Negative binomial mixed models for analyzing longitudinal CD4 count data

2024. 02. 15 BIBS seminar

Jiwon Park

Contents

BIBS

Introduction

Introduction

- It is important to correctly model the CD4 cell count or disease biomarkers of a patient in the presence of covariates or factors determining the disease progression over time.
- The Poisson mixed-effects models (PMM) can be an appropriate choice for repeated count data.
- However, this model is not realistic because of the restriction that the mean and variance are equal.
- PMM can be replaced by the negative binomial mixed-effects model (NBMM).

CD4 cell

- CD4 cell counts deliver a sign of the wellbeing of an individual immune system.
- It also provides information about disease progression that play an essential role in the immune system.

Materials and Method

Data

- Data from the CAPRISA 002 AI Study
- Between August 2004 and May 2005, CAPRISA introduced a cohort study recurring high-risk HIV negative women to a follow-up study
- monitored to examine disease progression and CD4 count/viral load evolution.

Method

- Statistical modeling method applied to count data
- **A linear model** consists of a response variable Y, which is assumed to be normally distributed, and several predictors $(x_1, x_2, \ldots, x_p)$.
- Multiple regression analysis studies the linear relationships among two or multiple independent variables and one dependent (response) variable. The multiple regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i = \beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i = \beta_0 + \boldsymbol{\beta}' \boldsymbol{x}_i + \varepsilon_i, i = 1, \ldots, n.$$

Method

- They extend these multiple linear regression model ideas to **generalized linear models (GLM)** where the distribution of the outcome variable can include distributions other than normal.
- Then the outcome y_i can be continuous, count, ordinal, categorical, and so on as long as its distribution is from the exponential family.
- **A Poisson process** is mainly used as an initial point for modeling the stochastic difference of count data around a theoretical expectation.

Method

- The Poisson regression is a commonly-used statistical model for *n* responses $y_1, \dots, y_n$ whose domain is non-negative integer values.
- Each y_i is modeled as an independent Poisson random variable.
- Thus, a model for the Poisson rate parameter λ_i is given by

$$y_i \overset{iid}{\sim} \text{Poisson}\ (\lambda_i)$$

$$\ln \lambda_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}$$

- $x_{i1}, \ldots, x_{ip}$: a set of *p* explanatory variables
- $\beta = \beta_0, \ldots, \beta_p$: regression coefficients

Method

- The probability mass function(pmf) of the poisson random variable with parameter (λ_i)

$$f\left(y_i, \lambda_i\right) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, y_i = 0, 1, 2, \cdots$$

$$= \sum_{i=1}^{n} \left[y_i \ln\left(\lambda_i\right) - \lambda_i - \ln y_i!\right]$$

Method

Likelihood function

- Likelihood function of Poisson regression model for getting model parameter estimates based on Data.

$$\ell(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} \left[y_i \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) - e^{\sum_{j=0}^{p} \beta_j x_{ij}} - \ln y_i! \right]$$

$$= \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i' \boldsymbol{\beta} - \exp\left(\boldsymbol{x}_i' \boldsymbol{\beta} \right) - \ln y_i! \right\}.$$

Method

Overdispersion

- $var(y_i)$ = $\Phi * \lambda_i$, Φ > 1 → over-dispersion (Φ : variance parameter)
- $var(y_i)$ = $\Phi * \lambda_i$, Φ =1
- $var(y_i)$ = $\Phi * \lambda_i$, Φ < 1 → under-dispersion

► Relation : $var(y_i) = \hat{\Phi} * \bar{y}$

$\therefore \hat{\Phi} = var(y_i) / \bar{y}$

<u>Overdispersion criteria for realistic data</u>

Pearson chi-square statistic (χ^2)/df >1 : suspicion

Pearson chi-square statistic (χ^2)/df >2 : definite

Method

Negative binomial distribution

- The parameter estimates based on the negative binomial model are not exceptionally different from those based on Poisson model.
- However, the Poisson model underestimates the SEs when over-dispersion is present, leading to improper inference.

Method

Negative binomial distribution

- In Negative Binomial distribution, the variance tends to increase more significantly as the mean increases, including the inherent additional variability.
- y_i ~ NB($\mu_i, \mu_i[1 + \mu_i/\alpha]$) where α ($\alpha > 0$)
  - $E(y) = \alpha\theta = \mu$
  - $var(y) = \alpha\theta(1 + \theta)$

Method

Gamma distribution

- The part where gamma distributions are used for overdispersion modeling plays a key role in defining the dispersion structure of the negative binomial distribution model.
- Gamma distribution is used to estimate this overdispersion parameter, which increases the flexibility of the model, allowing more accurate capture of the overdispersion observed in real data.

