proj5

Jiayi Shi

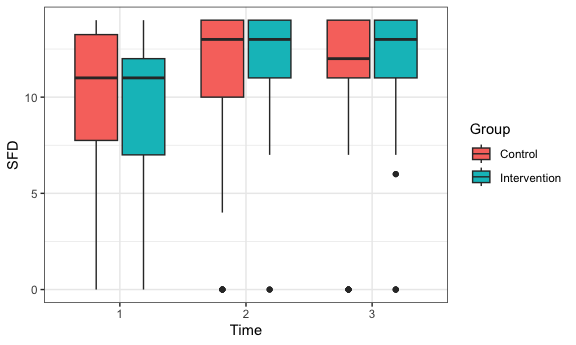
2025-04-18

# Load data & EDA

data <- read\_excel("data.xlsx") %>%   
 janitor::clean\_names() %>%   
 mutate(group = factor(group, levels = c(0,1), labels = c("Control", "Intervention")), time = factor(time), school = factor(school))  
  
table1::table1(~time+school+sfd|group, data)

|  | Control (N=147) | Intervention (N=177) | Overall (N=324) |
| --- | --- | --- | --- |
| **time** |  |  |  |
| 1 | 49 (33.3%) | 59 (33.3%) | 108 (33.3%) |
| 2 | 49 (33.3%) | 59 (33.3%) | 108 (33.3%) |
| 3 | 49 (33.3%) | 59 (33.3%) | 108 (33.3%) |
| **school** |  |  |  |
| 1 | 0 (0%) | 84 (47.5%) | 84 (25.9%) |
| 2 | 63 (42.9%) | 0 (0%) | 63 (19.4%) |
| 3 | 0 (0%) | 93 (52.5%) | 93 (28.7%) |
| 4 | 84 (57.1%) | 0 (0%) | 84 (25.9%) |
| **sfd** |  |  |  |
| Mean (SD) | 10.5 (4.57) | 10.8 (4.13) | 10.6 (4.33) |
| Median [Min, Max] | 12.0 [0, 14.0] | 12.0 [0, 14.0] | 12.0 [0, 14.0] |
| Missing | 10 (6.8%) | 18 (10.2%) | 28 (8.6%) |

ggplot(data, aes(x = time, y = sfd, fill = group)) +  
 geom\_boxplot() +  
 labs(x = "Time", y = "SFD") +  
 scale\_fill\_discrete(name = "Group", labels = c("Control", "Intervention"))+  
 theme\_bw()



#summarized\_data <- aggregate(sfd ~ time + group, data, mean)  
  
  
miss = data %>% filter(is.na(sfd))  
library(naniar)  
mcar\_test(data) #MAR/MNAR

## # A tibble: 1 × 4  
## statistic df p.value missing.patterns  
## <dbl> <dbl> <dbl> <int>  
## 1 9.96 4 0.0411 2

# change sfd  
df <- data %>%  
 pivot\_wider(names\_from = time, values\_from = sfd, names\_prefix = "sfd\_") %>%  
 mutate(  
 sfd\_6mo = sfd\_2 - sfd\_1,  
 sfd\_12mo = sfd\_3 - sfd\_1  
 ) %>%   
 select(id, group, school, sfd\_6mo, sfd\_12mo)   
  
d <- df %>%   
 pivot\_longer(c(sfd\_6mo, sfd\_12mo), names\_to = "time", values\_to = "change\_sfd", names\_prefix = "sfd\_") %>%   
 mutate(time = factor(time, levels = c("6mo", "12mo"), labels = c("6-month", "12-month")))  
  
  
# A model of nested random effects (block within site) could be 1|site/block if block labels are reused across multiple sites, or (1|site)+ (1|block) if the nesting structure is explicit in the data and each level of block only occurs within one site. A model of crossed random effects (block and year) would be (1|block)+(1|year).  
  
