select z-statistic values at interim to evaluate
z <- seq(0,4,.0125)
# compute adapted total n assuming observed effect size
# and previously derived group sequential design;
# also include enrollment overrun of 75 at interim
overrun <- 75
n2o <- ssrCP(z1=z, x=gsD, cpadj=c(.3,.8), overrun=overrun)
```

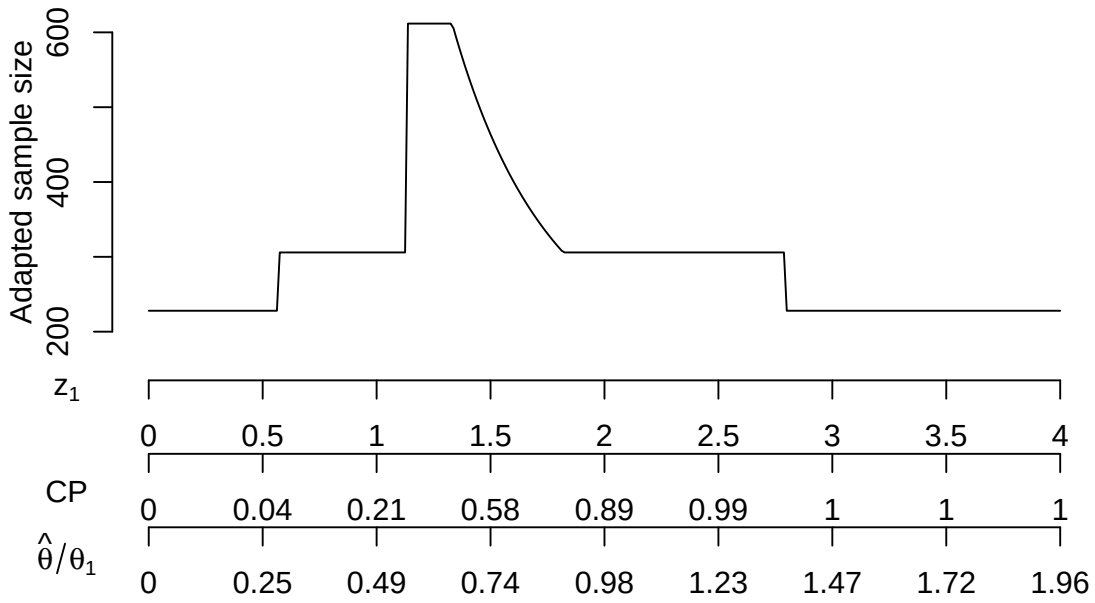

Figure 5: Sample size adaptation based on conditional power based on the observed interim treatment effect.

```
# print power and expected sample size for delta=.27
Power.ssrCP(n2o,theta=.27/2)

##   theta delta     Power       en
## 1 0.135  0.27 0.6868128 330.2952

plot(n2o, ylim=c(200,600))
```

Figure 5 applies the above code to adapt sample size based on the interim test statistic using an extension of the plot command for `ssrCP` objects. Note the three x-axis scales in the plot: 1) the Z-value (Z_1) at the interim analysis, 2) the conditional power at the interim analysis for the underlying group sequential design final sample size assuming the interim treatment effect size, and 3) the interim effect size based on the assumption $Z_1 = \sqrt{n_1}\hat{\theta}_{n_1}$ which will generally hold approximately for asymptotically efficient test statistics. There are 6 intervals on the x-axis to consider on this plot. The two regions on either end represent the interim sample size plus overrun (`ceiling(gsD$n.I[1]+overrun)`=228) when the trial stops for futility ($Z_1 < 0.57$) or efficacy ($Z_1 \geq 2.8$) using the group sequential bounds at the interim analysis:

```
round(c(gsD$lower$bound[1],gsD$upper$bound[1]),2)

