1. A study of children at risk of injuries

(a)

Table 1: Descriptive statistics of each variable

Age of the child	0	1	2	3	4			
	3264	3002	2997	2981	2942			
Number of injuries or death	0	1	2	3	4			
	14745	385	48	6	2			
Age of mother at birth of child	Age <20		20 Age 24		25 Age 29		Age >30	
	2250		4186		4340		4410	
Low Birth Weight	Weight 2.5kg				Weight <2.5kg			
	9736				5450			
Mom's year of education	Less than 12		12 years		13-15 years		16 or greater	
	4690		5114		3568		1814	
Average income quintile	1	2	3	4	5			
	3445	3140	3064	2929	2608			
Proximity to playground	Greater than a half mile away				Less than a half mile away			
	7614				7572			
Number of siblings	0	1	2	3	4 or more			
	3931	3800	3281	2332	1842			
Mother's race	White				Black			
	8099				7087			
Late or absent prenatal care	Care received				Care not received			
	9550				5636			

Table 1 shows descriptive statistics of each variable. Except for the age of the child and the number of injuries or death, all other variables are categorical variables. First, most of the children do not suffer from injury or death, as 14745 among 15186 covariate combinations (97.1%) do not include injured or dead children. Next, levels of children's age, age of mother, income level, proximity to playground, race of mom are equally distributed over combinations, so it would help avoid potential bias due to unbalance in the number of levels. Other covariates' levels are not equally distributed, but not on an extreme scale. Third, our interest is the number of injuries or death under certain covariate combinations, so we need to apply the count data analysis method. We can consider binomial or Poisson regression to fit the data. Finally, since this study is a prospective study, it is possible to estimate the relative risk of injury or deaths in children of a certain covariate combination relative to children of another covariate combination. In conclusion, we can see that injury or death of a young child is a very rare event, and by comparing the probability of a child suffers from injury or death with a certain covariate combination, we can investigate the effect of each variable in the relative risk of injury or deaths.

(b)

By collapsing the data based on proximity variable, we got two independent Poisson random variables given by

$$Y_1 \sim Poi(\lambda_1), Y_0 \sim Poi(\lambda_0)$$

where Y_1, Y_0 are random variables representing the number of injuries or deaths of children who live less than a half mile away and greater than or equal to a half mile away from a playground,

respectively, ignoring all other covariates. The mean of the Poisson distribution is given by the risk of injury or death (denoted as π_1, π_0) times total child years (denoted as n_1, n_0), given by

$$\lambda_i = n_i \pi_i, i = 0, 1$$

Finally, the only difference between λ_1 and λ_0 comes from the proximity status, and thus we can model them by

$$\begin{aligned} log(\lambda_1) &= log(n_1) + \beta_0 + \beta_{prox}, \\ log(\lambda_0) &= log(n_0) + \beta_0. \end{aligned}$$

It is a Poisson rate regression model with only one covariate, proximity, with offset of total child years. This gives the relative risk of injury or death in children living close to playgrounds relative to children living further away as $\exp(\beta_{prox})$, shown as

$$RR = \frac{P(\text{Injury or death} \mid \text{Living close to playgrounds})}{P(\text{Injury or death} \mid \text{Living further away})} = \frac{\pi_1}{\pi_0} = exp(\beta_{prox})$$

Thus, estimated relative risk and its 95% confidence interval is

$$\begin{aligned} &\widehat{RR} = \exp(\widehat{\beta}_{prox}) = 2.340 \\ &95\% \text{ Wald C.I} : RR \in (1.967, 2.752). \end{aligned}$$

It implies that the probability of a child being injured or dead who lives close to playgrounds is 2.340 times that of a child who lives further away, so thus living close to playgrounds is dangerous to young children.

(c)

The model is given by

$$Y_k | X_k \sim Poi(\lambda_k), E(Y_k | X_k) = \lambda_k = n_k \pi_k$$

where the assumption on the mean of the Poisson distribution is given by

$$\begin{aligned} log\left(\frac{E(Y_k|X_k)}{n_k}\right) = log\left(\frac{\lambda_k}{n_k}\right) =& X_k^T \beta \\ =& \beta_0 + \sum_{j=1}^{4} \beta_{age,j} X_{k,age}^j + \sum_{j=1}^{3} \beta_{agemom,j} X_{k,agemom}^j \\ &+ \beta_{lbw} X_{k,lbw} + \sum_{j=1}^{3} \beta_{educ,j} X_{k,educ}^j + \sum_{j=2}^{5} \beta_{income,j} X_{k,income}^j \\ &+ \beta_{prox} X_{k,prox} + \sum_{j=1}^{4} \beta_{othchld,j} X_{k,othchld}^j + \beta_{race} X_{k,race} + \beta_{pnclate} X_{k,pnclate} \end{aligned}$$

where Y_k is a random variables representing the number of injuries or deaths of children whose covariate combination is k^{th} combination, where the covariate combination is denoted by X_k ($k = 1, \ldots, 15186$), n_k is total child years, π_k is a probability of a child got injured or dead during one year whose covariate combination is X_k.

The mean of Poisson distribution is given by the risk of injury or death times total child years because the average number of injuries or death among children in a certain group over a year is proportional to the size of the group (or total child years) and to the probability of a child got injured or dead. This two factors solely determine the mean of the number of injuries or death. Finally, since the

outcome is non-negative integer, we model the outcome follows Poisson distribution with the mean discussed above.

$\beta_{age,j}$ are regression coefficients effects when the child's age is 1,2,3,4 ($\beta_{age,0}$ is set to 0 to indicate the reference level), $X^j_{k,age}$ is a dummy variable such that $X^j_{k,age} = \text{Indicator}(X_k\text{'s age category} = j)$, and other regression coefficients and dummy variables are defined similarly. Then, β is a vector of these regression coefficients, and X_k is a vector of these dummy variables.

Since we model the logarithm of $\lambda_k/n_k = \pi_k$ as a linear function of covariates, this is a Poisson rate regression model (modeling rate (or risk) π_k instead of mean λ_k).

Now, under this model, we can see that the logarithm of the relative risk is a linear function of the covariates, shown as

$$RR(X_a \text{ vs. } X_b) = \frac{\pi_a}{\pi_b} = \exp((X_a - X_b)^T\beta)$$

when comparing risks of two covariate combinations X_a and X_b.

The likelihood for the model is given by

$$\begin{aligned} L(\beta) = \prod_{k=1}^{15186} \frac{e^{-\lambda_k}\lambda_k^{y_k}}{y_k!} &= \exp(\sum_{k=1}^{15186} -\lambda_k + y_k log(\lambda_k) - log(y_k!)) \\ &= \exp(\sum_{k=1}^{15186} y_k X_k^T\beta - n_k exp(X_k^T\beta) + y_k log(n_k) - log(y_k!)). \end{aligned}$$

This belongs to a exponential family, and the original Poisson rate regression model reduces to ordinary GLM, and thus we can obtain the maximal likelihood estimator for the model.

Table 2: Relative Risk Estimation result based on Poisson rate regression model

Parameter	**Description**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**P value**
β_{prox}	Proximity to playgrounds	0.3270	0.1118	0.1078	0.5462	8.55	0.0035
$\exp(\beta_{prox})$	Relative Risk	1.387		1.113	1.726		

Thus, estimated relative risk and its 95% confidence interval is

$$\begin{aligned} &\widehat{RR} = \exp(\widehat{\beta}_{prox}) = 1.387 \\ &95\% \text{ Wald C.I} : RR \in (1.113, 1.726). \end{aligned}$$

Comparing this result with the result in (b) $\widehat{RR} = 2.340$, C.I: $RR \in (1.967, 2.752)$, when adjusting for all provided covariates, the relative risk of injury or death in children living close to playgrounds relative to children living further away is estimated smaller than un-adjusted one. It still implies that living close to playgrounds is dangerous to young children (since $\widehat{RR} > 1$), but the amount of risk is indeed not that large that we have seen in (b). Thus, by adjusting all covariates, we can conduct more accurate analysis on the relative risk, and indeed the relative risk was overestimated in (a) due to absence of adjustment.

(d)

The model is given by

$$Y_k|X_k \sim Poi(\lambda_k), E(Y_k|X_k) = \lambda_k = n_k\pi_k$$

where the assumption on the mean of the Poisson distribution is given by

$$log\left(\frac{E(Y_k|X_k)}{n_k}\right) = log(\frac{\lambda_k}{n_k}) = X_k^T\beta$$

which is the same as the model in (c), but the variance is given by

$$Var(Y_k|X_k) = \phi E(Y_k|X_k) = \phi\lambda_k$$

where $\phi = 1$ indicates ordinary Poisson regression model, while $\phi > 1$ indicates overdispersion in the data. This model is fitted by quasi likelihood estimation, and ϕ is estimated as a ratio of the deviance to its associated degrees of freedom.

Table 3: Estimation Result of Overdispersion Poisson regression model

Parameter	**Description**	**Level**	**Estimate**	**Standard Error**	**Confidence Limits**		**Wald Chi-Square**	**P value**
β_0	Intercept		-7.4025	0.833	-9.0351	-5.7699	78.97	<.0001
$\beta_{age,1}$	Age	1	-0.5334	0.6972	-1.8998	0.8331	0.59	0.4442
$\beta_{age,2}$		2	-0.582	0.7085	-1.9707	0.8066	0.67	0.4114
$\beta_{age,3}$		3	-0.9263	0.8018	-2.4978	0.6453	1.33	0.248
$\beta_{age,4}$		4	0.1009	0.5939	-1.0631	1.265	0.03	0.8651
$\beta_{agemom,1}$	Age mom	20-24	-0.3879	0.629	-1.6207	0.8449	0.38	0.5374
$\beta_{agemom,2}$		25-29	-0.6791	0.78	-2.2079	0.8497	0.76	0.384
$\beta_{agemom,3}$		30+	-1.0153	0.9482	-2.8737	0.8432	1.15	0.2843
β_{lbw}	Low birth weight		0.5273	0.6279	-0.7033	1.758	0.71	0.401
$\beta_{educ,1}$	Mom education years	12	-0.233	0.5378	-1.2872	0.8211	0.19	0.6648
$\beta_{educ,2}$		13-15	-0.3536	0.8161	-1.953	1.2459	0.19	0.6648
$\beta_{educ,3}$		16+	-1.4152	1.5562	-4.4653	1.635	0.83	0.3632
$\beta_{income,2}$	Income Quintile	2	-0.1568	0.6465	-1.4239	1.1102	0.06	0.8083
$\beta_{income,3}$		3	-0.2588	0.7002	-1.6311	1.1136	0.14	0.7117
$\beta_{income,4}$		4	-0.5148	0.7679	-2.0199	0.9903	0.45	0.5026
$\beta_{income,5}$		5	-0.3675	0.8515	-2.0365	1.3015	0.19	0.6661
β_{prox}	Proximity	1	0.327	0.5635	-0.7774	1.4314	0.34	0.5617
$\beta_{othchld,1}$	Other siblings	1	0.416	0.5757	-0.7124	1.5444	0.52	0.4699
$\beta_{othchld,2}$		2	0.5393	0.7271	-0.8859	1.9644	0.55	0.4583
$\beta_{othchld,3}$		3	1.0312	0.9124	-0.7572	2.8195	1.28	0.2584
$\beta_{othchld,4}$		4+	1.1697	1.1036	-0.9932	3.3326	1.12	0.2892
β_{race}	Mom race		-0.092	0.606	-1.2798	1.0957	0.02	0.8793
$\beta_{pnclate}$	Paternal Care late		0.0682	0.5774	-1.0635	1.1999	0.01	0.906
ϕ	Scale		25.387					

According to Table 3, estimated relative risk pertaining to the proximity and its 95% confidence interval is

$$\widehat{RR} = \exp(\widehat{\beta}_{prox}) = 1.387$$
$$95\% \text{ Wald C.I} : RR \in (0.460, 4.185).$$

All parameter estimates are shown to be statistically insignificant under $\alpha = 0.05$ significance level. That is, even though β_{prox} estimate is positive, since it is not statistically significant, there is no guarantee to say that children live close to playgrounds are at a higher risk of injuries relative to

children live further away. Therefore, unlike the results from (b), due to inflated standard error according to the introduction of overdispersion parameter, no covariates suggest an increased risk of injury or deaths, including the proximity to playgrounds. The results from (a) do not change because they come from descriptive statistics analysis, which do not depend on the model assumption.

