Risk of Developing Respiratory Infection with Vitamin-A Deficiency

A study (ICHS) was conducted in West Java, Indonesia, to determine the effects of Vitamin A deficiency in preschool children. The goal of this problem is to analyze these data using marginal model (GEE) to address the interests of the investigators. The investigators were particularly interested in **whether children with Vitamin A deficiency were at increased risk of developing respiratory infection**, which is one of the leading causes of death in this part of the world. 250 children were recruited in the study, and their (baseline) age in years (bage), gender(gender: 0 = male, 1 = female), and whether they suffered vitamin A deficiency (vita: 0 = no, 1 = yes) was recorded at an initial clinic visit (time 0). Also recorded was the response, whether the child was suffering from a respiratory infection (infect: 0 = no, 1 = yes). The children were then re-examined at 3 month intervals for 15 months (at 3, 6, 9, 12, and 15 months) after the first visit, and the presence or absence of respiratory infection was recorded at each of these visits.

1 Exploratory Data Analysis

Covariates and Response

Firstly, we explore the data with respect to the primary scientific aim of the study.

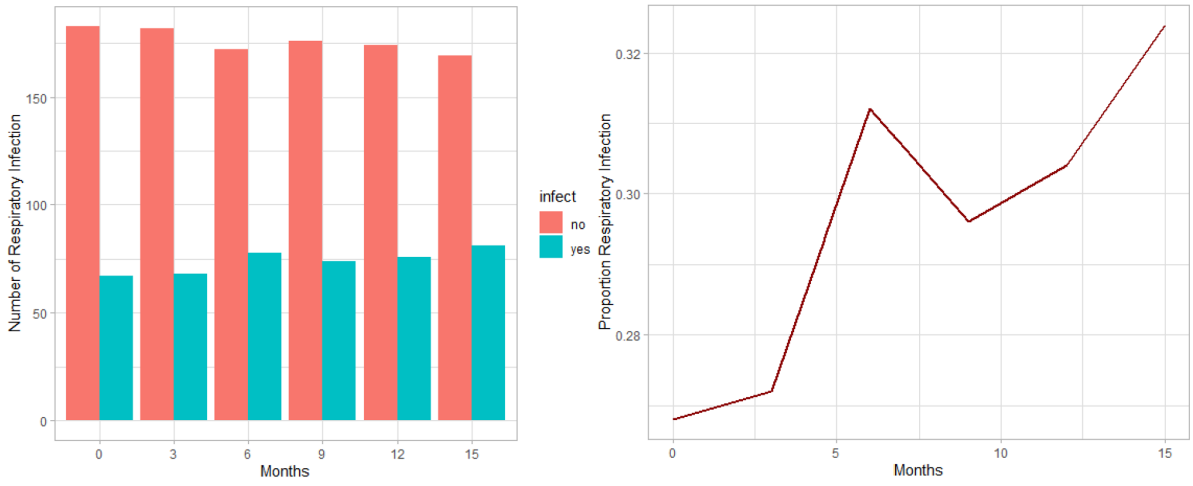

Figure 1: Distribution of the Response Variable for Different months

The Figure 1 (Left) shows the number of different situations of respiratory infection in 15 months. It seems that the distribution of respiratory infection are stable across time. We draw the proportion of respiratory infection across time in Figure 1 (Right), it shows that the percentage tends to increase across time but it increases just a little. And then we start to check the distribution of baseline

variables across subjects:

Baseline Age							
Months	1	2	3	4	5	6	7
Count	45	40	33	37	37	38	20

Variables	Gender	Respiratory Infection	Vitamin A deficiency
Count	male:female = 115:135	no:yes = 183:67	no:yes = 159:91

Table 1: Distribution of Baseline Variables Across Subjects

From Table 1, the gender of the patients is almost 1:1, which means the data is almost balanced for gender. We can see that most of patients at baseline are without respiratory infection and without vitamin A deficiency. Then we try to see the distribution of the response variable for different groups.

