Choosing the right strategy to model longitudinal count data in Epidemiology: An application with CD4 cell counts

Daniele de Brito Trindade(1), Raydonal Ospina(1), Leila D. Amorim(2),*

ABSTRACT

BACKGROUND: Statistical models for analysis of correlated count data are important for answering epidemiological issues that involve taking individual count measurements repeatedly over time through the use of longitudinal studies. Conventional regression models for this type of data are inadequate and can lead to inappropriate conclusions and inference. Longitudinal studies in Public Health involve evaluation and monitoring of patients with infectious diseases, such as HIV/AIDS, to assess their health status, to check the effectiveness of the treatment, and to make prognosis about the evolution of the disease, including interdependencies of clinical manifestations. The purpose of this article is to describe various statistical strategies for the analysis of longitudinal count data with emphasis on how to choose the most suitable model for the data and in the interpretation of the results.

METHODS: We illustrate the applicability of various statistical strategies by evaluating the effect of associated factors on lymphocyte CD4+T cell count in HIV seropositive patients in Salvador, Bahia, Brazil. We describe the Poisson and Negative Binomial models using the multilevel (ML) approach and the generalized estimations equations (GEE) for the analysis of longitudinal count data.

RESULTS: The interpretations of the results derived from ML and GEE differ and thus their direct comparison should be avoided.

CONCLUSION: We believe that the statistical methods for the analysis of longitudinal studies with correlated count data can be useful to address several important issues in public health, especially in helping to monitor patients and checking the effectiveness of treatments.

Key words: Multilevel Models; Generalized Estimating Equations; Poisson Model; Negative Binomial Model; HIV-AIDS.

(1) Departamento de Estatística, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, Cidade Universitária, s/n, Recife, Pernambuco, 50670-901.
(2) Departamento de Estatística, Instituto de Matemática, Universidade Federal da Bahia, Av.Adhemar de Barros, s/n, Salvador, Bahia, 40170-110.

CORRESPONDING AUTHOR: Leila D Amorim, Departamento de Estatística, Instituto de Matemática, Universidade Federal da Bahia, Av.Adhemar de Barros s/n, Campus de Ondina, Salvador-Bahia, Brazil, CEP 40170-115 e-mail: leiladen@ufba.br tel: 55-71-3283-6344

DOI: ***10.2427/11520***
Accepted on December 12, 2015

INTRODUCTION

The use of statistical models for the analysis of correlated count data has grown in public health research. In fact, a recent review shows that from a selection of 108 articles in the medical field between 2000 and 2012 using generalized linear mixed models (GLMMs) 20.4% considered models for count data [1]. The most common models for count data include the Poisson and Negative Binomial (NB) distributions. Christofides and collaborators, for example, used the Poisson random effects model to predict the incidence of pregnancy due to sexual violence [2] and a hospital surveillance study used the NB model to identify the associations between year, seasons and rate of infections to evaluate infection by *Clostridium difficile* [3].

Longitudinal Studies (LS) are crucial in dealing with epidemiological issues as they trace the behavior of the response variable (outcome) over time, investigating the effect of covariates on the profile of the outcome, making predictions, and assessing global or individual changes in the response over time [4]. In biostatistics these studies are known as cohort studies, whereas in areas such as sociology, economics and business, they are referred to as panel studies. Data from LS have unique characteristics, including their temporal ordering and dependence between consecutive measurements [4].

Past literature has focused mainly on the description of models for correlated data of continuous or binary responses for applied researchers, referring particularly to its characterization and application using linear or logistic models [5-9]. The theoretical developments of Poisson and NB models have led to methodological extensions for analysis of longitudinal data in the recent past. These lastest innovations are currently available in statistical software. Nevertheless, they have not been widely used and disseminated by applied researchers, which is partially due to their technical complexity. At the same time, there is a consensus about the inappropriacy of using conventional regression models to fit correlated data, which would provide incorrect standard errors and, consequently, could lead to misleading inference and conclusions [4-9]. Therefore, there is a need for a unified framework to present and describe these methods making them easily accessible to researchers in the field of Public Health. This paper aims to present distinct modeling strategies for the analysis of longitudinal count data, explaining their use and limitations so as to promote a better understanding of the usefulness of these tools in answering scientific questions in Epidemiology. To illustrate this, we analyze data on the number of CD4+T lymphocytes repeatedly measured in HIV seropositive individuals in Salvador, Bahia, Brazil.