Method

Gamma distribution

$$f(\lambda;\alpha,\theta)=\frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\theta^{\alpha}\Gamma(\alpha)},\quad \lambda>0,\quad \alpha>0,\quad \theta>0$$

$$f(Y|\lambda)=\frac{e^{-\lambda}\lambda^{k}}{k!}\frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\theta^{\alpha}\Gamma(\alpha)}$$

$$=\binom{\alpha+k-1}{k}\left(\frac{\theta}{1+\theta}\right)^{k}\left(\frac{1}{1+\theta}\right)^{\alpha}=\frac{\Gamma(\alpha+k)}{k!\Gamma(\alpha)}\left(\frac{\theta}{1+\theta}\right)^{k}\left(\frac{1}{1+\theta}\right)^{\alpha},$$

- $Y|\lambda \sim poisson\ (\lambda)$
- $\lambda \sim \text{Gamma}\ (\alpha,\theta)$
- $\boldsymbol{f(Y)=poisson\ (Y|\lambda) * f(\lambda) \sim NB\ \left(\mu_i, \mu_i(1+\frac{\mu_i}{\alpha})\right)}$

- θ : scale parameter
- α : shape parameter

$$\binom{\alpha+k-1}{k}=\frac{(\alpha+k-1)(\alpha+k-2)\ldots\alpha}{k!}=\frac{(\alpha+k-1)!}{k!(\alpha-1)!}$$

Method

Negative binomial distribution

$$L(\boldsymbol{\beta},\alpha)=\prod_{i=1}^{n}\frac{\Gamma(\alpha+k_i)}{k_i!\Gamma(\alpha)}\left(\frac{\theta_i}{1+\theta_i}\right)^{k_i}\left(\frac{1}{1+\theta_i}\right)^{\alpha}$$

i : # of observation
α : parameter of Gamma distribution
β : model parameter

$$=\sum_{i=1}^{n}\left(\sum_{j=0}^{k_i-1}\log(\alpha+j)-\log k_i!+k_i\log\theta_i-k_i\log(1+\theta_i)+\alpha\log 1-\alpha\log(1+\theta_i)\right)$$

$$\ell(\boldsymbol{\beta},\alpha)=\sum_{i=1}^{n}\left(\sum_{j=0}^{k_i-1}\log(\alpha+j)-\log k_i!+k_i\log\theta_i-(k_i+\alpha)\log(1+\theta_i)\right)$$

Method

Generalized linear mixed - effect model

- It is necessary to extend the GLM to generalized linear mixed-effects models, including a **subject-specific random effect** introduced in the *linear predictor* to seize the dependence.
- This involves modeling the inherent associations that exist between data measured multiple times for the same subject over time.
- This approach enables more accurate capture of each individual's patterns of change over time, thus increasing the accuracy of the analysis.

Method

Generalized linear mixed - effect model

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})X_{1ij} + \ldots + (\beta_p + b_{ip})X_{pij} + \varepsilon_{ij},$$

$$\log\left(E\left(y_{ij}\right)\right) = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_0 + b_1 x_{1ij} + \cdots + b_p x_{pij}$$

$$\log\left\{E(y_{ij}|\boldsymbol{b_i})\right\} = \eta_{ij} = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i.$$

β_0 , $\beta_1, \beta_3, \ldots, \beta_p$: fixed effect

$b_{i0}, b_{i1}, \ldots, b_{ip}$: random effect

ϵ_{ij} : residual

j : time point

BIBS
BioInformatics & BioStatistics Lab.

Method

$$\log\left(\mu_{ij}\right) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i + \varepsilon_{ij},$$

$$P\left(y_{ij} = y|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}\right) = \frac{e^{-\mu_{ij}}\mu_{ij}^{y}}{y!} = \frac{1}{y!}e^{-\exp\left(\boldsymbol{x}_{ij}^{\prime}\boldsymbol{\beta}+\boldsymbol{z}_{ij}^{\prime}\boldsymbol{b}_i\right)}\exp\left(\boldsymbol{x}_{ij}^{\prime}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\prime}\boldsymbol{b}_i\right)^{y}$$

$$= \frac{1}{y!}\exp\left[\left(\boldsymbol{x}_{ij}^{\prime}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\prime}\boldsymbol{b}_i\right)^{y} - \exp\left(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i\right)\right], y = 0, 1, 2, \ldots$$

b_i : random effect
β : fixed effect
x_{ij} : variable of interest
z_{ij} : explanatory variables for random effects
μ_{ij} : $E(y_{ij})$