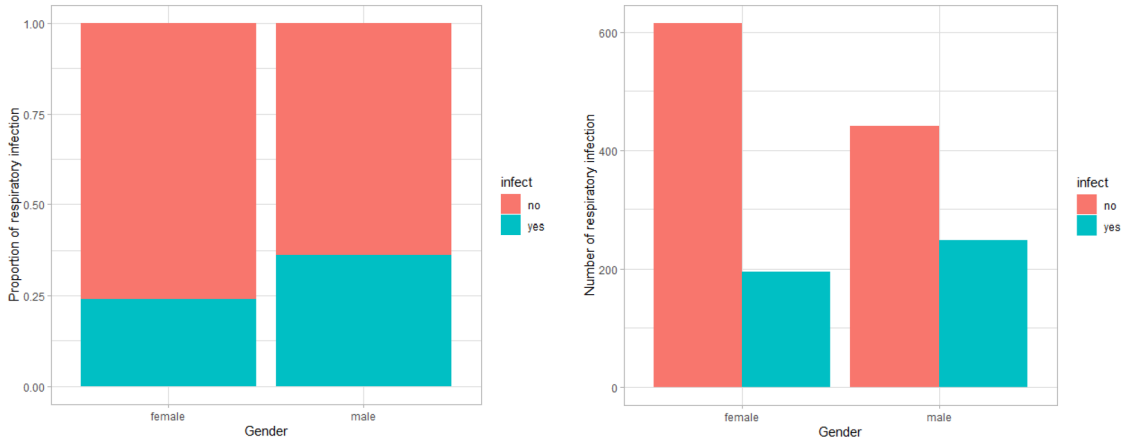

Figure 2: Distribution of the Response Variable for Different Gender Groups

The Figure 2 shows that the distribution of infection are similar between female and male groups. However, there are more proportion of infection in male group.

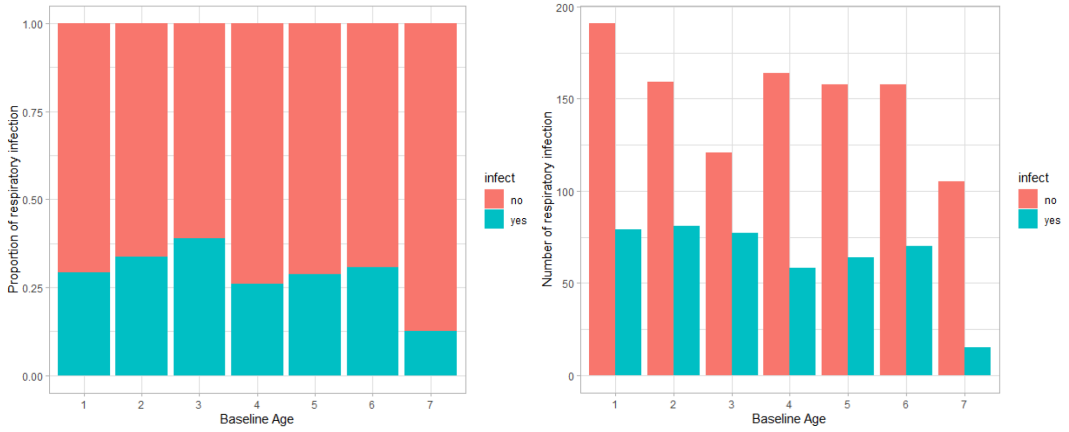

Figure 3: Distribution of the Response Variable for Different Baseline Age Groups

The Figure 3 shows that the distribution of respiratory infection are almost stable between groups

with different baseline age, but patients of 2-3 years old tend to have higher risk of respiratory infection.