(e)

To account for the fact that the assumed conditional variance is a quadratic function of the conditional mean, we introduce a two-level hierarchical model given by

$$\begin{aligned}&\text{Level 1: } Y_k|\lambda_k \sim Poi(\lambda_k),\ \lambda_k = n_k\pi_k\\&\text{Level 2: } \lambda_k|X_k \sim \text{Gamma}(\mu_k, 1/c),\ \text{where } E(\lambda_k|X_k) = \mu_k \text{ and } Var(\lambda_k|X_k) = c\mu_k^2.\end{aligned}$$

Then, the marginal distribution of Y_k is given by

$$\begin{aligned}&Y_k|X_k \sim NB(c, \mu_k),\ \text{where } E(Y_k|X_k) = \mu_k \text{ and } Var(Y_k|X_k) = \mu_k + c\mu_k^2,\\&P(Y_k = t|X_k) = \frac{\Gamma(t+1/c)}{\Gamma(t+1)\Gamma(1/c)}\frac{(c\mu_k)^t}{(1+c\mu_k)^{t+1/c}}\end{aligned}$$

which is a Negative Binomial distribution. By imposing the assumption on the mean of the Negative Binomial distribution given by

$$log\left(\frac{E(Y_k|X_k)}{n_k}\right) = log(\frac{\mu_k}{n_k}) = X_k^T\beta,$$

it becomes a Negative Binomial regression, which models the variance as a quadratic function of the mean, as

$$\begin{aligned}Var(Y_k|X_k) &= E(Y_k|X_k) + cE(Y_k|X_k)^2,\\E(Y_k|X_k) &= \mu_k = n_k e^{X_k^T\beta}.\end{aligned}$$

$c = 0$ indicates ordinary Poisson regression model, because there would be no distribution assumption on λ_k if $c = 0$, while $0 < c$ indicates overdispersion in the data.

The likelihood for the model is given by

$$\begin{aligned}L(\beta) &= \prod_{k=1}^{15186}\frac{\Gamma(y_k+1/c)}{\Gamma(y_k+1)\Gamma(1/c)}\frac{(c\mu_k)^{y_k}}{(1+c\mu_k)^{y_k+1/c}}\\&= \exp\left(\sum_{k=1}^{15186} y_k log\left(\frac{c\mu_k}{1+c\mu_k}\right) - \frac{1}{c}log\left(1+c\mu_k\right) + log\left(\frac{\Gamma(y_k+1/c)}{\Gamma(y_k+1)\Gamma(1/c)}\right)\right)\\&= \exp\left(\sum_{k=1}^{15186} y_k log\left(\frac{cn_k e^{X_k^T\beta}}{1+cn_k e^{X_k^T\beta}}\right) - \frac{1}{c}log\left(1+cn_k e^{X_k^T\beta}\right) + log\left(\frac{\Gamma(y_k+1/c)}{\Gamma(y_k+1)\Gamma(1/c)}\right)\right).\end{aligned}$$

If c is known, this belongs to a exponential family, and the model reduces to ordinary GLM, and thus we can obtain the maximal likelihood estimator for the model. c is also estimated by maximal likelihood estimator.

Table 4: Relative risk estimates and their confidence intervals from (b) through (e)

Model	Description	RR Estimate	SE(log($\widehat{RR}$))	Confidence Limits		Wald Chi-Square	P value
(b)	Univariate Poisson regression model	2.340	0.0888	1.967	2.752	91.62	<.0001
(c)	Multivariate Poisson rate regression model	1.387	0.1118	1.113	1.726	8.55	0.0035
(d)	Overdispersion Poisson regression model	1.387	0.5635	0.460	4.185	0.34	0.5617
(e)	Negative Binomial regression model	1.383	0.1136	1.107	1.728	8.15	0.0043

Table 4 shows all relative risk estimates and their confidence intervals from parts (b) through (e). Given the result, we can say that the proximity to a playground increases the risk for injury or deaths compared to living further away from one, as all relative risk estimates are positive and statistically significant except the model in (d). The model in (b) is oversimplified model, so it cannot estimate relative risk accurately, while model in (c) and (e) give similar inference results for relative risk. Even though overdispersion Poisson regression model in (d) failed to prove statistically significance of relative risk greater than 1, its estimate is still similar with (c) and (e).

(f)

The hypothesis of whether the number of siblings is linearly related to the logarithm of the relative risk is given by

$$H_0 : \beta_{othchld,2} = 2\beta_{othchld,1},\ \beta_{othchld,3} = 3\beta_{othchld,1},\ \beta_{othchld,4} = 4\beta_{othchld,1} \ \text{ vs } \ H_1 :\ \text{not } H_0.$$

Under the null hypothesis, the model reduces to

$$Y_k|X_k \sim Poi(\lambda_k), E(Y_k|X_k) = \lambda_k = n_k\pi_k$$

where the assumption on the mean of the Poisson distribution is given by

$$\begin{aligned} log\left(\frac{E(Y_k|X_k)}{n_k}\right) = log\left(\frac{\lambda_k}{n_k}\right) =& X_k^T\beta \\ =& \beta_0 + \sum_{j=1}^{4}\beta_{age,j}X_{k,age}^j + \sum_{j=1}^{3}\beta_{agemom,j}X_{k,agemom}^j \\ &+ \beta_{lbw}X_{k,lbw} + \sum_{j=1}^{3}\beta_{educ,j}X_{k,educ}^j + \sum_{j=2}^{5}\beta_{income,j}X_{k,income}^j \\ &+ \beta_{prox}X_{k,prox} + \boldsymbol{\beta}_{othchld}X_{k,othchld} + \beta_{race}X_{k,race} + \beta_{pnclate}X_{k,pnclate} \end{aligned}$$

where $\beta_{othchld}$ is a regression coefficient pertaining to the number of siblings, $X_{k,othchld} = X_{k,othchld}^1 + 2X_{k,othchld}^2 + 3X_{k,othchld}^3 + 4X_{k,othchld}^4 \in \{0,1,2,3,4\}$ is a number of siblings, and all the other coefficients and dummy variables are defined as previously. Note that under H_0,

$$\begin{aligned} \sum_{j=1}^{4}\beta_{othchld,j}X_{k,othchld}^j &= \beta_{othchld,1}X_{k,othchld}^1 + \beta_{othchld,2}X_{k,othchld}^2 + \beta_{othchld,3}X_{k,othchld}^3 + \beta_{othchld,4}X_{k,othchld}^4 \\ &\overset{H_0}{=} \beta_{othchld,1}X_{k,othchld}^1 + 2\beta_{othchld,1}X_{k,othchld}^2 + 3\beta_{othchld,1}X_{k,othchld}^3 + 4\beta_{othchld,1}X_{k,othchld}^4 \\ &= \beta_{othchld,1}\left(X_{k,othchld}^1 + 2X_{k,othchld}^2 + 3X_{k,othchld}^3 + 4X_{k,othchld}^4\right) \\ &= \beta_{othchld,1}X_{k,othchld}, \end{aligned}$$

thus the summation of regression coefficients times dummy variables is converted into the linear predictor of the number of siblings. H_0 is equivalent to the following form, given by

$$H_0 : L\beta = \begin{bmatrix} \beta_{othchld,2} - 2\beta_{othchld,1} \\ \beta_{othchld,3} - 3\beta_{othchld,1} \\ \beta_{othchld,4} - 4\beta_{othchld,1} \end{bmatrix} = 0$$

where L is a 3 by 23 contrast matrix utilized for the test such that

$$L = \begin{array}{c} \begin{matrix} \beta_0 & \beta_{age} & \dots & \beta_{othchld,1} & \beta_{othchld,2} & \beta_{othchld,3} & \beta_{othchld,4} & \dots & \beta_{pnclate} \end{matrix} \\ \begin{pmatrix} 0 & 0 & \dots & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & -3 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & -4 & 0 & 0 & 1 & \dots & 0 \end{pmatrix} \end{array}$$

The results of likelihood ratio, score and Wald tests for testing the null hypothesis are given in Table 5.

Table 5: Test results

Test	**Construction**	**Asymptotic Distribution**	**Statistic**	**P value**
LRT	$2\left(l(\hat{\beta}) - l(\hat{\beta}^0)\right)$	$\chi^2(3)$	2.59	0.4597
Score	$\dot{l}(\hat{\beta}^0)^T[-\ddot{l}(\hat{\beta}^0)]^{-1}\dot{l}(\hat{\beta}^0)$	$\chi^2(3)$	2.40	0.4927
Wald	$(L\hat{\beta})^T[L\widehat{\Sigma}_{\hat{\beta}}L^T]^{-1}(L\hat{\beta})$	$\chi^2(3)$	2.59	0.4588

l is a log likelihood of the full model, $\hat{\beta}$, $\hat{\beta}^0$ are MLE under full model and reduced model (under H_0) respectively, and $\widehat{\Sigma}_{\hat{\beta}}$ is an asymptotic covariance matrix of $\hat{\beta}$. All three test statistics have the same asymptotic distribution of χ^2 with the degrees of freedom equals 3. Their statistics are similar, and p values are also similar ($0.45 \sim 0.49$). None of the tests rejects H_0 at the $\alpha = 0.05$ significance level. Therefore, we can say that the effect of siblings is linear in the logarithm of the relative risk.

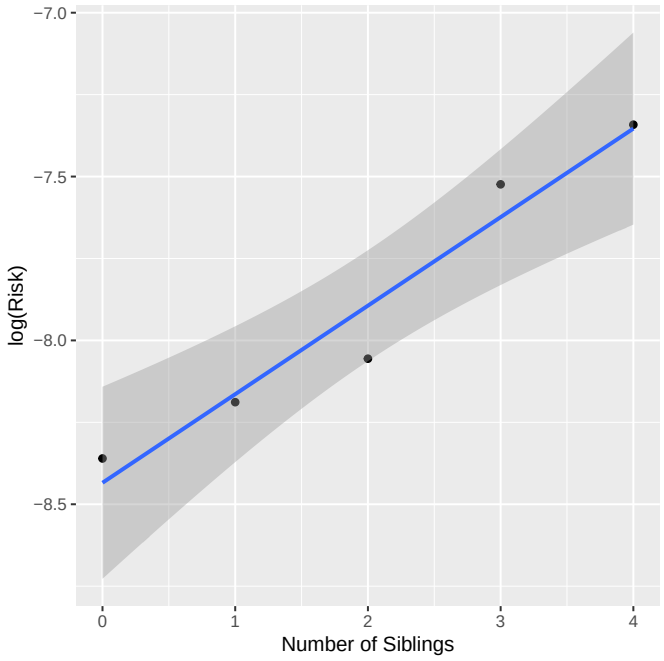

Figure 1: Number of siblings versus log(Risk)

Figure 1 plots the number of siblings versus logarithm of collapsed risk. By collapsing the data based on the number of siblings, we can estimate collapsed risk by

$$\hat{\pi}_i = \frac{Y_i}{n_i}$$

where $Y_i, i \in \{0, 1, 2, 3, 4\}$ are random variables representing the number of injuries or deaths of children whose number of siblings is i, ignoring all other covariates. n_i, $i \in \{0, 1, 2, 3, 4\}$ are total child years of children with i siblings. As shown in Figure 1, the logarithm of collapsed risk is in the linear relationship with the number of siblings. Therefore, it is reasonable to assume that the number of siblings can be measured as a numeric value, and its effect is linear in the logarithm of the relative risk.

