Health Anxiety Study - Power and Sample Size

Michael C Sachs

2016-08-31

Summary

To analyze the type of trial under consideration, we recommend a linear mixed effects model with the treatment group by visit time interaction being the main parameter of interest. A model with random intercepts and slopes would suffice to account for the correlated observations and missing data. The treatment by time interaction, multiplied by 12, can be interpreted as the average difference between treatment groups in the change in HAI score over 12 weeks. For instance, in the previous study, we might report that over the course of 12 weeks of therapy, BSM treatment led to a 2.4 point smaller decrease in HAI score as compared to KBI (95% CI: 0.1 to 4.6 points). Report the changes over 12 weeks in each treatment group with confidence intervals for ease of interpretation.

In the new study, for a sample size of 250 [*Erratum. The proper text should read "200" for a one-sided test; see page 7–i.e., the Results section–of this power analysis report. /Principal investigator*] individuals total, and a true treatment difference of 0 d, there is approximately 80% power to confirm noninferiority, using a 95% confidence interval to rule out the margin of 0.3 Cohen's d.

Background and Assumptions

The investigators are planning a randomized clinical trial comparing two modalities of cognitive behavioral therapy (CBT): internet-based self-guided and face-to-face. The treatments each last for 12 weeks. The primary outcome measure is the short version of the health anxiety inventory (HAI). The HAI will be measured at baseline before treatment (time 0) and 12 more times after that at the end of each week.

The main hypothesis is that internet-based CBT is no worse at improving health related anxiety compared to face to face CBT. In previous trials, face-to-face CBT has been shown to be efficacious at decreasing the HAI as compared to a control treatment. Several studies report an effect size of 1.4 - 1.5 (Cohen's d), while one study showed a much smaller effect size of 0.3 - 0.4.

The trial under consideration here is therefore planned as a non-inferiority study, with a noninferiority margin of 0.3. That is, we aim to demonstrate conclusively that internet-based CBT is no more than 0.3 standard deviations (20% worse, assuming an effect of

1.5) worse than face-to-face CBT in terms of reducing the HAI over the course of 12 weeks. This margin was determined based on clinical judgement in consideration of the previously reported effect sizes in similar studies.

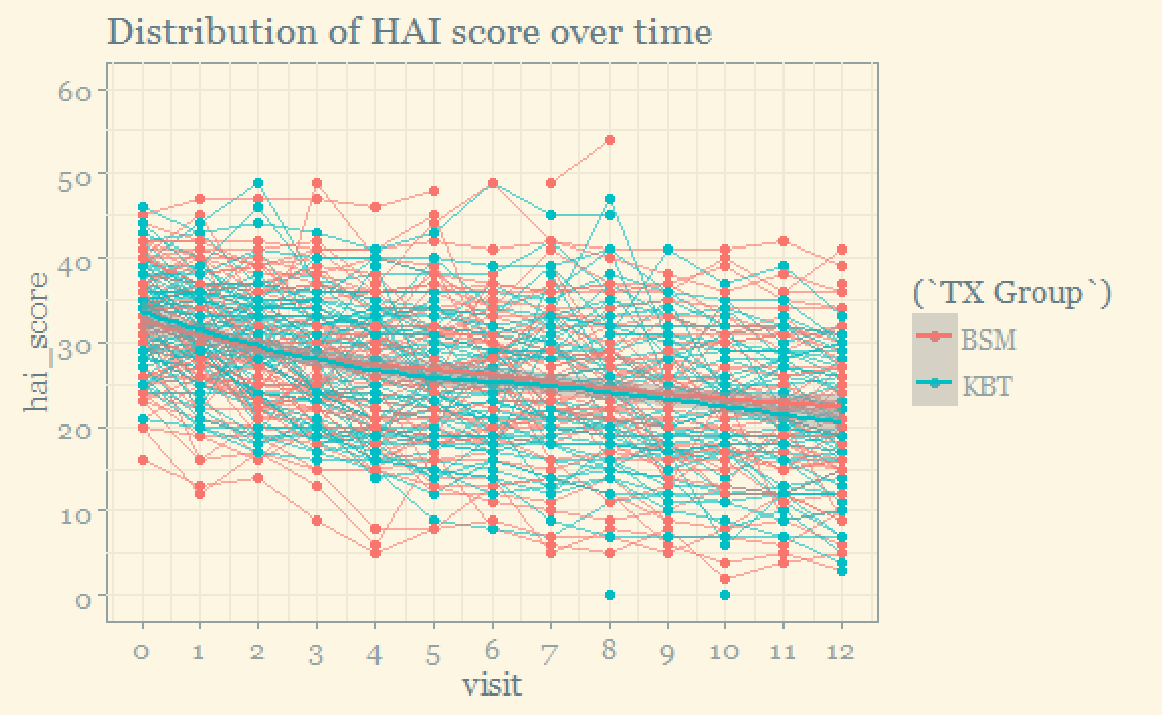
Based on preliminary data, the average change in HAI score over 12 weeks was approximately -10 points, and the standard deviation of that change was about 7.5 points. Therefore, a Cohen's d of 0.3 corresponds to a difference in slopes of 7.5 * 0.3 = 2.25. Thus 2.25 will be our non-inferiority margin on the scale of changes over 12 weeks.

Data Analysis and Interpretation

Previous studies have only compared the post treatment HAI score between groups. There is a wealth of information contained in the other 12 weeks worth of data that can be used to improved the statistical efficiency of the analysis. Here we will demonstrate how using data from a previous study.

