Zihang Wang
Mentor: Ellen Graham
Autumn 2023

Introduction of Longitudinal Data Analysis - HIV

Background

HIV (human immunodeficiency virus) is the virus which causes AIDS. It damages people's immune systems, making them vulnerable to other diseases (HIV gov). It can be transmitted through sex, pregnancy, sharing needles, to name a few. Today, HIV is still a serious medical crisis all over the world. According to the statistics in UNAIDS, there are about 39 million people who suffer from HIV all over the world in 2022. Even worse, based on current knowledge and medical technology, HIV is incurable, which means that there is currently no viable treatment for the virus. Once someone contracts HIV, the virus will stay in the body for the life (HIV gov). Thus, determination of the incidence rate of HIV is significant for HIV surveillance and for evaluating the effectiveness of HIV prevention efforts (Fei Gao & Marlena Bannick, 2022). A traditional and accurate way of estimating HIV incidence is following a group of HIV negative people over several years and counting how many contracts HIV. The number who contract HIV divided by the time observed gives an estimate of incidence. However, tracking a group of people for a such long period of time is very time consuming and expensive. Another effective and practical approach is to focus on the LAg Avidity which is a measure of the binding strength of a certain set of antibodies against HIV (Kernis et al., 2020). To be specific, we can utilize LAg Avidity as a substitution for an individual's duration of HIV infection if we comprehend how it varies over time.

To explore the incidence rate of HIV over time, I chose the second approach and used a cleaned dataset with 175 identifiers (infected individuals) from my mentor Ellen Graham. The origin of the cleaned data is from Duong's dataset which was collected from a longitudinal sample of 259 HIV-1-infected individuals (n = 2737). The basic variables of dataset including source, number of seroconverters, available specimens, and likely or confirmed HIV-1 subtypes. Then, I use the relationship between infection time and infection duration to estimate the distribution of infection time and model this relationship through linear mixed model. The estimated duration of infection can then be used to estimate the rate at which people become infected with HIV, known as HIV incidence.

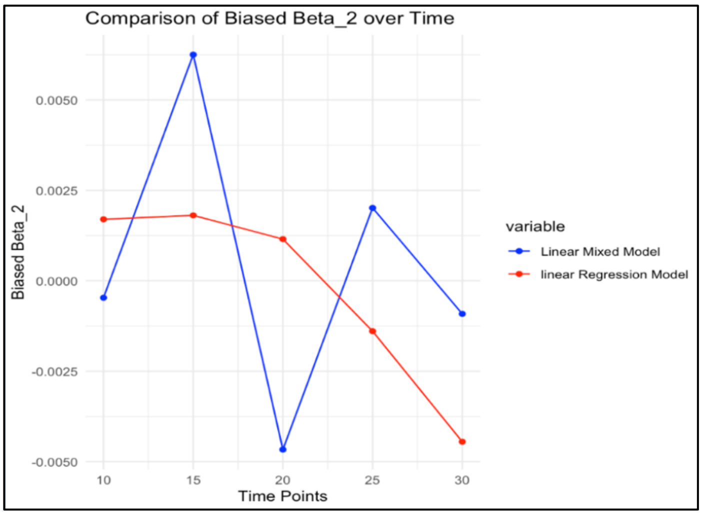
Data Simulation - Statistical method

Since the dataset is collected over years from the same identifiers, it can be defined as longitudinal data. Each data point in longitudinal data is dependent because of the repeated measurements, time-related effects, individual variability and so on. For analyzing the longitudinal data, it is significant to choose suitable model to fit the data. Traditional linear regression model fails because of the lack of time variable as well as the assumption of independence. Firstly, linear regression model ignores the time component, it typically does not account for the temporal order of observations. However, in longitudinal data, the time component is crucial as it captures the dynamics and trends over time.