2. A study of association between diabetes and periodontal disease

(a)

Table 6: Descriptive statistics of each factor

Factor	Level	Group			χ^2 test P value
		Disease	Normal	Overall	
Age	20-34	249 (0.24)	773 (0.76)	1022	<.0001
	35-44	437 (0.42)	600 (0.58)	1037	
	45-54	920 (0.56)	723 (0.44)	1643	
	55+	818 (0.63)	480 (0.37)	1298	
Smoker	Yes	1105 (0.57)	838 (0.43)	1943	<.0001
	No	1319 (0.43)	1738 (0.57)	3057	
Health Insurance	Yes	1087 (0.44)	1365 (0.56)	2452	<.0001
	No	1337 (0.52)	1211 (0.48)	2548	
Gender	Female	1231 (0.42)	1680 (0.58)	2911	<.0001
	Male	1193 (0.57)	896 (0.43)	2089	
Diabetes	Yes	1602 (0.55)	1293 (0.45)	2895	<.0001
	No	822 (0.39)	1283 (0.61)	2105	

Table 6 represents marginal observed counts and relative frequencies (the number in parenthesis) of periodontal disease for the levels of each factor. For each factor, χ^2 independence test of the factor and periodontal disease status showed highly significant result, meaning that every factor is related with periodontal disease marginally. For Age factor, the relative frequency of disease compared to normal increases as subject's age increases. The relative frequency of disease is higher for a subject who smokes, does not have a health insurance, suffers from diabetes, and is a male, but not on an extreme scale.

(b)

Estimated **marginal** odds ratio for periodontal disease with respect to diabetes status and 95% confidence interval is

$$\widehat{OR} = 1.933$$
$$95\% \text{ Wald C.I} : OR \in (1.725, 2.168).$$

Estimated odds ratio for periodontal disease with respect to diabetes status for **smokers** and 95% confidence interval is

$$\widehat{OR} = 1.498$$
$$95\% \text{ Wald C.I} : OR \in (1.246, 1.802).$$

Estimated odds ratio for periodontal disease with respect to diabetes status for **non-smokers** and 95% confidence interval is

$$\widehat{OR} = 2.195$$
$$95\% \text{ Wald C.I} : OR \in (1.894, 2.547).$$

(c)

The logistic regression model is given by

$$Y_i|X_i \sim Bernoulli(p_i)$$

where the assumption on the mean of the Bernoulli distribution is given by

$$\begin{aligned} logit\left(E(Y_i|X_i)\right) = log\left(\frac{p_i}{1-p_i}\right) =& X_i^T\beta \\ =& \beta_0 + \sum_{j=2}^{4}\beta_{age,j}X_{i,age}^j + \beta_{smoker}X_{i,smoker} + \beta_{hinsurance}X_{i,hinsurance} \\ &+ \beta_{female}X_{i,female} + \beta_{diab}X_{i,diab} + \beta_{smoker*diab}X_{i,smoker}X_{i,diab} \end{aligned}$$

where Y_i $(i = 1, \ldots, 5000)$ is a Bernoulli random variable indicating i^{th} subject has a periodontal disease, p_i is a probability of i^{th} subject has a periodontal disease. $\beta_{age,j}$ $(j = 2, 3, 4)$ is a regression coefficient pertaining to a subject's age category, and $\beta_{age,1}$ is set to 0 to indicate the reference level. $X_{i,age}^j$ is a dummy variable such that $X_{i,age}^j$ = Indicator(i^{th} subject's age category = j), and other regression coefficients and dummy variables are defined similarly. For example, β_{smoker} is a regression coefficient pertaining to a subject's smoking status, and $X_{i,smoker}$ is 1 if i^{th} subject is a smoker, and 0 if not. Then, β is a vector of these regression coefficients, and X_i is a vector of these dummy variables.

(d)

Full model under the alternative hypothesis has the only difference in the mean assumption, given by

$$\begin{aligned} logit\left(E(Y_i|X_i)\right) = log\left(\frac{p_i}{1-p_i}\right) =& X_i^T\beta \\ =& \beta_0 + \sum_{j=2}^{4}\beta_{age}^j X_{i,age}^j + \beta_{smoker}X_{i,smoker} + \beta_{hinsurance}X_{i,hinsurance} \\ &+ \beta_{female}X_{i,female} + \beta_{diab}X_{i,diab} + \beta_{smoker*diab}X_{i,smoker}X_{i,diab} \\ &+ \sum_{j=2}^{4}\beta_{age*diab}^j X_{i,age}^j X_{i,diab} + \beta_{hinsurance*diab}X_{i,hinsurance}X_{i,diab} \\ &+ \beta_{female*diab}X_{i,female}X_{i,diab} \end{aligned}$$

where $\beta_{age*diab}^j$, $\beta_{hinsurance*diab}$, $\beta_{female*diab}$ are newly added interaction effects not present in the null model in (c).

Assuming the full model is true, let consider the following hypothesis

$$H_0 : \beta_{age*diab}^j = 0 \ (j = 2, 3, 4),\ \beta_{hinsurance*diab} = 0,\ \beta_{female*diab} = 0 \ \text{ vs } \ H_1 : \text{not } H_0.$$

Under H_0, the full model reduces to the model in (c). Therefore, by comparing two models, we can conduct goodness-of-fit test based on the LRT, the score test, and the Wald test.

Note that the null hypothesis can be written by

$$H_0 : L\beta = \begin{bmatrix} \beta_{age*diab}^2 \\ \beta_{age*diab}^3 \\ \beta_{age*diab}^4 \\ \beta_{hinsurance*diab} \\ \beta_{female*diab} \end{bmatrix} = 0$$

where L is a 5 by 14 contrast matrix utilized for the test such that

$$L = \begin{array}{c} \begin{array}{cccccccccc} \beta_0 & \beta^2_{age} & \dots & \beta_{smoker*diab} & \beta^2_{age*diab} & \beta^3_{age*diab} & \beta^4_{age*diab} & \beta_{hinsurance*diab} & \beta_{female*diab} \end{array} \\ \begin{pmatrix} 0 & 0 & \dots & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

The results of likelihood ratio, score and Wald tests for testing the null hypothesis are given in Table 7.

Table 7: Test results

Test	**Construction**	**Asymptotic Distribution**	**Statistic**	**P value**
LRT	$2\left(l(\hat{\beta}) - l(\hat{\beta}^0)\right)$	$\chi^2(5)$	7.152	0.2095
Score	$\dot{l}(\hat{\beta}^0)^T[-\ddot{l}(\hat{\beta}^0)]^{-1}\dot{l}(\hat{\beta}^0)$	$\chi^2(5)$	7.131	0.2111
Wald	$(L\hat{\beta})^T[L\widehat{\Sigma}_{\hat{\beta}}L^T]^{-1}(L\hat{\beta})$	$\chi^2(5)$	7.122	0.2117

l is a log likelihood of the full model, $\hat{\beta}$, $\hat{\beta}^0$ are MLE under full model and reduced model (under H_0) respectively, and $\widehat{\Sigma}_{\hat{\beta}}$ is an asymptotic covariance matrix of $\hat{\beta}$. All three test statistics have the same asymptotic distribution of χ^2 with the degree of freedom equals 5. Their statistics are similar, and p values are also similar ($0.20 \sim 0.21$). None of the tests rejects H_0 at the $\alpha = 0.05$ significance level. Therefore, we can say that the null model explains the data well enough, comparable to the full model. In conclusion, the new four two-way interactions do not improve the model fit significantly, so it is possible to remove them from the model.

(e)

Table 8: Estimation Result of Logistic regression model

Parameter	Description	Estimate	Standard Error	Wald 95% Confidence Limits		Z statistic	P value
β_0	Intercept	-0.871	0.095	-1.058	-0.687	-9.213	<.0001
β^2_{age}	Age 35-44	0.831	0.100	0.636	1.027	8.332	<.0001
β^3_{age}	Age 45-54	1.380	0.094	1.196	1.566	14.618	<.0001
β^4_{age}	Age 55+	1.696	0.103	1.496	1.898	16.520	<.0001
$\beta_{hinsurance}$	Health Insurance	-0.475	0.062	-0.596	-0.355	-7.717	<.0001
β_{female}	Female	-0.615	0.064	-0.741	-0.490	-9.632	<.0001
β_{smoker}	Smoker	0.476	0.100	0.281	0.672	4.769	<.0001
β_{diab}	Diabetes	0.338	0.083	0.176	0.500	4.086	<.0001
$\beta_{smoker*diab}$	Smoker & Diabetes	-0.292	0.127	-0.540	-0.043	-2.302	0.021

The maximum likelihood estimates, standard errors, Wald 95% confidence limits, Z statistics, and P values are given in the Table 8. We can see that every parameter estimates are statistically significant.

Covariate-adjusted odds ratio for periodontal disease and diabetes for **smokers** and its 95% confi-

dence intervals is given from the relationship that

$$\begin{aligned}
&\text{OR(Diabetes vs Not| Smoker)} \\
&= \frac{P(\text{Disease} \mid \text{Diabetes, Smoker})/P(\text{Normal} \mid \text{Diabetes, Smoker})}{P(\text{Disease} \mid \text{Not Diabetes, Smoker})/P(\text{Normal} \mid \text{Not Diabetes, Smoker})} \\
&= \exp\left(\text{logit}(P(\text{Disease} \mid \text{Diabetes, Smoker})) - \text{logit}(P(\text{Disease} \mid \text{Not Diabetes, Smoker}))\right) \\
&= \exp(\beta_{diab} + \beta_{smoker*diab}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\widehat{OR}(\text{Diabetes vs Not} \mid \text{Smoker}) = 1.0470 \\
&95\% \text{ Wald C.I} : OR \in (0.8582, 1.2774).
\end{aligned}$$

Similarly, covariate-adjusted odds ratio for periodontal disease and diabetes for **non-smokers** and its 95% confidence intervals is given from the relationship that

$$\begin{aligned}
&\text{OR(Diabetes vs Not| Non-Smoker)} \\
&= \frac{P(\text{Disease} \mid \text{Diabetes, Non-Smoker})/P(\text{Normal} \mid \text{Diabetes, Non-Smoker})}{P(\text{Disease} \mid \text{Not Diabetes, Non-Smoker})/P(\text{Normal} \mid \text{Not Diabetes, Non-Smoker})} \\
&= \exp\left(\text{logit}(P(\text{Disease} \mid \text{Diabetes, Non-Smoker})) - \text{logit}(P(\text{Disease} \mid \text{Not Diabetes, Non-Smoker}))\right) \\
&= \exp(\beta_{diab}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\widehat{OR}(\text{Diabetes vs Not} \mid \text{Non-Smoker}) = 1.4018 \\
&95\% \text{ Wald C.I} : OR \in (1.1921, 1.6483).
\end{aligned}$$

Interpretation: Among smokers, the odds of having periodontal disease when a person suffers from diabetes is 1.0470 times the odds of having periodontal disease when a person does not suffer from diabetes. Thus, among smokers, having diabetes does not affect the probability of having periodontal disease. However, among non-smokers, the odds of having the disease when a person suffers from diabetes is 1.4018 times the odds of having the disease when a person does not suffer from diabetes. Therefore, among non-smokers, having diabetes is highly related with having periodontal disease.

