Question 1

(a)

Table 1: Summary statistics for the selected variables

Variables	Smoking Cessation Status		n	Statistics	P value
	Yes (n=428)	No (n=1201)			
Age	m=46.7, sd=12.5	m=42.9, sd=11.9	–	z = 5.460	< 0.001
Sex					
Male	237 (0.15)	562 (0.34)	799 (0.49)	$\chi^2_{(1)} = 8.954$	0.0028
Female	191 (0.12)	639 (0.39)	830 (0.51)		
Race					
White	390 (0.24)	1024 (0.63)	1414 (0.87)	$\chi^2_{(1)} = 8.951$	0.0028
Black or other	38 (0.02)	177 (0.11)	215 (0.13)		
Education					
College or more	63 (0.04)	119 (0.07)	182 (0.11)	$\chi^2_{(1)} = 6.883$	0.0087
Otherwise	365 (0.22)	1082 (0.66)	1447 (0.89)		
Weight at Baseline	m=72.6, sd=16.1	m=70.5, sd=15.6	–	z = 2.413	0.0158
Number of cigarettes smoked per day	m=18.8, sd=12.3	m=21.2, sd=11.6	–	z = -3.571	0.0004
Number of years spent smoking	m=26.6, sd=13.0	m=24.3, sd=11.8	–	z = 3.423	0.0006
Exercise					
Little or no	177 (0.11)	458 (0.28)	635 (0.39)	$\chi^2_{(1)} = 1.244$	0.2648
Otherwise	251 (0.15)	743 (0.46)	994 (0.61)		
Activity level					
Inactive	48 (0.03)	114 (0.07)	162 (0.10)	$\chi^2_{(1)} = 0.862$	0.3531
Otherwise	380 (0.23)	1087 (0.67)	1467 (0.09)		

The summary statistics table is shown in Table 1. The continuous variables are age, weight, number of cigarettes smoked per day, and number of years spent smoking. The mean (m) and standard deviation (sd) for these variables are given. Sex, race, education, exercise, and activity level are the categorical variables with two levels. The contingency tables are given for the binary variables.

The logistic regressions were fitted for each continuous variable to test whether there is a significant relationship with smoking cessation. It was assumed that the responses follow Bernoulli distribution with the univariate model and the logit link function of the smoking cessation probability. The test statistics and the P-values are given in Table 1. The P-values less than the significant level of 0.05 suggest that every continuous variable has a relationship with the probability of quitting smoking.

Pearson's Chi-squared tests with Yates' continuity correction were conducted for each binary variable. The number of samples in each cell was assumed to follow multinomial distribution given the total number of samples. The asymptotic distribution of the statistics is valid because the margins are large and the expected cell frequencies are larger than five. The P-values of exercise and activity level are larger than the significant level of 0.05, which means there is no significant evidence to conclude that these variables are associated with smoking cessation. It could be concluded that for each level of sex, race, education variables, there is considerable difference in the level of smoking cessation.

(b)

The model to evaluate the effect of smoking cessation on weight gain is $wt\_chg_i = \beta_0 + \beta_1 I(qsmk_i = 1) + \epsilon_i$, where i indicates the ith subject, wt_chg_i is weight gain, $I(qsmk_i = 1)$ is the indicator for smoking cessation, and $\epsilon_i \sim N(0, \sigma^2)$ independent of each subject.

The estimated coefficient of the parameter β_1 is 2.5406 with the 95% confidence interval of (1.6558, 3.4254). It is 95% confident that the interval (1.6558, 3.4254) contains the true value of the coefficient. For the simple linear model, the five assumptions which are Homogeneity of variances, Independence, Linearity, Existence, and Gaussian error assumptions are assumed (HILE Gauss). Among them, homogeneity of variance and Gaussian error and independence assumptions are necessary to be evaluated. Independence assumption can be verified by checking sampling scheme which is not available. The residual-predicted plot (R-P plot) for within and between cell in Figure 1 shows that homogeneity assumption is satisfied. The Q-Q plots within each cell in Figure 1 indicate heavy-tailed residual distribution, but it would be acceptable because ANOVA is robust to moderate violation of normality assumption with large sample size.

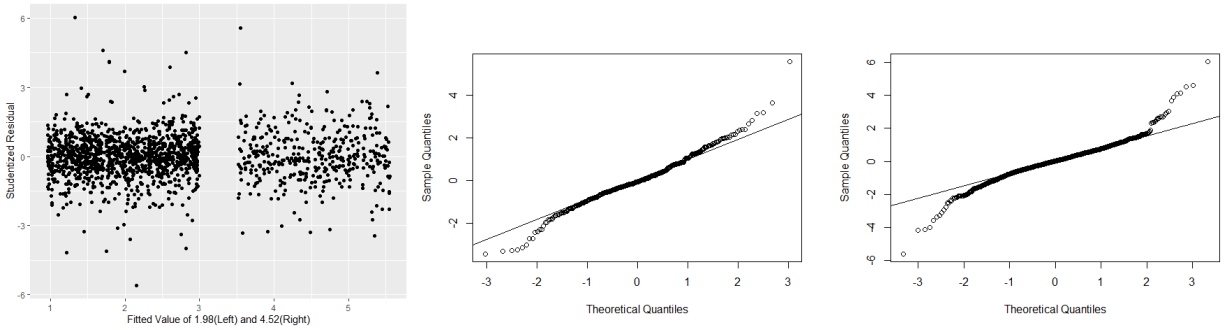

Figure 1: R-P Plot (Left), Q-Q Plot for qsmk:yes (Middle) and for qsmk:no (Right)

(c)

The multivariate linear model to analyze the effect of smoking cessation on weight gain adjusted by the selected variables is below.

$$\begin{aligned} wt\_chg_i = \beta_0 &+ \beta_1 I(qsmk_i = yes) + \beta_2 age_i + \beta_3 I(sex_i = female) + \beta_4 I(race_i = black\, or\, others) \\ &+ \beta_5 I(education_i = otherwise) + \beta_6 weight_i + \beta_7 smkintensity_i + \beta_8 smkyrs_i \\ &+ \beta_9 I(exercise_i = otherwise) + \beta_{10} I(active_i = otherwise) + \epsilon_i, \end{aligned} \tag{1}$$

where i indicates the ith subject, wt_chg_i is weight gain, $I(\cdot)$ is the indicator function, age_i is age variable, $weight_i$ is weight at baseline, $smkintensity$ is the number of smoke per day, and $smkyrs$ is the number of years spent smoking, respectively. $\epsilon_i \sim N(0, \sigma^2)$ independent of each subject. HILE Gauss are assumed and necessary to be verified for the model.