Preliminary data used to obtain estimates for key parameters we obtained from the file called "Health anxiety example data to Michael with password.xlsx" sent on 2016-08-19 from Erik Hedman. This trial is similar to the one being planned. We use these data to guide our assumptions on the parameters used to calculated power for the new trial.

The following plot shows the distributions of the HAI scores over time, by treatment group. The smooth curves are flexible fitted models that represent the average trend in each group. The trends are roughly linear over the observation period.



A simple way to analyze this trial would be to compare the mean HAI score at the end of the study by treatment group. Since the trial is randomized, we know that there is no true difference in HAI score at baseline. However, due to sampling variability, there may be an observed difference in mean HAI score at baseline between treatment groups. Furthermore, since the distribution of HAI score at baseline has a fairly wide spread, we can gain efficiency by calculating the change in HAI score from baseline to 12 weeks for each subject, and then comparing the mean change in HAI score by treatment group.

A logical extension to that is to use information from each visit to estimate the average change in HAI score for each subject. This is called a derived variable analysis, where for each subject, the derived variable is the slope for the change in HAI score over time. We then compare the average slopes by treatment group.

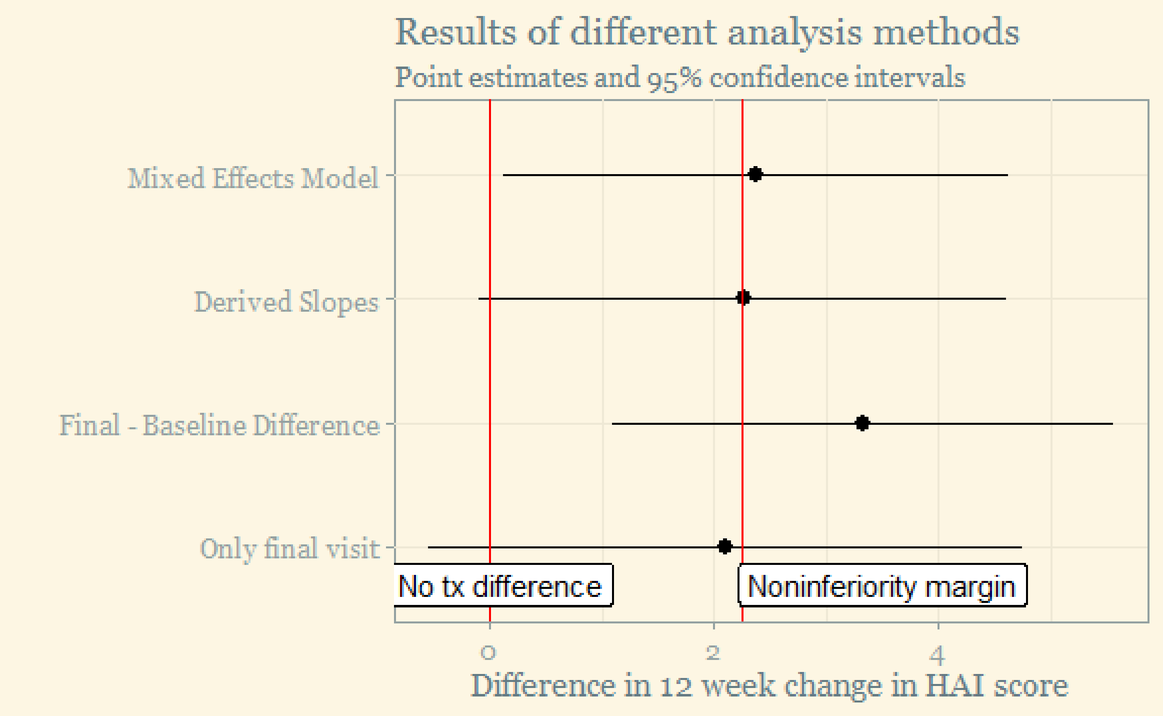
A further extension to this concept is to use linear mixed effects models to estimate the average difference between treatment groups in the change in HAI score over time. Such a model would account for the within-subject correlation over time, in addition to the intermittent missing data. Specifically, the model is

$$Y_{it} = (\alpha + a_i) + (\beta + b_i) * t + \delta * Z_i + \gamma * t * Z_i + \varepsilon_i,$$