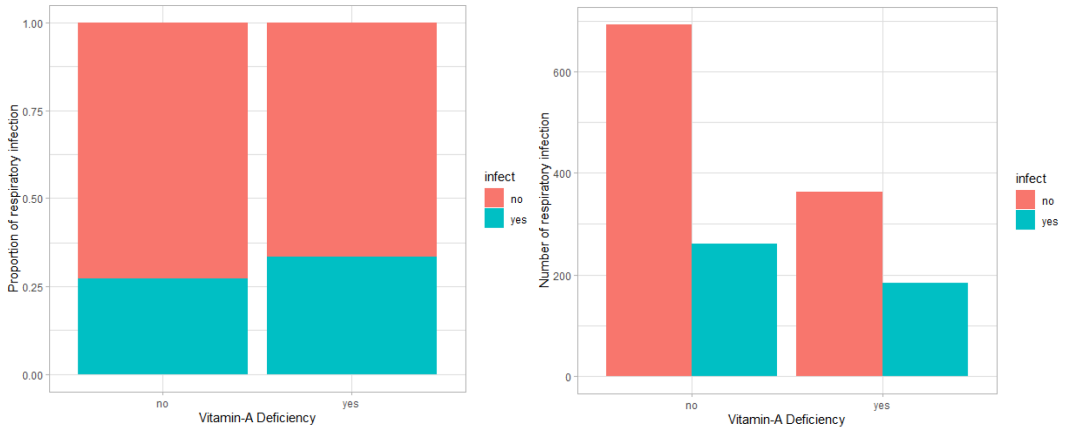

Figure 4: Distribution of the Response Variable for Different Groups of Vitamin-A Deficiency

The Figure 4 shows that people with Vitamin-A deficiency tend to have respiratory infection. And then we check the distribution of the vitamin-A deficiency for different groups.

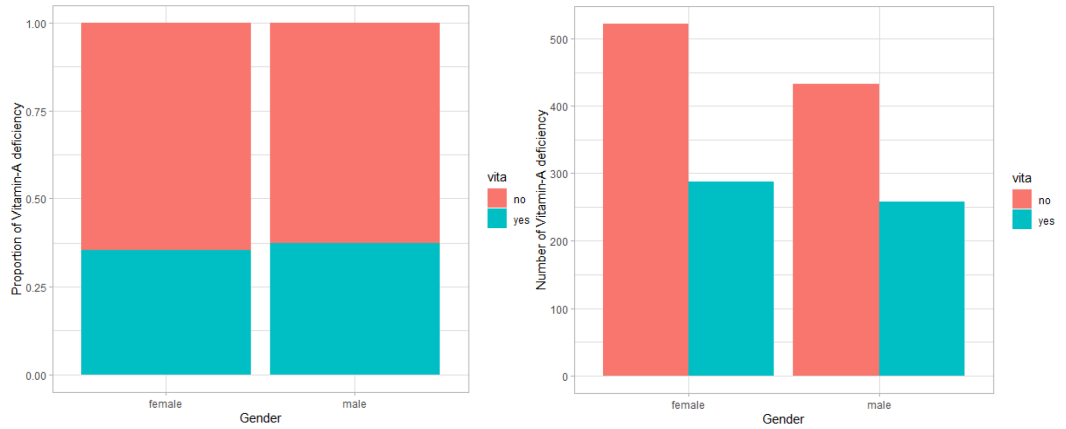

Figure 5: Distribution of the Vitamin-A Deficiency for Different Gender Groups

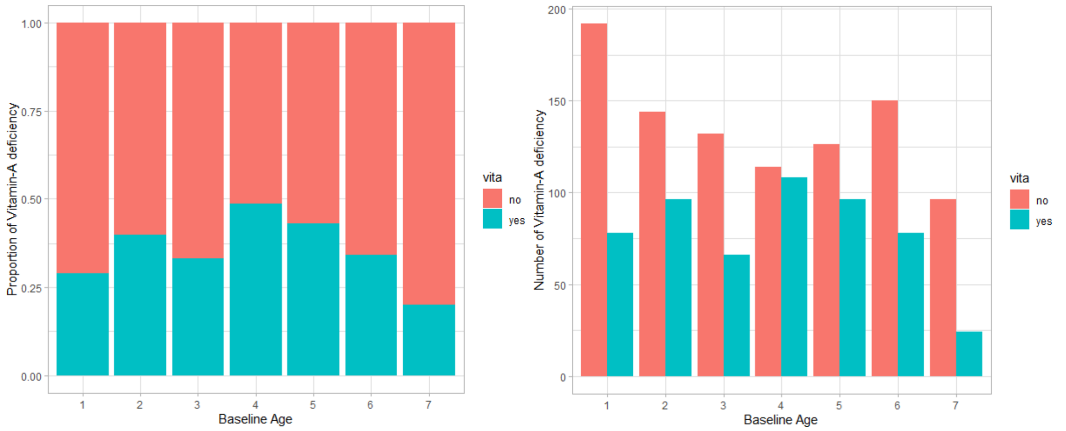

Figure 6: Distribution of the Vitamin-A Deficiency for Different Baseline Age Groups

In Figure 5, it seems that the vitamin-A deficiency for male and female groups are similar since the proportions are almost the same. In Figure 6, it seems that the distribution of vitamin-A deficiency are slightly stable between groups with different baseline age. However, patients of 4-5 years old tend to have higher risk of vitamin-A deficiency.
Now we will check how response changes by checking how many subjects with an infection at a given wave will still have an infection at the following wave. We define the changeable observations as the infection status during the time is different than the initial baseline situation.

