P9185 Statistical Practices and Research for Interdisciplinary Sciences (SPRIS)

Project V Report

Design and Analysis for Covid-19 Vaccine

Jungang Zou, jz3183

Department of Biostatistics,

Mailman School of Public Health, Columbia University

May 5, 2023

Contents

1 Introduction

The pandemic caused by the Covid-19 virus has resulted in significant human losses, as we are all aware. The disease has spread rapidly worldwide, leading to numerous deaths and hospitalizations, as well as economic and social disruptions. The impact of Covid-19 has been felt on a global scale and has highlighted the importance of public health preparedness. To prove the effectiveness of their developing vaccine, the pharmaceutical company plans to carry out a phase III trial that will be randomized, observer-blinded, and placebo-controlled, with 100 U.S. sites participating at a 1-to-1 ratio. The main goal is to evaluate the vaccine's efficacy in preventing the first symptomatic onset of Covid-19, 14 days after the second injection, in seronegative participants. The vaccine's efficacy is defined as (1 – hazard ratio) * 100.

This report aimed to verify the efficacy of Covid-19, focusing on the following aspects: a longitudinal study to investigate the difference in Serious Adverse Events (SAE) between the treatment group and the control group, and a survival analysis to explore the likelihood of acquiring Covid-19 within 12 months, as well as the median and mean duration until infection after the second vaccine dose.

2 Exploratory Data Analysis

In this study, we analyzed two separate datasets. The first one is a longitudinal dataset to analyze the SAEs. There are total 20625 individuals who received the vaccine and 20625 individuals received the placebo. Each individual has three follow-up measurements. The outcome SAE is a binary

variable that indicates the existence of SAEs during the specific follow-up period. SAE is a sparse variable, which means the SAEs are rare. Missingness also exists in SAE, where we assume the pattern is Miss at Random. The summary statistics of all variables stratified by follow-up periods are in Figure 1.

The second data is survival data, consisting of 2299 participants who received two shots of the vaccine and did not get infected by COVID before the recruitment. We created a new variable $time = LastFUTime * I(Censored) + InfectionTime * I(Infected)$. Stratified by the event data and censored data, the summary statistics can be found in Figure 2. To solve the left truncation caused by the recruitment process, we choose the $EnrollmentTime$ to account for the left truncation time.

3 Methods

3.1 Notation

Consider a dataset consisting of n individuals, each with 3 repeated measures denoted by $SAE_i = (SAE_{i1}, \ldots, SAE_{i3})$. 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 $SAE = (SAE_1, \ldots, SAE_n)$, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and are interested in modeling the conditional distribution $f(SAE|\mathbf{X}, \theta)$ and with some unknown parameters θ and using statistical methods. For the survival outcome, we use $\delta_i = 1$ to denote the event that happens for i-th individual and censored outcome if $\delta_i = 0$. The survival outcome for i-th individual is the pair $(time_i, \delta_i)$.

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. The GLMM model for SAE is specified as follows:

$$\begin{aligned} logit(SAE_{ij}) =& \beta_0 + \beta_1 * I(time = 2) + \beta_2 * I(time = 3) \\ &+ \beta_4 * I(time = 2)I(group = vaccine) \\ &+ \beta_5 * I(time = 3)I(group = vaccine) \\ &+ \beta_6 * I(sex = male) + \beta_7 age + \alpha_i + site_k \\ \alpha_i \sim& \ i.i.d \ N(0, \sigma^2) \\ site_k \sim& \ i.i.d \ N(0, \sigma_s^2) \end{aligned} \tag{1}$$

where $logit(.)$ is the logit link function for the binary outcome, α_i is the random intercept for i-th individual, $site_k$ is the corresponding site effects for i-th individual.

3.3 Mean and median survival time

The median survival time is the time at which 50% of the study population has experienced the event of interest (e.g., death, disease progression). The mean survival time is the average length of time until an event of interest (such as death or failure) occurs in a group of individuals. There are

some methods to estimate median survival time and mean survival time: parametric estimation: specifies a distribution for *time* then estimates the parameter using MLE; Non-parametric estimation is estimated from K-M curve. We compared the mean and median survival time under K-M curve, Exponential, Weibull (AFT), Gompertz, Gamma, Lognormal, Log-logistic distributions, as well as the predictive probability of acquiring Covid-19 within 12 months and corresponding 95% CI.

4 Results

The estimates for the GLMM model are in Figure 3. Compared with the placebo group, individuals in the vaccine group are more likely to have SAE ($p = 0.07$). However, if considering about time effect, there is no significant difference for vaccine group across time. The interaction terms between group and time are all insignificant.

The estimated mean and median survival time, as well as the predictive probability of acquiring Covid-19 within 12 months and corresponding 95% CI, are displayed in Figure 4. By comparing with the AIC values, we can find the model with Lognormal is the best model with the smallest AIC (AIC = 6673.26). The median and mean survival time from Lognormal are 588.01 and 628.67. The predictive probability of acquiring Covid-19 within 12 months is 0.1, with 95% CI [0.09, 0.11]. The survival curve of Lognormal is displayed in Figure 5.

5 Conclusion

Individuals in the vaccine group are more likely to have SAE (p=0.07). However, if considering about time effect, there is no significant difference for the vaccine group across time. Choosing the lognormal model to model survival function, we get the probability of contracting Covid-19 within 12 months is 0.1 (0.09, 0.11). The median survival time is 588.01, and 628.67 for the mean survival time.