#lmm = glmmTMB(change\_sfd~time\*group+(1|school/id), df)  
#summary(lmm)  
#anova(lmm)  
#table(df$change\_sfd)  
#var(df$change\_sfd,na.rm=T)/mean(df$change\_sfd, na.rm=T)  
  
#ggplot(df, aes(x = change\_sfd)) +  
# geom\_histogram(binwidth = 1, fill = "skyblue", color = "black") +  
# labs(x = "Change in SFD", y = "Frequency") +  
# scale\_x\_continuous(breaks = -14:14) +  
# theme\_minimal()  
  
shapiro.test(df$sfd\_6mo) #reject normality

##   
## Shapiro-Wilk normality test  
##   
## data: df$sfd\_6mo  
## W = 0.91918, p-value = 1.533e-05

lmm0 = glmmTMB(sfd~time\*group+(1|school/id), data=data)  
summary(lmm0)

## Family: gaussian ( identity )  
## Formula: sfd ~ time \* group + (1 | school/id)  
## Data: data  
##   
## AIC BIC logLik deviance df.resid   
## 1689.6 1722.8 -835.8 1671.6 287   
##   
## Random effects:  
##   
## Conditional model:  
## Groups Name Variance Std.Dev.   
## id:school (Intercept) 4.184e+00 2.0453884  
## school (Intercept) 1.950e-08 0.0001396  
## Residual 1.327e+01 3.6423153  
## Number of obs: 296, groups: id:school, 105; school, 4  
##   
## Dispersion estimate for gaussian family (sigma^2): 13.3   
##   
## Conditional model:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 9.6042 0.6029 15.929 <2e-16 \*\*\*  
## time2 1.3920 0.7533 1.848 0.0646 .   
## time3 1.3052 0.7690 1.697 0.0897 .   
## groupIntervention -0.6568 0.8183 -0.803 0.4222   
## time2:groupIntervention 1.6313 1.0298 1.584 0.1132   
## time3:groupIntervention 1.4047 1.0472 1.341 0.1798   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lmm1= glmmTMB(sfd~time+(1|school/id), data)  
summary(lmm1)

## Family: gaussian ( identity )  
## Formula: sfd ~ time + (1 | school/id)  
## Data: data  
##   
## AIC BIC logLik deviance df.resid   
## 1686.8 1709.0 -837.4 1674.8 290   
##   
## Random effects:  
##   
## Conditional model:  
## Groups Name Variance Std.Dev.   
## id:school (Intercept) 4.128e+00 2.0316561  
## school (Intercept) 1.626e-08 0.0001275  
## Residual 1.348e+01 3.6709347  
## Number of obs: 296, groups: id:school, 105; school, 4  
##   
## Dispersion estimate for gaussian family (sigma^2): 13.5   
##   
## Conditional model:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 9.2476 0.4095 22.585 < 2e-16 \*\*\*  
## time2 2.2632 0.5176 4.373 1.23e-05 \*\*\*  
## time3 2.0614 0.5260 3.919 8.89e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(lmm0,lmm1)

## Data: data  
## Models:  
## lmm1: sfd ~ time + (1 | school/id), zi=~0, disp=~1  
## lmm0: sfd ~ time \* group + (1 | school/id), zi=~0, disp=~1  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)  
## lmm1 6 1686.8 1709.0 -837.41 1674.8   
## lmm0 9 1689.6 1722.8 -835.80 1671.6 3.2196 3 0.359

dat <- data %>% drop\_na() %>%   
 mutate(sd = 14-sfd)

# count model comparison

## cross-sectional

dd <- dat %>%  
 group\_by(id) %>%  
 mutate(base = first(sfd[time == 1])) %>%  
 ungroup()  
  
ddd6 <- dd %>% filter(time == 2)  
  
## Negative‑binomial (NB)  
m\_nb <- glmmTMB(  
 sfd ~ group+base,  
 family = nbinom2, data = ddd6)  
  
## Beta‑binomial (BB)  
m\_bb <- glmmTMB(  
 cbind(sfd, 14-sfd) ~ group+base,  
 family = betabinomial, data = ddd6)  
  
## Zero‑inflated NB  
m\_zinb <- glmmTMB(  
 sfd ~ group+base,  
 ziformula = ~., family = nbinom2, ddd6)  
  
## Zero‑inflated Beta‑binomial  
m\_zibb <- glmmTMB(  
 cbind(sfd, 14-sfd) ~ group+base,  
 ziformula = ~1, family = betabinomial, data = ddd6)  
  
