P9185 Statistical Practices and Research for Interdisciplinary Sciences (SPRIS)

Project IV Report

Preliminary results from asthma-pass: a school-wide intervention for asthmatic children

Jungang Zou, jz3183

Department of Biostatistics,

Mailman School of Public Health, Columbia University

April 26, 2023

Contents

1 Introduction

Asthma is a greater burden for low-income African-American and Hispanic children in inner-city areas like the Bronx. Physical activity (PA) is essential for asthma management in children. Urban minority children with asthma, however, face multiple barriers to PA, ranging from personal to community levels. Investigators propose an intervention program in collaboration with primary care physicians, community health workers, and school personnel to promote the availability of guideline-based preventive and rescue medication, provide education to children and caregivers, and encourage physical activity. They conducted a pilot cluster-randomized controlled trial of Asthma-PASS in four Bronx elementary schools with 108 asthmatic children. The four elementary schools were randomly assigned to either the Asthma-PASS intervention group (2 schools) or the AM comparison group (2 schools), in which participants followed the standard routine provided by each school (i.e., the standard of care). The participating children were followed up at 6 and 12 months after baseline.

This report aimed to do a longitudinal analysis to determine if children in schools that received the Asthma-PASS intervention demonstrated greater improvement in the number of symptom-free days (SFD) at the 6- or 12-month follow-up than those in the AM comparison group.

2 Exploratory Data Analysis

The dataset analyzed in this report was longitudinal data. The outcome stands for the symptom-free days in the past two weeks prior to each follow-

up visit. Out of 108, 59 individuals from school 1 (N=28) and school 3 (N=31) were assigned to the intervention group, while 49 individuals from school 2 (N=21) and school 4 (N=28) were assigned to the control group. The summary statistics can be found in Figure 1. We can also find there is missingness in the data and there are 3 individuals who are missing in all three measurements. Formatted as a wide format, we ran Little's MCAR test [1] and got p-value = 0.89, so we conclude the missingness is MCAR. From the summary statistics and Spaghetti plot in Figure 2, we can find there are differences in SFD between the baseline and other follow-up periods for all schools. Further statistical analysis should be done to determine if the intervention has an improvement in SFD.

3 Methods

3.1 Notation

Consider a dataset consisting of n individuals, each with 3 repeated measures denoted by $SFD_i = (SFD_{i1}, \ldots, SFD_{i3})$. Additionally, at each time period j, each individual has p covariates represented by the vector $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ij3})^\top$. We denote $SFD = (SFD_1, \ldots, SFD_n)$, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and are interested in modeling the conditional distribution $f(SFD|\mathbf{X}, \theta)$ and with some unknown parameters θ and using statistical methods.

To make the problem more clear, we assume individuals from different schools can have different effects on the outcome. We use $School_i, i = 1, 2, 3, 4$ to denote the "school effects".

3.2 Generalized Linear Mixed Effects Model

The generalized linear mixed-effects model (GLMM) is a widely used regression model for non-normal outcome variables in the longitudinal study. By combining a linear mixed effects model and a generalized linear model, GLMMs can model the effects of both fixed and random predictors on the response variable, while also accounting for the correlation among observations within groups. In this study, since demographic information is not useful, we use the random-intercept model to adjust for individual effects and the "school effects". The GLMM model for *SFD* is specified as follows:

$$\begin{aligned} g(SFD_{ij}) &= \mathbf{X}_{ij}\beta + \alpha_i + School_i \\ School_i &\in \{School_1, School_2, School_3, School_4\} \\ \alpha_i &\sim i.i.d\ N(0, \sigma^2) \\ School_1, School_2, School_3, School_4 &\sim i.i.d\ N(0, \sigma_s^2) \end{aligned} \tag{1}$$

where $g(.)$ is the link function for the count outcome, α_i is the random intercept for i-th individual, $School_i$ is the corresponding school effects for i-th individual.

3.3 Over-dispersion and Zero-inflated models

When modeling count data, investigators commonly use the Poisson or Binomial link function. However, if there is over-dispersion, indicated by the regression deviance divided by (n - p) ¿ 1, the Beta-Binomial link function can be used. In this study, if we ran GLMM with a Binomial link, we can get the test statistic as $5.357503 > 1$, resulting in the over-dispersion. Alter-

natively, when dealing with an excess of zeros in count data, Zero-inflated models can be used. In this data, there is an excess of zeros for $14 - SFD$, shown in Figure 3. For the data in this study, with an excess of zeros for $14 - SFD$, we choose to model it using the Zero-inflated Beta-Binomial (ZIBB [2]) link function for GLMM.