The estimated coefficient of the parameters and their 95% confidence interval (CI) are given in Table 2.

The R-P plot and Q-Q Plot for this model are given in Figure 2. There seems no pattern

Table 2: Estimate of Parameters and CI

	Parameter	Estimate	Std. Error	t-value	P-value	Confidence Interval	
β_0	Intercept	16.43	1.669	9.848	< 0.001	13.16	19.71
β_1	Smoking Cessation	3.376	0.441	7.649	< 0.001	2.510	4.241
β_2	Age	-0.205	0.033	-6.213	< 0.001	-0.269	-0.140
β_3	Sex	-1.458	0.446	-3.268	0.001	-2.333	-0.583
β_4	Race	0.534	0.582	0.917	0.359	-0.608	1.675
β_5	Education	0.974	0.608	1.602	0.109	-0.218	2.166
β_6	Weight at Baseline	-0.102	0.014	-7.447	< 0.001	-0.128	2.166
β_7	Smoking Intensity	0.022	0.017	1.307	0.192	-0.011	0.056
β_8	Smoking Year	0.043	0.033	1.292	0.197	-0.022	0.109
β_9	Exercise	-0.094	0.407	-0.230	0.818	-0.892	0.705
β_{10}	Active	-0.203	0.656	-0.309	0.757	-1.489	1.083

in R-P plot. Although the Q-Q plot indicates heavy-tailed residuals, it would be acceptable because sample size is large.

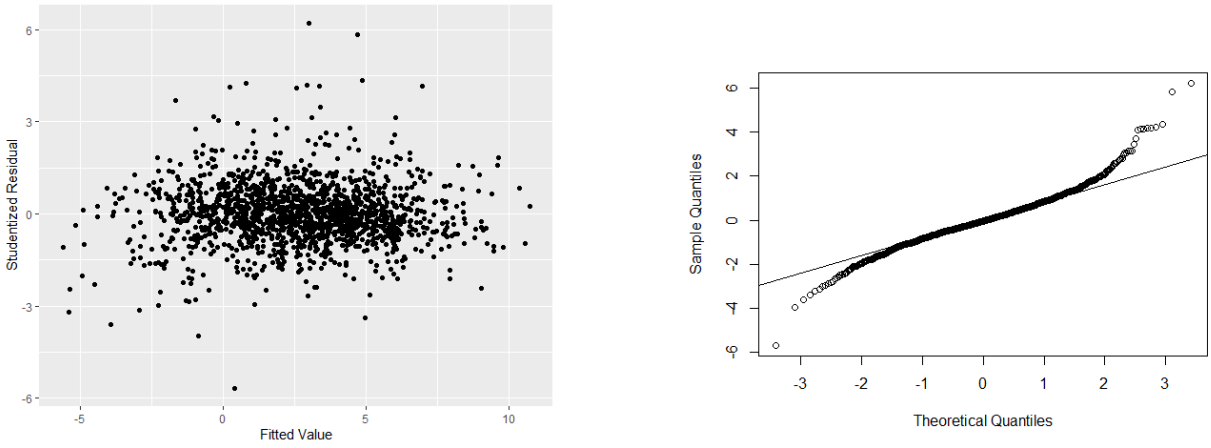

Figure 2: R-P Plot (Left) and Q-Q Plot (Right)

The 95% confidence interval for the effect of smoking cessation on weight gain is (2.510, 4.241), therefore it is 95% confident that the interval contains the true value of the effect. Both lower and upper limit of the interval increase from (1.6558, 3.454) in part (b) to (2.510, 4.241) in part (c). The confidence interval in part (c) is more reliable than the one in part (b) because the model in part (c) is adjusted by the key baselines. The effect of adjusting by the baseline variables is to eliminate possible confounding effects which can cause spurious association between smoking cessation and weight gain in part (b).

(d)

The logistic regression model assumes that the response follows Bernoulli distribution with the mean model in Equation (2).

$$\begin{aligned} logit(\pi_i) = \beta_0 &+ \beta_1 age_i + \beta_2 I(sex_i = female) + \beta_3 I(race_i = black\, or\, others) \\ &+ \beta_4 I(education_i = otherwise) + \beta_5 weight_i + \beta_6 smkintensity_i + \beta_7 smkyrs_i \\ &+ \beta_8 I(exercise_i = otherwise) + \beta_9 I(active_i = otherwise), \end{aligned} \quad (2)$$

where $\pi_i = E(Y_i)$ indicates the probability of smoking cessation. The other variables are defined the same as in Equation (1) of part (c).

The parameter estimates are given in the Table 3.

Table 3: Estimate of Parameters

	Parameter	Estimate	Std. Error	t-value	P-value
β_0	Intercept	-1.431	0.518	-2.763	0.006
β_1	Age	0.046	0.010	4.674	< 0.001
β_2	Sex	-0.454	0.143	-3.188	0.001
β_3	Race	-0.784	0.205	-3.822	< 0.001
β_4	Education	-0.443	0.178	-2.490	0.013
β_5	Weight at Baseline	0.007	0.004	1.519	0.129
β_6	Smoking Intensity	-0.026	0.006	-4.640	< 0.001
β_7	Smoking Year	-0.028	0.010	-2.784	0.005
β_8	Exercise	-0.154	0.129	-1.197	0.231
β_9	Active	-0.142	0.200	-0.709	0.478

The confusion matrix is shown in Table 4 by setting 0.5 as the cut-off value. The model accuracy is 0.749 and Cohan's kappa is 0.070. The ROC curve and AUC are given in Figure 3. Considering that the model was tested from the data used for model fitting, it cannot be concluded that the model predicts smoking cessation accurately. The value of kappa close to zero means the prediction is not significantly greater than random guessing.

Table 4: Confusion matrix

	Reference	
Prediction	0	1
0	1148	15
1	378	25

The estimated coefficient of age (β_1) and sex (β_2) are 0.0458 (0.0266, 0.0652) and -0.4544 (-0.7353, -0.1762), respectively. Therefore, it is 95% confident that the intervals of (0.0266, 0.0652) and (-0.7353, -0.1762) contain the true value of the coefficient for age (β_1) and sex (β_2). Since both 95% CI does not contain zero, it would be possible that there is a significant relationship of smoking cessation with age and sex.

Given the other variables, the odds of stopping smoking by the end of the study increases $exp(0.0458) = 1.0469$ times as likely when age increases by 1 unit. Also, females have $exp(-0.4544) = 0.6348$ times the odds of quitting smoking compared to males.