(f)

The linear model for the empirical logits is given by

$$\begin{aligned}
f_i = \text{logit}(p_i) =& \log\left(\frac{p_i}{1-p_i}\right) \\
=& X_i^T\beta + \varepsilon_i \\
=& \beta_0 + \sum_{j=2}^{4} \beta_{age,j} X_{i,age}^j + \beta_{smoker} X_{i,smoker} + \beta_{hinsurance} X_{i,hinsurance} \\
&+ \beta_{female} X_{i,female} + \beta_{diab} X_{i,diab} + \beta_{smoker*diab} X_{i,smoker} X_{i,diab} + \varepsilon_i
\end{aligned}$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$, $i = 1, \ldots, 64$ is a measurement error. Regression coefficients vector (β) and covariate vector (X_i) is defined previously in (c). This model is a linear regression model with new outcome f_i under heteroscedasticity assumption among errors. To fit the model, we perform weighted least squares estimation to deal with heteroscedasticity among errors using inverse variance weights based on var(f_i). To estimate var(f_i), we start from the natural modeling

$$y_i \sim B(n_i, \pi_i)$$

where π_i is a probability of a person having periodontal disease whose covariate combination is i^{th} combination (denoted by X_i). Then $Ey_i = n_i\pi_i$, $Var(y_i) = n_i\pi_i(1-\pi_i)$. Let $g(y) = log(\frac{y}{n_i - y})$, then

$$\begin{aligned}
\text{var}(f_i) &= \text{var}\left(log\left(\frac{y_i}{n_i - y_i}\right)\right) \\
&= \text{var}\left(g(y_i)\right) \\
&\approx \dot{g}(Ey_i)^2 Var(y_i) \\
&= \frac{1}{n_i\pi_i(1-\pi_i)} \\
&\approx \frac{1}{n_i p_i(1-p_i)}.
\end{aligned}$$

Therefore, weight for each i^{th} data is given by $var(f_i)^{-1} = n_i p_i(1-p_i)$. Now, the vector form of the model is given by

$$F = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{64} \end{bmatrix} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_{64}^T \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{smoker*diab} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{64} \end{bmatrix} = X\beta + \epsilon$$

where F is a vector of empirical logits, X is a design matrix, $\epsilon \sim N_{64}(0, \Sigma)$ is a vector of ε_i, $\Sigma = diag((n_i p_i(1-p_i))^{-1})$, and $W = \Sigma^{-1}$ is a inverse variance weights matrix. Then, weighted least squares estimator of β is given by

$$\begin{aligned}
\widehat{\beta} &= (X^T W X)^{-1} X^T W F \\
&= \left(\sum_{i=1}^{64} n_i p_i(1-p_i) X_i X_i^T\right)^{-1} \left(\sum_{i=1}^{64} n_i p_i(1-p_i) f_i X_i\right).
\end{aligned}$$

(g)

From the weighted least squares estimator of β, given by $\widehat{\beta}$, we can compute expected value of periodontal disease count for each group (denoted by $\widehat{y}_i$) as follows:

$$\begin{aligned}
\widehat{f}_i &= X_i^T \widehat{\beta} \\
\widehat{p}_i &= \frac{e^{\widehat{f}_i}}{1 + e^{\widehat{f}_i}} \\
\widehat{y}_i &= n_i \widehat{p}_i = \frac{n_i e^{\widehat{f}_i}}{1 + e^{\widehat{f}_i}}.
\end{aligned}$$

Then, the residual chi-squared goodness-of-fit test for the model is given by

$$\chi^2 = \sum_{i=1}^{64} \frac{(y_i - \widehat{y}_i)^2}{\widehat{y}_i} \sim \chi^2(3)$$

where the degree of freedom is computed by $(4-1)(2-1)(2-1)(2-1)(2-1) = 3$. The test statistic is 29.548, and p value is 1.72e-06, which implies that the model fitting does not explain the data well if we assume weighted least squares linear model for the empirical logits.

(h)

Table 9: Estimation Result of Weighted Least Squares linear model

Parameter	Description	Estimate	Standard Error	Wald 95% Confidence Limits		T statistic (DF=55)	P value
β_0	Intercept	-0.859	0.097	-1.054	-0.664	-8.821	<.0001
β_{age}^2	Age 35-44	0.813	0.102	0.608	1.017	7.956	<.0001
β_{age}^3	Age 45-54	1.348	0.097	1.154	1.541	13.962	<.0001
β_{age}^4	Age 55+	1.666	0.105	1.456	1.875	15.917	<.0001
$\beta_{hinsurance}$	Health Insurance	-0.478	0.063	-0.605	-0.352	-7.554	<.0001
β_{female}	Female	-0.602	0.066	-0.734	-0.469	-9.114	<.0001
β_{smoker}	Smoker	0.462	0.104	0.254	0.67	4.453	<.0001
β_{diab}	Diabetes	0.347	0.085	0.177	0.518	4.083	<.0001
$\beta_{smoker*diab}$	Smoker & Diabetes	-0.28	0.131	-0.543	-0.017	-2.131	0.038

Based on the weighted least squares estimation, the parameter estimates, standard errors, Wald 95% confidence limits, T statistics, and P values are given in the Table 9. We can see that every parameter estimates are statistically significant.

Covariate-adjusted odds ratio for periodontal disease and diabetes for **smokers** and its 95% confidence intervals is given from the relationship that

$$\begin{aligned}
&\text{OR(Diabetes vs Not| Smoker)} \\
&= \frac{P(\text{Disease} \mid \text{Diabetes, Smoker})/P(\text{Normal} \mid \text{Diabetes, Smoker})}{P(\text{Disease} \mid \text{Not Diabetes, Smoker})/P(\text{Normal} \mid \text{Not Diabetes, Smoker})} \\
&= \exp\left(\text{logit}(P(\text{Disease} \mid \text{Diabetes, Smoker})) - \text{logit}(P(\text{Disease} \mid \text{Not Diabetes, Smoker}))\right) \\
&= \exp(\beta_{diab} + \beta_{smoker*diab}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\widehat{OR}(\text{Diabetes vs Not} \mid \text{Smoker}) = 1.0700 \\
&95\% \text{ Wald C.I} : OR \in (0.8831, 1.2964).
\end{aligned}$$

Similarly, covariate-adjusted odds ratio for periodontal disease and diabetes for **non-smokers** and its 95% confidence intervals is given from the relationship that

$$\begin{aligned}
&\text{OR(Diabetes vs Not| Non-Smoker)} \\
&= \frac{P(\text{Disease} \mid \text{Diabetes, Non-Smoker})/P(\text{Normal} \mid \text{Diabetes, Non-Smoker})}{P(\text{Disease} \mid \text{Not Diabetes, Non-Smoker})/P(\text{Normal} \mid \text{Not Diabetes, Non-Smoker})} \\
&= \exp\left(\text{logit}(P(\text{Disease} \mid \text{Diabetes, Non-Smoker})) - \text{logit}(P(\text{Disease} \mid \text{Not Diabetes, Non-Smoker}))\right) \\
&= \exp(\beta_{diab}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\widehat{OR}(\text{Diabetes vs Not} \mid \text{Non-Smoker}) = 1.4153 \\
&95\% \text{ Wald C.I} : OR \in (1.2126, 1.6518).
\end{aligned}$$

Interpretation: Among smokers, the odds of having periodontal disease when a person suffers from diabetes is 1.0700 times the odds of having periodontal disease when a person does not suffer from diabetes. Thus, among smokers, having diabetes does not affect the probability of having periodontal disease. However, among non-smokers, the odds of having the disease when a person suffers from

diabetes is 1.4153 times the odds of having the disease when a person does not suffer from diabetes. Therefore, among non-smokers, having diabetes is highly related with having periodontal disease.

(i)

The maximum likelihood estimation for logistic regression has a relative strength that since the distribution is in the exponential family, we can obtain the asymptotic normality of maximum likelihood estimator. We utilize it for the testing hypothesis regarding the regression coefficients, as done in (d). In contrast, the weighted least square estimation requires proper choice of weights to expand asymptotic theory, and if the choice was wrong, the accuracy and efficiency of the estimation could be hampered.

On the other hand, the weighted least squares estimation is relatively easier than the maximum likelihood estimation. As shown in (f), we have a closed form of the weighted least squares estimator. However, we have to loop over iterative updating scheme to obtain maximum likelihood estimator.

3. Clinical trial design to test effectiveness of a novel blood pressure management based on a pilot study

(a)

The data is comprised of N = 256 subjects with various visits (0 $\sim$ 12 months), and there are 28 subjects who only visited at baseline. Removing these subjects, the pilot data is a longitudinal data of N = 228 subjects.

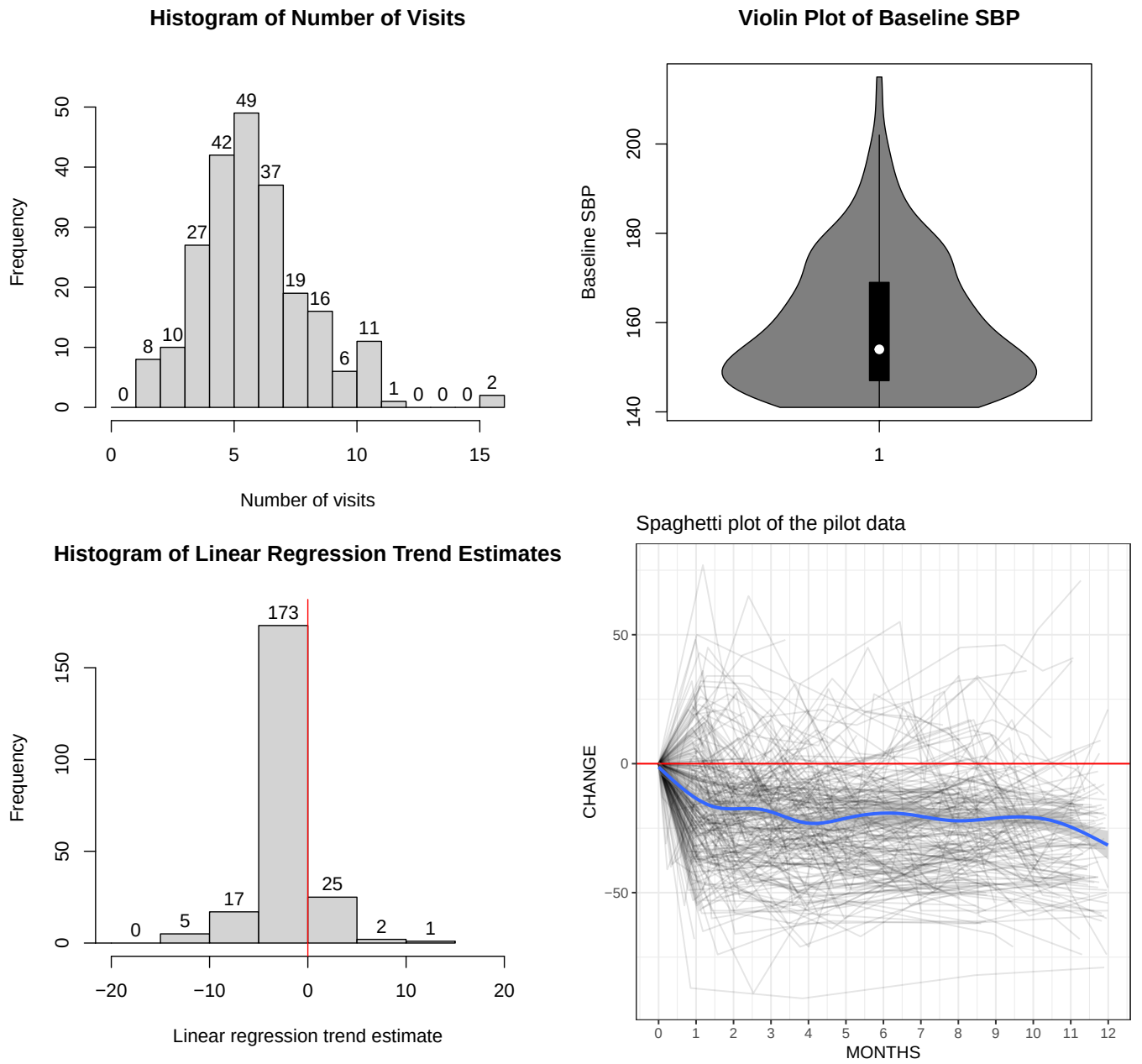

Figure 2: Histogram of number of visits (Top, Left), Violin plot of baseline SBP (Top, Right), Histogram of linear regression trend estimates (Bottom, Left), Spaghetti plot with GAM smoothing curve (Bottom, Right)

Figure 2 presents various descriptive statistics of the pilot data. The average and the most frequent number of visits was 6 times (sd: 2.31), and the distribution was bell shaped around the average. The average baseline SBP was 159.4 (sd: 14.82), and most of the baseline SBP ranged around 140 $\sim$ 160, with a long right tale. To check the heuristic trend of the SBP change, the linear regression model *SBP change by Month* is fitted for every subject. The bottom left panel of Figure 2 is the histogram of the regression coefficient estimates. 87.7% of the estimates were negative, suggesting that the SBP was decreasing for most of the subjects. Finally, the spaghetti plot and its Generalized Additive Model smoothing curve suggests that the SBP was decreasing for most of the subjects, and the trend of mean SBP change was approximately linear.