HIV/AIDS remains a global challenge and a major public health problem. The World Health Organization estimates that so far around 25 million men, women, and children have died from AIDS worldwide [10]. In Brazil, 544,846 AIDS cases were reported from the beginning of the epidemic until 2009. In the city of Salvador, the capital of the State Bahia in the Northeast of Brazil, 2,944 new cases were registered between 2000 and 2008. In the context of this epidemic, the assessment of the number of CD4+ T lymphocytes over time is important to monitor the history of HIV infection and its consequent progression to AIDS. The CD4+ T cell counts and the quantification of the viral load have been used both in the indication and evaluation of the need for modification of antiretroviral regimens [5]. Given the magnitude of this epidemic and the methodological challenges to implementing a more effective and robust analysis, we characterize different statistical strategies for analyzing longitudinal count data, illustrating their applicability and interpretation through the evaluation of the effect of factors associated with CD4+ counts in HIV patients.

METHODS

Data. The data refers to 587 HIV-seropositive patients in the city of Salvador, who were registered on the Laboratory Testing Control System (SISCEL, in Portuguese) of the Brazilian Ministry of Health between January 2002 and August 2012. This system was developed with the purpose of monitoring the laboratory procedures of lymphocyte T CD4/CD8 cell counting and to perform the quantification of HIV viral load both for treatment indication and for monitoring patients undergoing antiretroviral therapy.

The CD4+ cell counts have great intra and inter variability, specifically when values are above 200 cells/mm^3, hindering its identification in the early stages of the infection. So far there is no objective ideal value of the number of CD4+ cells to specify the beginning of the antiretroviral treatment for all patients because the rate of disease progression can vary widely among individuals. The viral load (VL) is defined as the number of virus copies in 1 milliliter of blood. Initial results in untreated patients can reach up to 1 million or more copies/ml. During treatment, high VL is between 5,000 and 10,000 copies/ml. A low VL (between 40 and 500 copies/ml) indicates slow disease progression [6]. Given the usual limits for the CD4+ count and VL, individuals who had CD4+ above 1,500 cells/mm³, either at baseline or during follow-up periods, and those with a VL greater than 1 million copies/mm at baseline were excluded from our analysis.

Statistical modeling considered the number of CD4+ cells as the outcome, which was measured at different points in time after receiving the antiretroviral therapy provided by the government program. The covariates include patient's gender (0 = male, 1 = female) and the following information at baseline: age (in years); a dummy variable treatment (0 = he/she was not in treatment before registration at SISCEL, 1 = he/she was in treatment); the categorized CD4+ in baseline (0 if CD4+ <350 cells/mm³, and 1 if CD4+ ≥ 350 cells/mm³) and categorized VL (0 if VL <500 copies/ml, 1 if VL between 500 (inclusive) and 5,000 copies/ml, and 2 if VL ≥ 5,000 copies/ml). The *time* variable, in years, was also included in the model to indicate when the CD4+ count was taken after registration at SISCEL. Furthermore, the individuals do not have the same number of repeated measurements (number of observations per patient ranged from 1 to 24) and the measurements were taken at different time points, i.e. the study is unbalanced and unequally spaced.

Statistical Models

Regression Models for Count Data. Count data are quite common in epidemiological studies. This type of data assumes only non-negative integer values (i.e. 0, 1, 2, …) and is usually modeled using the Poisson distribution, which is characterized by having equal mean and variance of the response variable (Y_i), hence, $E(Y_i) = Var(Y_i) = \mu_i$. However, when overdispersion is present, i.e. when $E(Y_i) < Var(Y_i)$, the Poisson model is no longer appropriate. In such situations, the Negative Binomial model can be used, and is denoted by $Y_i \sim NB(\mu_i, \mu_i + \alpha\mu_i^2)$, where α controls for the overdispersion [11]. To illustrate the use of these models, consider our CD4+ data. Let Y_i be the number of CD4+ cells recorded in the ith row of the dataset (i=1, 2, …, 8,072). Assuming that the observations are independent, the data can be described by a Poisson or by a NB model to evaluate the effects of the covariates on the CD4+ counts, which can be defined as:

$$log(\mu_i)=\beta_0 + \beta_1 x\ time + \beta_2 x\ treatment + \beta_3 x\ gender + \beta_4 x\ age + \beta_5 x\ CD4\_baseline + \beta_6 x\ VL\_dummy1 + \beta_7 x\ VL\_dummy2 \quad (1)$$

where μ_i denotes the mean number of CD4+ for the *i*th individual and 8,072 is the total number of observations, which refers to repeated measurements of 587 individuals in this study. The main difference between the Poisson and the NB models is the additional parameter and, consequently, the specification of the likelihood functions associated with them. The parameter estimation can be achieved via likelihood maximization by using a nonlinear optimization procedure [12].