Change	114 (First0 : First1 = 74:40)
No Change	136 (First0 : First1 = 109:27)

Number of Change	1	2	3	4	5
Count	41	27	14	15	17

Table 2: Summary of Changeable Observations

From Table 2, it also shows the ratio of people in initial status of respiratory infection. For the observations that change during the time, Table 2 also shows that most of people just change once however there are also many people change more than three times. It means that some patients with respiratory infection at baseline tends to recur again in the following time.

2 Correlation Structure

In this part, we continue to do exploratory analysis for correlation structure of the response variable. We define:

$$Y_{ij} = I(\text{The subject i has respiratory infection at time j})$$

and

$$\mu_{ij} = \mathbb{E}(Y_{ij}|x_{ij}) = P(Y_{ij} = 1)$$

We set up the 'logit' link function:

$$\text{logit}(\mu_{ij}) = \log(\frac{\mu_{ij}}{1-\mu_{ij}})$$

In the flexible generalized linear model, we include all covariates and all two-way interactions into the model. Since the Vitamin-A deficiency of subject i at time j is $vita_{ij}$(=1 if yes), the gender of subject i is $gender_i$(=1 if female) and the baseline age of subjet i is $bage_i$. We set the linear predictor:

$$\begin{aligned}\eta_{ij} = \beta_0 &+ \beta_1 time_{ij} + \beta_2 gender_i + \beta_3 bage_i + \beta_4 vita_{ij} \\ &+ \beta_5 time_{ij} * gender_i + \beta_6 time_{ij} * bage_i + \beta_7 time_{ij} * vita_{ij} \\ &+ \beta_8 gender_i * bage_i + \beta_9 gender_i * vita_{ij} + \beta_{10} bage_i * vita_{ij}\end{aligned}$$

In this case, the mean model is:

$$\text{logit}(\mu_{ij}) = \log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \eta_{ij}$$

The variance model is:

$$Var(Y_{ij}|x_{ij}) = \mu_{ij}(1 - \mu_{ij})$$
$$\phi = 1$$

Then we explore the correlation structure of the response variable. To further check the correlation structure of the residuals, the auto-correlation scatter plot is shown in Figure 7. We can see that the correlations of each time lag are relatively the same, which indicates the exchangeable correlation structure. And then we will examine autocorrelation function of the standardized residuals.

