A very interesting project – the NIA IMPACT Collaboratory – where a primary goal is to fund a large group of pragmatic pilot studies to investigate promising interventions to improve health care and quality of life for people living with Alzheimer’s disease and related dementias. One of my roles on the project team is to advise potential applicants on the development of their proposals. In order to provide helpful advice, it is important that we understand what we should actually expect to learn from a relatively small pilot study of a new intervention.

There is a rich literature on this topic. For example, these papers by [*Lancaster et al*](https://onlinelibrary.wiley.com/doi/full/10.1111/j..2002.384.doc.x) and [*Leon et al*](https://www.sciencedirect.com/science/article/pii/S002239561000292X) provide nice discussions about how pilot studies should fit into the context of larger randomized trials. The key point made by both groups of authors is that pilot studies are important sources of information about the *feasibility* of conducting a larger, more informative study: Can the intervention actually be implemented well enough to study it? Will it be possible to recruit and retain patients? How difficult will it be to measure the primary outcome? Indeed, what is the most appropriate outcome to be measuring?

Another thing the authors agree on is that the pilot study is *not* generally well-equipped to provide an estimate of the treatment effect. Because pilot studies are limited in resources (both time and money), sample sizes tend to be quite small. As a result, any estimate of the treatment effect is going to be quite noisy. If we accept the notion that there is some true underlying treatment effect for a particular intervention and population of interest, the pilot study estimate may very well fall relatively far from that true value. As a result, if we use that effect size estimate (rather than the true value) to estimate sample size requirements for the larger randomized trial, we run a substantial risk of designing an RCT that is too small, which may lead us to miss identifying a true effect. (Likewise, we may end up with a study that is too large, using up precious resources.)

My goal here is to use simulations to see how a small pilot study could potentially lead to poor design decisions with respect to sample size.

**A small, two-arm pilot study**

In these simulations, I will assume a two-arm study (intervention and control) with a true intervention effect \(\Delta = 50\). The outcome is a continuous measure with a within-arm standard deviation \(\sigma = 100\). In some fields of research, the effect size would be standardized as \(d = \Delta / \sigma\). (This is also known as [Cohen’s \(d\)](https://rpsychologist.com/d3/cohend/).) So, in this case the true standardized effect size \(d=0.5\).

If we knew the true effect size and variance, we could skip the pilot study and proceed directly to estimate the sample size required for 80% power and Type I error rate \(\alpha = 0.05\). Using the pwr.t.test function in the pwr library, we specify the treatment effect (as \(d\)), significance level \(\alpha\), and power to get the number of subjects needed for each study arm. In this case, it would be 64 (for a total of 128):

library(pwr)

pwr.t.test(n = NULL, d = 50/100, sig.level = 0.05,

power = 0.80, type = "two.sample")

##

## Two-sample t test power calculation

##

## n = 64

## d = 0.5

## sig.level = 0.05

## power = 0.8

## alternative = two.sided

##

## NOTE: n is number in \*each\* group

If we do not have an estimate of \(d\) or even of the individual components \(\Delta\) and \(\sigma\), we may decide to do a small pilot study. I simulate a single study with 30 subjects in each arm (for a total study sample size of 60). First, I generate the data set (representing this one version of the hypothetical study) with a treatment indicator \(rx\) and an outcome \(y\):

library(simstudy)

defd <- defDataAdd(varname = "y", formula = "rx \* 50", variance = 100^2)

ss <- 30

set.seed(22821)

dd <- genData(n = ss\*2)

dd <- trtAssign(dd, grpName = "rx")

dd <- addColumns(defd, dd)

head(dd)

## id rx y

## 1: 1 0 -150

## 2: 2 1 48

## 3: 3 0 -230

## 4: 4 1 116

## 5: 5 1 91

## 6: 6 1 105

Once we have collected the data from the pilot study, we probably would try to get sample size requirements for the larger RCT. The question is, what information can we use to inform \(d\)? We have a couple of options. In the first case, we can estimate both \(\Delta\) and \(\sigma\) from the data and use those results directly in power calculations:

lmfit <- lm(y ~ rx, data = dd)

Delta <- coef(lmfit)["rx"]

Delta

## rx

## 78

sd.rx <- dd[rx==1, sd(y)]

sd.ctl <- dd[rx==0, sd(y)]

[pool.sd](http://pool.sd) <- sqrt( (sd.rx^2 + sd.ctl^2) / 2 )