Note that the traditional regression models for counting responses assume that the observations are independent. However, when clustered or longitudinal designs are used this assumption is no longer reasonable [13].

Models for Longitudinal Data. The models for longitudinal data are required when there are repeated measurements of the outcome for the same individual over time, which leads to a dependence structure in the data. The two approaches commonly used to analyze longitudinal data are the conditional and the marginal models [14]. One of the most important conditional models for longitudinal data is the linear mixed or multilevel model, in which the coefficients have an individual or cluster-specific interpretation. This model is conditional on random effects that describe the behavior of a response that varies for a specific individual. In marginal models on the other hand, the dependent variable (outcome) is modeled separately from the correlation between the measurements of each sample unit (denoted as intra-unit or intra-individual correlation). Consider a generic notation, where *m* individuals that are followed-up may have n_i repeated measures which can vary between individuals, and consider that index *i* denotes individuals and *j* indicates the observations. Using generalized estimating equations (GEE) as a marginal strategy, the expected marginal mean, $E(Y_{ij})$ $i=1,\ldots,m$ and $j=1,\ldots,n_i$, is modeled as a function of the explanatory variables [4]. The dependence structure among repeated measurements of the same individual are dealt

TABLE 1

WORKING CORRELATION MATRICES COMMONLY USED IN MARGINAL MODELS CONSIDERING AN EXAMPLE WITH THREE REPEATED MEASURES FOR ALL SUBJECTS

TYPES OF WORKING CORRELATION MATRIX	MATRICIAL FORM
Independent: Assume that correlations for distinct measurements of the same individual are zero. This form is not adequate for longitudinal studies because their data are generally highly correlated.	$R_i(\alpha)=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$
Exchangeable: Assume that correlations between all repeated measurements of the same individuals are equal.	$R_i(\alpha)=\begin{bmatrix}1&\alpha&\alpha\\\alpha&1&\alpha\\\alpha&\alpha&1\end{bmatrix}$
Autoregressive of order 1 (AR1): Assume that adjacent correlations are greater in magnitude. The intra-individual correlation over time is an exponential function of its length. For longitudinal data, this is the most parsimonious correlation structure because it depends on one single parameter and yet it enables the correlations to diminish over time.	$R_i(\alpha)=\begin{bmatrix}1&\alpha&\alpha^2\\\alpha&1&\alpha\\\alpha^2&\alpha&1\end{bmatrix}$
Unstructured: All *n(n-1)/2* correlations of R_i are estimated. This structure is more efficient and useful when there are only few time points. When there are several repeated measurements, the estimation of this structure is very complicated. Besides that, missing data makes it difficult to estimate R_i.	$R_i(\alpha)=\begin{bmatrix}1&\alpha_1&\alpha_2\\\alpha_1&1&\alpha_1\\\alpha_2&\alpha_1&1\end{bmatrix}$

with via the definition of the “working” correlation matrices, which are shown in Table 1.

Multilevel Approach. The simplest multilevel model for count data considers a single random intercept effect that represents the differences between the individuals. Let $X^t_{ij}= (X_{ij2}, X_{ij3},\dots,X_{ij(p-1)})$ be the covariate matrix, t_{ij} the time when the *j*th measure of the *i*th individual was taken, $\beta= (\beta_2, \beta_3, \dots, \beta_{(p-1)})^T$, and $b_{oi} \sim N(0, t_o)$ assumed to be independent of X^t_{ij}. Then, the linear predictor is given by:

$$\log(\mu_{ij})= n_{ij} =y_{0i}+y_{1i}t_{ij}+ X^t_{ij}\beta, \qquad (2)$$

where
$$y_{0i} = \beta_{0} + b_{oi}$$
$$y_{1i} = \beta_1$$

The model parameters are estimated by maximum likelihood using iterative methods such as the Fisher Scoring or the Newton-Raphson [13].