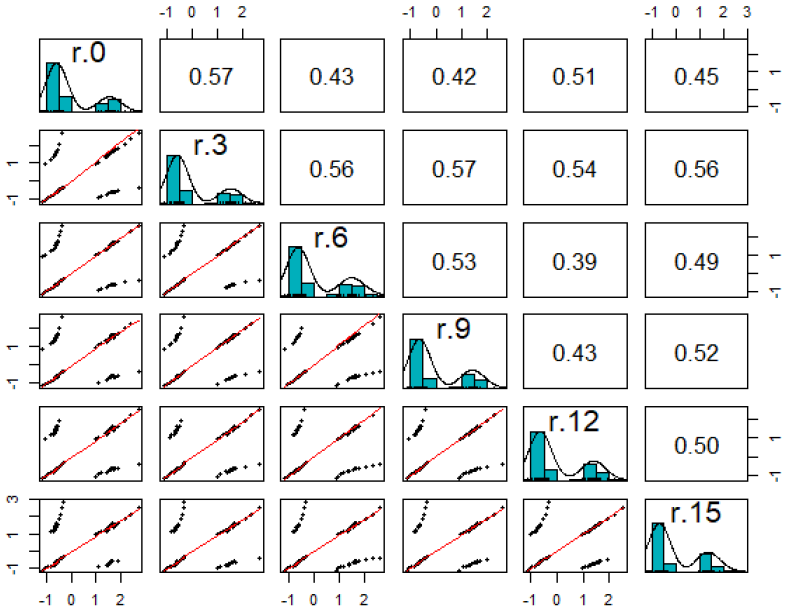

Figure 7: Auto-correlation Scatter Plot

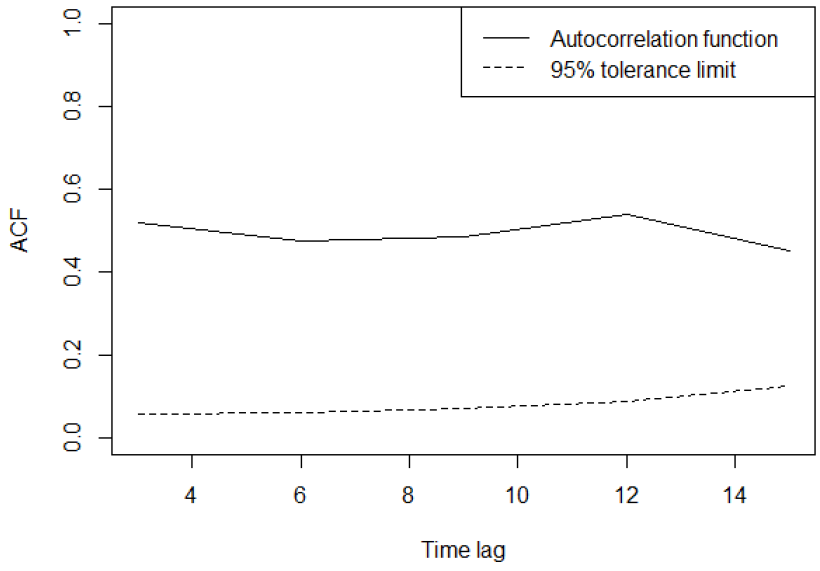

Figure 8: Correlogram

And then we will examine autocorrelation function of the standardized residuals. The Figure 8

shows the correlogram with relatively smooth autocorrelation function curve. The curve is upper than the tolerance limit, which means the correlations are significantly non-zero. Since the ACF are relatively the same for each time lag, we can use exchangeable correlation structure as working correlation model to further modeling.

3 GEE Modeling

In this part, we use GEE with robust variance estimator, and the exchangeable correlation model as working correlation structure, to fit a series of marginal models with the goal of finding a single model to address the investigators' questions of interest. The QIC (Quasi-likelihood under the Independence model Criterion) statistic proposed by Pan and the related QICu statistic can be used to compare GEE models. In this case, we design a backward model selection method with QICu starting from full model in section 2:

1. We start from full model including all two-way interaction terms and then use asymptotically chi-square test on each parameter β_i where $i = 1, ..., 10$:

   $$H_0 : \beta_i = 0 \text{ and } H_a : \beta_i \neq 0$$

   we calculate statistic $\chi^2 = (\hat{\beta}_i)^T (Var(\hat{\beta}_i))^{-1} (\hat{\beta}_i)$. The test-statistic is χ^2 distributed with degree of freedom 1. Then we calculate the p-value $P(\chi_1^2 > \chi^2)$, we will drop the term with highest p-value and move to next step.(Since we are mainly focus on the relationship between Vitamin-A deficiency and respiratory infection, if the the first-order term vita has highest p-value then we choose to drop the second-highest one.)

2. After dropping the term with highest p-value, we will re-fit the model with rest terms. We track the QIC and QICu in this model, if the QICu is lower than the previous model then we keep this method and move to next step. If the QICu is higher than the previous model, then we break the loop and then the previous model with lower QICu is preferred.

3. With the model we get in Step 2, we do the same testing method as Step 1 in this model. we will drop the term with highest p-value and move to step 2.

As the model selection method mentioned above, we record the model selection history in Table 3:

Step	Term with highest p-value	p-value	QIC	QICu
0	gender:vita	0.957	1778.2	1777.8
1	time:gender	0.861	1776.2	1775.8
2	gender	0.775	1774.04	1773.80
3	time:vita	0.6168	1772.14	1772.01
4	bage	0.5328	1770.25	1770.14
5	time:bage	0.15656	**1769.55**	**1769.30**
6	bage:vita	0.06704	1771.445	1771.34
7	vita	0.2042	1792.10	1791.93

Table 3: History of Model Selection Method

The Table 3 shows that in the step 5, we have obtained the lowest QIC (= 1769.55) and QICu (= 1769.30), in this case, we will use the following GEE model, and we set that:

$$\eta_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 vita_{ij} + \beta_3 time_{ij} * bage_i + \beta_4 bage_i * vita_{ij} + \beta_5 gender_i * bage_i$$

Then we fit this model in R and the results of fitting is shown in Table 4, and the 95% confidence intervals with asymptotic variance for each parameter are shown in Table 5.