## [1] 0.57 2.80
```

If we had wished to plot without accounting for the overrun, setting the parameter `overrun=0` would be required. The intervals just inside these two intervals use the planned final group sequential sample size because the conditional power is outside the interval where sample size is adapted specified by (`cpadj=c(.3,.8)`) and yet the interim test statistic is in the continuation region for the group sequential design. In the above code, we set a maximum increase from the planned sample size of no more than 2 times the originally planned sample size (the default `maxinc=2`); this is seen in the region with peak sample size in the plot. With a default of `z2=z2NC` we assume a combination test based on the weighted Z-statistics with weights based on the pre-planned sample sizes as in equation (4). With the CP scale, you can see that the group sequential sample size is adapted up when the conditional power is between .3 and .8 at the interim analysis. With the $\hat{\theta}/\theta_1$ scale, you can see that if the interim treatment effect is approximately as powered for or better ($\hat{\theta}/\theta_1 \geq 1$), no sample size adaptation is done.

Next we consider sample size adaptation based conditional power assuming the originally planned alternate hypothesis treatment effect (Liu and Chi, 2001). Note that if we were to use the plot command from above, the CP scale would still be based on the observed interim treatment effect. The range of conditional power in which sample size will be adapted has been altered to `cpadj=c(.385,.82)` in order to make the power comparable to the conditional design above as will be seen later in Figure 7. We also set the targeted conditional Type 2 error to .177 to match the maximum conditional power for which the sample size is adapted (1-.177). Note that the maximum increase in sample size is never required for this design (Figure 6).

```
# compute adapted total n assuming H1 effect size
n2h1 <- ssrCP(z1=z, x=gsD, cpadj=c(.385,.823), beta=.177,
              overrun=75, theta=gsD$theta[2])
# print power and expected sample size for delta=.27
Power.ssrCP(n2h1,theta=.27/2)

##   theta delta     Power      en
## 1 0.135  0.27 0.6869699 317.037
```

For the final conditional power design, we set the targeted Type 2 error at the interim analysis to be small in order to set a single possible sample size to adapt to (Figure 6). The maximum increase is set to 1.55 times the group sequential maximum sample size (`maxinc=1.522`), again to match power for the other conditional power designs (Figure 7).

```
# compute adapted sample size at constant level
n2c <- ssrCP(z1=z, x=gsD, cpadj=c(.3,.8), overrun=75,beta=.02,maxinc=1.522)
# print power and expected sample size for delta=.27
Power.ssrCP(n2c,theta=.27/2)

##   theta delta     Power       en
## 1 0.135  0.27 0.6868198 327.0911
```

Figure 6 compares the sample size adaption for the three strategies shown above. In this case, `ggplot` is used to plot after combining design characteristics in a data frame as demonstrated below.