AIC(m\_nb, m\_bb, m\_zinb, m\_zibb) %>%   
 as\_tibble(rownames = "model") %>%   
 arrange(AIC)

## # A tibble: 4 × 3  
## model df AIC  
## <chr> <dbl> <dbl>  
## 1 m\_zibb 5 365.  
## 2 m\_bb 4 385.  
## 3 m\_nb 4 602.  
## 4 m\_zinb 7 NA

summary(m\_bb)

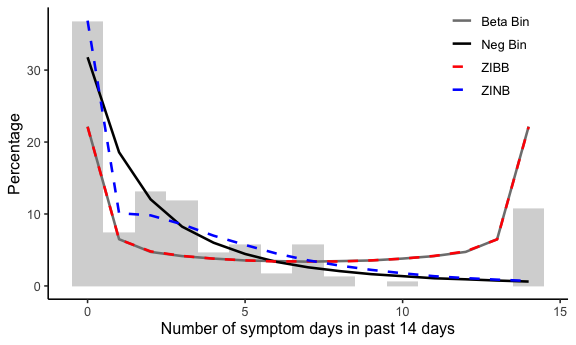
## Family: betabinomial ( logit )  
## Formula: cbind(sfd, 14 - sfd) ~ group + base  
## Data: ddd6  
##   
## AIC BIC logLik deviance df.resid   
## 385.0 395.4 -188.5 377.0 94   
##   
##   
## Dispersion parameter for betabinomial family (): 1.01   
##   
## Conditional model:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.79733 0.39420 2.023 0.0431 \*  
## groupIntervention 0.19768 0.32488 0.609 0.5429   
## base 0.04943 0.03314 1.492 0.1358   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## longitudinal

## Negative‑binomial (NB)  
m\_nb <- glmmTMB(  
 sd ~ time\*group + (1|school/id),  
 family = nbinom2, data = dat)  
  
## Beta‑binomial (BB)  
m\_bb <- glmmTMB(  
 cbind(sd, 14-sd) ~ time\*group + (1|school/id),  
 family = betabinomial, data = dat)  
  
## Zero‑inflated NB  
m\_zinb <- glmmTMB(  
 sd ~ time\*group + (1|school/id),  
 ziformula = ~., family = nbinom2, dat)  
  
## Zero‑inflated Beta‑binomial  
m\_zibb <- glmmTMB(  
 cbind(sd, 14-sd) ~ time\*group + (1|school/id),  
 ziformula = ~1, family = betabinomial, data = dat)  
  
AIC(m\_nb, m\_bb, m\_zinb, m\_zibb) %>%   
 as\_tibble(rownames = "model") %>%   
 arrange(AIC)

## # A tibble: 4 × 3  
## model df AIC  
## <chr> <dbl> <dbl>  
## 1 m\_bb 9 1280.  
## 2 m\_zibb 10 1282.  
## 3 m\_zinb 17 1338.  
## 4 m\_nb 9 1356.

clr <- c("Neg Bin" = "black",  
 "Beta Bin" = "grey50",  
 "ZINB" = "blue",  
 "ZIBB"= "red")  
  
lt <- c("Neg Bin" = "solid",  
 "Beta Bin" = "solid",  
 "ZINB" = "dashed",  
 "ZIBB" = "dashed")  
  
  
B <- 1000   
  
avg\_sim\_pct <- function(fit, label, B = 1000) {  
 sims\_mat <- simulate(fit, nsim = B)   
 sims\_vec <- as.vector(as.matrix(sims\_mat))   
  
 tibble(sd = sims\_vec) %>%  
 filter(sd <= 14, sd >= 0) %>%   
 count(sd) %>%   
 complete(sd = 0:14, fill = list(n = 0)) %>%  
 mutate(  
 pct = 100 \* n / length(sims\_vec),   
 model = label  
 )  
}  
  
  
sim\_all <- bind\_rows(  
 avg\_sim\_pct(m\_nb, "Neg Bin", B),  
 avg\_sim\_pct(m\_bb, "Beta Bin", B),  
 avg\_sim\_pct(m\_zinb, "ZINB", B),  
 avg\_sim\_pct(m\_zibb, "ZIBB", B)  
)  
  