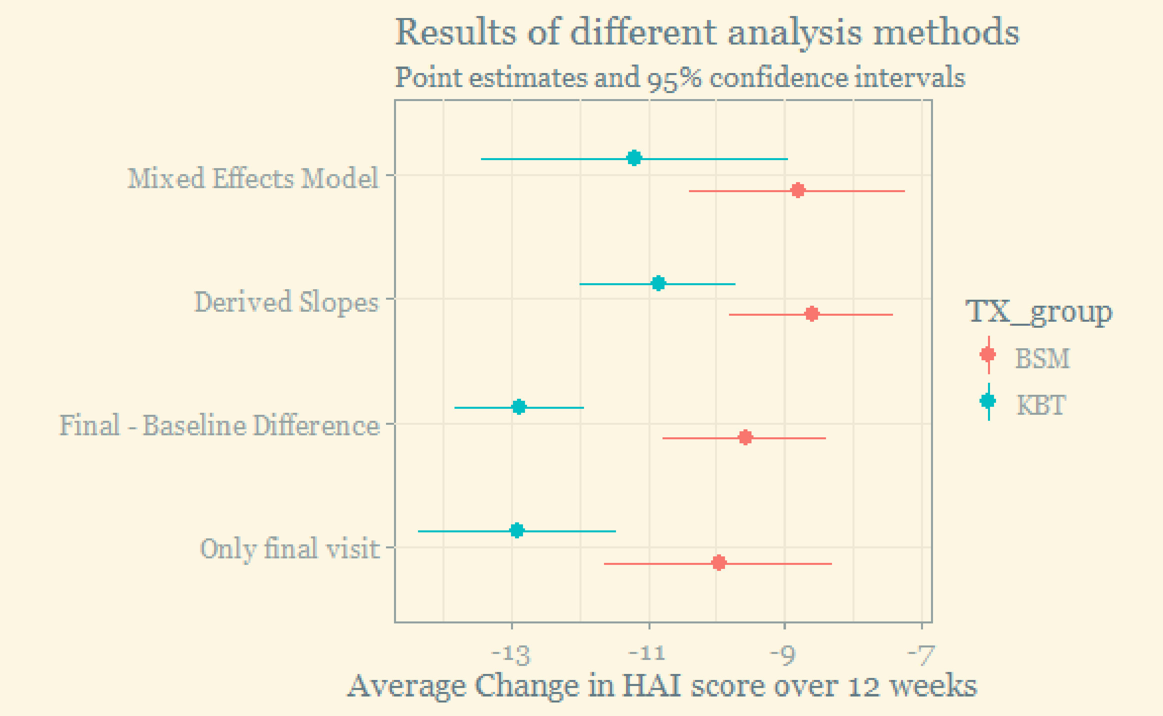
where Y_{it} is the HAI score for subject i at visit t, Z_i is the treatment group for subject i, and a_i, b_i, ε_i are normally distributed random variables with mean 0 (random effects). This model can be referred to as a linear mixed effects model with random intercepts and random slopes. The fixed effects are parameters for the time effect, the treatment effect, and their interaction. The parameter of interest is the interaction parameter γ, which represents the average difference between treatment groups in the change in HAI score per week. The model can be estimated with maximum likelihood using a variety of software. One example using the lme4 package in R is given in the appendix. We also provide a link to the SPSS documentation for mixed effects models.

The results of these different analysis methods are shown in the next figure. The parameters shown all have the same interpretation: the difference between treatment groups in the standard deviation change in HAI score after 12 weeks of therapy. In this study all confidence intervals overlap with the noninferiority margin. The final visit and derived slopes analysis confidence intervals also overlap with 0, meaning that those results are inconclusive. The other two confidence intervals exclude 0, indicating that one treatment is superior.

The width of the confidence intervals is an indicator of the precision of the parameter estimates. The confidence interval width of the "Only final visit" method is 5.3, compared to the others that range from 4.45 to 4.7. This supports our belief that the "Only final visit" approach is clearly inferior.



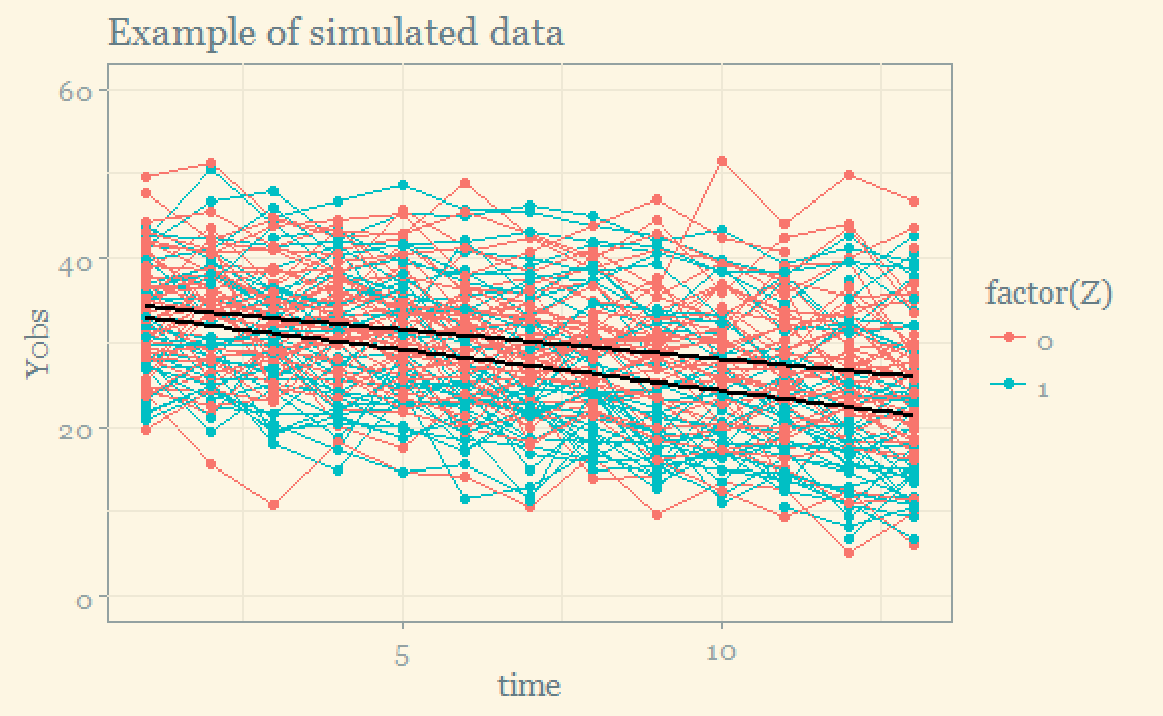
These parameters (interaction terms) can be difficult to interpret, so it is also good to present the effects within each treatment group to determine which one is more effective. The next plot shows the average change in HAI score over 12 weeks by treatment group, where the average change is estimated using the different methods as above. It appears that the KBT treatment group, on average, had a greater decline in HAI score over 12 weeks.