Coefficients	β	Estimate	Std.err	Wald	Pr($>\|W\|$)
(Intercept)	β_0	-0.78367	0.17321	20.47	6.1e-06
time	β_1	0.03669	0.01573	5.44	0.01971
vita	β_2	-0.53646	0.45521	1.39	0.23859
time:bage	β_3	-0.00553	0.00390	2.01	0.15656
bage:vita	β_4	0.22944	0.10489	4.78	0.02872
bage:gender	β_5	-0.17615	0.04917	12.83	0.00034

Table 4: Results of Model Fitting

β	left	right
β_0	-1.12	-0.444
β_1	0.00585	0.0675
β_2	-1.43	0.356
β_3	-0.0132	0.00212
β_4	0.0238	0.435
β_5	-0.273	-0.0798

Table 5: The Wald 95% Confidence Intervals

We can also get the estimated parameter $\alpha = 0.499$ in exchangeable correlation structure.

4 Model Confirmation

In this part, we confirm the mean response is well-captured by the fitted mean model in section 3, by comparing the model fit to the empirical proportion of subjects with respiratory infections. In our model, the term $time_{ij}$ and $vita_{ij}$ are of linearity with the η_{ij}. In this case, we plot the empirical and predicted proportions of respiratory infections across the covariates:

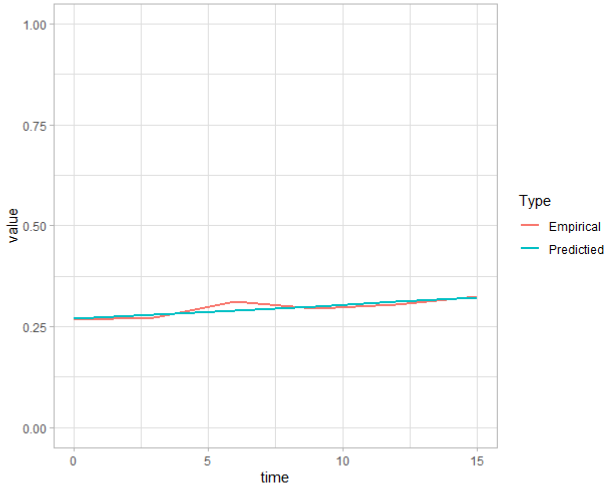

Figure 9: Empirical and Predicted Proportions across Time

In Figure 9, it shows that the empirical and predicted proportions of respiratory infections are very close across time, which indicates that the mean response is well-captured by the fitted mean model with linear covariate $time_{ij}$. And then we check $vita_{ij}$, the Table 6 shows that the empirical and predicted proportions of respiratory infections are very close for two groups of different Vitamin-A deficiency.

vita	Predicted	Empirical
0=no	0.273	0.274
1=yes	0.335	0.335

Table 6: Across Vitamin-A Deficiency

Then we further check the linearity of baseline age $bage_i$:

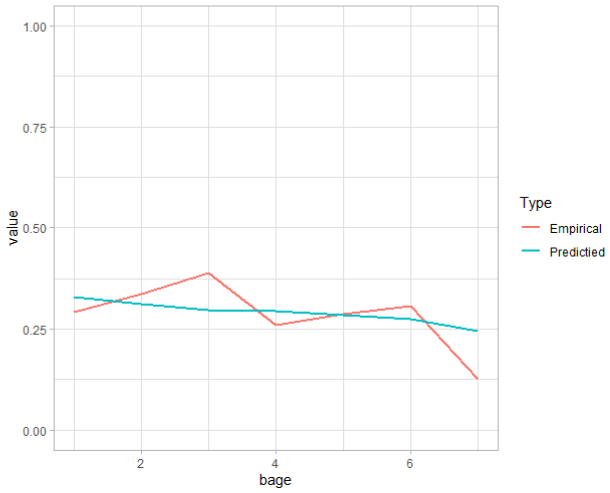

Figure 10: Empirical and Predicted Proportions for Baseline Age

We can see that the empirical proportion curve is fluctuated with predicted proportion curve, which means that the term $bage_i$ might be influenced by other covariates like $time_{ij}$ or $gender_i$. However, these two curves are pretty close so that we can use it to predict the risk. In conclusion, the mean response is well-captured by the fitted mean model.