Result

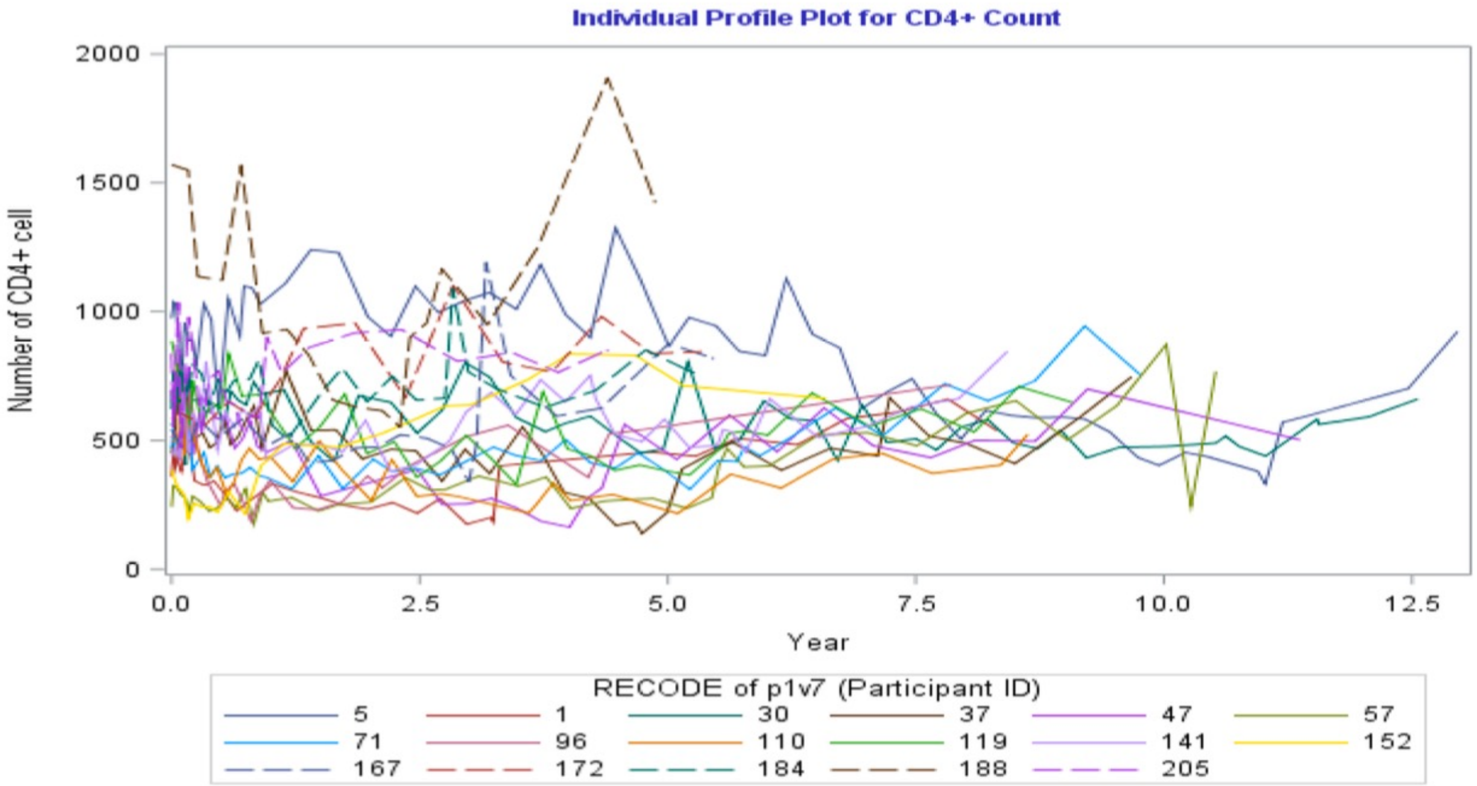
Result

Covariates	Level	CD4 count N (%)			p-value	% Missing
		< 200	200–500	> 500		
Baseline BMI category	Underweight	2 (0.03)	219 (3.12)	254 (3.62)	< 0.0001	0.0
	Normal weight	114 (1.62)	2305 (32.84)	2690 (38.32)		
	Overweight	18 (0.26)	512 (7.29)	657 (9.36)		
	Obese	0	17 (0.24)	231 (3.29)		
Baseline viral load	Undetected	0	0	16 (0.23)	< 0.0001	0.0
	Low	20 (0.28)	791 (11.27)	1532 (21.83)		
	Medium	45 (0.64)	1209 (17.22)	1497 (21.23)		
	High	69 (0.98)	1053 (15)	787 (11.21)		
Number of sexual partners	No partner	29 (0.41)	565 (8.05)	579 (8.25)	< 0.0001	0.0
	Stable partner	85 (1.21)	2274 (32.4)	3078 (43.85)		
	Many partners	20 (0.28)	214 (3.05)	175 (2.49)		
Age group	< 20	1 (0.01)	130 (1.82)	121 (1.72)	< 0.0001	0.0
	20–29	97 (1.38)	1872 (26.67)	1977 (28.17)		
	30–39	17 (0.24)	813 (11.58)	1255 (17.88)		
	40–49	19 (0.27)	203 (2.89)	369 (5.26)		
	50–59	0	35 (0.5)	91 (1.3)		
	≥ 60	0	0	19 (0.27)		
Educational level	Primary school	3 (0.04)	104 (1.48)	181 (2.58)	0.0129	0.0
	Secondary school	131 (1.87)	2949 (42.01)	3651 (52.02)		
Place of residence	Rural	62 (0.88)	1467 (20.90)	1806 (25.73)	0.7176	0.06
	Urban	72 (1.03)	1586 (22.6)	2026 (28.86)		
ART initiation group	Pre ART	110 (1.57)	2566 (36.56)	2783 (39.65)	< 0.0001	0.0
	Post ART	20 (24)	487 (6.94)	1049 (14.95)		

(chi-square test)

- It shows the relationship between the number of CD4 cells in HIV-infected patients and various factors.
- CD4 cell count is an indicator of immune system status, and this table shows the patient's weight, viral load, number of age, education level, residence, and ART treatment compared to the CD4 cell count range.

Result



- This graph shows how patients' immune systems change over time