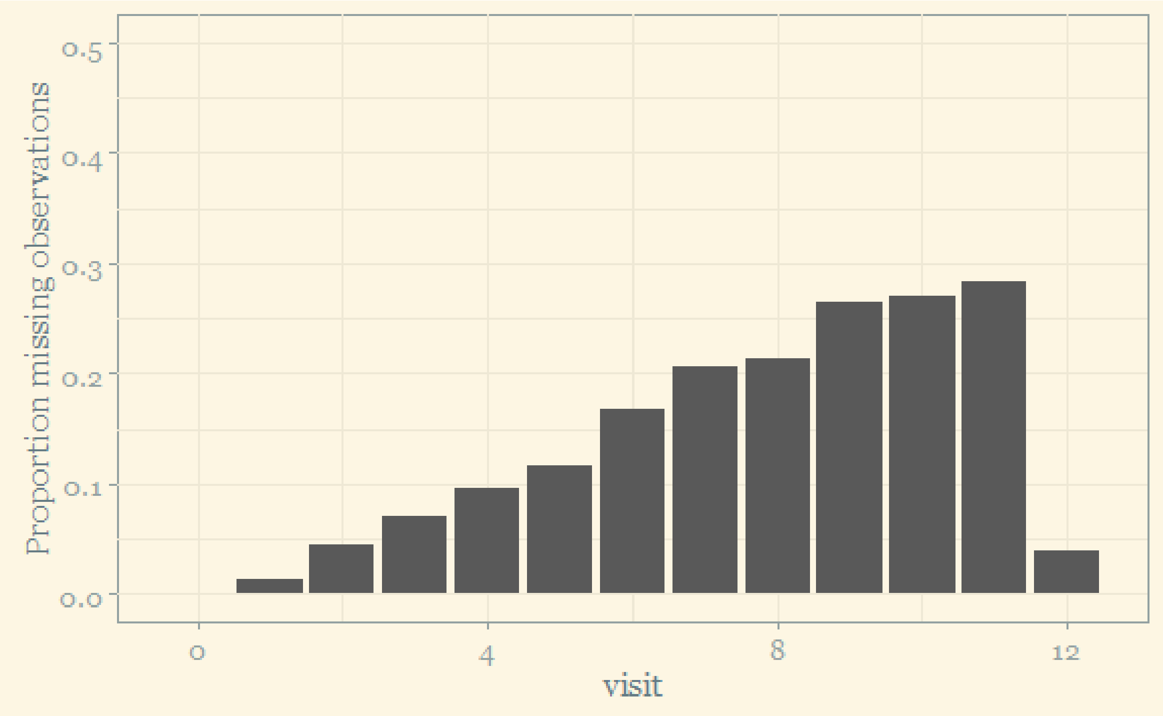
Power and Sample Size

A simulation based approach was used to evaluate power and sample size. The aforementioned data was used as guide on the data-generation mechanism. Briefly, the linear mixed effects model described in the equation above was used to generate hypothetical observations, using the estimated values from the previous study as parameter values. The data were generated using different values of the sample size, the value of γ (our parameter of interest), and different missing data mechanisms.

For each set of parameters, we generated data for a hypothetical trial. It was analyzed using the mixed effects model as described above, with Wald based confidence intervals, and we looked to see if the upper limit of the confidence interval excluded the noninferiority margin. If it excluded the margin, then the trial was considered a success. This procedure was replicated 5000 times, and the proportion of successes gives us an estimate of the power under those conditions. The next plot shows data from a single replicate of the experiment. Compare it to the real data set shown above.