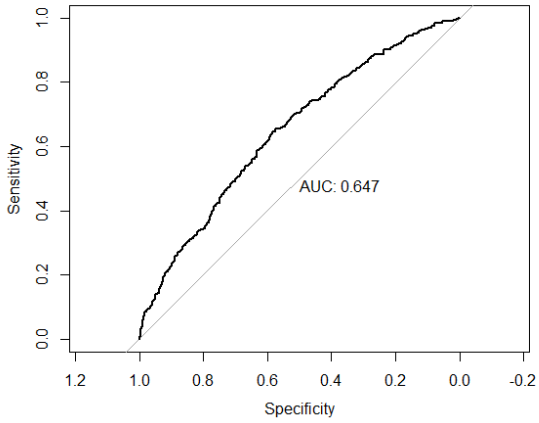

Figure 3: ROC Curve

(e)

By the properties of MLE, the asymptotic distribution is below in logistic regression.

$$\sqrt{n}(\hat{\beta} - \beta) \longrightarrow N(0, I(\beta)^{-1})$$
$$I(\beta) = \lim_{n\to\infty} \frac{1}{n} X^T W X$$
$$X = (x_1^T, ..., x_n^T)^T,\ W = diag(\frac{exp(x_1\beta)}{(1+exp(x_1\beta))^2}, ..., \frac{exp(x_n\beta)}{(1+exp(x_n\beta))^2}).$$

By Delta method with the following equations, the estimator $\hat{q} = \frac{1+e^{-x_0\hat{\beta}}}{1+e^{-x_1\hat{\beta}}}$ is a consistent estimator for q under the model in part (d), and the asymptotic distribution of the estimator is $\sqrt{n}(\hat{q} - q) \longrightarrow N(0, \bigtriangledown g(\beta)^t I(\beta)^{-1} \bigtriangledown g(\beta))$.

$$g(\beta) = \frac{1+e^{-x_0\beta}}{1+e^{-x_1\beta}}$$
$$\frac{\partial g(\beta)}{\partial \beta_j} = \frac{e^{-x_1\beta}}{1+e^{-x_1\beta}} \frac{1}{1+e^{-x_1\beta}} (1+e^{-x_0\beta}) x_{1j} - \frac{1}{1+e^{-x_1\beta}} e^{-x_0\beta} x_{0j}$$
$$x_0 = (1,\ 41.51,\ 1,\ 0,\ 0,\ 61.41,\ 18.31,\ 21.25,\ 1,\ 1)$$
$$x_1 = (1,\ 41.51,\ 1,\ 0,\ 1,\ 61.41,\ 18.31,\ 21.25,\ 1,\ 1).$$

The estimated $\hat{q}$ is 0.7138 and its standard error is 0.0929. The null hypothesis for a test of whether post-secondary education is associated with smoking cessation is $H_0 : q = 1$. Under the null, the test statistic is $\hat{q}$, and it follows the above asymptotic distribution with $q = 1$. The asymptotic 95% confidence interval is (0.5317, 0.8959), which does not contain the null value of 1. Therefore, there is a significant evidence to conclude that given the value of the other variables, the probability of stopping smoking is associated with education status.

Question 2

(a)

The data are from the 12-week randomized study which is designed to measure Patient Global Impression of Change (PGIC) at week 4, week 8, and week 12. The possible values of PGIC are (a): Very much, much improved (PGIC: 1), (b):Minimally improved, no improvement (PGIC: 2), and (c):Minimally worse, much worse, or very much worse (PGIC: 3), which is an ordinal variable with (a) being the most preferable result. The data contains total of 586 measurements of PGIC from 225 subjects. The subjects were grouped into three treatment groups which consists of A, B, and C. There are 89 missing observations of PGIC from 45 subjects. The type of missingness is summarized in Figure 4. The missing data pattern is monotone and the percentage of the subjects having at least one missing is 20%. The data includes intensity and logcBLPdur data as baseline variables.

Missing Data Patterns

	Occasion				
Group	1	2	3	Freq	Percent
1	X	X	X	180	80.00
2	X	X	.	13	5.78
3	X	.	.	20	8.89
4	.	.	.	12	5.33

Figure 4: Missing type

The sample cumulative log odds with (c) being the smallest level by treatment is given in Figure 5. Overall, treatment C is likely to increase the probability of being (a) against (b) and (c), and treatment A and B tend to increase the probability of being (b). In addition, the slope of the trajectory in the cumulative log odds graph seems to vary across treatment. From the mean cumulative log odds graph, it seems that a linear trajectory can be assumed in a proportional odds model.

The sample correlation among PGIC values from week 4, 8, and 12 within each subject is given in Figure 6. The relationship between the PGIC values are all positive, but not large, which means PGIC values from each measurement within a subject may not be strongly correlated.

(b)

The proportional odds model assumes that the response follows Multinomial distribution with the mean model given in Equation (3).

$$
\begin{aligned}
log\left(\frac{\pi_{i(c)}}{\pi_{i(b)}+\pi_{i(a)}}\right) &= \alpha_{(c)} + \beta_1 I(trt_i = B) + \beta_2 I(trt_i = C) + \beta_3 Intensity_i + \beta_4 logcLBPdur_i \\
log\left(\frac{\pi_{i(c)}+\pi_{i(b)}}{\pi_{i(a)}}\right) &= \alpha_{(b)} + \beta_1 I(trt_i = B) + \beta_2 I(trt_i = C) + \beta_3 Intensity_i + \beta_4 logcLBPdur_i,
\end{aligned}
\tag{3}
$$

where i indicates the ith subject, $\pi_{i(j)}$ is the probability of PGIC being in the jth category with $j = a, b, and\, c$ at week 12 given the other variables. For the treatment variable, treatment A is set to the reference group, and $I(\cdot)$ is the indicator function. $Intensity_i$ and