(b)

(i) The linear mixed model with a random intercept is given by

$$Y_{ij} = \beta_0 + \beta_{base}B_i + \beta_{month}t_{ij} + \beta_{inter}B_i t_{ij} + b_i + \varepsilon_{ij}, \quad i = 1, \ldots, 228, \quad j = 1, \ldots, n_i \tag{1}$$

where Y_{ij} is subject i's SBP change from the baseline at j^{th} visit, n_i is subject i's number of visits, B_i is the standardized baseline SBP of subject i, t_{ij} is the number of months from baseline at j^{th} measurement time of subject i, $\boldsymbol{\beta} = (\beta_0, \beta_{base}, \beta_{month}, \beta_{inter})'$ are fixed effects, $b_i \sim N(0, \sigma_b^2)$ are random intercepts specific to subjects, and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ are the normal random errors. Further, we assume b_i are identically and independently distributed among subjects, ε_{ij} are identically and independently distributed within subjects, and b_i and ε_{ij} are independent.

The linear mixed model with a random intercept and slope is given by

$$Y_{ij} = \beta_0 + \beta_{base}B_i + \beta_{month}t_{ij} + \beta_{inter}B_i t_{ij} + b_{i0} + b_{i1}t_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, 228, \quad j = 1, \ldots, n_i \tag{2}$$

where $\boldsymbol{b_i} = (b_{i0}, b_{i1})' \sim N(0, \boldsymbol{D})$ are random effects specific to subjects with unstructured positive definite covariance matrix $\boldsymbol{D}$, and all the other covariates and coefficients are the same with the previous model.

To test whether the random slope is needed in the model, we test

$$H_0 : \boldsymbol{D}_{2,2} = 0 \text{ vs } H_1 : \boldsymbol{D}_{2,2} > 0,$$

where $\boldsymbol{D}_{2,2}$ is a (2,2) element of $\boldsymbol{D}$. The LRT statistic is known to follow 50-50 mixture of χ_1^2 and χ_2^2 distributions, and the critical value at significance level $\alpha = 0.05$ is 5.14. After fitting the models, the LRT statistic is given by 2.96, and thus we cannot reject no random slope hypothesis. Indeed, AIC criteria also supports the random intercept model (Random Intercept & Slope model: 10360.2, Random Intercept model: 10359.2). In conclusion, the random slope is not needed in the model.

(ii) The random intercept model fitting result suggests that there is no interaction effect between baseline SBP and time (T statistic = -0.77, DF (Kenward and Roger) = 1063, p-value = 0.4388). Furthermore, AIC criteria also supports the model without interaction effect (With Interaction: 10359.2, Without Interaction: 10353.0). Thus, we select the final model that has a linear trajectory over time, adjusted for baseline, without interaction, and includes a random intercept only, given by

$$Y_{ij} = \beta_0 + \beta_{base}B_i + \beta_{month}t_{ij} + b_i + \varepsilon_{ij}, \quad i = 1, \ldots, 228, \quad j = 1, \ldots, n_i$$

where $\boldsymbol{\beta} = (\beta_0, \beta_{base}, \beta_{month})'$ are fixed effects, $b_i \sim N(0, \sigma_b^2)$ are random intercepts specific to subjects, and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ are the normal random errors. Restricted maximum likelihood estimation is performed for the parameter estimation, and degrees of freedom is computed based on Kenward and Roger's method. The estimation result is given in Table 10.

Table 10: Estimation Result of the Linear Mixed Model with a random intercept by REML

Parameter	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**DF**	**T statistic**	**P value**
β_0	-15.6421	1.0841	-17.7720	-13.5123	515	-14.43	<.0001
β_{base}	-4.2375	0.2948	-4.8184	-3.6566	232	-14.37	<.0001
β_{month}	-1.1263	0.1340	-1.3892	-0.8634	1046	-8.41	<.0001
σ_b^2	118.02	-	-	-	-	-	-
σ_e^2	240.88	-	-	-	-	-	-

All regression parameters are statistically significant under significance level $\alpha = 0.05$. Negative baseline effect estimate suggests that a subject with a higher baseline SBP would exhibit a larger change

from baseline on average, as expected. Negative month effect estimate suggests that a subject's SBP decreased by 1.12 on average every month, when this person's baseline SBP was on the sample average. The estimated variance of random intercept is 118.02, and the estimated error variance is 240.88.

(iii)

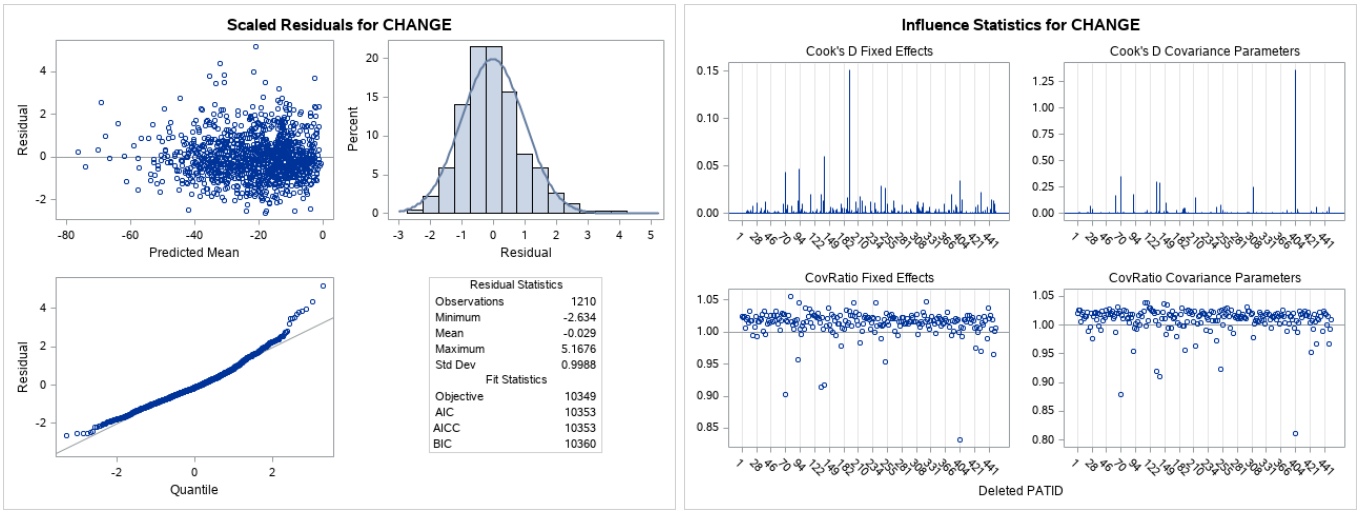

Figure 3: Scaled Residuals Diagnostic Plots (Left), Influence Statistics Plots (Right)

From the diagnostic plots (Figure 3), the scaled residuals are randomly distributed, and they do not show any systematic structure. No serious violation of the normality assumption was found according to the QQplot. Only one data are identified to have large Cook's Distance for fixed effects, but not on the extreme scale (~ 0.05 for most of the subjects). Therefore, the model was fitted well.

(c)

Assume we have total N subjects to randomize. Based on (b), we consider the following model as the subjects' SBP change by time, having different trajectory by treatment group,

$$Y_{ij} = (\beta_0^{(z_i)} + b_i) + \beta_{base}^{(z_i)} B_i + \beta_{month}^{(z_i)} t_j + \varepsilon_{ij}, \ i = 1, \ldots, N, \ j = 1, 2, 3 \tag{3}$$

where z_i is an indicator such that $z_i = 1$ when i^{th} subject is randomized to IPTM, and $z_i = 0$ when i^{th} subject is randomized to ICBM. There are $\frac{N}{2}$ ones and $\frac{N}{2}$ zeros among z_i. $\beta_0^{(z_i)}$, $\beta_{base}^{(z_i)}$, and $\beta_{month}^{(z_i)}$ are the treatment group-specific fixed intercept, baseline SBP effect, and slope effects, respectively. b_i is a subject-specific random intercept such that $b_i \sim N(0, \sigma_b^2)$. ε_{ij} are the within-subject random errors at j^{th} visits such that $\varepsilon_{ij} \sim N(0, \sigma_e^2)$. We assume iid random intercept and iid random error, and all the random terms are independent. t_j is the number of months from baseline at j^{th} visit, where $t_1 = 2$, $t_2 = 4$, and $t_3 = 8$. B_i is the standardized baseline SBP.

Let $\delta = \beta_{month}^{(1)} - \beta_{month}^{(0)}$, which measures difference in the SBP change trajectories of IPTM and that of ICBM over the 2 to 8 months post-baseline. Under the null hypothesis of no treatment difference, $H_0 : \delta = 0$. We are interested in the sample size (N) that gives 90% power when testing the null hypothesis at significance level $\alpha = 0.05$.

From collaborators' hypothesis that the patient receiving IPTM will have an additional SBP reduction of 4.0 mmHg on average at 8 months, and the reduction will increase linearly up to 4.0 mmHg at 8 months, we have the following relationship,

$$E(Y_{ij}^{IPTM} - Y_{ij}^{ICBM}) = (\beta_0^{(1)} - \beta_0^{(0)}) + (\beta_{base}^{(1)} - \beta_{base}^{(0)}) B_i + (\beta_{month}^{(1)} - \beta_{month}^{(0)}) t_j = -\frac{1}{2} t_j$$

for every $t_j \in [0, 8]$ and B_i. It gives $\beta_0^{(1)} = \beta_0^{(0)}$, $\beta_{base}^{(1)} = \beta_{base}^{(0)}$, and $\beta_{month}^{(1)} = \beta_{month}^{(0)} - \frac{1}{2}$. We assume the true parameters in the model are given by the fitted model in part(b), including variance of random intercept and random error.

Finally, we compute the power of the test at true value $\delta = \beta_{month}^{(1)} - \beta_{month}^{(0)} = -\frac{1}{2}$. For this, we use empirical power instead. Assume we generated M dataset based on the model. For each dataset, we fit the model, and perform the test of no treatment difference, $H_0 : \delta = 0$ at significance level $\alpha = 0.05$. Then the empirical power is the ratio of the number tests that rejected H_0 divided by M. We search on the the sample size N to find the sample size that achieves 90% empirical power.

The implementation steps are as follow:

1. Set candidate sample size N
2. Generate $\frac{N}{2}$ subjects' data based on ICBM mean change model, given by

$$Y_{ij} = -15.6481 + b_i - 4.2375B_i - 1.1263t_j + \varepsilon_{ij}$$

   where b_i, ε_{ij} are random sample from $N(0, 118.02)$, $N(0, 240.88)$, respectively, and B_i is a random sample from the standardized baseline SBP distribution at the pilot study.
3. Generate $\frac{N}{2}$ subjects' data based on IPTM mean change model, given by

$$Y_{i'j} = -15.6481 + b_{i'} - 4.2375B_{i'} - 1.6263t_j + \varepsilon_{i'j}$$

   where $b_{i'}$, $\varepsilon_{i'j}$ are random sample from $N(0, 118.02)$, $N(0, 240.88)$, respectively, and $B_{i'}$ is a random sample from the standardized baseline SBP distribution at the pilot study.
4. Based on total N subjects' data, fit the model

$$Y_{ij} = (\beta_0^{(z_i)} + b_i) + \beta_{base}^{(z_i)} B_i + \beta_{month}^{(z_i)} t_j + \varepsilon_{ij}, \ i = 1, \ldots, N, \ j = 1, 2, 3,$$

   and test $H_0 : \delta = \beta_{month}^{(1)} - \beta_{month}^{(0)} = 0$ at significance level $\alpha = 0.05$.
5. Repeat Step 2∼4 $M = 1000$ times. Empirical power is the ratio of rejected tests among M trials.
6. Adjust N to achieve 90% empirical power.