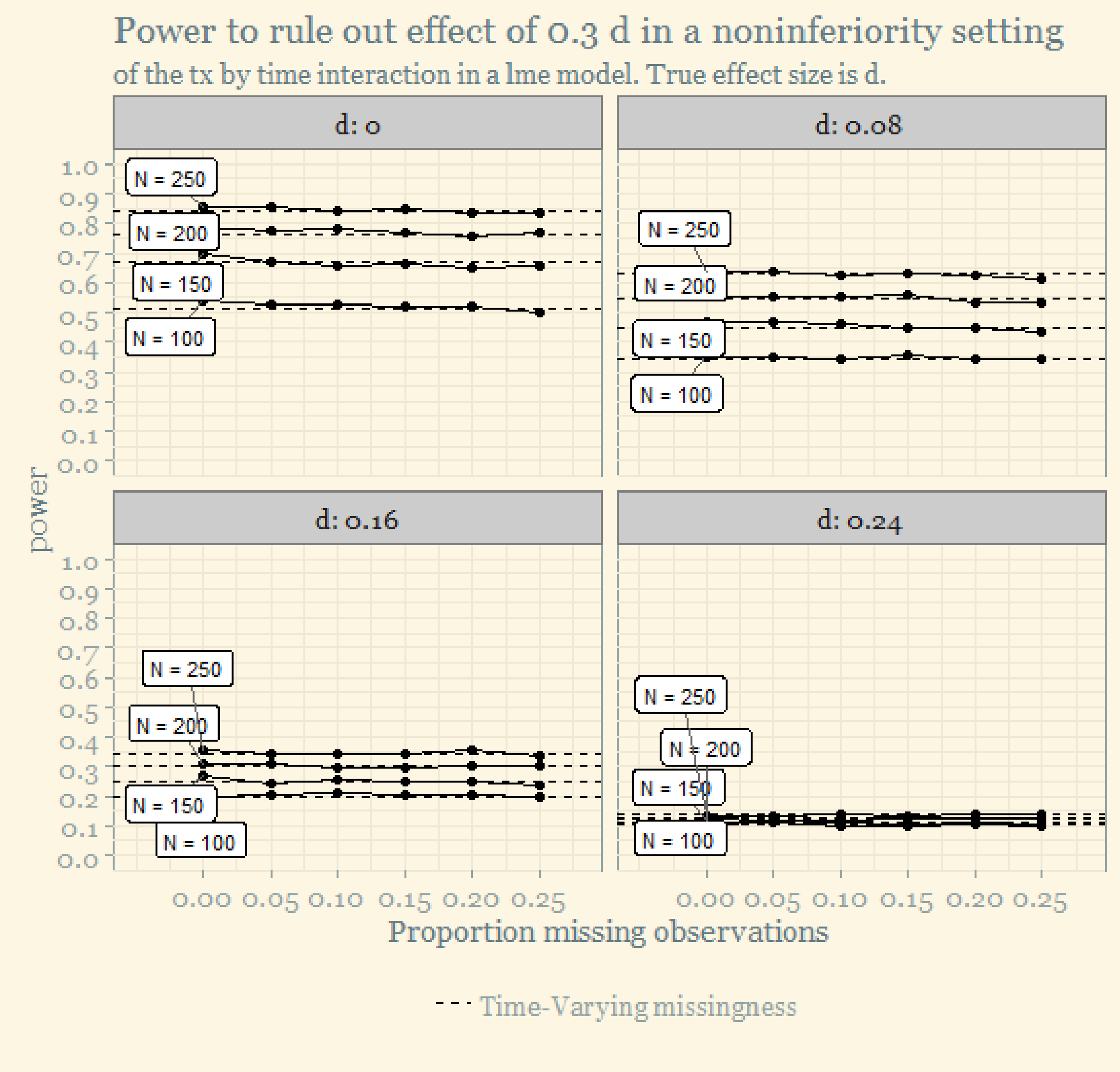
Several different missing data mechanisms were explored. First, different proportions of missing data were assumed to be uniform over the visit times. Second, we used the distribution of missingness from the prior study, which was clearly not uniform over the visit times (see below).



Results

Each line below represents a different sample size, ranging from 100 to 250. The 4 panels correspond to 4 values of the treatment difference of interest (the interaction term) on the Cohen's d scale. A larger value of d means that the experimental treatment is inferior to the standard treatment, but below the noninferiority margin. The dotted lines represent the scenario with missing data generated the same way it was observed in the previous trial.

For a sample size of 200 individuals total, and a true treatment difference of 0, there is approximately 80% power to confirm noninferiority, using a one-sided 95% confidence interval to rule out the margin. To achieve the same power with a true treatment difference of 0.08 would require greater than 250 individuals. As d approaches the noninferiority margin, the probability of declaring noninferiority approaches the type I error rate. Compared to a two-sided test, 200 individuals and a treatment difference of 0 only achieves 70% power.



Power to rule out effect of 0.3 d in a noninferiority setting

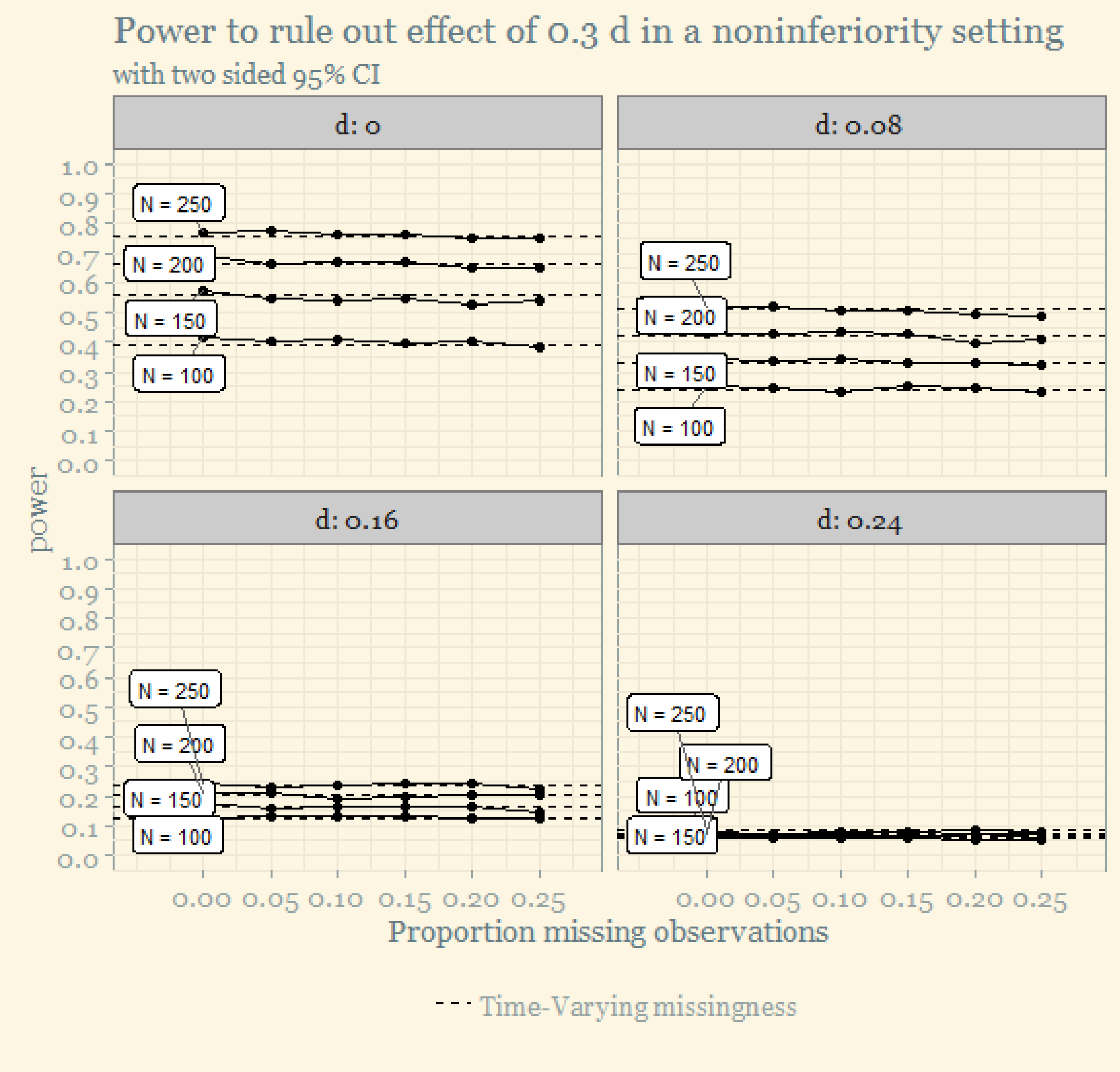
of the tx by time interaction in a lme model. True effect size is d.