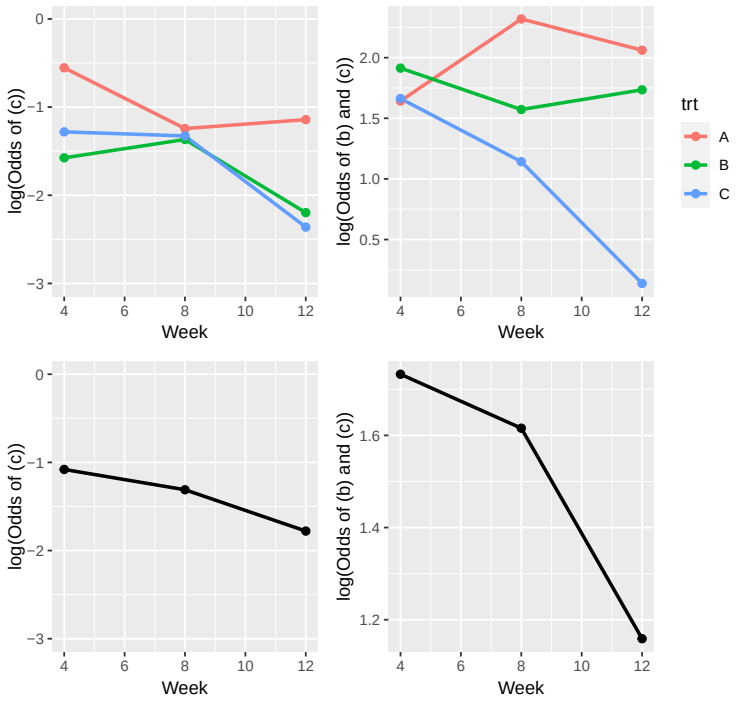

Figure 5: Cumulative log odds by treatment over 12 weeks by treatment (Top) and overall l (Bottom)

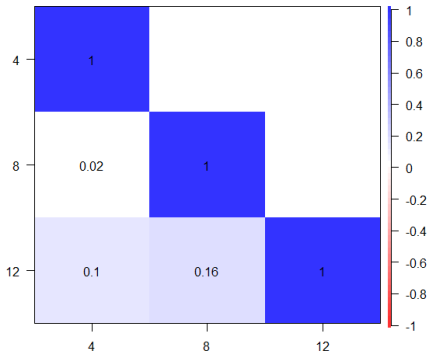

Figure 6: Correlation between week and PGIC

$logcLBPdur_i$ are considered as continuous variables. The mean models in Equation (3) have the shared parameters of βs.

Based on the assumption of Multinomial distribution with the mean model above, Newton-Raphson method can be used to calculate the maximum likelihood estimators. The null hypotheses to compare the effect of the treatment are $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, and $H_0^3 : \beta_2 - \beta_3 = 0$. Since there are multiple testings, Bonferroni correction can be used to guarantee the type 1 error rate at 0.05. Since the model is based on likelihood method,

Missing Completely At Random (MCAR) and Missing At Random (MAR) are the missing mechanisms that make the complete case analysis be valid.

Table 5: Estimate of Parameters

	Parameter	Estimate	Std. Error	Wald Chi-square	P-value
$\alpha_{(c)}$	Intercept1	-0.9888	0.6501	2.31	0.1283
$\alpha_{(b)}$	Intercept2	2.3156	0.6792	11.62	0.0007
β_1	Treatment B	-0.6712	0.3812	3.10	0.0783
β_2	Treatment C	-1.8799	0.4095	21.07	< 0.0001
β_3	Intensity	0.0736	0.0985	0.56	0.4546
β_4	logcLBPdur	-0.4310	0.3459	1.55	0.2128

Table 6: Contrasts Results

	Contrast	DF	Chi-Square	P-value	Type
H_0^1	Treatment A versus B	1	3.16	0.0757	LR
H_0^2	Treatment A versus C	1	23.01	<0.0001	LR
H_0^3	Treatment B versus C	1	10.19	0.0014	LR

The parameter estimates are given in Table 5, and the results of the test for the null hypotheses are shown in Table 6. The P-values for H_0^2 and H_0^3 are less than the corrected significance level of $\frac{0.05}{3} = 0.016$, thereby it can be concluded that treatment C has a significantly different effect on PGIC. Considering the parameter estimate of β_2 is -1.8799, the odds of being in (c) group or (b) and (c) groups for those who received treatment C is 0.1526 times as likely compared to those who received treatment A and B. That is, the treatment C showed a significant improvement on PGIC in the study compared to the other treatment.

(c)

One of the assumptions in the proportional odds model is the same βs for two mean models. This means that the two odds ratios for comparing two treatments within each model of the equation (3) should be equal each other. Figure 7 shows the odds ratio of the overall and cumulative contingency table between treatment and PGIC. Given two treatments, the two odds ratios are not similar one another. Therefore, it is possible that the assumption for the proportional odds model would not be hold in the data.

The generalized logistic model assumes that the response follows Multinomial distribution with the mean model given in Equation (4).

$$\begin{aligned} log\left(\frac{\pi_{i(c)}}{\pi_{i(a)}}\right) &= \beta_{c0} + \beta_{c1} I(trt_i = B) + \beta_{c2} I(trt_i = C) + \beta_{c3} Intensity_i + \beta_{c4} logcLBPdur_i \\ log\left(\frac{\pi_{i(b)}}{\pi_{i(a)}}\right) &= \beta_{b0} + \beta_{b1} I(trt_i = B) + \beta_{b2} I(trt_i = C) + \beta_{b3} Intensity_i + \beta_{b4} logcLBPdur_i, \end{aligned} \tag{4}$$

The Table 7 shows the parameter estimates in the mean model. The null hypotheses to compare the effect of the treatment are $H_0^1 : \beta_{b1} = \beta_{c1} = 0$, $H_0^2 : \beta_{b2} = \beta_{c2} = 0$, and $H_0^3 : \beta_{b2} - \beta_{b1} = \beta_{c2} - \beta_{c1} = 0$. The hypotheses test results are shown in Table 8. The P-values

	A	B	C
(c)	15	6	5
(b)	40	45	26
(a)	7	9	27

	A	B
(c)	15	6
(b), (a)	47	54
OR: 2.8723		

	A	B
(c), (b)	55	51
(a)	7	9
OR: 1.2866		

	A	C
(c)	15	5
(b), (a)	47	53
OR: 3.3030		

	A	C
(c), (b)	55	31
(a)	7	27
OR: 6.8433		

	B	C
(c)	6	5
(b), (a)	54	53
OR: 1.1778		

	B	C
(c), (b)	51	31
(a)	9	27
OR: 4.9355		

Figure 7: Contingency Table and Odds Ratios

for H_0^2 and H_0^3 are less than the corrected significance level of $\frac{0.05}{3} = 0.016$, thereby it can be concluded that treatment C has a significantly different effect on PGIC. Considering the parameter estimate of β_{c2} and β_{b2} are -2.5258 and -1.8082 respectively, treatment C shows more improvement on PGIC than the other treatment overall. For example, the ratio of the probabilities of being in the highest effective group (a) over (c) for treatment C group is 12.50 times as likely the ratio for the treatment A.