3.4 Power analysis

Power analysis in randomized controlled trials (RCTs) is a statistical method used to calculate the sample size required to detect a significant treatment effect with a certain degree of confidence. In this study, we conduct a simple power analysis, assuming we perform 4 two-sided t-tests with Bonferroni correction at 3-, 6-, 9-, and 12-month follow-up periods. Approximately, let the outcome SFD be continuous, and compare the mean SFDs in each test stratified by the trial period. By the power analysis, we first calculated individualized trial sample size per period: $n_I = 2 * (\frac{Z_{0.99375}+Z_{0.2}}{1/3})^2$. Then sample size per period per cluster would be $N = n_I * (1 + 3ICC_{student} * (1 + 29 * ICC_{school}))/30$, where $ICC_{student}$ and ICC_{school} are the intraclass correlation coefficient for students and schools respectively.

4 Results

The estimates for the GLMM model are in Figure 4. Compared with the baseline, there are differences for *SFD* at 6-month follow-up and 12-month follow-up. Holding other variables, 0.7668 increase in log odds for a 6-month follow-up and 0.6230 increase in log odds for a 12-month follow-up for the

same subject, compared with baseline measurement. However, there is no significant evidence to support there is a treatment difference between the treatment group and the control group. Also, the estimate for the zero-inflated intercept is significant ($p - value < 2e - 16$), indicating the correct specification for the zero-inflated model.

Calculating ICC for schools and students, we can get $ICC_{student} = 1$ and $ICC_{school} = 1.778e - 8$. It shows there is very little variability among different schools but large variability among different individuals. Using the formula of the sample size mentioned in 3.4, we know the sample size per period per school is 26.8.

5 Conclusion

This report focuses on the longitudinal analysis for *SFD*. By adjusting for the over-dispersion and zero-inflation problems, we choose the zero-inflated Beta-Binomial linear mixed effects model to test the association between *SFD* and covariates. The results show there *SFD* is increasing when the follow-up period increases but there is no significant difference between the treatment group and the control group. By calculating the power, we find the minimal sample size for an effective study is 26.8 for each time period for each school.

Appendices

The appendix includes all supplementary tables, formulas, and figures that are referred to in this report.

.1 ZIBB regression

The ZIBB regression in this study is as follows:

$$\begin{aligned} g(SFD_{ij}) = \beta_0 + \beta_1 * I(6 - monthfollow - up) + \\ \beta_2 * I(12 - monthfollow - up) + \beta_3 * I(Interventiongroup) + \\ \beta_4 * I(6 - monthfollow - up) * I(Interventiongroup) + \\ \beta_5 * I(12 - monthfollow - up) * I(Interventiongroup) \end{aligned} \tag{2}$$

.2 Figures

	school 1 (N=28)	school 2 (N=21)	school 3 (N=31)	school 4 (N=28)	Overall (N=108)
group					
Control group	0 (0%)	21 (100%)	0 (0%)	28 (100%)	49 (45.4%)
Intervention group	28 (100%)	0 (0%)	31 (100%)	0 (0%)	59 (54.6%)
sfd_baseline					
Mean (SD)	9.96 (4.34)	9.38 (5.07)	8.03 (5.02)	9.78 (4.37)	9.25 (4.70)
Median [Min, Max]	12.0 [0, 14.0]	11.0 [0, 14.0]	9.00 [0, 14.0]	11.0 [0, 14.0]	11.0 [0, 14.0]
Missing	1 (3.6%)	0 (0%)	1 (3.2%)	1 (3.6%)	3 (2.8%)
sfd_6_month_follow_up					
Mean (SD)	11.7 (3.97)	11.8 (2.53)	12.1 (1.92)	10.4 (5.57)	11.5 (3.86)
Median [Min, Max]	14.0 [0, 14.0]	13.0 [7.00, 14.0]	12.0 [7.00, 14.0]	14.0 [0, 14.0]	13.0 [0, 14.0]
Missing	3 (10.7%)	2 (9.5%)	4 (12.9%)	1 (3.6%)	10 (9.3%)
sfd_12_month_follow_up					
Mean (SD)	11.7 (3.44)	11.0 (4.32)	11.6 (3.84)	10.9 (4.63)	11.3 (4.01)
Median [Min, Max]	13.0 [0, 14.0]	12.0 [0, 14.0]	13.0 [0, 14.0]	12.0 [0, 14.0]	13.0 [0, 14.0]
Missing	5 (17.9%)	1 (4.8%)	4 (12.9%)	5 (17.9%)	15 (13.9%)