```
# add design descriptor to output data frames
n2o$dat$Design <- "Obs. treatment effect"
n2h1$dat$Design <- "Planned treatment effect"
n2c$dat$Design <- "Single adapted N"
# place z-values and total sample size into a data frame
# for each of these designs
d1 <- rbind(n2o$dat, n2h1$dat, n2c$dat)
# plot total sample size as a function of interim z-value  for both designs
ggplot(data=d1,aes(x=z1,y=n2,col=Design))+geom_line() +
  ylab("Sample size")+xlab("Interim z-value")
```

Alternative group sequential designs

In addition to the group sequential design used as a basis for the above conditional power designs, we consider two alternatives:

- Group sequential (2): Powered for larger effect size ($\delta_1 = .309$) to match overall power of "Observed treatment effect" conditional power design at $\delta = .27$.
- Group sequential (3): Powered for minimum effect size of interest, $\delta_1 = 0.27$.

For both of these alternatives, the interim analysis sample size and bounds are set to be the same as the original group sequential design which means the relative interim timing was decreased from 50% of observations. Spending function parameters are set to match the original design Z_1-cutoff values. The maximum sample sizes for the 3 group sequential designs are 306, 366, and 570, respectively. We leave the code for the comparisons in Figures (7) and (8) to the appendix since there is some detail involved, but there is not much new in terms of understanding the routines of interest.

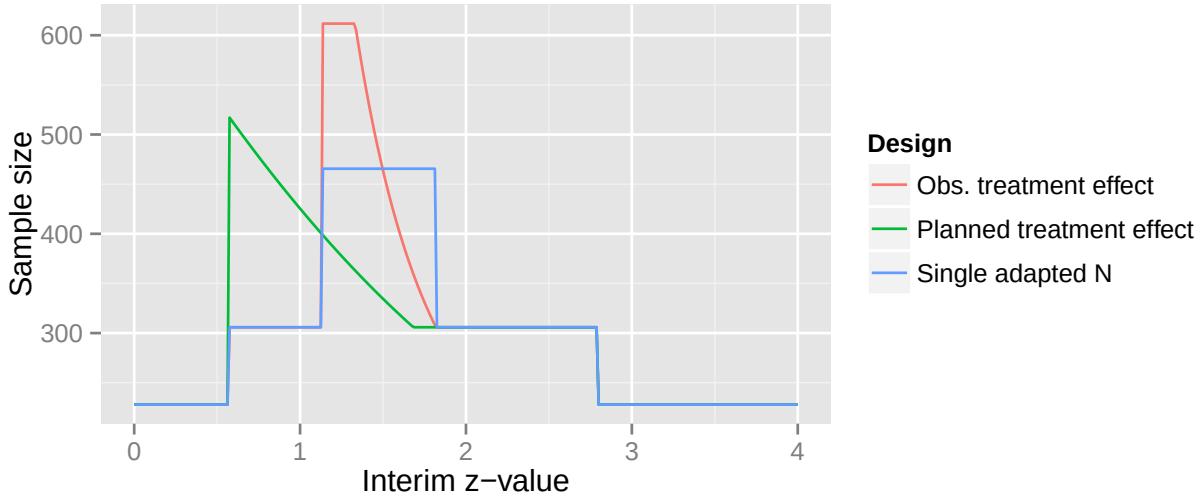

Figure 6: Comparison of 3 sample size adaptation strategies.

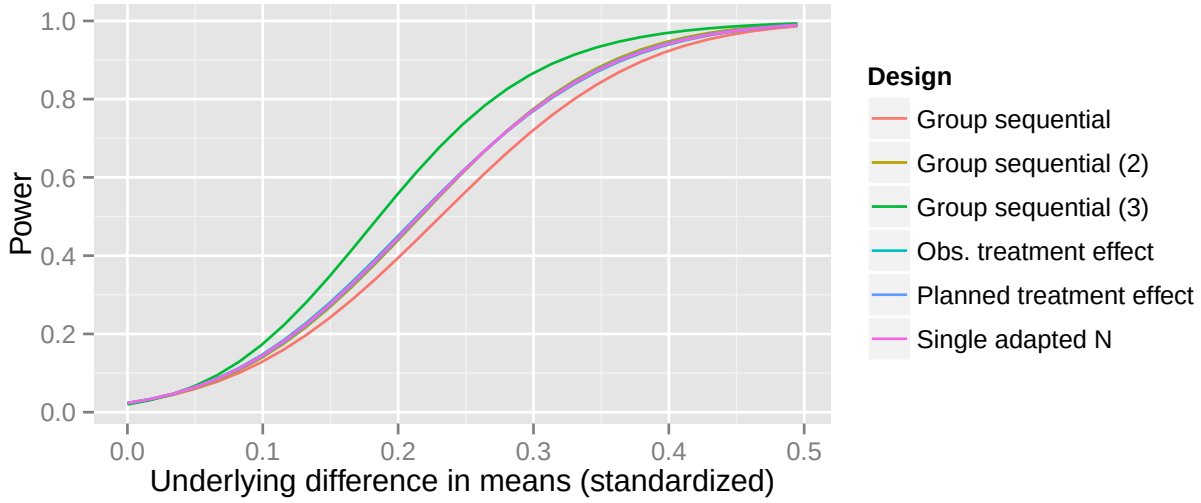

Figure 7: Power by underlying difference in treatment means.

The power for these designs and a group sequential design powered at 80% for $\delta_1 = .27$ (Group sequential (3)) is shown by underlying treatment effect in Figure 7. Like the conditional power designs, the"Group sequential (2)" design matches the power of the "Observed treatment effect" conditional power adaptive design when the treatment effect is $\delta = .27$, the minimum effect size of interest. You can immediately see that all of these designs have nearly identical power curves, slightly improved over the original group sequential design. Moving on to Figure (8), we see that while there are subtle differences in in the expected sample size curves, none of the designs with similar power curves is uniformly better for expected sample size across the range of treatment effects displayed. To get a design adequately powered for the lesser treatment effect of $\delta = .27$ requires a fairly substantial increase in expected sample size as displayed using "Group sequential design (3)."

Summary

We have summarized the theory behind 2-stage group sequential design and an extension to adapt sample size based on conditional power. Code for deriving and comparing these designs was provided. Plots of designs taking advantage of the ggplot2 functionality are supplied with the gsDesign package and we have shown how to derive and plot conditional power.

For the particular designs shown where the group sequential design is based on an optimistic treatment

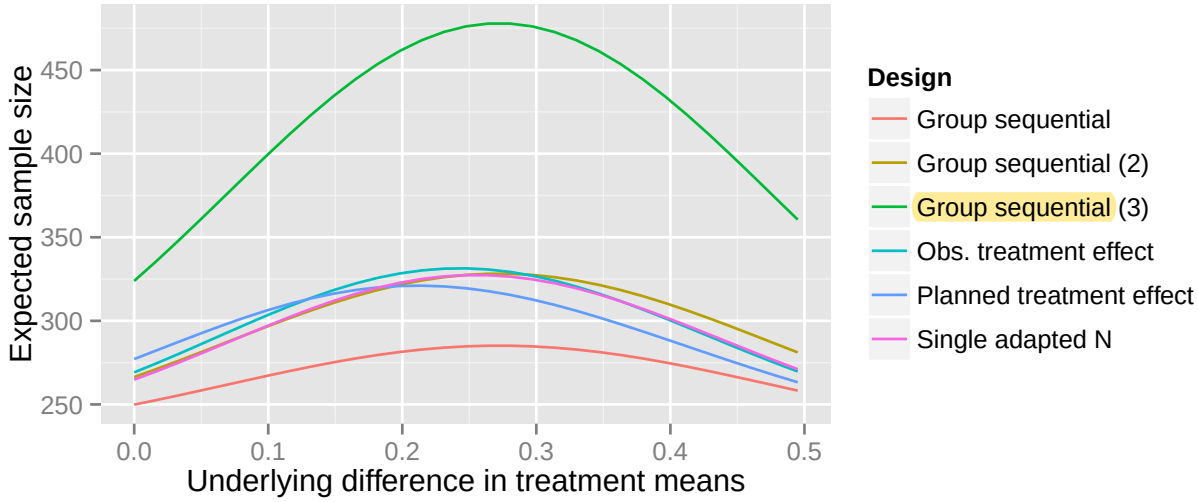

Figure 8: Expected sample size for conditional power and group sequential designs.

effect, the conditional power adaptation does not improve overall power to the desired level; this requires a substantially larger sample size as demonstrated using a fully-powered group sequential design. In terms of efficiency of designs with comparable power curves, expected sample sizes are similar and no design is uniformly better than the others across the range of treatment effects examined. Among designs with similar power and expected sample size, the maximum sample size for the group sequential approach and the "Single adapted N" approach seem more attractive than the more traditional conditional power designs. While we have not generalized these comparisons to other situations, the comparative methods used are easily applied using the provided software. Certainly the comparisons suggest that naively taking one approach or the other without examining design characteristics carefully may be a mistake.

The fact that design characteristics are approximated well without simulation based on asymptotic theory is convenient in terms of the time that may be required to do a full evaluation of alternative design approaches. The software easily describes the implications of design choices by plotting adaptations based on interim outcomes and also makes it easy to compare power and expected sample sizes for different designs across a range of effect sizes.