Table 7: Estimate of Parameters

	Parameter	Estimate	Std. Error	Wald Chi-square	P-value
β_{c0}	Intercept	0.8891	1.1699	0.5776	0.4473
β_{b0}	Intercept	2.3315	0.8775	7.0590	0.0079
β_{c1}	Treatment B	-1.2642	0.7062	3.2050	0.0734
β_{b1}	Treatment B	-0.1794	0.5540	0.1048	0.7461
β_{c2}	Treatment C	-2.5258	0.6782	13.8713	0.002
β_{b2}	Treatment C	-1.8082	0.4987	13.1473	0.003
β_{c3}	Intensity	0.1798	0.1756	1.0486	0.3058
β_{b3}	Intensity	0.0345	0.1251	0.0759	0.7829
β_{c4}	logcLBPdur	-0.7498	0.5976	1.5741	0.2096
β_{b4}	logcLBPdur	-0.5687	0.4514	1.5874	0.2077

Table 8: Contrasts Results

	Contrast	DF	Wald Chi-Square	P-value
H_0^1	Treatment A versus B	2	4.5522	0.1027
H_0^2	Treatment A versus C	2	17.4146	0.0002
H_0^3	Treatment B versus C	2	12.6104	0.0018

(d)

The marginal model for a longitudinal analysis consists of the mean model, the covariance function and the correlation structure. The mean model with the cumulative logit function is given in Equation (5). The covariance function is given in Equation (6), and the correlation structure is assumed to be independent. Since the sample size is large enough, empirical

variance estimator can be used. The analysis using a marginal model with GEE is valid only when missing mechanism is MCAR.

$$\begin{aligned} log\left(\frac{\pi_{ij(c)}}{\pi_{ij(b)}+\pi_{ij(a)}}\right) &= \alpha_{(c)} + \beta_1 Time_{ij} + \beta_2 I(trt_i = B) + \beta_3 I(trt_i = C) + \beta_4 I(trt_i = B) Time_{ij} \\ &\quad + \beta_5 I(trt_i = C) Time_{ij} + \beta_6 Intensity_i + \beta_7 logcLBPdur_i \\ log\left(\frac{\pi_{ij(c)}+\pi_{ij(b)}}{\pi_{ij(a)}}\right) &= \alpha_{(b)} + \beta_1 Time_{ij} + \beta_2 I(trt_i = B) + \beta_3 I(trt_i = C) + \beta_4 I(trt_i = B) Time_{ij} \\ &\quad + \beta_5 I(trt_i = C) Time_{ij} + \beta_6 Intensity_i + \beta_7 logcLBPdur_i \end{aligned} \tag{5}$$

where ij indicates the jth measurement in the ith subject. $E(Y_{ij(k)}|x_{ij}) = \pi_{ij(k)}$ is the probability of PGIC being in the kth category for ij with $k = a, b, and\, c$. $Time_{ij}$ is a continuous variable meaning a measurement week. The other variables are defined the same as in part (b).

$$Cov(Y_{ij})_{2\times 2} = \phi \begin{bmatrix} \pi_{ij(c)}(1-\pi_{ij(c)}) & -\pi_{ij(c)}\pi_{ij(b)} \\ -\pi_{ij(b)}\pi_{ij(c)} & \pi_{ij(b)}(1-\pi_{ij(b)}) \end{bmatrix} \tag{6}$$

where $Y_{ij} = (y_{ij(c)}\ y_{ij(b)})^T$, and $y_{ij(k)} = 1$ when PGIC is k in ij, and 0 otherwise.

The null hypotheses for the comparison of the treatments are $H_0^1 : \beta_2 = \beta_4 = 0$, $H_0^2 : \beta_3 = \beta_5 = 0$, and $H_0^3 : \beta_3 - \beta_2 = \beta_5 - \beta_4 = 0$. The estimated parameters and the results of the test are given in Table 9 and 10.

Table 9: Estimate of Parameters

	Parameter	Estimate	Std. Error	Z	P-value
$\alpha_{(c)}$	Intercept1	-0.3885	0.5443	-0.71	0.4754
$\alpha_{(b)}$	Intercept2	2.6141	0.5738	4.56	< 0.0001
β_1	Time	-0.0390	0.0466	-0.84	0.4033
β_2	Treatment B	-0.5188	0.5230	-0.99	0.3211
β_3	Treatment C	0.2383	0.5257	0.45	0.6503
β_4	Treatment B*Time	0.0023	0.0595	0.04	0.9692
β_5	Treatment C*Time	-0.1456	0.0622	-2.34	0.0192
β_6	Intensity	0.0096	0.0574	0.17	0.8671
β_7	logcLBPdur	-0.2589	0.1922	-1.35	0.1780

Table 10: Contrasts Results

	Contrast	DF	Chi-Square	P-value	Type
H_0^1	Treatment A versus B	2	6.73	0.0345	Score
H_0^2	Treatment A versus C	2	17.31	0.0002	Score
H_0^3	Treatment B versus C	2	8.97	0.0113	Score

The P-value for the null hypotheses of H_0^2 and H_0^3 are less than the corrected significant level of $\frac{0.05}{3} = 0.016$, which means there is a strong evidence that the linear trajectories of the two mean model for treatment C are different from the ones for treatment B and

A. Overall, treatment C shows better treatment effect with more subjects being in PGIC group (a) than the other treatments, especially at week 12. Although the intercept of the trajectory for treatment C is higher than the others, its slope is negative with large absolute value resulting that the probability of being in higher effective group ((a) or (b)) is larger for those who received treatment C than the other treatment at week 8 and week 12. Therefore, one can conclude that the treatment C has a more positive effect on PGIC.

(e)

The generalized estimating equation (GEE) method with the mean model in Equation (7), the variance function of $Var(Y_{ij}) = \pi_{ij}(1-\pi_{ij})$, and the compound symmetry correlation structure was used for the test.

$$\begin{aligned} logit(\pi_{ij}) &= \beta_0 + \beta_1 Time_{ij} + \beta_2 I(trt_i = B) + \beta_3 I(trt_i = C) + \beta_4 I(trt_i = B) Time_{ij} \\ &\quad + \beta_5 I(trt_i = C) Time_{ij} + \beta_6 Intensity_i + \beta_7 logcLBPdur_i, \end{aligned} \tag{7}$$

where ij indicates the jth measurement in the ith subject. $E(Y_{ij}|x_{ij}) = \pi_{ij}$ is the probability of PGICbin being 1 (Very much, much improved) for ij. The other variables are defined the same as part (b) and (d).