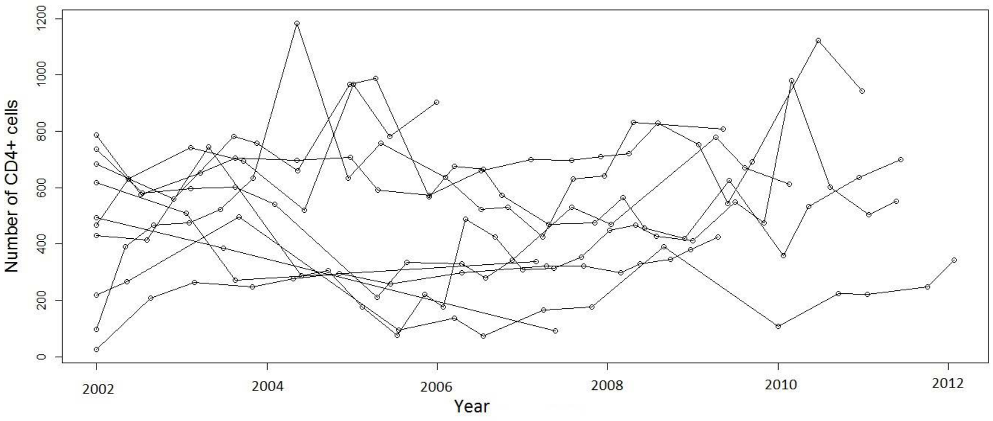
In many practical situations, it is reasonable to assume not only an average per individual, but also that the effects on repeated measurements of the response are dependent on random effects. Therefore more complex multilevel models that include two random effects can be suitable. For instance, the random coefficients given in (2) can be defined as: $y_{0i}=\beta_0+\beta_{0\iota}$ and $Y_{1i}= \beta_\iota +b_{1i}$, where the random vector $b_i=[b_{0i}\, b_{1i}]$ follows a bivariate normal distribution with mean 0 and covariance matrix $\begin{bmatrix}\tau_0&\tau_{01}\\\tau_{10}&\tau_1\end{bmatrix}\begin{bmatrix}\tau_0&\tau_{01}\\\tau_{10}&\tau_1\end{bmatrix}$. Again, upon Poisson or NB distributional assumptions for the response variable the parameters are estimated iteratively using the Newton-Raphson algorithm to maximize the likelihood function. The random effects can in theory take any probability distribution. However, for ease of computation, control and robustness of inferential processes the statistical packages restrict its use to particular cases.

Model selection is based on consistent Akaike information criterion (CAIC) defined as CAIC = -2L + *p*[log(*mn*)+1], where *L* is the log-likelihood function, *n* is the average number of repeated measurements, and *p* is the number of parameters [15]. The model with the lowest CAIC is chosen.

Marginal Approach. GEE are extensions of GLMM’s for correlated data and require only the correct specification of univariate marginal distributions provided one is willing to adopt a “working” correlation matrix [11]. The linear predictor is specified as $n_{ij}=Z^T_{ij}\beta^*$, where $\beta^*= (\beta_o,\beta_1,\dots,\beta_{(p-1)})^T$ is a *p*-dimensional

FIGURE 1

INDIVIDUAL PROFILES FOR CD4+ COUNT IN 10 RANDOMLY SELECTED INDIVIDUALS. SALVADOR-BA, 2002-2012



vector of fixed parameters associated with the covariate vector $Z^T_{ij=}(1,t_{ij},X_{ij2},X_{ij3},\ldots,X_{ij(p-1)})$. A link function that relates the marginal mean to the linear predictor is specified. In the case of the Poisson and NB distributions the canonical link function is the logarithm (log), i.e., $\mu_{ij}=exp\ (Z^T_{ij}\,\beta^*)$. In this approach, the variance is written as a function of the mean [13]. The estimates of β are obtained by the Newton-Raphson iterative method. Model selection is carried out based on the criterion of quasi-likelihood under the independence model (QIC) [16]. This criterion compares models with different correlation structures, such that the smallest QIC identifies the best model [17].

Computational Support. We used software R version 3.2.0[18] and Stata[19] version 10 to implement the methods described here. Details on syntax are presented in the Appendix.