[pool.sd](http://pool.sd)

## [1] 94

The estimated standard deviation (94) is less than the true value, and the effect size is inflated (78), so that the estimated \(\hat{d}\) is also too large, close to 0.83. This is going to lead us to recruit fewer participants (24 in each group) than the number we actually require (64 in each group):

pwr.t.test(n = NULL, d = Delta/[pool.sd](http://pool.sd), sig.level = 0.05,

power = 0.80, type = "two.sample")

##

## Two-sample t test power calculation

##

## n = 24

## d = 0.83

## sig.level = 0.05

## power = 0.8

## alternative = two.sided

##

## NOTE: n is number in \*each\* group

Alternatively, if we had external information that provided some insight into the true effect size, or, absent that, we use a minimally clinically significant effect size, we might get a better result. In this case, we are quite fortunate to use an effect size of 50. However, we will continue to use the variance estimate from the pilot study. Using this approach, the resulting sample size (56) happens to be much closer to the required value (64):

pwr.t.test(n = NULL, d = 50/[pool.sd](http://pool.sd), sig.level = 0.05,

power = 0.80, type = "two.sample")

##

## Two-sample t test power calculation

##

## n = 56

## d = 0.53

## sig.level = 0.05

## power = 0.8

## alternative = two.sided

##

## NOTE: n is number in \*each\* group

**Speak truth to power**

Now the question becomes, what is the true expected power of the RCT based on the sample size estimated in the pilot study. To estimate this true power, we use the true effect size and the true variance (i.e. the true \(d\))?

In the first case, where we actually used the true \(d\) to get the sample size estimate, we just recover the 80% power estimate. No surprise there:

pwr.t.test(n = 64, d = 0.50, sig.level = 0.05, type = "two.sample")$power

## [1] 0.8

In the second case, where we used \(\hat{d} = \hat{\Delta} / \hat{\sigma}\) to get the sample size \(n=24\), the true power of the larger RCT would be 40%:

pwr.t.test(n = 24, d = 0.50, sig.level = 0.05, type = "two.sample")$power

## [1] 0.4

And if we had used \(\hat{d} = 50 / \hat{\sigma}\) to get the sample size estimate \(n=56\), the true power would have been 75%:

pwr.t.test(n = 56, d = 0.50, sig.level = 0.05, type = "two.sample")$power

## [1] 0.75

**Conservative estimate of standard deviation**

While the two papers I cited earlier suggest that it is not appropriate to use effect sizes estimated from a pilot study (and more on that in the next and last section), this [1995 paper](https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780141709) by R.H. Browne presents the idea that we *can* use the estimated standard deviation from the pilot study. Or rather, to be conservative, we can use the upper limit of a one-sided confidence interval for the standard deviation estimated from the pilot study.

The confidence interval for the standard deviation is not routinely provided in R. Another [paper](https://www.sciencedirect.com/science/article/pii/S0378375810005070?via%3Dihub) analyzes one-sided confidence intervals quite generally under different conditions, and provides a formula in the most straightforward case under assumptions of normality to estimate the \(\gamma\*100\%\) one-sided confidence interval for \(\sigma^2\):

\[  
\left( 0,\frac{(N-2)s\_{pooled}^2}{\chi^2\_{N-2;\gamma}} \right)  
\]

where \(\chi^2\_{N-2;\gamma}\) is determined by \(P(\chi^2\_{N-2} > \chi^2\_{N-2;\gamma}) = \gamma\). So, if \(\gamma = 0.95\) then we can get a one-sided 95% confidence interval for the standard deviation using that formulation:

gamma <- 0.95

qchi <- qchisq(gamma, df = 2\*ss - 2, lower.tail = FALSE)

ucl <- sqrt( ( (2\*ss - 2) \* [pool.sd](http://pool.sd)^2 ) / qchi )

ucl

## [1] 111

The point estimate \(\hat{\sigma}\) is 94, and the one-sided 95% confidence interval is \((0, 111)\). (I’m happy to provide a simulation to demonstrate that this is in fact the case, but won’t do it here in the interest of space.)