The logistic imputation model in Equation (8) was used for missing data. The number of the imputed dataset is 50 and the seed for reproducibility was 12345678.

$$\begin{aligned} logit(\pi_{i1}) &= \gamma_{10} + \gamma_{11} Y_{i2} + \gamma_{12} I(trt_i = B) + \gamma_{13} I(trt_i = C) + \gamma_{14} I(trt_i = B) Y_{i2} \\ &\quad + \gamma_{15} I(trt_i = C) Y_{i2} + \gamma_{16} Intensity_i + \gamma_{17} logcLBPdur_i \\ logit(\pi_{i2}) &= \gamma_{20} + \gamma_{21} Y_{i1} + \gamma_{22} Y_{i3} + \gamma_{23} I(trt_i = B) + \gamma_{24} I(trt_i = C) + \gamma_{25} I(trt_i = B) Y_{i1} \\ &\quad + \gamma_{26} I(trt_i = B) Y_{i1} + \gamma_{27} I(trt_i = C) Y_{i3} + \gamma_{28} I(trt_i = C) Y_{i3} \\ &\quad + \gamma_{29} Intensity_i + \gamma_{2(10)} logcLBPdur_i, \\ logit(\pi_{i3}) &= \gamma_{30} + \gamma_{31} Y_{i2} + \gamma_{32} I(trt_i = B) + \gamma_{33} I(trt_i = C) + \gamma_{34} I(trt_i = B) Y_{i2} \\ &\quad + \gamma_{35} I(trt_i = C) Y_{i2} + \gamma_{36} Intensity_i + \gamma_{37} logcLBPdur_i, \end{aligned} \tag{8}$$

where $\pi_{ij} = E(Y_{ij}|X_{ij})$, $Y_{ij} = PGICbin_{ij}$. $j = 1, 2, and\, 3$ means week 4, 8, and 12, respectively. The other variables are defined the same as above.

The estimated parameters are given in Table 11. The empirical sandwich estimator was used for the covariance of the estimates. The estimates of the parameters were computed by using the mean of each estimate from the 50 imputed datasets. The variance estimates was calculated by using the formula $\hat{V}_{MI} = \bar{V} + (1 + 1/M)\hat{B}$, where $\bar{V}$ is the average variance estimate and $\hat{B}$ is the between imputation variance.

The null hypotheses for the comparison of the treatments are $H_0^1 : \beta_2 = \beta_4 = 0$, $H_0^2 : \beta_3 = \beta_5 = 0$, and $H_0^3 : \beta_3 - \beta_2 = \beta_5 - \beta_4 = 0$. The results of the test are given in Table 12. The P-value for the null hypotheses of H_0^2 and H_0^3 are less than the corrected significant level of $\frac{0.05}{3} = 0.016$, which means there is a strong evidence that the linear trajectories for treatment C in the mean model is different from that for treatment B and A. Although the intercept of the trajectory for treatment C is smaller than the ones for the other treatment, its slope is strongly positive resulting that the probability of being in higher effective group ($PGICbin = 1$) is larger for those who received treatment C than the other treatment at week 8 and week 12. For example, the odds of PGICbin being 1 for those who received

Table 11: Estimate of Parameters

	Parameter	Estimate	Std. Error	t-value	P-value
β_0	Intercept	-2.1225	0.7514	-2.82	0.0047
β_1	Time	-0.0621	0.0719	-0.86	0.3882
β_2	Treatment B	-0.3801	0.7420	-0.51	0.6085
β_3	Treatment C	-1.0780	0.7248	-1.49	0.1370
β_4	Treatment B*Time	0.0828	0.0915	0.91	0.3653
β_5	Treatment C*Time	0.2612	0.0860	3.04	0.0024
β_6	Intensity	0.0398	0.0843	0.47	0.6372
β_7	logcLBPdur	0.3331	0.2818	1.18	0.2374

Table 12: Contrasts Results

	Contrast	Num DF	Den DF	F-value	P-value
H_0^1	Treatment A versus B	2	9137.4	0.76	0.4685
H_0^2	Treatment A versus C	2	8190.1	9.59	<0.0001
H_0^3	Treatment B versus C	2	4584.1	7.00	0.0009

treatment C at week 12 is 4.232954 ($= exp(-1.0780 + 0.3801 + 12 \times (0.2612 - 0.0828))$ times as likely the odds for those who received treatment B. Therefore, one can conclude that the treatment C has a more positive effect on PGIC.

Question 3

(a)

Table 13 is the table of the rates of the congenital anomaly in Minnesota by maternal age and birth order. The marginal rates of each level of maternal age shows that the rate of the congenital anomaly increases as maternal age increases. The increase in the rate is considerable when age group goes from five at 0.01592 to six at 0.06793. Although there seems no distinct trend in the marginal rates over birth order, birth order group 4 and 5 with age group 6 shows the highest rates among all combination with the value of 0.01688 and 0.01945, respectively.

Table 13: Table of Incidence rates by maternal age and birth order

	Maternal Age						
Birth Order	1	2	3	4	5	6	Total
1	0.00043	0.00033	0.00044	0.00096	0.00415	0.00786	0.01418
2	0.00036	0.00060	0.00049	0.00078	0.00309	0.01507	0.02039
3	0.00040	0.00059	0.00047	0.00133	0.00267	0.00866	0.01411
4	0	0.00086	0.00072	0.00140	0.00310	0.01688	0.02296
5	0	0.00026	0.00093	0.00205	0.00291	0.01945	0.02561
Total	0.00119	0.00264	0.00306	0.00653	0.01592	0.06793	0.09726

(b)

The Poisson regression modeling event rate assumes that the response follows Poisson distribution with the mean model $E(Y_i) = \mu_i = n_i\lambda(X_i\beta)$, where $\lambda(X_i\beta) = exp(X_i\beta)$. Thus, $log(\mu_i) = log(n_i) + X_i\beta$. Three mean models were compared in which $X_i\beta$ are defined as Equation (9) (Model 1), (10) (Model 2), and (11) (Model 3).

$$X_i\beta = \beta_1 + \sum_{j=2}^{6} \beta_j I(m\_age = j) + \sum_{j=2}^{4} \gamma_j I(b\_order = j) \tag{9}$$

$$X_i\beta = \beta_0 + \beta_1 m\_age + \beta_2 b\_order + \beta_3 m\_age \times b\_order \tag{10}$$

$$X_i\beta = \beta_0 + \beta_1 m\_age + \beta_2 b\_order \tag{11}$$

Where $I(\cdot)$ is the indicator function, therefore $m\_age = 1$ and $b\_order = 1$ are the reference groups in Model 1. m_age and b_order are on the continuous scale in Model 2 and Model 3.