- By tracking these changes, they can monitor the progression or treatment effectiveness of HIV.

Result

Random effect models	Information criteria					
	$-2\log \ell$	AIC	AICC	BIC	CAIC	HQIC
Model 1	87,781.28	87,833.28	87,833.48	87,923.23	87,949.23	87,869.54
Model 2	88,603.50	88,649.50	88,649.66	88,729.07	88,752.07	88,681.58
Model 3	88,591.64	88,637.64	88,637.80	88,717.21	88,740.21	88,669.72
Model 4	89,156.39	89,202.39	89,202.55	89,281.96	89,304.96	89,234.47
Model 5	89,837.18	89,879.18	89,879.31	89,951.83	89,972.83	89,908.47
Model 6	92,302.08	92,344.08	92,344.21	92,416.73	92,437.73	92,373.37
Model 7	91,190.61	91,232.61	91,232.74	91,305.26	91,326.26	91,261.90

Table 4. Comparison of random effect models.

Model 1: *Intercept, Time,* $\sqrt{Time}$.
Model 2: *Intercept, Time.*
Model 3: *Intercept,* $\sqrt{Time}$.
Model 4: *Time,* $\sqrt{Time}$.
Model 5: *Intercept only.*
Model 6: *Time only.*
Model 7: $\sqrt{Time}$ *only.*

Result

Effect	Num DF	Den DF	NB		Poisson	
			F value	Pr > F	F value	Pr > F
Time in month	1	235	62.53	< 0.0001	14.80	0.0002
Sqrt_Time	1	234	86.36	< 0.0001	48.41	< 0.0001
Baseline BMI category	3	6307	6.26	0.0003	6.31	0.0003
ART initiation	1	6307	345.45	< 0.0001	5890.28	< 0.0001
Baseline VL	3	6307	7.48	< 0.0001	12.79	< 0.0001
No. of sexual partners	2	6307	1.64	0.1935	1.85	0.1578
Age group	5	6307	1.46	0.1987	27.34	< 0.0001
Education level	1	6307	0.25	0.6196	0.15	0.6990
Place of residence	1	6307	0.01	0.9246	0.11	0.7406

Table 5. Type III Analysis of fixed effects for Poisson and NB distribution.

- An analysis of variance that assesses the significance of fixed effects.
- *Time in month, Sqrt_time, ART, VL* appear to be a significant fixed effect in both models, suggesting that these variables are likely to have a significant influence on CD4 cell number.