Figure 1: Summary statistics for each variable stratified by school

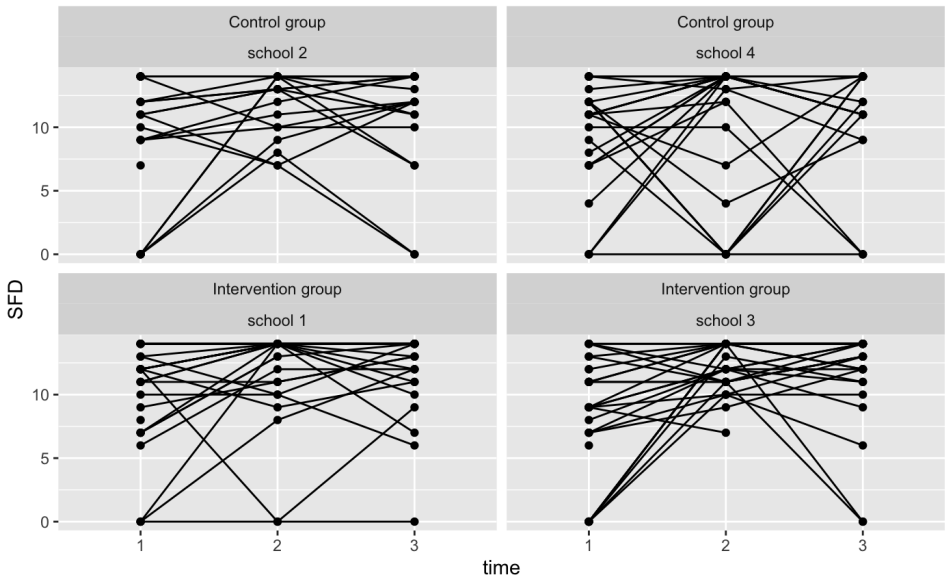

Figure 2: Spaghetti plot for SFD

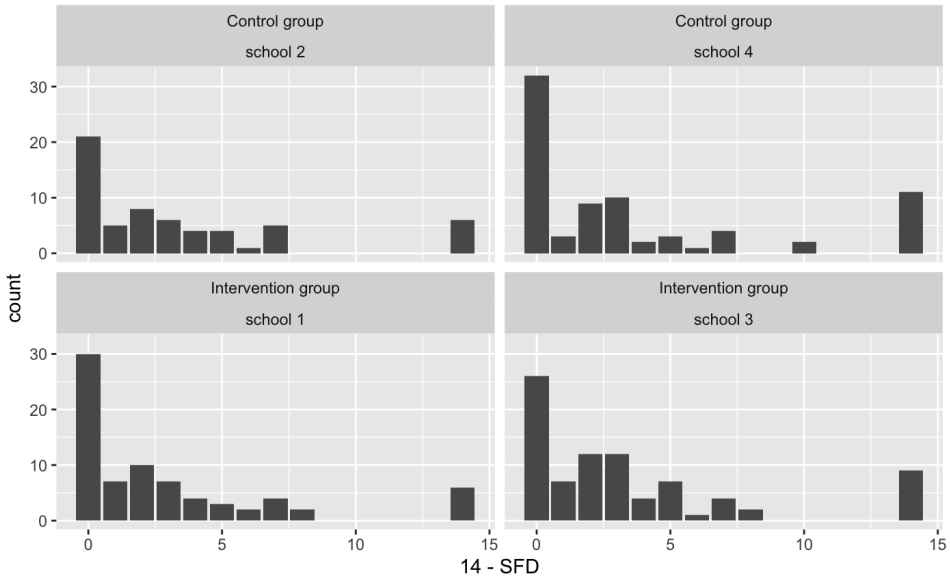

Figure 3: Zero-inflated issue

```
 Family: betabinomial  ( logit )
Formula:          cbind(SFD, 14 - SFD) ~ time + group + group * time + (1 | school/ID)
Zero inflation:                      ~1
Data: data

     AIC      BIC   logLik deviance df.resid
  1199.6   1236.5   -589.8   1179.6      292

Random effects:

Conditional model:
 Groups    Name        Variance  Std.Dev.
 ID:school (Intercept) 5.068e-01 7.119e-01
 school    (Intercept) 9.011e-09 9.493e-05
Number of obs: 296, groups:  ID:school, 105; school, 4

Dispersion parameter for betabinomial family (): 6.21

Conditional model:
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                                   1.4239     0.2039   6.984 2.88e-12 ***
time6-month follow up                         0.7668     0.2811   2.728  0.00638 **
time12-month follow up                        0.6230     0.2814   2.214  0.02684 *
groupIntervention group                      -0.1802     0.2721  -0.662  0.50790
time6-month follow up:groupIntervention group 0.1056     0.3742   0.282  0.77785
time12-month follow up:groupIntervention group 0.2233    0.3761   0.594  0.55273
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Zero-inflation model:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.1174     0.1885  -11.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: Estimates of the GLMM model

References

[1] Roderick JA Little. "A test of missing completely at random for multivariate data with missing values". In: *Journal of the American statistical Association* 83.404 (1988), pp. 1198–1202.

[2] Brandie Wagner, Paula Riggs, and Susan Mikulich-Gilbertson. "The importance of distribution-choice in modeling substance use data: a comparison of negative binomial, beta binomial, and zero-inflated distributions". In: *The American journal of drug and alcohol abuse* 41.6 (2015), pp. 489–497.