Appendices

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

.1 Figures

	Time1 (N=41250)	Time2 (N=41250)	Time3 (N=41250)
GROUP			
Control	20625 (50.0%)	20625 (50.0%)	20625 (50.0%)
Vaccine	20625 (50.0%)	20625 (50.0%)	20625 (50.0%)
SEX			
Female	20503 (49.7%)	20503 (49.7%)	20503 (49.7%)
Male	20747 (50.3%)	20747 (50.3%)	20747 (50.3%)
AGE			
Mean (SD)	44.6 (11.0)	44.6 (11.0)	44.6 (11.0)
Median [Min, Max]	45.0 [16.0, 92.0]	45.0 [16.0, 92.0]	45.0 [16.0, 92.0]
SAE			
0	38526 (93.4%)	37154 (90.1%)	35597 (86.3%)
1	62 (0.2%)	58 (0.1%)	55 (0.1%)
Missing	2662 (6.5%)	4038 (9.8%)	5598 (13.6%)

Figure 1: Summary statistics for SAE data

	Censored (N=1866)	Infected (N=433)	Overall (N=2299)
EnrollmentTime			
Mean (SD)	137 (25.3)	146 (24.7)	139 (25.4)
Median [Min, Max]	133 [86.0, 205]	146 [95.0, 205]	137 [86.0, 205]
time			
Mean (SD)	412 (81.7)	356 (24.9)	402 (77.6)
Median [Min, Max]	383 [276, 570]	357 [289, 422]	371 [276, 570]

Figure 2: Summary statistics for survival data

```
Random effects:
 Groups  Name        Variance  Std.Dev.
 ID:SITE (Intercept) 9.459e-05 0.009726
 SITE    (Intercept) 1.088e-01 0.329883
Number of obs: 111452, groups:  ID:SITE, 41194; SITE, 100

Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.337776   0.381146 -19.252   <2e-16 ***
TIMETime2                0.061085   0.279250   0.219   0.8268
TIMETime3                0.192615   0.272151   0.708   0.4791
GROUPVaccine             0.467185   0.258171   1.810   0.0704 .
SEXMale                 -0.098638   0.151003  -0.653   0.5136
AGE                      0.014266   0.006789   2.101   0.0356 *
TIMETime2:GROUPVaccine -0.154932   0.368686  -0.420   0.6743
TIMETime3:GROUPVaccine -0.428425   0.373080  -1.148   0.2508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: Estimates of the GLMM model

model	AIC	median	mean	infection_p	infection_p_lower	infection_p_upper
Exponential	7138.91	967.50	1395.81	0.23	0.21	0.25
Weibull (AFT)	6800.72	589.31	586.07	0.10	0.09	0.11
Gompertz	6893.36	584.57	560.78	0.12	0.11	0.14
Gamma	6716.76	582.36	603.72	0.10	0.08	0.11
Lognormal	6673.26	588.01	628.67	0.10	0.09	0.11
Log-logistic	6757.03	589.22	640.23	0.10	0.09	0.12
KM	NA	NA	512.00	0.16	0.14	0.18

Figure 4: Estimated parameters for survival analysis

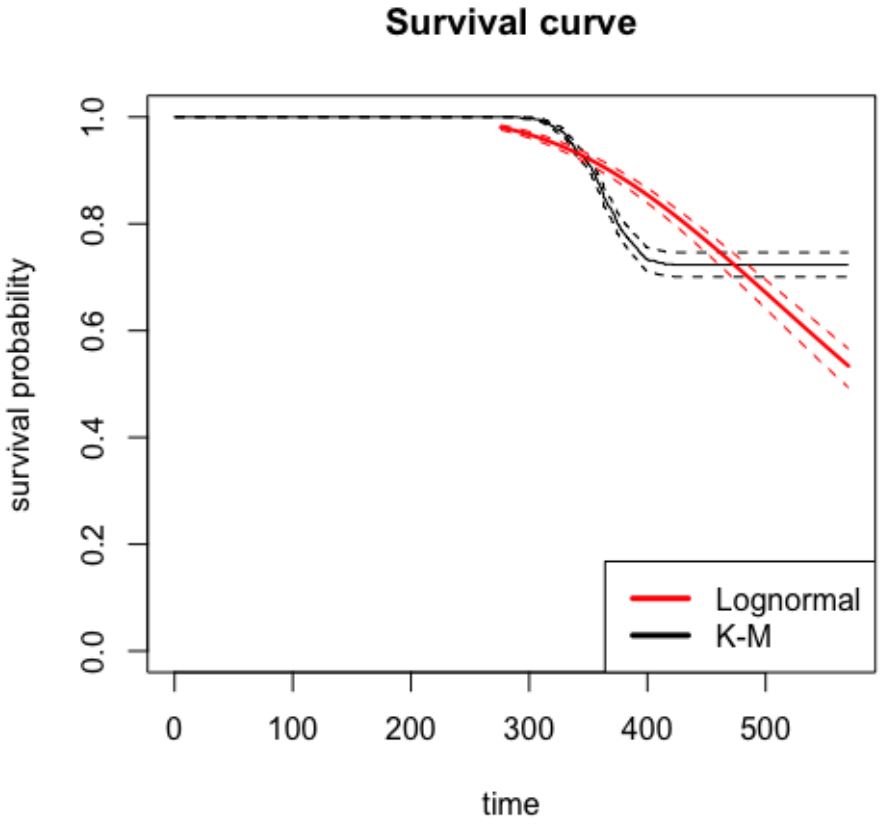

Figure 5: Survival curve for Lognormal distribution