d: 0

d: 0.08

d: 0.16

d: 0.24



Appendix

Additional Resources (Links)

1. Reporting of Noninferiority Studies
2. Tutorial on linear mixed effects models, Part 1
3. Tutorial on linear mixed effects models, Part 2
4. Mixed models in SPSS

Code

```r
library(rio)
library(tidyr)
library(stringr)
library(lme4)
library(dplyr)
library(purrr)
library(ggplot2)
library(ggthemes)
library(ggrepel)
library(extrafont)
loadfonts(device = "win")

theme_set(theme_solarized(base_family = "Georgia"))

has <- import("../Data/Health anxiety example data to Michael with password.c
sv")

has$`HAI-M_shai_Veckom.[0]` <- has$`HAI-L_shai_Pre HAX`
has$`HAI-M_shai_Veckom.[12]` <- has$`HAI-L_shai_POST`
haslong <- has %>% gather(visitcode, hai_score, starts_with("HAI-M_shai_Vecko
m."))
haslong$visit <- as.numeric(str_match(haslong$visitcode, ".*\\[([0-9]+)\\]")[
, 2])
ggplot(haslong, aes(x = visit, y = hai_score, color = (`TX Group`), group = `
Internt ID`)) +
  geom_line(alpha = 0.6) + geom_point() + ggtitle("Distribution of HAI score
over time") +
  stat_smooth(method = "loess", se = TRUE, aes(group = `TX Group`)) + scale_x
_continuous(breaks = 0:12) +
  scale_y_continuous(limits = c(0, 60), breaks = seq(0, 60, by = 10))
sd.raw <- sd(has$`HAI-L_shai_POST`, na.rm = TRUE)
sd.diff <- sd(has$HAI_DIFF, na.rm = TRUE)

test.raw <- t.test(has$`HAI-L_shai_POST` ~ has$`TX Group`)

has$HAI_DIFF <- has$`HAI-L_shai_POST` - has$`HAI-L_shai_Pre HAX`
test.diff <- t.test(has$HAI_DIFF ~ has$`TX Group`)

hasslope <- haslong %>% group_by(`Internt ID`) %>%
  do({
    data.frame(haislope = lm(hai_score ~ visit, data = .)$coefficients[2] * 1
2,
               TX_group = .$`TX Group`[1], stringsAsFactors = FALSE)
  })

sd.slope <- sd(hasslope$haislope, na.rm = TRUE)

test.slope <- t.test(haislope ~ TX_group, data = hasslope)
```

```r
## mixed effects model
fitlme <- lmer(hai_score ~ visit * I(`TX Group` == "BSM") + (1 + visit | `Int
ernt ID`), data = haslong)
lme.est <- data.frame(point = fixef(fitlme)[4] * 12,
             lower = (fixef(fitlme)[4] * 12 - 1.96 * sqrt(diag(vcov(fitlme)))
[4] * (12)) ,
             upper = (fixef(fitlme)[4] * 12 + 1.96 * sqrt(diag(vcov(fitlme)))
[4] * (12)) )

pte <- function(test.obj) {

  data.frame(point = diff(rev(test.obj$estimate)),
             lower = test.obj$conf.int[1],
             upper = test.obj$conf.int[2])

}

resplo <- list(test.raw, test.diff, test.slope) %>% map_df(pte) %>%
  bind_rows(lme.est)

resplo$desc <- c("Only final visit", "Final - Baseline Difference", "Derived
Slopes", "Mixed Effects Model")

ggplot(resplo, aes(x = desc, y = point, ymin = lower, ymax = upper)) +
  geom_pointrange() + geom_hline(yintercept = c(0.0, 2.25), color = "red") +
  annotate("label", x = c(.7, .7), y = c(0.1, 2.25 + 1.25),
           label = c("No tx difference", "Noninferiority margin")) +
  scale_x_discrete(limits = resplo$desc) +
  xlab("") + ylab("Difference in 12 week change in HAI score") + ggtitle("Res
ults of different analysis methods", subtitle = "Point estimates and 95% conf
idence intervals") +
  coord_flip()

bytrt <- has %>% group_by(`TX Group`) %>%
  summarize(mnpost = mean(`HAI-L_shai_POST`, na.rm = TRUE) - mean(`HAI-L_shai
_Pre HAX`, na.rm = TRUE),
            sdpost = sqrt(var(`HAI-L_shai_POST`, na.rm = TRUE) + var(`HAI-L_s
hai_Pre HAX`, na.rm = TRUE) -
              2 * cor(`HAI-L_shai_POST`, `HAI-L_shai_Pre HAX`, use = "pairwis
e")),
            mndiff = mean(HAI_DIFF, na.rm = TRUE),
            sddiff = sd(HAI_DIFF, na.rm = TRUE))
```

```
res1 <- bytrt %>% gather("type", "mean", mnpost, mndiff) %>% gather("type2",
"sd", sdpost, sddiff) %>%
  mutate(typefin = substr(type, 3, 6), typefin2 = substr(type2, 3, 6)) %>%
  filter(typefin == typefin2) %>% select(TX_group =`TX Group`, mean = mean, s
d = sd, type = typefin) %>%
  mutate(lower = mean - 1.96 * sd / sqrt(nrow(has)), upper = mean + 1.96 * sd
/ sqrt(nrow(has)))

res2 <- hasslope %>% group_by(TX_group) %>%