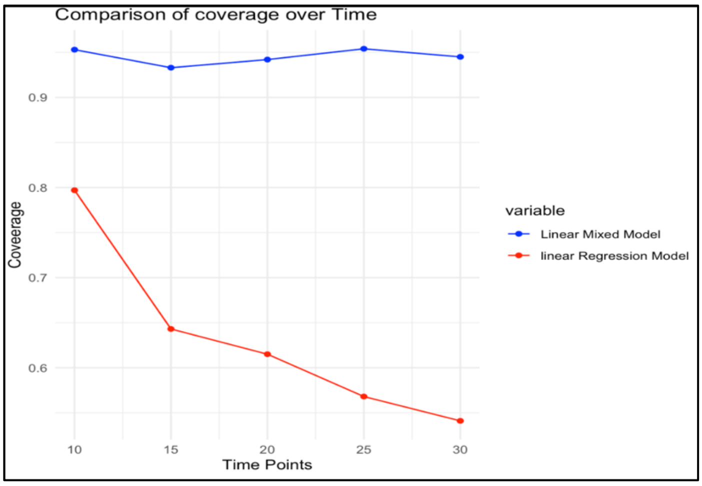
Additionally, the linear regression model assumes that each data point is independent to each other. On the contrary, in longitudinal studies, observations from the same individual over time are likely to be correlated. This violation of the independence assumption can lead to biased estimates and incorrect inference.

To furtherer explain, I did a data simulation between linear regression model and linear mixed model to find which model is more suitable. My simulated dataset comprises measurements from 100 subjects across 5 time points including 10, 15, 20, 25 and 30. Each subject is associated with a covariate X, drawn from a uniform distribution, to reflect the variation in real-life data collection. The core of our simulation lies in the model $Y_{it} = \beta_0 + \beta_1 X + \beta_2 t + b_{0i} + b_{1i} t + \epsilon_{it}$, where Y_{it} is the outcome for each subject i at time t. X is the covariate, b_{0i} is the random intercept, b_{1i} is the random slope, and ϵ_{it} is the random error. This model captures both the fixed effects of the covariate and time, and the random effects that vary by individual, thus allowing for the correlation within an individual's repeated measures.

By iterating this simulation process 1000 times and fitting both linear regression and linear mixed models to the generated data, I extracted and analyzed the bias in the parameter estimates (β_2) and the coverage probability of the confidence intervals.



Graph 1



Graph 2

From the first graph, each time point has a different biased β_2 in both models, and we can see that the biased value is very small, so it is a normal condition. However, for the second graph, we can clearly see that linear regression model has a relative smaller coverage ranged from 0.8 to 0.47. If we want the model to properly perform, the coverage should be around 0.95 and 95% confidence interval. For the linear mixed model, it has the coverage about 0.95. This provides a clear comparison of the models' performance, highlighting the necessity of using linear mixed models when dealing with data that exhibit inherent correlations. In this way, I choose to use linear mixed model to model the relationship in HIV incidence over time in HIV research.