5 Model Interpretation

We recall our final GEE model as following linear predictor:

$$\eta_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 vita_{ij} + \beta_3 time_{ij} * bage_i + \beta_4 bage_i * vita_{ij} + \beta_5 gender_i * bage_i$$

Then we give the marginal model interpretation for some of the typical parameter estimates in the final model:

1. β_0: the average intercept for the log odds for having respiratory infection at baseline time among male people without vitamin-A Deficiency.

2. $\beta_1 + k\beta_3$: the mean log odds ratio for having respiratory infection for a one-unit (ie, 3 months) difference in visiting time among male people with baseline age k (in years) and without vitamin-A Deficiency.

3. $\beta_2 + k\beta_4$: the mean difference of log odds ratio for having respiratory infection among male people with vitamin-A deficiency of baseline age k at first visit, compared with male people without vitamin-A deficiency of baseline age k at first visit.

4. β_4: the mean log odds ratio for having respiratory infection for a one-unit (in years) difference in baseline age among male people with vitamin-A deficiency at first visit.

5. β_5: the mean log odds ratio for having respiratory infection for one-unit (in years) difference in baseline age among female people without vitamin-A deficiency at first visit.

From section 3, the estimated parameter $\alpha = 0.499$ in exchangeable correlation structure, which is:

$$\begin{pmatrix} 1 & 0.499 & ... & 0.499 & 0.499 \\ 0.499 & 1 & 0.499 & ... & 0.499 \\ ... & ... & ... & ... & ... \\ 0.499 & ... & ... & 0.499 & 1 \end{pmatrix}_{6\times 6}$$

In this case, it means that the correlation of having respiratory infection within the same subject is 0.499 for any two separate visiting time. The Table 7 shows the model fitting results:

Coefficients	β	Estimate	Std.err	Wald	Pr($>\|W\|$)
(Intercept)	β_0	-0.78367	0.17321	20.47	6.1e-06
time	β_1	0.03669	0.01573	5.44	0.01971
vita	β_2	-0.53646	0.45521	1.39	0.23859
time:bage	β_3	-0.00553	0.00390	2.01	0.15656
bage:vita	β_4	0.22944	0.10489	4.78	0.02872
bage:gender	β_5	-0.17615	0.04917	12.83	0.00034

Table 7: Results of Model Fitting

Given the significant level 0.05, we can see that:

- The p-value for term $vita_{ij}$ is $0.23859 > 0.05$, which means that we cannot reject the hypothesis that $\beta_2 = 0$. However, we cannot say the vitamin-A deficiency has nothing to do with the infection. In this case, we will keep testing other β.

- Then we turn to the interaction term $bage : vita$, since the p-value of this term is $0.02872 < 0.05$, which means that we can reject the hypothesis that $\beta_4 = 0$. In this case, the mean log odds ratio for having respiratory infection for a one-unit (in years) difference in baseline age among male people with vitamin-A deficiency at first visit is statistically significant. Since $\beta_4 = 0.22944 > 0$, it suggests that for male people with vitamin-A deficiency, the risk of having respiratory infection tends to increase with the increment of age.

- Finally, we turn to the interaction term $bage : gender$, since the p-value of this term is $0.00034 < 0.05$, which means that we can reject the hypothesis that $\beta_5 = 0$. In this case, the the mean log odds ratio for having respiratory infection for one-unit (in years) difference in baseline age among female people without vitamin-A deficiency at first visit is statistically significant. Since $\beta_5 = -0.17615 < 0$, it suggests that for female people without vitamin-A deficiency, the risk of having respiratory infection tends to decrease with the increment of age.

And then we try to see whether the increasing age would affect the risk of infection in female people with vitamin-A deficiency. The effect can be represented as $\beta_4 + \beta_5$. Since $\beta_4 + \beta_5 > 0$, it suggests that for female people with vitamin-A deficiency, the risk of having respiratory infection tends to increase with the increment of age.
In conclusion, for all people with vitamin-A deficiency, the risk of having respiratory infection tends to increase with the increment of age. And this is consistent with the exploratory data analysis.
Furthermore, since $k\beta_5$ can be also interpreted as the mean difference log odds ratio for having respiratory infection from female group of age k without vitamin-A deficiency at first visit to male group with same conditions. Since $k\beta_5 < 0$, we can conclude that for the group without vitamin-A deficiency of given age, the male tend to have higher risk of respiratory infection than the female group. And this is consistent with the exploratory data analysis.