  summarize(mean = mean(haislope, na.rm = TRUE),
            sd = sd(haislope, na.rm = TRUE),
            type = "slope", lower = mean - 1.96 * sd / sqrt(nrow(has)),
            upper = mean + 1.96 * sd / sqrt(nrow(has)))


bet <- fixef(fitlme)
cov <- vcov(fitlme)

res3 <- data.frame(TX_group = c("BSM", "KBT"), mean = c(((c(0, 1, 0, 1) %*% b
et) * 12)[1, 1],
  ((c(0, 1, 0, 0) %*% bet) * 12)[1, 1]),
  sd = c(12 * sqrt(c(0, 1, 0, 1) %*% cov %*% c(0, 1, 0, 1))[1, 1],
         12 * sqrt(c(0, 0, 0, 1) %*% cov %*% c(0, 0, 0, 1))[1, 1]),
  type = c("mixef", "mixef"))

res3 <- res3 %>% mutate(lower = mean - 1.96 * sd, upper = mean + 1.96 * sd)

allres <- bind_rows(res1, res2, res3)

desc <- c(post = "Only final visit",
          diff = "Final - Baseline Difference", slope = "Derived Slopes",
          mixef = "Mixed Effects Model")

ggplot(allres, aes(x = type, y = mean, ymin = lower, ymax = upper, color = TX
_group)) +
  geom_pointrange(position = position_dodge(width = .5)) +
  xlab("") + scale_x_discrete(limits = c("post", "diff", "slope", "mixef"),
                              labels = desc[c("post", "diff", "slope", "mixef
")]) +
  ylab("Average Change in HAI score over 12 weeks") +
  ggtitle("Results of different analysis methods",
          subtitle = "Point estimates and 95% confidence intervals") +
  coord_flip()
load("example-plot.RData")
pex + ggtitle("Example of simulated data")
haslong %>% group_by(visit) %>% summarize(propmiss = mean(is.na(hai_score)))
%>%
  ggplot(aes(x = visit, y = propmiss)) + geom_bar(stat = "identity") +
  scale_y_continuous("Proportion missing observations", limits = c(0, .5))
load("sim-results-onesided-2016-08-30.RData")


res.power$d <- res.power$gamma * 12 / 7.5
labs <- res.power %>% group_by(gsize, d) %>% summarize(labyy = max(power), la
bxx = 0)
res.power$single <- sapply(res.power$propmiss, length) == 1

ggplot(subset(res.power, single), aes(x = unlist(propmiss), group = factor(gs
ize), y = power)) +
  geom_hline(data = subset(res.power, !single), aes(yintercept = power, linet
ype = "Time-Varying missingness")) +
  geom_line() + geom_point()  +
  geom_label_repel(data = labs, aes(x = labxx, y = labyy, label = paste("N ="
, gsize), group = NULL),
                   size = 3, nudge_x = -.1) +
  labs(title = "Power to rule out effect of 0.3 d in a noninferiority setting
",
          subtitle = "of the tx by time interaction in a lme model. True effe
ct size is d.") +
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, by = .1)) +
  scale_x_continuous("Proportion missing observations", limits = c(-0.05, .28
),
                     breaks = seq(0, 0.25, by = 0.05)) +
  scale_linetype_manual(values = c(2), guide = guide_legend(title = NULL)) +
  theme(legend.position = "bottom") + facet_wrap(~ d, labeller = "label_both"
)

load("sim-results-final-2016-08-27.RData")


res.power$d <- res.power$gamma * 12 / 7.5
labs <- res.power %>% group_by(gsize, d) %>% summarize(labyy = max(power), la
bxx = 0)
res.power$single <- sapply(res.power$propmiss, length) == 1
```

```
ggplot(subset(res.power, single), aes(x = unlist(propmiss), group = factor(gs
ize), y = power)) +
  geom_hline(data = subset(res.power, !single), aes(yintercept = power, linet
ype = "Time-Varying missingness")) +
  geom_line() + geom_point()  +
  geom_label_repel(data = labs, aes(x = labxx, y = labyy, label = paste("N ="
, gsize), group = NULL),
                   size = 3, nudge_x = -.1) +
  labs(title = "Power to rule out effect of 0.3 d in a noninferiority setting
",
          subtitle = "with two sided 95% CI") +
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, by = .1)) +
  scale_x_continuous("Proportion missing observations", limits = c(-0.05, .28
),
                     breaks = seq(0, 0.25, by = 0.05)) +
  scale_linetype_manual(values = c(2), guide = guide_legend(title = NULL)) +
  theme(legend.position = "bottom") + facet_wrap(~ d, labeller = "label_both"
)
```