Result

Covariates	Negative binomial mixed-effects model				Poisson mixed-effects model		
	Estimate	SE	Pr>|t|	95% CI for NB estimate	Estimate	SE	Pr>|t|
Intercept	6.4697	0.04982	<0.0001	(6.3715, 6.5679)	6.4625	0.04264	<0.0001
Time in month	0.007824	0.000989	<0.0001	(0.005875, 0.009774)	0.006564	0.001706	0.0002
Sqrt_Time	– 0.08649	0.009307	<0.0001	(– 0.1048, – 0.06815)	– 0.06839	0.009830	<0.0001
ART initiation (post)	0.2301	0.01238	<0.0001	(0.2058, 0.2543)	0.1947	0.002537	<0.0001
Baseline BMI category (ref. = normal weight)							
Obese	0.4815	0.1113	<0.0001	(0.2633, 0.6996)	0.4985	0.1147	<0.0001
Overweight	0.02561	0.04975	0.6067	(– 0.07191, 0.1231)	0.03131	0.05148	0.5431
Underweight	0.005901	0.07927	0.9407	(– 0.1495, 0.1613)	0.01691	0.08264	0.8379
Baseline HIV viral load category (ref. = low VL)							
High VL	– 0.2393	0.05157	<0.0001	(– 0.3404, – 0.1382)	– 0.3074	0.05065	<0.0001
Medium VL	– 0.1258	0.04587	0.0061	(– 0.2157, – 0.03585)	– 0.1121	0.04686	0.0168
Undetectable	0.1377	0.2901	0.6351	(– 0.4310, 0.7064)	0.1199	0.2978	0.6872
Number of sexual partners (ref. = stable partner)							
Many partners	– 0.1560	0.09394	0.0967	(– 0.3402, 0.02811)	– 0.1674	0.09908	0.0911
No partner	– 0.04821	0.04993	0.3343	(– 0.1461, 0.04967)	– 0.05913	0.05164	0.2522
Age group in years (ref. = < 20)							
20–29	0.01166	0.03104	0.7072	(– 0.04919, 0.07251)	– 0.00791	0.007830	0.3125
30–39	0.02852	0.03432	0.4060	(– 0.03876, 0.09580)	– 0.01239	0.008474	0.1438
40–49	– 0.00719	0.04545	0.8743	(– 0.09629, 0.08191)	– 0.03422	0.01112	0.0021
50–59	– 0.05694	0.06662	0.3927	(– 0.1875, 0.07365)	– 0.1399	0.01549	<0.0001
≥60	0.2082	0.1532	0.1741	(– 0.09205, 0.5084)	– 0.3107	0.03519	<0.0001
Education attainment (ref. = secondary or high school)							
Primary school	– 0.04509	0.09084	0.6196	(– 0.2232, 0.1330)	– 0.03582	0.09263	0.6990
Residence of participant (ref. = urban)							
Rural	– 0.00373	0.03947	0.9246	(– 0.08112, 0.07365)	0.01337	0.04038	0.7406

Table 6. Parameter estimates using Poisson and NB mixed-effects model.