ggplot() +  
 geom\_col(data = dat %>% count(sd),  
 aes(x = sd, y = 100 \* n / sum(n)), fill = "grey80", colour = NA, width = 1, alpha = .8) +  
  
 geom\_line(data = sim\_all,  
 aes(x = sd, y = pct, colour = model, linetype = model, group = model),  
 linewidth = .9, direction = "mid") +  
  
 scale\_colour\_manual(name = NULL, values = clr) +  
 scale\_linetype\_manual(name = NULL, values = lt) +  
  
 labs(x = "Number of symptom days in past 14 days", y = "Percentage") +  
 theme\_classic(base\_size = 12) +  
 theme(legend.position = c(.85, .85))



summary(m\_bb)

## Family: betabinomial ( logit )  
## Formula: cbind(sd, 14 - sd) ~ time \* group + (1 | school/id)  
## Data: dat  
##   
## AIC BIC logLik deviance df.resid   
## 1280.0 1313.2 -631.0 1262.0 287   
##   
## Random effects:  
##   
## Conditional model:  
## Groups Name Variance Std.Dev.   
## id:school (Intercept) 8.855e-01 9.410e-01  
## school (Intercept) 2.623e-09 5.122e-05  
## Number of obs: 296, groups: id:school, 105; school, 4  
##   
## Dispersion parameter for betabinomial family (): 1.44   
##   
## Conditional model:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.6950 0.2549 -2.727 0.0064 \*\*  
## time2 -0.7961 0.3224 -2.469 0.0135 \*   
## time3 -0.6171 0.3223 -1.915 0.0555 .   
## groupIntervention 0.2886 0.3389 0.852 0.3944   
## time2:groupIntervention -0.5142 0.4370 -1.177 0.2393   
## time3:groupIntervention -0.4972 0.4346 -1.144 0.2526   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

library(emmeans)  
emmeans(m\_bb, pairwise ~ group | time) #estimate

## $emmeans  
## time = 1:  
## group emmean SE df asymp.LCL asymp.UCL  
## Control -0.695 0.255 Inf -1.195 -0.1955  
## Intervention -0.406 0.228 Inf -0.854 0.0407  
##   
## time = 2:  
## group emmean SE df asymp.LCL asymp.UCL  
## Control -1.491 0.285 Inf -2.049 -0.9331  
## Intervention -1.717 0.272 Inf -2.249 -1.1845  
##   
## time = 3:  
## group emmean SE df asymp.LCL asymp.UCL  
## Control -1.312 0.284 Inf -1.868 -0.7563  
## Intervention -1.521 0.264 Inf -2.039 -1.0025  
##   
## Results are given on the logit (not the response) scale.   
## Confidence level used: 0.95   
##   
## $contrasts  
## time = 1:  
## contrast estimate SE df z.ratio p.value  
## Control - Intervention -0.289 0.339 Inf -0.852 0.3944  
##   
## time = 2:  
## contrast estimate SE df z.ratio p.value  
## Control - Intervention 0.226 0.380 Inf 0.594 0.5523  
##   
## time = 3:  
## contrast estimate SE df z.ratio p.value  
## Control - Intervention 0.209 0.377 Inf 0.554 0.5799  
##   
## Results are given on the log odds ratio (not the response) scale.

emmeans(m\_bb, pairwise ~ group | time, type = "response") #odds ratio

## $emmeans  
## time = 1:  
## group prob SE df asymp.LCL asymp.UCL  
## Control 0.333 0.0566 Inf 0.2324 0.451  
## Intervention 0.400 0.0547 Inf 0.2987 0.510  
##   
## time = 2:  
## group prob SE df asymp.LCL asymp.UCL  
## Control 0.184 0.0427 Inf 0.1141 0.282  
## Intervention 0.152 0.0351 Inf 0.0954 0.234  
##   
## time = 3:  
## group prob SE df asymp.LCL asymp.UCL  
## Control 0.212 0.0474 Inf 0.1338 0.319  
## Intervention 0.179 0.0389 Inf 0.1152 0.268  
##   
## Confidence level used: 0.95   
## Intervals are back-transformed from the logit scale   
##   
## $contrasts  
## time = 1:  
## contrast odds.ratio SE df null z.ratio p.value  
## Control / Intervention 0.749 0.254 Inf 1 -0.852 0.3944  
##   
## time = 2:  
## contrast odds.ratio SE df null z.ratio p.value  
## Control / Intervention 1.253 0.476 Inf 1 0.594 0.5523  
##   
## time = 3:  
## contrast odds.ratio SE df null z.ratio p.value  
## Control / Intervention 1.232 0.464 Inf 1 0.554 0.5799  
##   
## Tests are performed on the log odds ratio scale