While the example we have provided used independent normal observations with a known variance, the theory extends to many endpoint types due to asymptotic theory. For independent, identically distributed observations with an efficient test statistic the theory is straightforward. Thus, the methods demonstrated here can generally be applied for many situations encountered in typical clinical trial design. While survival analysis does not fall into this category, the approach can still be taken with some modification (Wassmer, 2006). The functions `gsDesign`, `ssrCP` and `condPower` simplify generation and characterization of 2-stage adaptive sample size using exact calculations and not requiring simulation. The gsDesign package also provides routines that could adapt group sequential designs with multiple interim analyses using either conditional power (`gsCP`) or predictive power (`gsPP`). In particular, the methods of Muller and Schafer (2001) can be applied by computing the conditional error at an interim analysis using the gsDesign function `gsCP` and the remainder of the trial can be designed as an independent trial with Type I error equal to that conditional error with any appropriate design. These designs are not as easily characterized as the 2-stage designs presented here; other R software for such designs in provided by Hack et al. (2013).

Bibliography

P. Bauer. Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie*, 20:130–148, 1989. [p1]

Y. H. J. Chen, D. L. DeMets, and K. K. G. Lan. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*, 23:1023–1038, 2004. doi: 10.1002/sim.1688. [p5, 6]

L. Cui, H. M. J. Hung, and S. J. Wang. Modifications of sample size in group sequential clinical trials. *Biometrics*, 55:853–857, 1999. [p1, 5, 6]

P. Gao, J. H. Ware, and C. Mehta. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, 18:1184–1196, 2008. [p1, 5, 6]

N. Hack, W. Brannath, and M. Brueckner. *AGSDest: Estimation in adaptive group sequential trials*, 2013. URL http://CRAN.R-project.org/package=AGSDest. R package version 2.1. [p11]

C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, Boca Raton, FL, 2000. [p2]

C. Jennison and B. W. Turnbull. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22:971–993, 2003. [p1]

K. Kim and D. L. DeMets. Design and analysis of group sequential tests based on type i error spending rate functions. *Biometrika*, 74:149–154, 1987. [p2]

K. K. G. Lan and D. L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663, 1983. [p2]

W. Lehmacher and G. Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55: 1286–1290, 1999. [p5]

Q. Liu and G. Y. Chi. On sample size and inference for two-stage adaptive designs. *Biometrics*, 57:172–177, 2001. [p1, 5, 6, 9]

Y. Lokhnygina and A. A. Tsiatis. Optimal two-stage group-sequential designs. *Journal of Statistical Planning and Inference*, 138:489–499, 2008. doi: 10.1016. [p1, 5]

C. Mehta and S. Pocock. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*, 30:3267–3284, 2011. doi: 10.1002/sim.4102. [p1, 5, 6]

H. H. Muller and H. Schafer. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57:886–891, 2001. [p11]

P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrika*, 35:549–556, 1979. [p2]

M. Posch, P. Bauer, and W. Brannath. Issues in designing flexible trials. *Statistics in Medicine*, 22:953–969, 2003. [p1, 5]

M. A. Proschan and S. A. Hunsberger. Designed extension of studies based on conditional power. *Biometrics*, 51:1315–1324, 1995. [p1, 5, 6]

N. Schmitz. *Optimal Sequentially Planned Decision Procedures.*, volume 79 of *Lecture Notes in Statistics*. Springer, New York, 1993. [p1]

S.-J. Wang, H. M. J. Hung, and R. O'Neill. Paradigms for adaptive statistical information designs: practical experiences and strategies. *Statistics in Medicine*, 31:3011–3023, 2012. doi: 10.1002/sim.5410. [p1]

G. Wassmer. Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal*, 48:714–729, 2006. [p11]

Keaven M. Anderson and Alison Pedley
Merck Research Laboratories
351 N. Sumneytown Pike, UG1C-46
Upper Gwynedd, Pennsylvania 19454
United States of America
keaven_anderson@merck.com

Appendix: code for Figures 7 and 8

The following code is commented fairly heavily in an attempt to explain the details of computations.