If we use \(\hat{\sigma}\_{ucl} = 111\) to estimate the sample size, we get a more conservative sample size requirement (78) than if we used the point estimate \(\hat{\sigma} = 94\) (where the sample size requirement was 56):

pwr.t.test(n = NULL, d = 50/ucl, sig.level = 0.05,

power = 0.80, type = "two.sample")

##

## Two-sample t test power calculation

##

## n = 78

## d = 0.45

## sig.level = 0.05

## power = 0.8

## alternative = two.sided

##

## NOTE: n is number in \*each\* group

Ultimately, using \(\gamma = 0.95\) might be too conservative in that it might lead to an excessively large sample size requirement. Browne’s paper uses simulation to to evaluate a range of \(\gamma\)’s, from 0.5 to 0.9, which I also do in the next section.

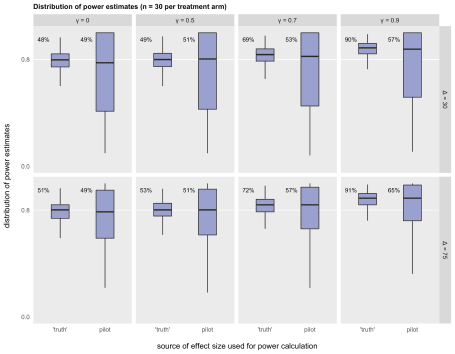
**Simulation of different approaches**

At this point, we need to generate multiple iterations to see how the various approaches perform over *repeated* pilot studies based on the same data generating process, rather than looking at a single instance as I did in the simulations above.

As Browne does in his paper, I would like to evaluate the distribution of power estimates that arise from the various approaches. I compare using an external source or minimally clinically meaningful effect size to estimate \(\Delta\) (in the figures below, this would be the columns labeled *‘truth’*) with using the effect size point estimate from the pilot (labeled *pilot*). I also compare using a point estimate of \(\sigma\) from the pilot (where \(\gamma=0\)), with using the upper limit of a one-sided confidence interval defined by \(\gamma\). In these simulations I compare three levels of \(\gamma\): \(\gamma \in (0.5, 0.7, 0.9)\).

In each of the simulations, I assume 30 subjects per arm, and evaluate true effect sizes of 30 and 75. In all cases, the true standard error \(\sigma = 100\) so that true \(d\) is 0.30 or 0.75.

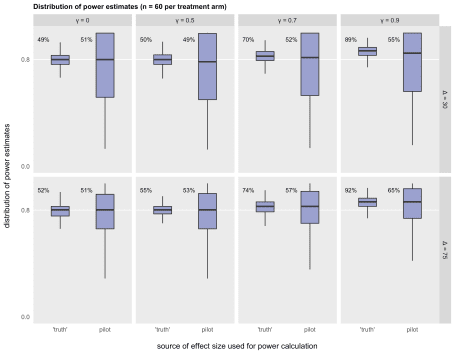
The box plots in the figure represent the distribution of power estimates for the larger RCT under different scenarios. Each scenario was simulated 5000 times each. Ideally, the power estimates should cluster close to 80%, the targeted level of power. In the figure, the percentage next to each box plot reports the percent of simulations with power estimates at or above the target of 80%.



Two things jump out at me. First, using the true effect size in the power calculation gives us a much better chance of designing an RCT with close to 80% power, even when a point estimate is used for \(\hat{\sigma}\). In Browne’s paper, the focus is on the fact that even when using the true effect size, there is a high probability of power falling below 80%. This may be the case, but it may be more important to note that when power is lower than the target, it is actually likely to fall relatively close to the 80% target. If the researcher is very concerned about falling below that threshold, perhaps using \(\gamma\) higher than 0.6 or 0.7 might provide an adequate cushion.

Second, it appears *using the effect size estimate from the pilot as the basis for an RCT power analysis is risky*. The box plots labeled as *pilot* exhibit much more variation than the *‘true’* box plots. As a result, there is a high probability that the true power will fall considerably below 80%. And in many other cases, the true power will be unnecessarily large, due to the fact that they have been designed to be larger than they need to be.

The situation improves somewhat with larger pilot studies, as shown below with 60 patients per arm, where variation seems to be reduced. Still, an argument can be made that using effect sizes from pilot studies is too risky, leading to an under-powered or overpowered study, neither of which is ideal.



A question remains about how best to determine what effect size to use for the power calculation if using the estimate from the pilot is risky. I think a principled approach, such as drawing effect size estimates from the existing literature or using clinically meaningful effect sizes, is a much better way to go. And the pilot study should focus on other important feasibility issues that *can* help improve the design of the RCT.

References:

Lancaster, G.A., Dodd, S. and Williamson, P.R., 2004. Design and analysis of pilot studies: recommendations for good practice. Journal of evaluation in clinical practice, 10(2), pp.307-312.