Result

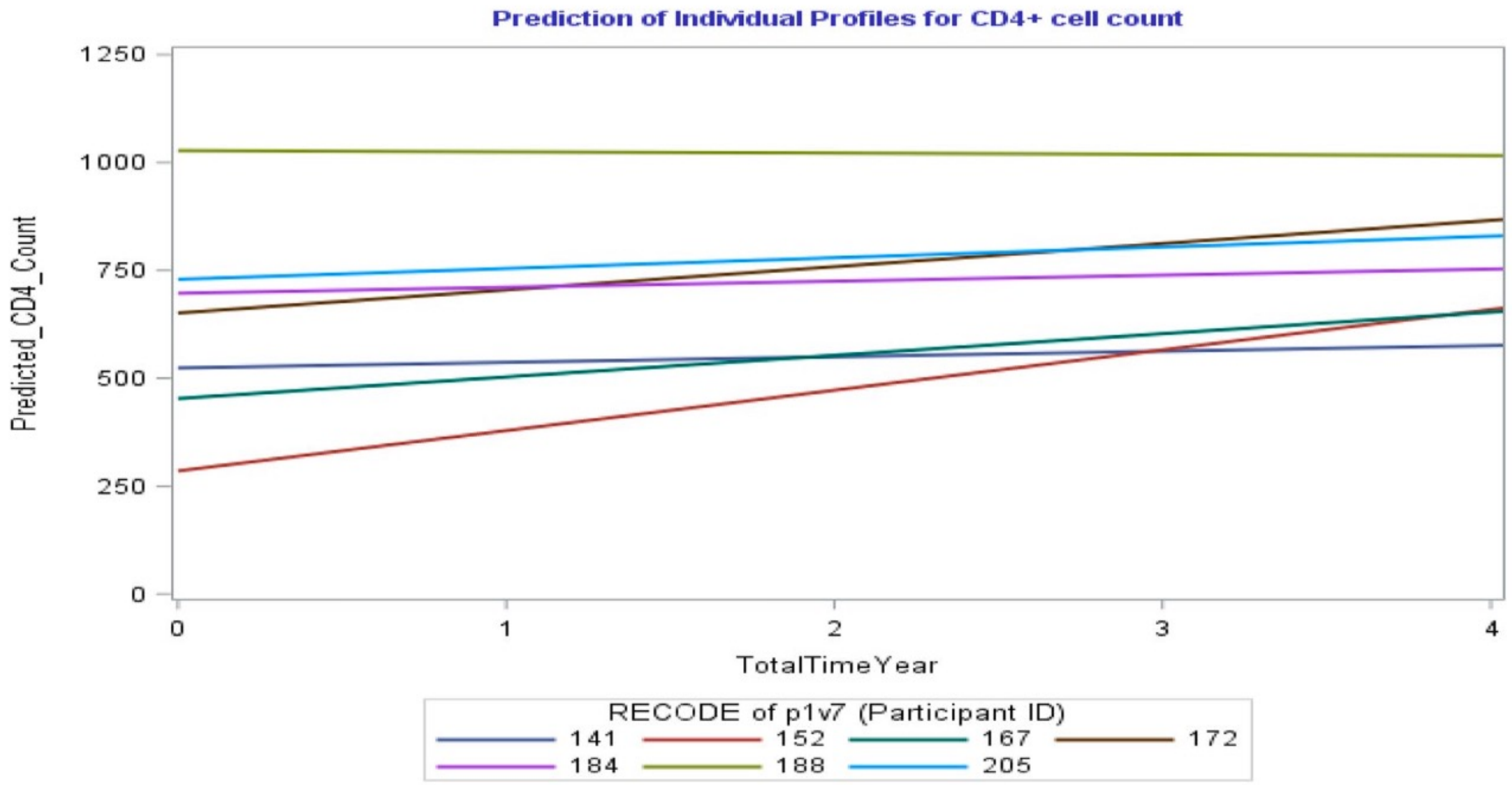


Figure 3. Prediction of 7 randomly selected individual profiles plot of CD4 count for 4 years.

A prediction of changes in the number of CD4+ cells in an arbitrarily selected individual patient

Result

$$log(\hat{\mu}_i) = 6.4697 + 0.007824 \times time - 0.08649 \times \sqrt{time} + 0.2301 \times postHAARTtreatment + 0.4815 \times obese - 0.2393 \times highVL - 0.1258 \times mediumVL.$$

- $\hat{\mu}_i$: Number of predicted CD4+ cells log-transformed
- $\log(\hat{\mu}_i)$: Predicted number of CD4+ cells in the i^th patient

$$\hat{\mu}_i = \exp\left(6.4697 + 0.007824 \times time - 0.08649 \times \sqrt{time} + 0.2301 \times postHAARTtreatment + 0.4815 \times obese - 0.2393 \times highVL - 0.1258 \times mediumVL\right).$$

- The actual number of CD4+ cells in which log-transformed predictions are converted back to the original scale

Result

Parameter	Parameter estimates (10 imputations)					
	Estimate	SE	Pr>|t|	95% confidence limits	Minimum	Maximum
Intercept	6.459413	0.049830	<0.0001	(6.36175, 6.55708)	6.458658	6.460775
Time in month	0.007475	0.000975	<0.0001	(0.00556, 0.00939)	0.007450	0.007508
Sqrt_Time	− 0.083647	0.009266	<0.0001	(− 0.10181, − 0.06549)	− 0.083982	− 0.083434
ART initiation (Post)	0.224037	0.012594	<0.0001	(0.19935, 0.24872)	0.223216	0.225014
Baseline BMI category (ref. = normal weight)						
Obese	0.474714	0.109902	<0.0001	(0.25931, 0.69012)	0.473892	0.475630
Overweight	0.024208	0.048971	0.6211	(− 0.07177, 0.12019)	0.023820	0.024529
Underweight	0.002070	0.078101	0.9789	(− 0.15101, 0.15515)	0.001321	0.003137
Baseline HIV viral load category (ref. = Low VL)						
High VL	− 0.239102	0.051294	<0.0001	(− 0.33964, − 0.13857)	− 0.239735	− 0.238839
Medium VL	− 0.122078	0.045390	0.0072	(− 0.21104, − 0.03311)	− 0.122251	− 0.121642
Undetectable	0.142848	0.286259	0.6178	(− 0.41821, 0.70391)	0.142510	0.143351
Number of sexual partners (ref. = stable partner)						
Many partners	− 0.153632	0.092090	0.0953	(− 0.33412, 0.02686)	− 0.154667	− 0.152911
No partner	− 0.046962	0.049227	0.3401	(− 0.14344, 0.04952)	− 0.047267	− 0.046691
Age group in years (ref. = < 20)						
20–29	0.013477	0.031659	0.6703	(− 0.04857, 0.07553)	0.012306	0.014325
30–39	0.033725	0.034974	0.3349	(− 0.03482, 0.10227)	0.032678	0.034744
40–49	− 0.005842	0.046177	0.8993	(− 0.09635, 0.08466)	− 0.007790	− 0.004745
50–59	− 0.052070	0.067501	0.4405	(− 0.18437, 0.08023)	− 0.054207	− 0.051024
≥ 60	0.206708	0.156046	0.1853	(− 0.09914, 0.51255)	0.205360	0.207553
Education attainment (ref. = secondary or high school)						
Primary school	− 0.046292	0.089605	0.6054	(− 0.22191, 0.12933)	− 0.046602	− 0.046009
Residence of participant (ref. = urban)						
Rural	− 0.001916	0.038813	0.9606	(− 0.07799, 0.07416)	− 0.002146	− 0.001596

- This table shows parameter estimates obtained after using the multiple imputation method to deal with missing data.
- This is a statistical technique for handling missing data, creating multiple alternative datasets, performing analyses on each, and then combining the results to give an overall estimate.

Discussion and Conclusion

Discussion & Conclusion

- GLMs extend the standard concept of linear models to outcome variables whose distribution is from a member of the exponential family.
- GLM consists of three components
  ① a stochastic component that characterizes the likelihood distribution of the response variable;
  ② a linear predictor that is a systematic component portraying the linear model characterized by the explanatory variables;
  ③ a link function that connect the mean of the response variable to a linear combination of the explanatory variables.

Discussion & Conclusion

- Longitudinal studies, also called mixed-effects models, are used to study changes in the response variable over a relevant interval of time or space and the effects of different factors on these changes.
- The two fundamental issues in longitudinal studies are constructing an appropriate model for the mean and choosing a reasonable but parsimonious model for the covariance structure of longitudinal data.
- Due to the presence of overdispersion in the dataset, a **Negative Binomial Mixed Model (NBMM)** with an unstructured (UN) covariance structure is suitable.