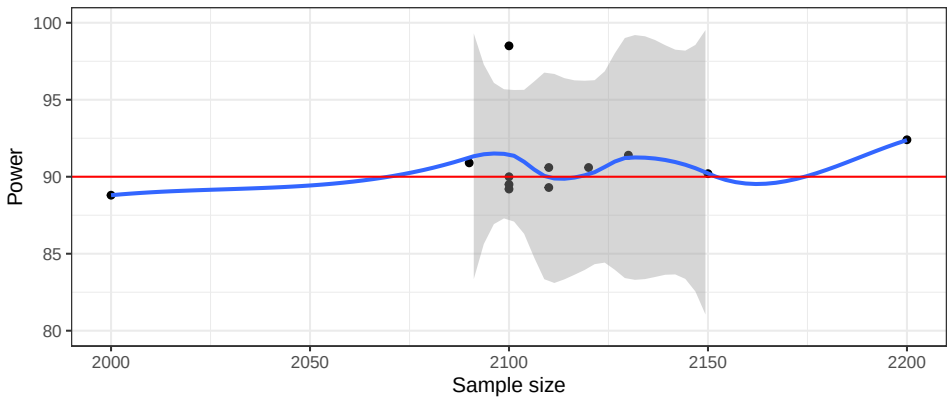

Figure 4: Sample size by Empirical Power

The simulation result is given in Figure 4. With degree 1 LOESS curve, we expect that sample size of 2100 would bring 90% power to detect a difference in the SBP change trajectory of IPTM and that of ICBM.

(d)

First, since the drop out is related with the unobserved data, the missing mechanism is not missing at random. It can highly affect the accuracy and efficiency of the estimation. To verify this, assume the missing mechanism is

$$\text{logit}(P(R_{ij} = 1)) = 2 - (1 - z_i) - 1(t_j \geq 4) * 1(y_{ij} > -2.0),\ j = 1, 2, 3,$$

where $1(A)$ is an indicator of whether event A occurred or not. The above mechanism comes from the assumption that drop out is more likely in subjects receiving ICBM ($z_i = 0$) and who fails to achieve a 2.0 mmHg or greater decrease in SBP by 4 months. By including the missing mechanism in the data generation step in (c), the same simulation can be conducted. $N = 2100$ and $M = 1000$ were selected for the simulation. The simulation results are in the Table 11.

Table 11: Estimation Result Comparison based on simulation

Effect	**True Model**		**Complete Data**		**Missing Data**	
	Estimate	Std Err	Mean	Std Dev	Mean	Std Dev
$\beta_0^{(0)}$	-15.6421	1.0841	-15.6844	0.6592	-15.4458	0.8353
$\beta_{base}^{(0)}$	-4.2375	0.2948	-4.2339	0.1511	-4.1217	0.1977
$\beta_{month}^{(0)}$	-1.1263	0.1340	-1.1221	0.1074	-1.3595	0.1708
Δ_{trt}	0	-	0.0716	0.9518	-0.1313	1.1241
Δ_{base}	0	-	-0.0016	0.2195	-0.0430	0.2654
Δ_{month}	-0.5	-	-0.5087	0.1522	-0.3704	0.2088

Here, $\Delta_{trt} = \beta_0^{(1)} - \beta_0^{(0)}$, $\Delta_{base} = \beta_{base}^{(1)} - \beta_{base}^{(0)}$, and $\Delta_{month} = \beta_{month}^{(1)} - \beta_{month}^{(0)}$ are differences of the parameters of the two SBP change trajectories (IPTM trajectory from ICBM trajectory). When the LMM model fitting was done with the complete data, the parameter estimation was done with high accuracy. The achieved empirical power was 90.90%, as shown in (c). However, when the LMM model fitting was done with the incomplete data with missingness, the parameter estimation was inaccurate for the parameter differences ($\Delta_{trt}, \Delta_{base}, \Delta_{month}$). The main parameter of interest, Δ_{month} was estimated by -0.37, whereas the true value is -0.5. The achieved empirical power was 39.4%, indicating that NMAR mechanism seriously hampered the estimation and inference quality.

(e)

The same simulation as (d) was performed, assuming the missing mechanism of

$$\text{logit}(P(R_{ij} = 1)) = 2 - (1 - z_i) - 1(t_j \geq 4) * 1(y_{ij} > -2.5),\ j = 1, 2, 3.$$

The only difference from (d) was the number inside of the indicator in the missing mechanism (-2.0 to -2.5). The simulation results are in the Table 12.

The results are very similar with the results in (d). Changing the missing mechanism slightly did not change the result dramatically because on average, a patient with average baseline SBP would exhibit a 20.1476 mmHg decrease in SBP from the baseline at month 4. Therefore, to see significant difference, we may assume the missing mechanism of $\text{logit}(P(R_{ij} = 1)) = 2 - (1 - z_i) - 1(t_j \geq 4) * 1(y_{ij} > -20)$.

(f)

By including 10% non-compliant rate in the data generation step in (c), the same simulation can be conducted. $N = 2100$ and $M = 1000$ were selected for the simulation. The simulation results are in Table 13.

Table 12: Estimation Result Comparison based on simulation

Effect	**True Model**		**Complete Data**		**Missing Data**	
	Estimate	Std Err	Mean	Std Dev	Mean	Std Dev
$\beta_0^{(0)}$	-15.6421	1.0841	-15.6837	0.6586	-15.4337	0.8335
$\beta_{base}^{(0)}$	-4.2375	0.2948	-4.2359	0.1469	-4.1155	0.1932
$\beta_{month}^{(0)}$	-1.1263	0.1340	-1.1221	0.1074	-1.3659	0.1709
Δ_{trt}	0	-	0.0712	0.9514	-0.1360	1.1216
Δ_{base}	0	-	-0.0045	0.2063	-0.0534	0.2561
Δ_{month}	-0.5	-	-0.5087	0.1522	-0.3683	0.2085

Table 13: Estimation Result Comparison based on simulation

Effect	**True Model**		**Complete Data**		**Non-compliant Data**	
	Estimate	Std Err	Mean	Std Dev	Mean	Std Dev
$\beta_0^{(0)}$	-15.6421	1.0841	-15.6844	0.6592	-14.1021	0.6620
$\beta_{base}^{(0)}$	-4.2375	0.2948	-4.2339	0.1511	-3.8133	0.1872
$\beta_{month}^{(0)}$	-1.1263	0.1340	-1.1221	0.1074	-1.0134	0.1063
Δ_{trt}	0	-	0.0716	0.9518	0.0221	0.8880
Δ_{base}	0	-	-0.0016	0.2195	-0.0024	0.2707
Δ_{month}	-0.5	-	-0.5087	0.1522	-0.4526	0.1465

The non-compliant data largely affects the accuracy of estimation as it worsens the reliability of the data. However, the main parameter of interest Δ_{month} was estimated with relatively high accuracy (estimated: -0.4526, true: -0.5), and indeed the power under the non-compliant assumption was computed by 86.2%, which is fairly high. Therefore, we can conclude that although non-compliance may introduce biases to the parameter estimation, the main parameter of interest can be estimated well with a high power, comparable to the power under the complete data.

4. Randomized clinical trial for a novel intervention to lower hemoglobin A1c

(a) Table 14 represents the number of sample, mean, standard deviation, minimum, and maximum value of baseline A1C by intervention group and overall sample, for parents (up) and children (bottom), respectively. The ANOVA P Value in the last column is computed from the ANOVA model $A1C_i = \alpha + \beta X_i + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, where $A1C_i$ $(i = 1, \ldots, 188)$ is i^{th} parent (or child)'s Baseline A1C, X_i is 1 if this parent (or child) is in the intervention group, and 0 if not. The ANOVA P value comes from the test of no group effect, $H_0 : \beta = 0$. We can see that the distributions of baseline A1C are similar over groups, for both parents and children. Insignificant ANOVA P Values indicate that for parents (or children), there is no evidence to say that baseline A1C varied by group.

Table 14: Descriptive Statistics of **Baseline** A1C values for Intervention, Control, and Overall samples

Baseline A1C		N	Mean	SD	Min	Max	ANOVA P Value
Parent	Intervention	94	5.53	0.39	4.9	6.4	0.656
	Control	94	5.56	0.43	4.7	6.6	
	Overall	188	5.55	0.41	4.7	6.6	
Child	Intervention	94	5.52	0.36	4.9	7.2	0.217
	Control	94	5.45	0.42	4.8	6.4	
	Overall	188	5.49	0.39	4.8	7.2	

Table 15 represents the same descriptive statistics as Table 14, but for follow-up A1C. We can see that the mean of A1C decreased compared to the baseline, but the amount was larger for the intervention group than the control group. For example, for parents, mean A1C decrease was 0.16 (5.53 $\rightarrow$ 5.37) in the intervention group, while it was 0.10 (5.56 $\rightarrow$ 5.46) in the control group. In addition, the amount of decrease was larger for children compared to parents. Decreased ANOVA P values also indicate that for parents (or children), we can say that follow-up A1C varied by group, if we use a lenient significance level. Table 16 is a contingency table of (Pre) Diabetes status (A1C $\geq$ 5.7%)

Table 15: Descriptive Statistics of **Follow-up** A1C values for Intervention, Control, and Overall samples

Follow-up A1C		N	Mean	SD	Min	Max	ANOVA P Value
Parent	Intervention	94	5.37	0.38	4.6	6.3	0.135
	Control	94	5.46	0.41	4.7	6.5	
	Overall	188	5.41	0.40	4.6	6.5	
Child	Intervention	94	5.21	0.38	4.6	6.7	0.116
	Control	94	5.29	0.39	4.5	6.1	
	Overall	188	5.25	0.39	4.5	6.7	

by intervention group, for parents (up) and children (bottom), respectively. χ^2 independence test of the group and diabetes status are given in the last column. We can see that the count distributions are similar over groups, for both parents and children. Insignificant χ^2 P Values indicate that for parents (or children), there is no evidence to say that the proportion of (pre) diabetes subjects at baseline varied by group.

Table 17 represents the same descriptive statistics as Table 16, but for follow-up counts. We can see that the number of diabetes subjects decreased compared to the baseline (except parents control group), but the amount was larger for the intervention group than the control group. For example, for parents, the number of diabetes subjects decreased by 15 (34 $\rightarrow$ 19) in the intervention group, while it increased by 1 (31 $\rightarrow$ 32) in the control group. In addition, the count decrease was larger for

Table 16: Contingency table of (Pre) Diabetes status by intervention group at **Baseline**

Baseline Count		(Pre) Diabetes	Normal	Total	χ^2 test P Value
Parent	Intervention	34	60	94	0.7591
	Control	31	63	94	
	Overall	65	123	188	
Child	Intervention	31	63	94	0.8757
	Control	29	65	94	
	Overall	60	128	188	

children compared to parents. Decreased and significant χ^2 P values also indicate that for parents (or children), we can say that the proportion of (pre) diabetes subjects at follow-up varied by group.

Table 17: Contingency table of (Pre) Diabetes status by intervention group at **Follow-up**

Follow-up Count		(Pre) Diabetes	Normal	Total	χ^2 test P Value
Parent	Intervention	19	75	94	0.0490
	Control	32	62	94	
	Overall	51	137	188	
Child	Intervention	11	83	94	0.0807
	Control	21	73	94	
	Overall	32	156	188	

(b) The model is given by

$$Y_{ij} = \alpha_j + \beta_j B_{ij} + \gamma_j G_i + \varepsilon_{ij} \ (i = 1, \ldots, 188, \ j = 1, 2)$$

where i indicates the ID of parent-child pair, and $j = 1$ indicates parent, while $j = 2$ indicates child. Y_{i1}, Y_{i2} are A1C change of i^{th} parent and child, respectively. B_{i1}, B_{i2} are baseline A1C value of i^{th} parent and child, respectively. G_i is an indicator variable such that $G_i = 1$ if i^{th} pair is in the intervention group, $G_i = 0$ if not. α_j is a intercept, β_j is an effect of baseline, and γ_j is an intervention effect, and all coefficients vary by whether the subject is a parent or a child. Finally, ε_{ij} is a measurement error such that $\epsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})^T \overset{iid}{\sim} N_2(0, \Sigma), \ i = 1, \ldots, 188$, where Σ is an unstructured positive definite matrix. Since we can expect that the data from a parent and their child are correlated, we introduce correlated measurement error to implement it. Thus, this model is a General Linear Model.