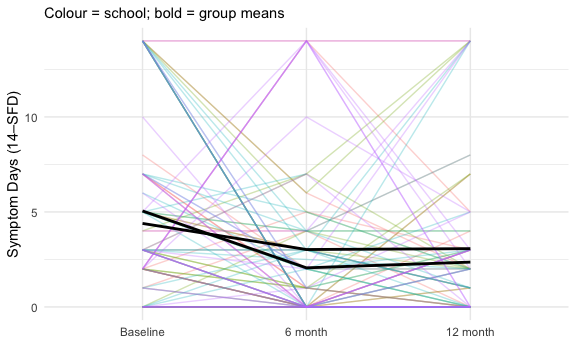
vc\_list <- VarCorr(m\_bb)$cond  
  
var\_components <- sapply(vc\_list, function(mat)  
 (attr(mat, "stddev"))^2)  
  
var\_school <- var\_components[2]  
var\_subject <- var\_components[1]  
  
  
# Approximate residual variance for beta-binomial (as binomial variance × φ)  
# For binomial: Var(Y) = n \* p \* (1 - p)  
# For beta-binomial: Var(Y) = n \* p \* (1 - p) \* (1 + (n - 1) \* φ)  
# => overdispersion increases residual variance  
  
# We approximate residual variance on the latent scale as:  
# (following glmm practice for latent variance in logit models)  
var\_resid <- (pi^2) / 3 # Latent scale approximation for logit link  
  
# Total variance = sum of random effects + residual  
var\_total <- var\_school + var\_subject + var\_resid  
  
# ICC for school level (proportion of total variance from school)  
icc\_school <- var\_school / var\_total  
  
# ICC for subject level (repeatability)  
icc\_subject <- var\_subject / var\_total  
  
# Print  
cat("ICC\_school: ", round(icc\_school, 4), "\n")

## ICC\_school: 0

cat("ICC\_subject: ", round(icc\_subject, 4), "\n")

## ICC\_subject: 0.2121

ggplot(dat, aes(x = time, y = sd, group = id, colour = school)) +  
 geom\_line(alpha = 0.3) +  
 stat\_summary(aes(group = group), fun = mean,  
 geom = "line", linewidth = 1, colour = "black") +  
 scale\_x\_discrete(labels = c("Baseline", "6 month", "12 month")) +  
 labs(x = NULL, y = "Symptom Days (14–SFD)",  
 subtitle = "Colour = school; bold = group means") +  
 theme\_minimal(base\_size = 11) +  
 theme(legend.position = "none")



###############################################################################  
## 3. Sample‑size for 30‑school cluster‑RCT ##  
###############################################################################  
## Target: detect ≥ 1/3 SD improvement at ANY of the 4 follow‑ups with 80% power  
alpha\_one\_sided <- 0.0125 # Bonferroni for 4 comparisons (0.05/4)  
z\_alpha <- qnorm(1 - alpha\_one\_sided/2)  
z\_beta <- qnorm(0.80)  
delta <- 1/3  
## 3‑b Independent‑samples size (no clustering, no repeated measures)  
n <- 2 \* (z\_alpha + z\_beta)^2 / delta^2   
c <- 30 # number of clusters  
m\_r <- 4 # number of repeated measures  
# Design effect:  
DE <- function(m) 1 + (m - 1) \* icc\_school + (1 + (m\_r - 1) \* icc\_subject) / m\_r  
m\_seq <- 1:100  
m\_star <- m\_seq[ which( c \* m\_seq >= DE(m\_seq) \* n )[1] ]  
  
cat("=> Need", m\_star, "children per school (total 30 x",  
 m\_star, "=", 30\*m\_star, ")\n")

## => Need 10 children per school (total 30 x 10 = 300 )