```
# select standardized effect sizes from 0 to 1.5 x H1 effect size
thetaplot <- (0:30)/20 * gsD$theta[2]
# compute corresponding parameters on delta (natural parameter) scale
deltaplot <- thetaplot * 2
# compute expected sample size and power for conditional power designs
en1 <- Power.ssrCP(n2o,theta=thetaplot)
en2 <- Power.ssrCP(n2h1,theta=thetaplot)
en3 <- Power.ssrCP(n2c,theta=thetaplot)
# add additional effect sizes to original group sequential design
# using gsProbability
gsDx <- gsProbability(d=gsD,theta=thetaplot)
# upper$prob and lower$prob are matrices containing boundary
# crossing probabilities for each effect size
en4 <- data.frame(theta=thetaplot,
                  en=gsDx$en,
                  Power=as.vector(gsDx$upper$prob[1,]+gsDx$upper$prob[2,]),
                  delta=deltaplot,
                  Design="Group sequential")

# compute group sequential design powered for a larger effect size
# and repeat above calculations; timing set for IA at 150
# start by getting updated standardized effect size;
# following value was computed with guess and test
deltanew <- .3095
n.fix<- nNormal(delta1=deltanew, sd=1, beta=.2, alpha=.025)
overrun <- 75
# initial timing < .5 of final due to larger final
# sample size; exact value should not be critical here
timing <- .4255
# need lower boundary crossing probability for new effect size
# under original design to know how to set lower spending
# when iterations start
gsDa <- gsProbability(d=gsD,theta=deltanew/2)
sflprob <- gsDa$lower$prob[1,1]