Leon, A.C., Davis, L.L. and Kraemer, H.C., 2011. The role and interpretation of pilot studies in clinical research. Journal of psychiatric research, 45(5), pp.626-629.

Browne, R.H., 1995. On the use of a pilot sample for sample size determination. Statistics in medicine, 14(17), pp.1933-1940.

Cojbasic, V. and Loncar, D., 2011. One-sided confidence intervals for population variances of skewed distributions. Journal of Statistical Planning and Inference, 141(5), pp.1667-1672.

Support:

This research is supported by the National Institutes of Health National Institute on Aging U54AG063546. The views expressed are those of the author and do not necessarily represent the official position of the funding organizations.

**Addendum**

Below is the code I used to run the simulations and generate the plots

getPower <- function(ssize, esize, gamma = 0, use.est = FALSE) {

estring <- paste0("rx \* ", esize)

defd <- defDataAdd(varname = "y", formula = estring, variance = 100^2)

N <- ssize \* 2

dd <- genData(n = N)

dd <- trtAssign(dd, grpName = "rx")

dd <- addColumns(defd, dd)

lmfit <- lm(y~rx, data = dd)

sd.rx <- dd[rx==1, sd(y)]

sd.ctl <- dd[rx==0, sd(y)]

[pool.sd](http://pool.sd) <- sqrt( (sd.rx^2 + sd.ctl^2) / 2 )

qchi <- qchisq(gamma, df = N - 2, lower.tail = FALSE)

ucl <- sqrt( ( (N-2) \* [pool.sd](http://pool.sd)^2 ) / qchi )

[p.sd](http://p.sd) <- estsd \* (gamma == 0) + ucl \* (gamma > 0)

p.eff <- esize \* (use.est == FALSE) +

coef(lmfit)["rx"] \* (use.est == TRUE)

if (abs(p.eff/[p.sd](http://p.sd)) < 0.0002) p.eff <- sign(p.eff) \* .0002 \* [p.sd](http://p.sd)

nstar <- round(pwr.t.test(n = NULL, d = p.eff/[p.sd](http://p.sd), sig.level = 0.05,

power = 0.80, type = "two.sample")$n,0)

power <- pwr.t.test(n=nstar, d = esize/100, sig.level = 0.05,

type = "two.sample")

return(data.table(ssize, esize, gamma, use.est,

estsd = estsd, ucl = ucl, nstar, power = power$power,

est = coef(lmfit)["rx"],

lcl.est = confint(lmfit)["rx",1] ,

ucl.est = confint(lmfit)["rx",2])

)

}

dres <- data.table()

for (i in c(30, 60)) {

for (j in c(30, 75)) {

for (k in c(0, .5, .7)) {

for (l in c(FALSE, TRUE)) {

dd <- rbindlist(lapply(1:5000,

function(x) getPower(ssize = i, esize = j, gamma = k, use.est = l))

)

dres <- rbind(dres, dd)

}}}}

above80 <- dres[, .(x80 = mean(power >= 0.80)),

keyby = .(ssize, esize, gamma, use.est)]

above80[, l80 := scales::percent(x80, accuracy = 1)]

g\_labeller <- function(value) {

paste("\U03B3", "=", value) # unicode for gamma

}

e\_labeller <- function(value) {

paste("\U0394", "=", value) # unicdoe for Delta

}

ggplot(data = dres[ssize == 30],

aes(x=factor(use.est, labels=c("'truth'", "pilot")), y=power)) +

geom\_hline(yintercept = 0.8, color = "white") +

geom\_boxplot(outlier.shape = NA, fill = "#9ba1cf", width = .4) +

theme(panel.grid = element\_blank(),

panel.background = element\_rect(fill = "grey92"),

axis.ticks = element\_blank(),

plot.title = element\_text(size = 9, face = "bold")) +

facet\_grid(esize ~ gamma,

labeller = labeller(gamma = g\_labeller, esize = e\_labeller)) +

scale\_x\_discrete(

name = "\n source of effect size used for power calculation") +

scale\_y\_continuous(limits = c(0,1), breaks = c(0, .8),

name = "distribution of power estimates \n") +

ggtitle("Distribution of power estimates (n = 30 per treatment arm)") +

geom\_text(data = above80[ssize == 30],

aes(label = l80), x=rep(c(0.63, 1.59), 6), y = 0.95,

size = 2.5)