It can be shown that Model 3 is nested within Model 1 by putting the null hypothesis is $H_0^1 : 2\beta_2 - \beta_3 = 3\beta_2 - \beta_4 = 4\beta_2 - \beta_5 = 5\beta_2 - \beta_6 = 0$ *and* $2\gamma_2 - \gamma_3 = 3\gamma_2 - \gamma_4 = 4\gamma_2 - \gamma_5 = 0$. Also, Model 3 is nested in Model 2 with the null hypothesis of $H_0^2 : \beta_3 = 0$. Therefore, those hypotheses can be tested by using the difference of deviance and χ^2_{df} with df being the difference of the degree of freedom. Then, Model 1 and Model 2 can be compared by value of AIC to find an appropriate model. Table 14 Shows the value of deviance and AIC. The goodness-of-fit test between Model 1 and Model 3 consists of the null hypothesis, H_0^1, and χ^2_7. The test statistic is $781.0520 - 128.1500 = 652.902$ and the P-value is less than 0.001. Thus, Model 1 shows a significant improvement over Model 3. The goodness-of-fit test between Model 2 and Model 3 consists of the null hypothesis, H_0^2, and χ^2_1. The test

statistic is $781.0520 - 604.2128 = 176.8392$ and the P-value is less than 0.001. Thus, Model 2 also shows a significant improvement over Model 3. Between Model 1 and Model 2, AIC of Model 1 is smaller than that of Model 2, thereby Model 1 can be the final model.

Table 14: Goodness-of-fit

	Deviance	DF	AIC
Model 1	128.1500	20	323.6117
Model 2	604.2128	26	787.6744
Model 3	781.0520	27	962.5136

The estimated coefficient, standard errors, and z-scores from Model 1 are given in Table 15.

Table 15: Estimate of Parameters

	Parameter	Estimate	Std. Error	z-value	P-value	Scaled Std. Error	z-value
β_1	Intercept	-7.8641	0.0899	-87.488	<0.0001	0.2274	<0.0001
β_2	m_age 2	0.1149	0.1014	1.134	0.2568	0.2564	0.6587
β_3	m_age 3	0.1336	0.1055	1.266	0.2053	0.2667	0.6220
β_4	m_age 4	0.9538	0.1039	9.181	<0.0001	0.2628	0.0017
β_5	m_age 5	1.6314	0.1049	15.546	<0.0001	0.2654	<0.0001
β_6	m_age 6	3.3001	0.1038	31.781	<0.0001	0.2626	<0.0001
γ_2	b_order 2	0.2055	0.0689	2.982	0.0029	0.1743	0.2522
γ_3	b_order 3	0.1756	0.0718	2.444	0.0145	0.1817	0.3454
γ_4	b_order 4	0.4592	0.0728	6.312	<0.0001	0.1840	0.0214
γ_5	b_order 5	0.5872	0.0683	8.599	<0.0001	0.1727	0.0028

(c)

The quasi-likelihood approach assumes that the mean model is Model 1 from part (b) and $Var(Y_i) = \phi\mu_i$. The estimated parameters and its scaled standard errors are given in Table 15. The standard error of the parameters with over-dispersion becomes larger than non-scaled standard errors. As a result, contrary to part (b), γ_2 and γ_3 are not statistically significantly different from zero with the significant level of 0.05.

(d)

The negative binomial regression model assumes that the response Y_i follows the negative binomial distribution with the parameter α and μ_i, where $E(Y_i) = \mu_i$, $Var(Y_i) = \mu_i + \alpha\mu_i^2$ and $\mu_i = n_i\lambda(X_i\beta) = n_i exp(X_i\beta)$. The AIC table for the models listed in part (b) is given in Table 16. Although Model 1 still has the smallest AIC, the AIC of Model 3 is close to that of Model 1 and Model 2. For interpretation purpose, Model 3 is used for the negative binomial regression.

The estimated coefficient and the parameter α are given in Table 17. From the estimated parameters, the incidence rate ratio for the effect of maternal age is 2.0582 (1.8587, 2.2790), which means that the incidence rate of the congenital anomaly becomes 2.0582 times as likely when *m_age* increases by one-unit. The incidence rate ratio for the effect of birth order on congenital anomaly is 1.0590 (0.9281, 1.2083), which means that the incidence rate of the

Table 16: AIC table

	AIC
Model 1	274.2376
Model 2	297.2056
Model 3	298.3932

congenital anomaly is 1.0590 times as likely when *b_order* increases by one-unit. Also, β_2 is not statistically significantly different from zero with the significance level of 0.05. Thus, one can conclude that only maternal age has a relationship with the development of congenital anomaly.

Table 17: Estimate of Parameters

	Parameter	Estimate	Std. Error	Wald Chi-Square	P-value
β_0	Intercept	-9.2598	0.2563	1304.77	<0.0001
β_1	m_age	0.7218	0.0520	192.63	<0.0001
β_2	b_order	0.0573	0.0673	0.73	0.3943
α	Dispersion	0.2192	0.0608		

(e)

The estimated dispersion parameter α is 0.2192. The null hypothesis $H_0 : \alpha = 0$ can be tested by using a score test. The test statistic is the score statistic which follows $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ under the null hypothesis. The calculated test statistic is 119.8326 and its p-value is less than 0.001. Thus, it can be concluded that there is a significant evidence that the dispersion parameter would not be zero. With $\alpha = 0$, the negative binomial regression is the same as Poisson regression without dispersion. Therefore, the testing result means there exist over-dispersion and it is possible that Poisson regression model would not fit well.

Question 4

(a)

The summary statistics for the variables in the clinical data is given in Table 18. Subtype variable is imbalanced with the percentage of the minority class being 12.4%. This imbalance can lead to difficulty in identifying the minority class in a prediction model because it is possible that the number of samples having Subtype 2 would not be enough to train a model.