Thank you for listening

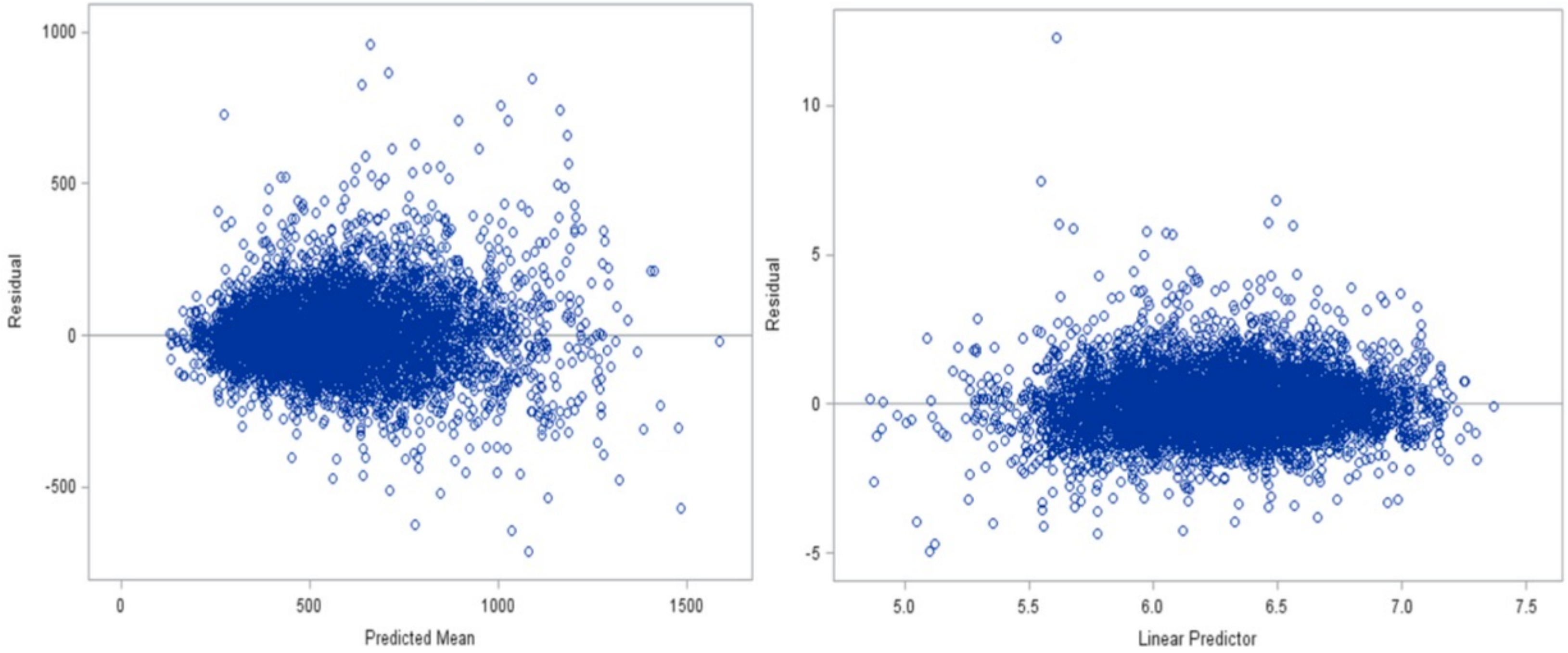
Supplementary

UN covariance structure

Supplementary Table 2 Comparison of fixed effects results across different covariance structure using Model 1

	UN		AR(1)		CS		Toep	
Covariates	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	6.4697	0.04982	6.4724	0.03423	6.4861	0.03410	6.4799	0.03439
Time in month	0.007824	0.000989	0.008516	0.01060	0.01439	0.01051	0.008272	0.01082
Sqrt_Time	-0.08649	0.009307	-0.08950	0.01180	-0.08434	0.01170	-0.08886	0.01201
ART Initiation (Post)	0.2301	0.01238	0.2284	0.01263	0.2363	0.01265	0.2277	0.01264
Baseline BMI category (ref.=Normal weight)								
Obese	0.4815	0.1113	0.6076	0.07836	0.5097	0.07765	0.6350	0.07813
Overweight	0.02561	0.04975	0.02687	0.03466	0.02072	0.03441	0.02970	0.03448
Underweight	0.005901	0.07927	0.09673	0.05503	0.03837	0.05470	0.09359	0.05481
Baseline HIV viral load category (ref.= Low VL)								
High VL	-0.2393	0.05157	-0.3307	0.03345	-0.3234	0.03321	-0.3377	0.03330
Medium VL	-0.1258	0.04587	-0.1527	0.03130	-0.1254	0.03112	-0.1567	0.03116
Undetectable	0.1377	0.2901	-0.04788	0.2242	0.1338	0.2256	-0.01985	0.2218
Number of sexual partners (ref.= Stable partner)								
Many partners	-0.1560	0.09394	-0.05213	0.06388	-0.1506	0.06352	-0.04274	0.06393
No partner	-0.04821	0.04993	-0.03423	0.03459	-0.05490	0.03434	-0.03322	0.03438
Age group in years(ref.= < 20)								
20-29	0.01166	0.03104	0.02553	0.02516	0.006652	0.02519	0.02065	0.02543
30-39	0.02852	0.03432	0.04911	0.02849	0.03351	0.02850	0.04303	0.02871
40-49	-0.00719	0.04545	0.007849	0.04070	0.01926	0.04068	-0.00114	0.04084
50-59	-0.05694	0.06662	-0.06551	0.06134	-0.03957	0.06135	-0.06503	0.06143
≥ 60	0.2082	0.1532	-0.2185	0.1606	0.2020	0.1601	-0.1844	0.1612
Education attainment (ref.= Secondary or high school)								
Primary school	-0.04509	0.09084	0.1126	0.06341	-0.00666	0.06299	0.09430	0.06306
Residence of participant (ref.= Urban)								
Rural	-0.00373	0.03947	0.003881	0.02707	0.01729	0.02689	0.003076	0.02694

Overdispersion



Visual evidence of overdispersion

Overdispersion

Fit Statistics for Conditional Distribution	Poisson	NB
– 2 log L(CD4 counts/r. effects)	199,670.3	85,320.39
Pearson χ^2	145,017.0	6396.89
Pearson χ^2/DF	20.66	**0.91**

Table 3. Measure of over-dispersion between Poisson and negative binomial distribution.