# values to start iterations below
n <- (deltanew/.27)^2*gsD$n.I[2]
sfupar <- gsD$upper$param
sflpar <- gsD$lower$param
nI <- n # an arbitrary 'bad' starting value
# do fixed point interation to match interim bounds
# and timing with original group sequential design
while(abs(nI-gsD$n.I[1])>.01){
  timing <- gsD$n.I[1]/n
  gsD2 <- gsDesign(k=2, timing=timing, beta=.2, alpha=.025, delta1=deltanew,
                   n.fix=n.fix,
                   overrun=overrun,sfu=sfPower,sfl=sfPower,sfupar=sfupar,sflpar=sflpar)
  sfupar <- (log(gsD$upper$prob[1,1])-log(.025))/log(timing)
  gsDa <- gsProbability(d=gsD2,theta=deltanew/2)
  sflpar <- (log(sflprob)-log(.2))/log(timing)
  n <- gsD2$n.I[2]
  nI <- gsD2$n.I[1]
}
# sum(gsProbability(d=gsD2,theta=.27/2)$upper$prob)
gsDx2 <- gsProbability(d=gsD2, theta=thetaplot)
en5 <- data.frame(theta=thetaplot,
                  en=gsDx2$en,
                  Power=as.vector(gsDx2$upper$prob[1,]+gsDx2$upper$prob[2,]),
```

```
                  delta=deltaplot,
                  Design="Group sequential (2)")

# do computations for timing and spending functions
# for design fully powered for delta=.27
nnor3 <- nNormal(delta1=.27, sd=1, beta=.2, alpha=.025)
thetanew <- gsDesign(beta=.2,n.fix=nnor3)$delta
gsDa <- gsProbability(d=gsD,theta=thetanew)
timing <- .269 # this was done by guess and test
sfupar <- (log(gsD$upper$prob[1,1])-log(.025))/log(timing)
sflpar <- (log(gsDa$lower$prob[1,1])-log(.2))/log(timing)

# now for a fully-powered group sequential design
gsD3 <- gsDesign(k=2, timing=timing, beta=.2, alpha=.025, delta1=.27,
                 n.fix=nNormal(delta1=.27, sd=1, beta=.2, alpha=.025),
                 overrun=75,sfu=sfPower,sfl=sfPower,sfupar=sfupar,sflpar=sflpar)
gsDx3 <- gsProbability(d=gsD3, theta=thetaplot)
en6 <- data.frame(theta=thetaplot,
                  en=gsDx3$en,
                  Power=as.vector(gsDx3$upper$prob[1,]+gsDx3$upper$prob[2,]),
                  delta=deltaplot,
                  Design="Group sequential (3)")
# add design names for conditional power designs
en1$Design <- "Obs. treatment effect"
en2$Design <- "Planned treatment effect"
en3$Design <- "Single adapted N"
# put characteristics for all designs in a single data frame
en1 <- rbind(en1, en2, en3, en4, en5, en6)
# power plot
ggplot(data=en1,aes(x=delta,y=Power,col=Design))+geom_line() +
  ylab("Power")+xlab("Underlying difference in means (standardized)")+
  scale_y_continuous(breaks=seq(0,1,.2))
# expected sample size plot
ggplot(data=en1,aes(x=delta,y=en,col=Design))+geom_line() +
  ylab("Expected sample size")+xlab("Underlying difference in treatment means")
```