Table 18: Summary of clinical data

Stage	1	2	3	4
	51 (0.102)	116 (0.232)	187 (0.374)	146 (0.292)
Grade	1	2	3	
	47 (0.094)	158 (0.316)	295 (0.59)	
MSI	1	2		
	153 (0.306)	347 (0.694)		
TMB	m:6.04	sd:2.40		
Subtype	1	2		
	438 (0.876)	62 (0.124)		

(b)

To investigate the collaborator's concern, one can evaluate the effect of the genes adjusting the clinical information in the logistic regression model. The two mean model to be compared are given in Equation (12) and (13).

$$\begin{aligned} logit(\pi_i) = \beta_0 + \sum_{j=1}^{3} \beta_j I(Stage_i = j+1) + \sum_{j=4}^{5} \beta_j I(Grade_i = j-2) + \beta_6 I(MSI_i = 2) \\ + \beta_7 TMB_i + \sum_{j=8}^{17} \beta_j Gene(j-7)_i \end{aligned} \tag{12}$$

$$\begin{aligned} logit(\pi_i) = \beta_0 + \sum_{j=1}^{3} \beta_j I(Stage_i = j+1) + \sum_{j=4}^{5} \beta_j I(Grade_i = j-2) + \beta_6 I(MSI_i = 2) \\ + \beta_7 TMB_i, \end{aligned} \tag{13}$$

where i indicates the ith subject, $\pi_i = E(Y_i|X_i)$ is the probability of having Subtype 2 given the other variables, and $I(\cdot)$ is the indicator funtion. The first 10 genes from Gene_1 to Gene_10 are used for the first model.

The null hypothesis is $H_0 : \beta_j = 0$ where $j = 8, ..., 17$. The alternative hypothesis is $H_1 :$ *at least one of* $\beta_j s$ *is not* 0. The test statistic is the difference of the deviances between the two model, and the test statistic follows Chi-squared distribution with the degree of the freedom, 10. The calculated statistic is 223.34, and the P-value is less than 0.001. Therefore, there is a significant evidence to reject the null hypothesis. The test result indicates that compared to the second model, the first model has a better fit to the data. Gene variables gives an additional information on explaining patient Subtype after adjusting clinical information.

(c)

To train a model, elastic net with 10 fold-cross validation was used. The tuning parameter space to be searched was the Cartesian product of the two set, $\alpha = \{0.1\, 0.2 \ldots 0.9,\ 1.0\}$ and $\lambda = \{0.002598259,\ 0.003949123,\ 0.006002316,\ 0.009122986,\ 0.013866128,\ 0.021075282,\ 0.032032556,\ 0.048686637,\ 0.073999360,\ 0.112472448\}$. The performance metric used for choosing optimal tuning parameter was Cohan's Kappa because the response is binary and the distribution of the response in the sample is not well-balanced.

The selected tuning parameter is $\alpha = 1.0$ and $\lambda = 0.01386613$. The result of the performance for the test set is summarized in Table 19 and 20. The Kappa value is around 0.3, which means a poor agreement. Although sensitivity is 0.9770 close to 1, specificity is relatively low with the value of 0.2500. Thus, it can be concluded that the model cannot appropriately identify true Subtype 2 in the test set.

Table 19: Confusion matrix

	Reference	
Prediction	0	1
0	85	9
1	2	3

Table 20: Performance Summary

Accuracy	0.8889
Kappa	0.3033
Sensitivity	0.9770
Sensitivity	0.2500

The Table 21 shows the list of the variables with non-zero coefficient in descending order with regard to the absolute value. Overall, the 10 genes included in the manuscript are selected in the reproduced model, and have the large estimated coefficients than the other variables with non-zero coefficient. However, the selected genes do not exactly match to the manuscript. It is possible that the author of the manuscript would have used a different penalization method. Since elastic net has a grouping effect, it is possible that some correlated genes were selected together. Also, data splitting and re-sampling method in cross validation generated randomness which affect the results. Lastly, different tuning parameter space to be searched could result in a different selection of the variables.

(d)

The advantage of pre-screening candidate genes by using univariate logistic regression is that it reduces computing time which is caused by high dimensionality of data when fitting penalized logistic regression. In addition, when there are many predictor variables in a model, the number of samples increases to train a model guaranteeing good prediction on new data. Pre-screening can reduce the number of predictors in the penalized logistic regression model. Refitting a logistic regression model with predictors having non-zero coefficient can also make it possible to find unnecessary predictors in a model.

Table 21: Non-zero Estimate of Parameters

	Coefficient		Coefficient		Coefficient
Intercept	-2.8244	Gene_7	0.7235	Gene_6	0.6948
Gene_4	0.5316	Gene_5	0.4859	Gene_10	0.4845
Gene_3	0.4340	Gene_9	0.3708	Gene_335	0.2860
Gene_433	-0.258	Gene_8	0.2280	Gene_2	0.2054
Gene_361	0.1704	Gene_374	-0.1700	Gene_277	-0.1662
Gene_461	-0.1648	Gene_1	0.1613	Gene_282	0.1610
Gene_160	-0.1525	Gene_268	-0.1407	Gene_18	0.1233
Gene_11	0.1233	Gene_199	-0.1187	Gene_357	-0.1027
Gene_231	0.1012	Gene_155	0.0992	Gene_272	0.0975
Gene_42	0.0970	Gene_152	0.0969	Gene_496	0.0958
Gene_240	0.0941	Gene_215	0.0930	Gene_468	-0.0918
Gene_458	-0.0839	Gene_116	-0.0805	Gene_120	-0.0798
Gene_202	0.0766	Gene_253	0.0719	Gene_28	-0.0708
Gene_287	0.0675	Gene_158	0.0648	Gene_409	0.0633
Gene_291	-0.0616	Gene_492	0.0538	Gene_101	-0.0524
Gene_64	-0.0474	Gene_135	-0.0456	Gene_34	0.0430
Gene_445	0.0380	Stage4	-0.0343	Gene_398	0.0339
Gene_238	0.0313	Gene_192	-0.0309	Gene_487	-0.0242
Gene_99	0.0218	Gene_459	0.0212	Gene_476	-0.0172
Gene_200	0.0162	Gene_29	-0.0143	Gene_395	0.0130
Gene_186	-0.0086	Gene_229	0.0078	Gene_362	0.0031
Gene_79	-0.0023	Gene_273	-0.0016	Gene_284	0.0011
Gene_125	-0.0006	Gene_453	-0.00004		

(e)

The performance of the fitted model on the UNC cohort is given Table 22 and 23. Sensitivity is still close to one with the value of 0.9333. However, accuracy and specificity are much smaller than the values in part (c). Kappa value is even worse with the value of -0.0357, which means there is no difference from random guessing. The reason of the poor performance comes from the fact that the model poorly identifies Subtype 2. In addition, the proportion of Subtype 2 is 25% in the UNC cohort which is larger than 12.4% in the clinical data. Therefore, the model performance is worse in UNC data. Considering these results, it is not confident to extend the model for patient at UNC because it is often important to correctly identify a minority class in biological setting.

Table 22: Confusion matrix

	Reference	
Prediction	0	1
0	70	24
1	5	1

Table 23: Performance Summary

Accuracy	0.71
Kappa	-0.0357
Sensitivity	0.9333
Sensitivity	0.0400