The assumptions under this model are (1) Intercept varies by parent or child. (2) Baseline effect varies by parent or child. (3) Intervention effect varies by parent or child. (4) Measurement errors independently and identically follow Normal distribution.

The restricted maximum likelihood estimation result is given in the Table 18. All parameter estimates are statistically significant under $\alpha = 0.05$ significance level. Negative baseline effect estimates indicate that a subject with high baseline A1C value would like to obtain a greater decrease in A1C value. Negative intervention effect estimates indicate that relative to the control group, a subject in intervention group would like to experience a greater decrease in A1C value. Therefore, for both parents and children, intervention effect of lowering A1C value exists.

Testing homogeneity of novel intervention effects across the parents and children is testing

$$H_0 : \gamma_1 = \gamma_2 \text{ vs } H_1 : \text{Not } H_0.$$

Table 18: Estimation Result of A1C change General Linear Model

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		DF	T statistic	P value
α_1	0.6997	0.1646	0.3749	1.0245	188	4.25	<.0001
α_2	0.4690	0.1644	0.1447	0.7932	188	2.85	0.0048
β_1	-0.1444	0.0294	-0.2024	-0.0864	188	-4.91	<.0001
β_2	-0.1152	0.0299	-0.1743	-0.0562	188	-3.85	0.0002
γ_1	-0.0634	0.0284	-0.1194	-0.0073	188	-2.23	0.0267
γ_2	-0.1504	0.0275	-0.2048	-0.0960	188	-5.45	<.0001
$\Sigma_{1,1}$	0.0373	0.0038					
$\Sigma_{1,2}$	0.0197	0.0030					
$\Sigma_{2,2}$	0.0350	0.0036					

Asymptotic likelihood ratio test statistic follows $\chi^2(1)$ distribution. The statistic is 10.5, and p value is 0.0011. Therefore, we reject the homogeneous novel intervention effects hypothesis and conclude that the intervention effect varies across parents and children. Indeed, since the intervention effect for child was greater in absolute value than the effect for parent (0.1504 for child and 0.0634 for parent), we can conclude that the novel intervention is more effective for children than parent to lower A1C value over 6 months.

The first model assumption (heterogeneous intercept in the model) can be evaluated by similar procedure. Likelihood ratio test of $H_0 : \alpha_1 = \alpha_2$ vs $H_1 : \text{Not } H_0$ gives $\chi^2(1)$ test statistic, which is 1.11, and p value is 0.2920. Therefore, we cannot reject the homogeneous intercept hypothesis and conclude that the intercept remains the same across parents and children.

The second model assumption (heterogeneous baseline effect) can be evaluated by similar procedure. Likelihood ratio test of $H_0 : \beta_1 = \beta_2$ vs $H_1 : \text{Not } H_0$ gives $\chi^2(1)$ test statistic, which is 0.6, and p value is 0.4385. Therefore, we cannot reject the homogeneous baseline effects hypothesis and conclude that the baseline effect remains the same across parents and children.

The fourth model assumption (Normality of error) can be evaluated by diagnostics and influence analysis. From the diagnostic plots (Figure 5), the Studentized residuals are equally and uniformly distributed, and they do not show any systematic structure. No serious violation of the normality assumption was found according to the QQplot. The distributions of residuals are similar over variables combinations, so it is reasonable to assume independently and identically distributed measurement errors. No data are identified to have extremely large Cook's Distance (~ 0.04 for all subjects), and PRESS residuals are well controlled. Therefore, the iid normality of measurement error assumption is valid.

In conclusion, the third and fourth assumptions (heterogeneous intervention effect and iid normal measurement error) has proven to be valid, while the first and second assumptions (heterogeneous intercept and heterogeneous baseline effect) can be ignored. There is no problem to utilize the given model because the model under homogeneous intercept and homogeneous baseline effect assumption is nested in the given model, meaning that the given model can explain the data better. However, for the parsimony of model, one can assume the model under homogeneity assumptions given by

$$Y_{ij} = \alpha + \beta B_{ij} + \gamma_j G_i + \varepsilon_{ij} \ (i = 1, \ldots, 188, \ j = 1, 2),$$

but estimation results will not be largely different from the original model.

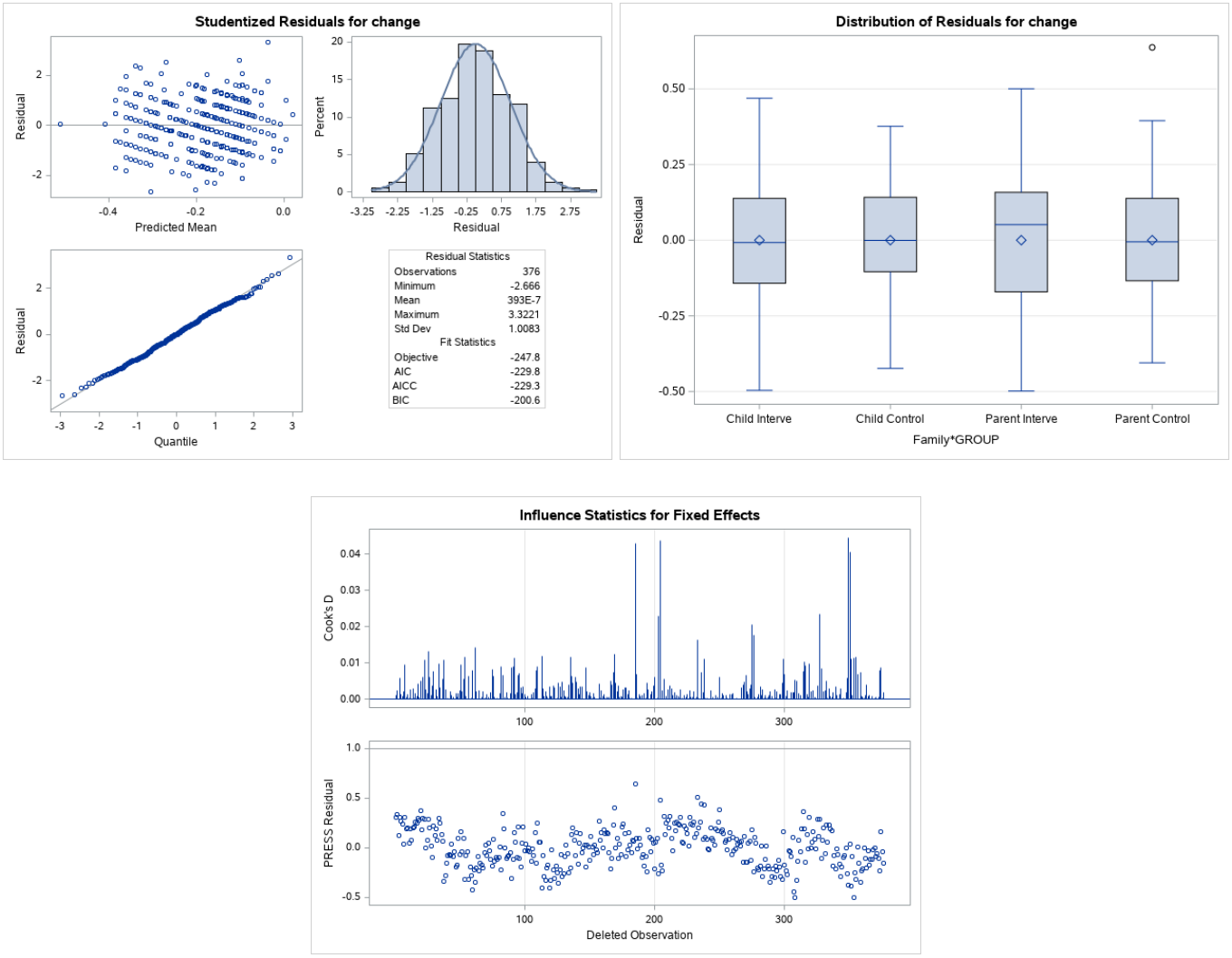

Figure 5: Studentized Residuals Plot (Left, Up), Distribution of Residuals by covariates combination (Right, Up), Influence analysis plots (Bottom)

(c) The model is given by

$$\begin{aligned} \text{Mean Model (logistic model)} : \text{logit}(\mu_{ij}) &= \alpha_j + \beta_j A_{ij} + \gamma G_i \\ \text{Variance Model} : \ \text{Var}(Z_{ij}) &= \mu_{ij}(1-\mu_{ij}) \\ \text{Working Correlation} : \text{Corr}(Z_{i1}, Z_{i2}) &= \tau \end{aligned}$$

where i indicates the ID of parent-child pair, and $j = 1$ indicates parent, while $j = 2$ indicates child. Z_{i1}, Z_{i2} are Bernoulli random variables of mean μ_{i1}, μ_{i2}, and $Z_{1i} = 1$ (or $Z_{i2} = 1$) indicates that the follow-up A1C value is greater than 5.7% for i^{th} parent (or child). A_{i1}, A_{i2} are baseline values of this dichotomized variable for i^{th} parent and child, respectively. G_i is an indicator variable such that $G_i = 1$ if i^{th} pair is in the intervention group, $G_i = 0$ if not. α_j is an intercept, and β_j is an effect of baseline, and they vary by parents or child. γ is a homogeneous intervention effect across parents and children. Since we can expect that the data from a parent and their child are correlated, we introduce correlation structure on Z_{i1} and Z_{i2} to deal with it. Thus, this model is a Binary response marginal model.

The assumptions under this model are (1) Intercept varies by parent of child. (2) Baseline effect varies by parent or child. (3) Intervention effect does not vary by parent or child.

Table 19: Estimation Result of (Pre) Diabetes status Marginal model by GEE

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Z statistic	P value
α_1	-2.3234	0.4119	-3.1308	-1.5160	-5.64	<.0001
α_2	-3.6803	0.7343	-5.1195	-2.2411	-5.01	<.0001
β_1	3.9655	0.5048	2.9761	4.9548	7.86	<.0001
β_2	4.4600	0.8091	2.8741	6.0459	5.51	<.0001
γ	-1.4915	0.3712	-2.2189	-0.7640	-4.02	<.0001
τ	-0.0375	-	-	-	-	-

The Generalized Estimating Equation result is given in the Table 19. All parameter estimates are statistically significant under $\alpha = 0.05$ significance level. Positive baseline effect estimates indicate that a subject in (Pre) Diabetes status at baseline is highly likely in (Pre) Diabetes status at follow-up. In specific, if a parent was in (Pre) Diabetes status at baseline, this person's odds of being in (Pre) Diabetes status at follow-up is 52.74 (= exp(3.9655)) times odds of another parent in the same intervention group who was normal at baseline. Negative intervention effect estimate indicates that relative to the control group, a subject in the intervention group would like to experience a higher chance of being normal. In specific, if a parent in the intervention group was in (Pre) Diabetes status at baseline, this person's odds of being in (Pre) Diabetes status at follow-up is only 0.225 (= exp(-1.4915)) times odds of another parent in the control group who was in (Pre) Diabetes status at baseline. Therefore, for both parents and children, intervention effect of lowering A1C value exists.

To allow for heterogeneous novel intervention effects, we consider the full (saturated) model

$$\begin{aligned} \text{Mean Model (logistic model)} &: \text{logit}(\mu_{ij}) = \alpha_j + \beta_j A_{ij} + \gamma_j G_i \\ \text{Variance Model} &: \text{Var}(Z_{ij}) = \mu_{ij}(1-\mu_{ij}) \\ \text{Working Correlation} &: \text{Corr}(Z_{i1}, Z_{i2}) = \tau \end{aligned}$$

where all settings are the same except the newly added heterogeneous novel intervention effects (γ_1, γ_2). This model is also a Binary response marginal model, and the parameter estimation is conducted by utilizing Generalized Estimating Equations.