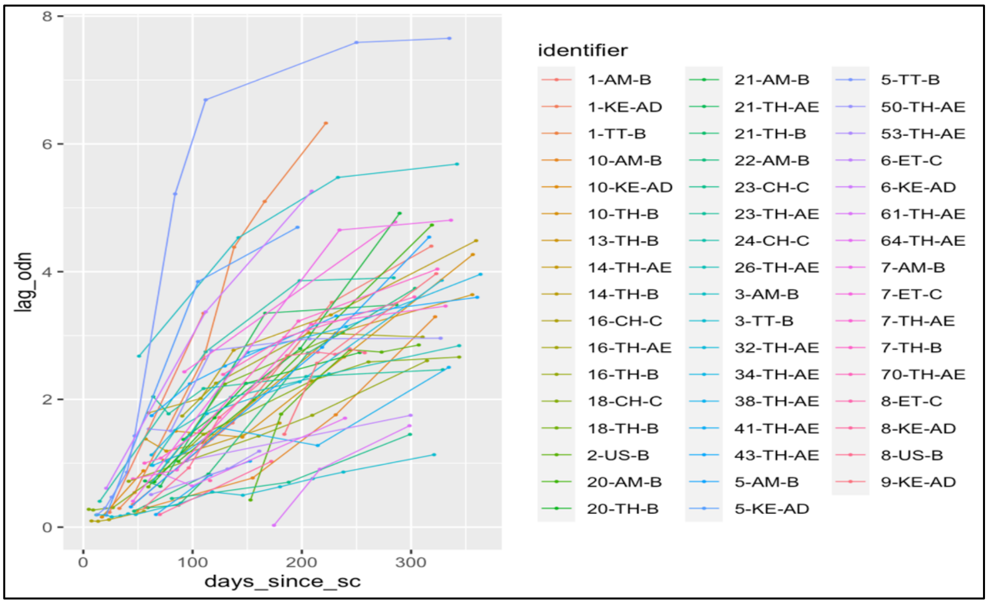
HIV Research

In my HIV research paper, I firstly recleaned my dataset through filtering the dataset to focus on data points collected within the first 365 days, ensuring relevance to the early phase of the timeline under investigation.

To gain insights into the structure of the dataset, I identified all unique case identifiers, and count the number of identifiers under study. Subsequently, the dataset was grouped by these identifiers to tally the number of observations recorded for each participant. The distribution of these counts was then visualized in a histogram, affording an immediate understanding of the frequency of data collection across participants. To make a targeted analysis and clear figure, I randomly 50 identifiers from the data to form the relationship between days since HIV seroconversion (as estimated by the midpoint between the last negative and first positive test) and LAg Avidity, the biomarker of interest.

Model and Results

The model I created is $LAgODn = \beta_0 + \beta_1 Days + b_0 + \epsilon$. I want to fit a model with the random slope b_1, but we run the computational issues, because of the data.



Graph 3

With the model, I found that the estimated coefficient for days_since_sc is approximately 0.0102. Also, the coverage is about 0.95. With 95% confidence, the true average increase in logged lag_odn per day is between approximately 0.0097 and 0.0107.

Conclusions

The findings of this study offer significant evidence of a progressive increase in LAg Avidity, as time since HIV infection increases. Through the application of a linear mixed model, which accounts for individual variability and inherent data correlation, we have ascertained that for every 100-day increment post-seroconversion, there is a notable increase in the lag Avidity levels. This increment is quantified at an average of 1.02 on the logarithmic scale, which translates to an exponential growth in the actual biomarker levels.

The model's precision, as reflected by the 95% confidence intervals, suggests a high degree of reliability in these estimates, situating the true average increase in the biomarker between 0.0097 and 0.0107 per day. This subtle yet consistent rise underscores the biomarker's potential utility in monitoring disease progression or response to therapy.

Furthermore, these results illuminate the dynamic nature of HIV as it interacts with the human immune system over time. Once someone gets HIV, the lag Avidity will increase over time, even if someone may use medical to control. HIV is still a serious and dangerous medical issue all over the world. The implications of these findings are profound, suggesting that continuous monitoring of biomarker levels could be integral to optimizing therapeutic strategies, assessing treatment efficacy, and potentially improving patient outcomes.

References

Chauhan, C. K., Lakshmi, P. V. M., Sagar, V., Sharma, A., Arora, S. K., & Kumar, R. (2020). Immunological markers for identifying recent HIV infection in North-West India. The Indian journal of medical research, 152(3), 227–233. https://doi.org/10.4103/ijmr.IJMR_2007_18

Gao, F., & Bannick, M. (2022). Statistical considerations for cross-sectional HIV incidence estimation based on recency test. Statistics in medicine, 41(8), 1446–1461. https://doi.org/10.1002/sim.9296

Global HIV &amp; AIDS statistics - fact sheet. UNAIDS. (n.d.). https://www.unaids.org/en/resources/fact-sheet#:~:text=Global%20HIV%20statistics,AIDS%2Drelated%20illnesses%20in%202022

HIV.gov. (2023, January 13). What are HIV and AIDS? https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids/