What do we learn from an equivalence study without statistical power?

(Forthcoming letter in the Annals of Emergency Medicine)

Kyle Peyton, MA, MPhil; Rachel Solnick, MD, MSc[1]

We were interested to read the article by Linden et al (2017) that claims to provide evidence to "support the safety of omitting a pelvic examination in women with a confirmed intrauterine pregnancy," (p. 833).[1] In light of previous studies questioning the utility of the pelvic exam, this trial addresses a critical question for patients with threatened abortion. We commend the authors for conducting the largest randomized trial on pelvic exams to date, but several methodological issues raise questions about whether clinicians can make reliable inferences from this study.

First, the implemented experiment does not seem to have adequate statistical power to detect equivalence. The results suggest the intervention (no pelvic exam) reduced "composite morbidity" by -0.024, with a 90% confidence interval of (-0.118, 0.071). This is a noisy estimate, and the 90% interval is not contained inside the chosen equivalence range (-0.08, 0.08). This does not provide evidence of equivalence.

Although the study's power calculations suggest 0.80 power with 720 participants, only 221 of the 1,280 patients deemed eligible for participation were randomly assigned. Our simulations (Figure 1), suggest the idealized study would detect equivalence about 84% of the time, whereas the implemented study never would. Why does the failure to detect equivalence support the safety of omitting the pelvic exam?

Although the authors suggest this result supports omitting a pelvic exam, the power of the reported experiment was approximately zero. Although reported power calculations based on 720 participants suggest 0.80 power, only 221 of the 1,280 patients deemed eligible for participation were randomized. As illustrated in Figure 1, the implemented experiment (N = 221) would fail to detect equivalence in all of the 10,000 simulations we conducted.

As shown in Figure 1, the equivalence range determines the decision rule for the test – we conclude equivalence if and only if the 90% interval falls inside this range. The wider the equivalence range, the more likely we are to declare equivalence. The choice of this margin is widely recognized as one of the most important decisions in the design of an equivalence study[2–4]. Why was an 8% change from a 15% baseline chosen as the appropriate equivalence range?

---

[1] Kyle Peyton is PhD candidate in Political Science, Statistical Consultant at the Center for Science and Social Science Information, Policy Fellow at the Institution for Social and Policy Studies, and Research Affiliate at the Human Cooperation Lab, Yale University (kyle.peyton@yale.edu). Solnick is Director of Health Policy at the Emergency Medicine Residents' Association, and Resident Physician in Emergency Medicine, Yale New Haven Hospital (rachel.solnick@yale.edu).

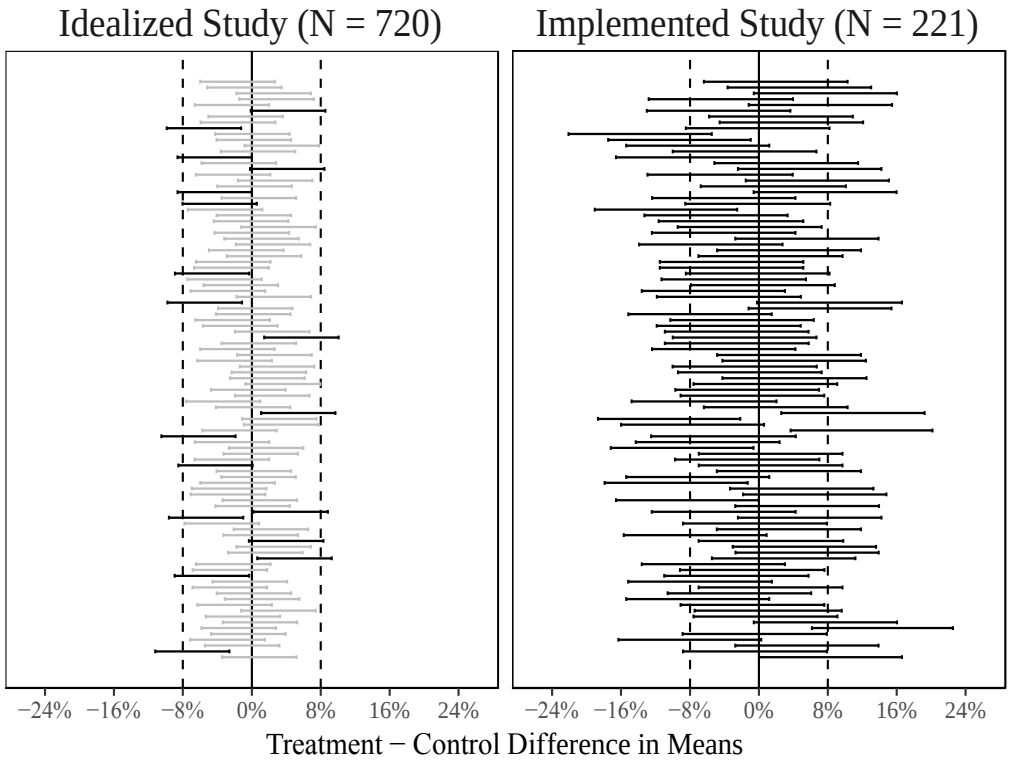


Figure 1: Two simulated versions of Linden et al (2017) with idealized versus implemented sample sizes. Each plot shows a random sample of 100 results from 10,000 simulations of each design. Estimated differences in composite morbidity are displayed as 90% confidence intervals. Vertical dotted lines are the author's selected equivalence range of 8%. Since equivalence is true in these simulations, grey intervals are correct rejections and black intervals are incorrect rejections. The idealized study has 0.84 power to detect equivalence, but the implemented study has 0 power.

A final issue is that the primary outcome was calculated by truncating morbidity data in a way that seems to attenuate differences between treatment and control groups. For example, patients who may have reported multiple morbidities are included as having one "composite morbidity". Does this increase the likelihood of declaring equivalence when there are nonetheless meaningful differences between groups?

We were excited to see a randomized experiment on such an important clinical topic, but have strong reservations about using this study to support omitting exams that are otherwise routinely recommended. We hope that future studies in this space are adequately powered to detect equivalence within a clinically appropriate margin. Such studies could be informative for emergency clinicians who care for pregnant women.

References

1. Linden JA, Grimmnitz B, Hagopian L, et al. Is the pelvic examination still crucial in patients presenting to the emergency department with vaginal bleeding or abdominal pain when an intrauterine pregnancy is identified on ultrasonography? A randomized controlled trial. *Annals of emergency medicine*. 2017;70(6):825-834.

2. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled clinical trials*. 2002;23(1):2-14.

3. Wellek S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press; 2010.

4. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. *Journal of general internal medicine*. 2011;26(2):192-196.

Supplementary code for "What Do We Learn from an Equivalence Study Without Statistical Power?"

*Kyle Peyton and Rachel Solnick**

First, let's reproduce the main result in Linden et al. (2017), hereafter LEA, as reported in Table 2,

```
# Reproduce main result from study:
n_trt <- 102
n_ctrl <- 100
trt <- c(rep(1, 20), rep(0, n_trt-20))
ctrl <- c(rep(1, 22), rep(0, n_ctrl-22))

s <- sqrt(((n_trt - 1)*var(trt) + (n_ctrl-1)*var(ctrl))/(n_trt + n_ctrl - 2))
se_diff <- s*sqrt(1/n_trt + 1/n_ctrl)

dim <- mean(trt) - mean(ctrl)
c(dim-1.645*se_diff, dim+1.645*se_diff)
```

```
## [1] -0.11829435  0.07045121
```

Next, let's define a helper function that assigns treatment and computes the 90% confidence interval for the difference in means estimator.

```
# Helper function for simulations
ci_fun <- function(N = 200, m = NULL, Y1 = NULL, Y0 = NULL, simple = TRUE,
                   truncate = FALSE, cutpoint = NULL){
  # Do a random assignment
  if(simple == FALSE){
    require(randomizr)
    Z <- complete_ra(N = N, m = m)
  } else{
    Z <- rbinom(n = N, size = 1, prob = 1/2)
  }

  # Diff in means for observed data
  Y <- Y1*Z + Y0*(1-Z)

  if(truncate == TRUE){
    Y <- ifelse(Y >= cutpoint, 1, 0)
  }

  fit <- lm(Y ~ Z)
  c(fit$coefficients[2], confint(fit, level = 0.90)[2,])
}
```

The helper functions allows for LEA's estimation strategy of truncating the outcome variable for patients with more than 1 adverse event via the `truncate` argument in the helper function above. We do not believe this is good practice for an equivalence study since it generates estimates that are attenuated toward zero.

Let's set up the simulations to estimate the power of LEA's study under a scenario where the null hypothesis is false so that the pelvic exam and omitting the pelvic exam are indeed equivalent. Let $R = 10,000$ denote the number of repetitions and $\delta = 0.08$ denote the equivalence margin.

*Peyton: kyle.peyton@yale.edu. Solnick: rachel.solnick@yale.edu.

Let $Y_i(0)$ denote individual i's potential outcome under control (pelvic exam), and $Y_i(1)$ denote individual i's potential outcome under treatment (no pelvic exam). Following LEA, we assume the adverse event rate is 0.15 among the untreated. For the simulations, we assume a true (constant, unobservable) unit level treatment effect $Y_i(1) - Y_i(0)$ of approximately zero so that our estimand of interest, the average treatment effect (ATE) $\mathbb{E}[\tau] := \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$, is in fact zero. To simulate the fixed potential outcomes under control, we take N draws from the Poisson distribution with $\lambda = 0.15$.

```
R <- 10000
delta <- 0.08

# Adverse event rate is 0.15 in control, and there is no effect of witholding
#   the pelvic exam (treatment):
N <- 720
Y0 <- rpois(n = N, lambda = 0.15)
Y1 <- Y0
```

Let Z_i denote the binary treatment assigment indicator so that m units are treated and $N - m$ units are in control. The estimator here is the simple difference in means, $\hat{\tau}_{DM} = \frac{1}{m}\sum_{i=1}^{m} Y_i \cdot Z_i - \frac{1}{N-m}\sum_{i=m+1}^{N-m} Y_i \cdot (1 - Z_i)$. The null hypothesis (non-equivalence) is then $H_0 : |\hat{\tau}_{DM}| > \delta$ and the alternative hypothesis (equivalence) is $H_A : |\hat{\tau}_{DM}| < \delta$. The power of this test is

$$\begin{aligned} 1 - \Pr(\textit{Type II Error}) &= 1 - \Pr(\textit{fail to reject } H_0 \mid H_0 \textit{ is false}) \\ &= \Pr(\textit{reject } H_0 \mid H_0 \textit{ false}) \\ &= \Pr\left(\textit{declare } |\hat{\tau}_{DM}| < \delta \;\middle|\; |\hat{\tau}_{DM}| < \delta\right) \end{aligned}$$

The TOST procedure used by LEA rejects if and only if the lower bound of the 90% confidence interval is inside the equivalence range, $LB(\hat{\tau}_{DM}) > -\delta$, **and** the upper bound is also inside the range, $UB(\hat{\tau}_{DM}) < \delta$. Failing to reject the null of non-equivalence (because the 90% interval is not bounded by $\pm\ \delta$) is a Type II error when the null is actually false. To estimate the power of this test, we simulate 10,000 intervals under LEA's design. Let's start with the idealized version:

```
# Simulate equivalence intervals when null is false (equivalence true) under
#   ideal design with ~ .80 power.
sim_cis_ideal <- t(replicate(R, ci_fun(N = N, Y1 = Y1, Y0 = Y0,
                                       truncate = TRUE, cutpoint = 1)))

# Confirm power is >= 0.83
sum(sim_cis_ideal[,2] > -delta & sim_cis_ideal[,3] < delta)/R
```

```
## [1] 0.8385
```

As expected, the idealized version is well powered. When the null is actually false (as specified above) the test correctly rejects about 84% of the time. The code below performs the same exercise for the implemented version of LEA's experiment under charitable assumptions that ignore potential problems due to the application of exclusion rules, treatment non-compliance, and attrition.

```
# Now simulate equivalence intervals under design actually implemented.
#   Be generous and assume nobody lost to followup.
#   Assume rate is still 0.15 after patients are excluded, etc.
n <- 221
Y0 <- rpois(n = n, lambda = 0.15)
Y1 <- Y0

sim_cis_real <- t(replicate(R, ci_fun(N = n, Y1 = Y1, Y0 = Y0,
                                      truncate = TRUE, cutpoint = 1)))
```

```r
# What's the power of this design?
sum(sim_cis_real[,2] > -delta & sim_cis_real[,3] < delta)/R
```

```
## [1] 0
```

The power of the test under this design is approximately zero. To illustrate the differences, we plot a random sample of 100 CIs from each scenario. A study with a Type II error rate of 0 would (correctly) reject the null every time. LEA's idealized study (left panel) has a Type II error rate of about 0.16 (1-0.84), so it (correctly) rejects the null of non-equivalence about 84% of the time. Correct rejections are indicated by the grey 90% confidence intervals. LEA's implemented study, with a Type II error rate of approx. 1, never (correctly) rejects the null. Incorrect rejections are indicated by the black 90% confidence intervals.

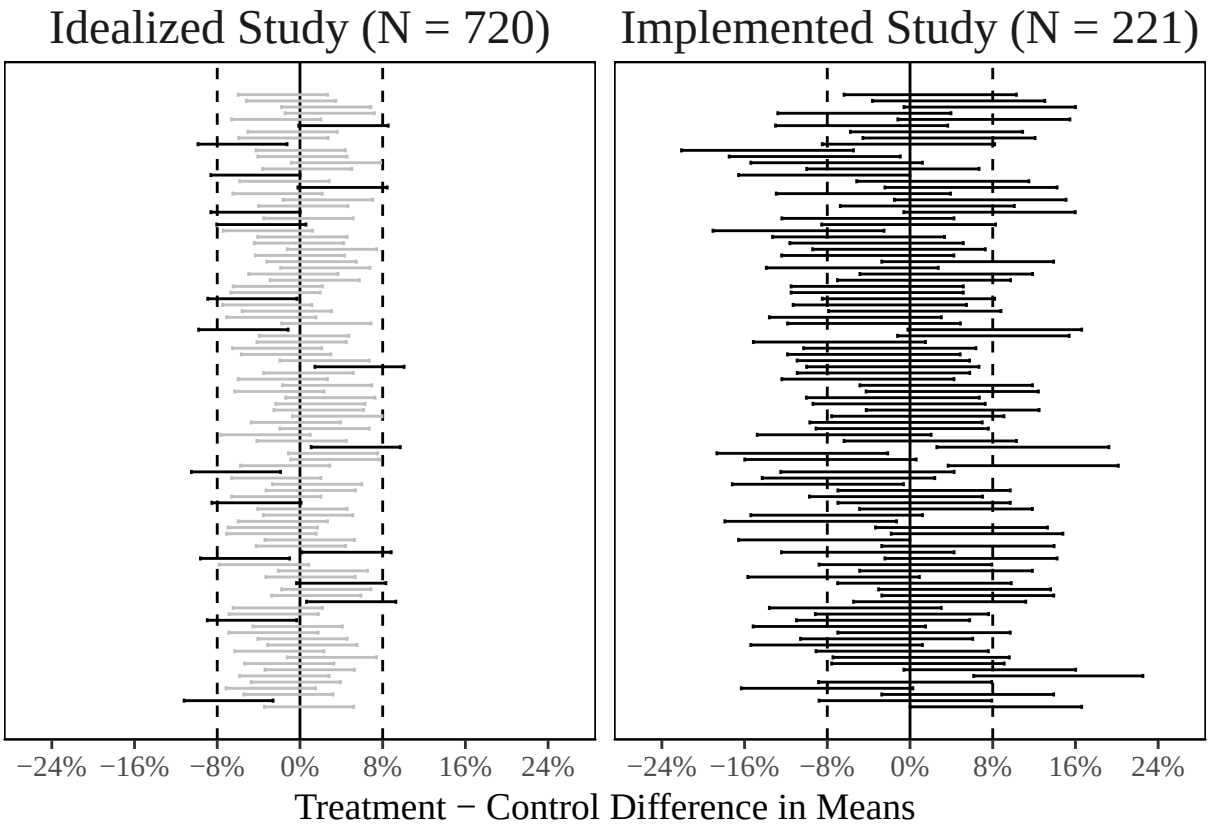
```r
# Take random sample of 100 CIs from each scenario and plot.
sim_cis_null <- rbind(sim_cis_ideal[sample(1:R, 100),],
                      sim_cis_real[sample(1:R, 100),])

# Plot random sample of 100 CIs when the null is false
sim_df_null <- data.frame(dhat = sim_cis_null[, 1], L = sim_cis_null[ ,2],
                          U = sim_cis_null[ ,3],
                          type = c(rep("Idealized Study (N = 720)", 100),
                                   rep("Implemented Study (N = 221)", 100)))

# Flag CIs that would lead to rejection of non-equivalence null, this is
#   the correct choice in this setting. Failing to reject null of
#   non-equivalence is a type 2 error in this setting
index <- which(sim_df_null$L > -delta & sim_df_null$U < delta)

sim_df_null$error <- TRUE
sim_df_null[index,]$error <- FALSE

equiv_plot_null <-
  ggplot(sim_df_null, aes(x = rep(1:100, 2), y = dhat)) +
  geom_hline(yintercept = 0, col = "black", lty = 1, size = 0.5) +
  geom_hline(yintercept = 0.08, col = "black", lty = 2, size = 0.5) +
  geom_hline(yintercept = -0.08, col = "black", lty = 2, size = 0.5) +
  geom_errorbar(aes(ymax = U, ymin = L, color = error), width = 0.65,
                size = 0.5) +
  scale_color_manual(name = "", values = c("grey", "black")) +
  facet_wrap( ~ type) +
  xlab("") + ylab("") +
  scale_y_continuous(name = "Treatment - Control Difference in Means",
                     labels = percent_format(), limits = c(-.26,.26),
                     breaks = c(-0.24, -0.16, -0.08, 0, 0.16, 0.08, 0.24)) +
  coord_flip() +
  theme_tufte(base_size = 14) +
  theme(legend.position="none",
        axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        strip.text.x = element_text(size = 18),
        axis.line.x = element_line(color="black"),
        panel.border = element_rect(colour = "black", fill=NA))
equiv_plot_null
```


Idealized Study (N = 720)
Implemented Study (N = 221)
−24%
−16%
−8%
0%
8%
16%
24%
Treatment − Control Difference in Means