Table 20: Estimation Result of (Pre) Diabetes status Marginal model under hetero intervention effect assumption by GEE

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Z statistic	P value
α_1	-2.3014	0.4239	-3.1321	-1.4706	-5.43	<.0001
α_2	-3.7003	0.7190	-5.1095	-2.2910	-5.15	<.0001
β_1	4.0107	0.5170	2.9974	5.0240	7.76	<.0001
β_2	4.4215	0.8355	2.7838	6.0591	5.29	<.0001
γ_1	-1.5965	0.4969	-2.5704	-0.6227	-3.21	0.0013
γ_2	-1.3810	0.5442	-2.4476	-0.3143	-2.54	0.0112
τ	-0.0328	-	-	-	-	-

The Generalized Estimating Equation result under the heterogeneous novel intervention effects assumption is given in the Table 20. All parameter estimates are statistically significant under $\alpha = 0.05$ significance level, and similar to the previous model (model under the homogeneous novel intervention effects assumption) estimates.

Testing the existence of novel intervention effects, we test the following hypothesis

$$H_0 : \gamma_1 = \gamma_2 = 0 \text{ vs } H_1 : \text{Not } H_0,$$

or equivalently $H_0 : \boldsymbol{L}\theta = 0$ vs $H_1 :$ Not H_0, where $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2)'$ is a vector of regression parameters and the contrast matrix $\boldsymbol{L}$ given by

$$\boldsymbol{L} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

In detail, for testing $H_0 : \boldsymbol{L}\theta = 0$, we use the following Wald chi-square statistic

$$T = (\boldsymbol{L}\hat{\theta})^T [\boldsymbol{L}\hat{\Sigma}_{\hat{\theta}}\boldsymbol{L}^T](\boldsymbol{L}\hat{\theta}) \dot{\sim} \chi^2_r, \ r = \text{rank}(\boldsymbol{L}),$$

where $\hat{\theta}$ is the GEE estimator for θ and $\hat{\Sigma}_{\hat{\theta}}$ is its asymptotic covariance matrix.

In this case, the Wald chi-square statistic is 16.26, the degrees of freedom is 2, and p value is 0.0003. Therefore, we reject the hypothesis of no novel intervention effect and conclude that there is a novel intervention effect (relative to the control) of lowering the probability of being (Pre) Diabetes.

Next, testing homogeneity of novel intervention effects across the parents and children is testing

$$H_0 : \gamma_1 = \gamma_2 \text{ vs } H_1 : \text{Not } H_0,$$

or equivalently $H_0 : \boldsymbol{L}\theta = 0$ vs $H_1 :$ Not H_0, where the contrast matrix $\boldsymbol{L} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$. In this case, the Wald chi-square statistic is 0.09, the degrees of freedom is 1, and p value is 0.7663. Therefore, we cannot reject the homogeneous novel intervention effects hypothesis and conclude that the intervention effect does not vary across parents and children.

(d)

Methods: To test whether there is the novel intervention effects (relative to the control) on lowering A1C values for at-risk parents and their at-risk children, we assume two models for the continuous and categorical versions of the A1C variable, respectively. Since we expect that the data from a parent and their child are correlated, the models should deal with the correlation structure.

To test the novel intervention effects on the mean decrease in A1C value, we fit the General linear model given by $Y_{ij} = \alpha_j + \beta_j B_{ij} + \gamma_j G_i + \varepsilon_{ij}$ $(i = 1, \ldots, 188, \ j = 1, 2)$, where i indicates the ID of parent-child pair, and $j = 1$ indicates parent, while $j = 2$ indicates child. Y_{i1}, Y_{i2} are A1C change of i^{th} parent and child, respectively. B_{i1}, B_{i2} are baseline A1C value of i^{th} parent and child, respectively. G_i is an indicator variable such that $G_i = 1$ if i^{th} pair is in the intervention group, $G_i = 0$ if not. α_j is a intercept, β_j is an effect of baseline, and γ_j is the novel intervention effect, and all coefficients vary by parent or child. Finally, ε_{ij} is a measurement error such that $\epsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})^T \sim N_2(0, \Sigma), \ i = 1, \ldots, 188$, where Σ is an unstructured positive definite matrix. Restricted Maximum likelihood estimation is performed for parameter estimation. We test the hypothesis of the existence of the novel intervention effect, given by $H_{01} : \gamma_1 = \gamma_2 = 0$ versus $H_{11} :$ Not H_{01} at significance level $\alpha = 0.05$ using likelihood ratio test. We also test the hypothesis of the homogeneous novel intervention effect, given by $H_{02} : \gamma_1 = \gamma_2$ versus $H_{12} :$ not H_{02} at significance level $\alpha = 0.05$ using likelihood ratio test.

To test the novel intervention effects on the probability of being (Pre) Diabetes (i.e. A1C $\geq 5.7\%$), we fit the Binary response marginal model given by

$$\begin{aligned} \text{Mean Model (logistic model)} : \text{logit}(\mu_{ij}) &= \alpha_j + \beta_j A_{ij} + \gamma_j G_i \\ \text{Variance Model} : \ \text{Var}(Z_{ij}) &= \mu_{ij}(1 - \mu_{ij}) \\ \text{Working Correlation} : \text{Corr}(Z_{i1}, Z_{i2}) &= \tau \end{aligned}$$

where Z_{i1}, Z_{i2} are Bernoulli random variables of mean μ_{i1}, μ_{i2}, and $Z_{1i} = 1$ (or $Z_{i2} = 1$) indicates that the follow-up A1C value is greater than 5.7% for i^{th} parent (or child). A_{i1}, A_{i2} are baseline values of this dichotomized variable for i^{th} parent and child, respectively. α_j is an intercept, β_j is an effect of baseline, and γ_j is the novel intervention effect, and all coefficients vary by parent or child.

Generalized Estimating Equation is performed for parameter estimation. We test the hypothesis of the existence of the novel intervention effect, given by $H_{01} : \gamma_1 = \gamma_2 = 0$ versus H_{11} : Not H_{01} at significance level $\alpha = 0.05$ using Wald chi-square statistic for the linear combinations of the regression parameters. We also test the hypothesis of the homogeneous novel intervention effect, given by $H_{02} : \gamma_1 = \gamma_2$ versus H_{12} : Not H_{02} at significance level $\alpha = 0.05$ using the same test.

Results: The regression coefficients estimates based on General linear model are given in Table 21.

Table 21: Estimation Result of A1C change General Linear Model

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		DF	T statistic	P value
α_1	0.6997	0.1646	0.3749	1.0245	188	4.25	<.0001
α_2	0.4690	0.1644	0.1447	0.7932	188	2.85	0.0048
β_1	-0.1444	0.0294	-0.2024	-0.0864	188	-4.91	<.0001
β_2	-0.1152	0.0299	-0.1743	-0.0562	188	-3.85	0.0002
γ_1	-0.0634	0.0284	-0.1194	-0.0073	188	-2.23	0.0267
γ_2	-0.1504	0.0275	-0.2048	-0.0960	188	-5.45	<.0001
$\Sigma_{1,1}$	0.0373	0.0038					
$\Sigma_{1,2}$	0.0197	0.0030					
$\Sigma_{2,2}$	0.0350	0.0036					

The novel intervention effects for the parents (γ_1) is estimated by -0.0634 (se: 0.0284), meaning that mean decrease in A1C of parents in the intervention group was 0.0634 higher than that of parents in the control group. The novel intervention effects for the children (γ_2) is estimated by -0.1504 (se: 0.0275). We reject no novel intervention effect hypothesis ($H_{01} : \gamma_1 = \gamma_2 = 0$) based on asymptotic likelihood ratio chi-square test (LRT statistic = 15.48, degrees of freedom = 2, p-value $< .0001$) under significance level $\alpha = 0.05$. We also reject the homogeneous novel intervention effect hypothesis ($H_{02} : \gamma_1 = \gamma_2$) based on asymptotic likelihood ratio chi-square test (LRT statistic = 10.5, degrees of freedom = 1, p-value = 0.0011) under significance level $\alpha = 0.05$. Other than that, negative baseline effect estimates indicate that a subject with a high baseline A1C value would like to experience a greater decrease in A1C value after 6 months. No serious violation of the model assumption was identified through diagnostics and influence analysis.

Table 22: GEE Estimation Result of (Pre) Diabetes status Marginal model

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Z statistic	P value
α_1	-2.3014	0.4239	-3.1321	-1.4706	-5.43	<.0001
α_2	-3.7003	0.7190	-5.1095	-2.2910	-5.15	<.0001
β_1	4.0107	0.5170	2.9974	5.0240	7.76	<.0001
β_2	4.4215	0.8355	2.7838	6.0591	5.29	<.0001
γ_1	-1.5965	0.4969	-2.5704	-0.6227	-3.21	0.0013
γ_2	-1.3810	0.5442	-2.4476	-0.3143	-2.54	0.0112
τ	-0.0328	-	-	-	-	-

The regression coefficients estimates based on Binary response marginal model are given in Table 22. The novel intervention effects for the parents (γ_1) is estimated by -1.5965 (se: 0.4967), meaning that the odds of a parent in intervention group being (Pre) diabetes at follow-up is only 0.2026 (= exp(-1.5965)) times that of a parent in control group. The novel intervention effects for the children (γ_2) is estimated by -1.3810 (se: 0.5442). We reject no novel intervention effect hypothesis

($H_{01} : \gamma_1 = \gamma_2 = 0$) based on the Wald chi-square test (statistic = 16.26, degrees of freedom = 2, p-value = 0.0003) under significance level $\alpha = 0.05$. However, we cannot reject the homogeneous novel intervention effect hypothesis ($H_{02} : \gamma_1 = \gamma_2$) based on the Wald chi-square test (statistic = 0.09, degrees of freedom = 1, p-value = 0.7663) under significance level $\alpha = 0.05$. Other than that, positive baseline effect estimates indicate that a subject in (Pre) Diabetes status at baseline is highly likely in (Pre) Diabetes status at follow-up after 6 months. No serious violation of the model assumption was identified through diagnostics and influence analysis (data not shown).

In conclusion, the novel intervention is effective for lowering mean A1C value for parents and children, while the amounts of decrease are different across parents and children. The intervention is also effective for lowering the probability of being (Pre) diabetes (i.e. A1C $\geq 5.7\%$), and the amounts of decrease are same across parents and children. One side remark is that even though the novel intervention effects on lowering the probability of being (Pre) diabetes are same for parents and children, the intercepts are statistically different. It explains the data that the decrease in total (Pre) Diabetes patients was larger for children compared to parents ($65 \rightarrow 51$ for parents, $60 \rightarrow 32$ for children), meaning that regardless of intervention group, children experienced greater decrease in the probability of being (Pre) Diabetes. It may be due to the fact that a child's young body possesses higher power of returning to normal (homeostatic), and thus any care can be more effective.

(e)

- Novel intervention successfully lowers the average A1C of subjects more than the standard care
- The additional mean decreases (relative to the standar care) are different for parents and children: 0.0634 for parents and 0.1504 for children
- Novel intervention successfully lowers the possibility of being (pre) diabetes (A1C $\geq 5.7\%$) than the standard care does
- The additional proportional decrease in odds are same for parents and children: intervention group's odds is only 22.5% of control group's odds.
- A subject with high baseline A1C would exhibit a larger decrease in A1C value
- A subject in (pre) diabetes status at baseline tends to be in the same status at follow-up

(f)

Dyadic RCT design

Advantages: (1) By capturing the correlation structure in the dyads, parameter estimation can be more accurate (2) Able to test the homogeneity of baseline effect or intervention effect with higher power due to it uses the information that data from a parent and their child are correlated

Disadvantages: (1) A parent - child pair is required, may be harder to collect data (2) Model becomes complicated if followed up multiple times

Parents & Children separate RCT design

Advantages: (1) The data collection is easier, any parents at risk and any children at risk can be included in the study (2) It is possible to predict a new parent's future outcome based on separate parents RCT analysis result. It does not require their child's data

Disadvantages: (1) The parameter estimation and inference should be done separately; joint inference on parents' and children's data might not be possible or be done with lower accuracy (2) When a parent and their child are assigned in the different intervention group, it is impossible to identify whether the difference in the parent's and the child's response came from the parents-children difference, or from the novel intervention-control difference.