Simulation Study

```
##

library(dplyr)
library(ggplot2)
library(ggthemes)
library(ggrepel)
library(lme4)
library(purrr)



gen_person <- function(neach = 13, alp = 34, gam0 = 0, bet = -1.0, gamma = 0.
0) {

  ttt <- seq(1, neach)
  Z <- rbinom(1, 1, .5)
  aaa <- rnorm(1, mean = 0, sd = 6)
  bbb <- rnorm(1, sd = .5)

  Y <- alp + aaa + gam0 * Z +
    (bet + bbb) * ttt +
    (gamma) * Z * ttt + rnorm(neach, sd = 3)

  data.frame(time = ttt, Z = Z, Y = Y)
}

gen_group <- function(gsize = 100, neach = 13,
                      alp = 33, gam0 = 0, bet = -.75, gamma = 0.0,
                      propmiss = 0.05) {   ## propmiss

  dat0 <- data.frame(pid = 1:gsize) %>% group_by(pid) %>%
    do(gen_person(neach, alp, gam0, bet, gamma))
  dat0 %>% group_by(pid) %>% mutate(miss = rbinom(neach, 1, propmiss),
                                    Yobs = ifelse(miss == 1, NA, Y))

}

panalyze <- function(dat0, marg = 2.25 / 12, outcome = "Yobs") {

  f1 <- " ~ time * Z + (1 + time | pid)"

  fit <- lmer(as.formula(paste(outcome, f1)), data = dat0, REML = FALSE,
              start = c(6, .5, 3), control = lmerControl(calc.derivs = FALSE)
)
  ci <- confint(fit, parm = 8, method = "Wald", level = 0.90)
  marg > ci[1, 2]

}


simulate <- function(B = 1000, param) {

  p0 <- sapply(param, unlist)
  replicate(B, {
    dat0 <- do.call(gen_group, as.list(p0))
    panalyze(dat0)
  }) %>% mean

}

set.seed(410)
dat0 <- gen_group(gam0 = 0)
system.time(panalyze(dat0))
```

```r
pex <- ggplot(dat0, aes(x = time, y = Yobs, color = factor(Z), group = pid))
+ geom_line() + geom_point() +
  stat_smooth(method = 'lm', se = FALSE, aes(group = factor(Z)), color = 'bla
ck') + ylim(c(0, 60))

#save(pex, file = "example-plot.RData")

params <- cross_d(list(propmiss = c(0, 0.05, 0.1, 0.15, 0.2, 0.25),
                       gsize = c(100, 150, 200, 250),
                       gamma = c(0, 0.05, 0.1, 0.15)
                       ))

params$propmiss <- as.list(params$propmiss)

params.timevary <- data_frame(propmiss = lapply(1:16, function(i) {
  c(0, 0.01, 0.05, 0.07, 0.1, 0.12, 0.17,
    0.21, 0.21, 0.26, 0.27, 0.28, 0.04)}),
  gsize = sort(rep(c(100, 150, 200, 250), 4)),
  gamma = rep(c(0, 0.05, 0.1, 0.15), 4))

params.all <- bind_rows(params, params.timevary)

res.power <- params.all %>% by_row(simulate, B = 5000, .to = "power", .collat
e = "rows")

save(res.power, file = paste0("sim-results-onesided-", Sys.Date(), ".RData"))
```

Reproducibility Note

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## locale:
## [1] LC_COLLATE=Swedish_Sweden.1252  LC_CTYPE=Swedish_Sweden.1252
## [3] LC_MONETARY=Swedish_Sweden.1252 LC_NUMERIC=C
## [5] LC_TIME=Swedish_Sweden.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] extrafont_0.17     ggrepel_0.5        ggthemes_3.2.0
##  [4] ggplot2_2.1.0.9000 purrr_0.2.2        dplyr_0.5.0
##  [7] lme4_1.1-12        Matrix_1.2-6       stringr_1.0.0
## [10] tidyr_0.6.0        rio_0.4.12         knitr_1.14
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.6      plyr_1.8.4       cellranger_1.1.0 formatR_1.4
##  [5] nloptr_1.0.4     tools_3.3.1      digest_0.6.10    gtable_0.2.0
##  [9] jsonlite_1.0     evaluate_0.9     tibble_1.1       nlme_3.1-128
## [13] lattice_0.20-33  openxlsx_3.0.0   csvy_0.1.3       DBI_0.4-1
## [17] curl_1.1         yaml_2.1.13      haven_0.2.1      Rttf2pt1_1.3.4
## [21] xml2_1.0.0       readODS_1.6.2    triebeard_0.3.0  grid_3.3.1
## [25] data.table_1.9.6 R6_2.1.2         readxl_0.1.1     foreign_0.8-66
## [29] rmarkdown_1.0    minqa_1.2.4      extrafontdb_1.0  readr_1.0.0
## [33] magrittr_1.5     scales_0.4.0     urltools_1.5.0   htmltools_0.3.5
## [37] splines_3.3.1    MASS_7.3-45      assertthat_0.1   colorspace_1.2-6
## [41] labeling_0.3     stringi_1.1.1    lazyeval_0.2.0   munsell_0.4.3
## [45] chron_2.3-47
```