6 GLMM Modeling

In this part,we fit a conditional model (GLMM with a random intercept), using the covariates chosen in previous final marginal model. The random intercept model is:

$$P(Y_{ij} = 1|U_i, X_i) = \frac{1}{1 + e^{-\eta_{ij}}}$$

The GLMM linear predictor is:

$$\eta_{ij} = \beta_0 + U_i + \beta_1 time_{ij} + \beta_2 vita_{ij} + \beta_3 time_{ij} * bage_i + \beta_4 bage_i * vita_{ij} + \beta_5 gender_i * bage_i$$

where $U_i \sim N(0, v^2)$ and X_i are independent. Given U_i, Y_{ij} are independent. The results of model fitting is shown in Table 8:

Coefficients	β	Estimate	Std.err	Wald	Pr($>|W|$)
(Intercept)	β_0	-1.60976	0.34909	-4.61	4e-06
time	β_1	0.07795	0.03227	2.42	0.01571
vita	β_2	-1.34934	0.87762	-1.54	0.12417
time:bage	β_3	-0.01211	0.00767	-1.58	0.11424
bage:vita	β_4	0.53563	0.20677	2.59	0.00958
bage:gender	β_5	-0.36249	0.09615	-3.77	0.00016

Table 8: Results of Model Fitting

The estimated $v^2 = 7.29$ is the between-subject variance of subject-specific intercepts; it reflects the variation in the propensity of people for having respiratory infection. Then we give the interpretation of the estimated coefficients:

1. β_0: the average subject-specific intercept for the log odds for having respiratory infection at baseline time among male people without vitamin-A Deficiency.

2. $\beta_1 + k\beta_3$: the subject-specific log odds ratio for having respiratory infection for a one-unit (ie, 3 months) difference in visiting time among male people with baseline age k (in years) and without vitamin-A Deficiency.

3. $\beta_2 + k\beta_4$: the subject-specific difference of log odds ratio for having respiratory infection among male people with vitamin-A deficiency of baseline age k at first visit, compared with male people without vitamin-A deficiency of baseline age k at first visit.

4. β_4: the subject-specific log odds ratio for having respiratory infection for a one-unit (in years) difference in baseline age among male people with vitamin-A deficiency at first visit.

5. β_5: the subject-specific log odds ratio for having respiratory infection for one-unit (in years) difference in baseline age among female people without vitamin-A deficiency at first visit.

Given the significant level 0.05, we can see that the β_0, β_1, β_4 and β_5 are statistically significant since their p-values are smaller than 0.05. And it is consistent with the previous GEE model.
And then we compare the coefficient estimates from the conditional model and marginal model. In this case by theory, they are different since we have:

$$\beta_{GEE} \approx (c^2v^2 + 1)^{-\frac{1}{2}}\beta_{GLMM}$$

where $c = 16\sqrt{3}/(15\pi)$ and v^2 is the variance of random intercept.
The Table 9 shows in our model fitting results, which is consistent with the theory:

Coefficients	β	GLMM Estimate	GEE Estimate	Ratio(GEE/GLMM)
(Intercept)	β_0	-1.60976	-0.78367	0.487
time	β_1	0.07795	0.03669	0.471
vita	β_2	-1.34934	-0.53646	0.398
time:bage	β_3	-0.01211	-0.00553	0.456
bage:vita	β_4	0.53563	0.22944	0.428
bage:gender	β_5	-0.36249	-0.17615	0.486

Table 9: Results of Model Fitting

The β_{GEE} is for marginal model which describes the ratio of population odds, but the β_{GLMM} is for conditional model which describes the ratio of an individual's odds. Therefore, β_{GLMM} does not inherit the population average interpretation in a logistic model. The marginalized version of the random intercept model is approximately logistic, but the β's are not the same; rather they are attenuated toward zero. Intuitively, the randomness of the GEE model is from variability with working correlation structure and the randomness of the GLMM model is from the variability of random intercept. In this case, their differences are in agreement with what the theory predicts.