RESULTS

In this study 63% of HIV seropositive individuals were males, 90% were under treatment at baseline and the average age of patients was 38 years (1 to 83 years). The mean follow-up was 4.6 years (3 months to 10.6 years). At baseline, 59% of patients had CD4 counts below 350 cells/mm³ and 64% had VL above 5,000 copies/ml. Overdispersion was detected in the data, indicating NB as the most appropriate model. However, for comparison and illustration of the methods described here, we present results for the NB and Poisson models.

The individual profiles graph for 10 randomly chosen patients is shown in Figure 1. Analyzing information displayed in Figure 1 we can gain insights regarding the variability between individual units at a given point in time; the variance within units over time; and the trends over time. Note that the space between the lines represents between unit variability and the change in each line (slope) represents within variability. We observe a wide variability in the number of CD4 and in the number of repeated measurements.

The relative risk estimates using GEE-Poisson and NB models, with different correlation structures, are presented in Table 3. According to the QIC, the best marginal model to fit this data is the NB with exchangeable correlation structure. It can be observed that patients with CD4 + counts above 350 cells/mm³ at baseline had a mean number of CD4 cells which was 43% greater than those with counts below that (RR=1.427; 95%CI=1.326-1.552). Those patients who were undergoing treatment at baseline had an average number of CD4+ 38.0% greater than patients who were not undergoing treatment, controlling for the other covariates in the model (RR=1.379; 95%CI=1.172-1.614). It is important to highlight that the interpretation of the estimates from Poisson and NB models are similar when using the same modelling strategy (marginal or conditional).

Figure 2 presents the estimated trajectories

TABLE 2

RELATIVE RISK ESTIMATES FOR ANALYSIS OF LONGITUDINAL CD4+ CELL COUNTS IN HIV-SEROPOSITIVE INDIVIDUALS USING MULTILEVEL POISSON AND NEGATIVE BINOMIAL MODELS. SALVADOR-BAHIA. 2002-2012

ASSOCIATED FACTORS	POISSON MODEL		NEGATIVE BINOMIAL MODEL	
	INDEPENDENT	MULTILEVEL MODEL WITH A RANDOM INTERCEPT (WITH GAMMA DIST.)	INDEPENDENT	MULTILEVEL MODEL WITH A RANDOM INTERCEPT (WITH BETA DIST.)
	RR (CI 95%)	RR (CI 95%)	RR (CI 95%)	RR (CI 95%)
TIME	1.033* (1.033;1.034)	1.030* (1.029;1.030)	1.040* (1.032;1.041)	1.023* (1.022;1.028)
TREATMENT	1.181* (1.174;1.184)	1.434* (1.233;1.664)	1.169* (1.103;1.236)	1.391* (1.245;1.545)
GENDER				
MALE	1.000	1.000	1.000	1.000
FEMALE	1.076*	1.049	1.083*	1.089*
	(1.074;1.078)	(0.955;1.150)	(1.046;1.108)	(1.027;1.156)
AGE IN BASELINE	1.000* (1.001;1.001)	1.000 (0.995;1.003)	1.000 (0.999;1.002)	1.021* (1.017;1.023)
CATEGORIZED CD4+ COUNT				
< 350 CELLS/MM³	1.000	1.000	1.000	1.000
≥ 350 CELLS/MM³	1.372*	1.429*	1.362*	1.501*
	(1.371;1.377)	(1.308;1.570)	(1.348;1.425)	(1.415;1.586)
VIRAL LOAD				
VL < 500 COPIES/ML	1.00	1.00	1.00	1.00
500≤ VL<5,000 COPIES/ML	0.923* (0.922;0.928)	0.978 (0.843;1.145)	0.942* (0.892;0.980)	0.951 (0.881;1.035)
VL≥5,000 COPIES/ML	1.000* (1.003;1.007)	1.099 (0.903;1.124)	1.099 (0.975;1.045)	0.901* (0.850;0.957)
CAIC	1,313,136.0	513,370.4	113,784.8	109,053.4

*p-value < 5%

Abbreviations: RR, relative risk; CI, confidence interval; CAIC, Consistente Akaike Information Criteria.

for the average number of CD4+ in accordance with the estimates obtained by GEE-NB model with exchangeable correlation structure for four patients with specific profiles for a period of 10 years. The first patient was not receiving treatment at baseline, male, 70 years old, with CD4 + at baseline below 350 cells/mm³ and VL above 5,000 copies/ml (Patient 1). The second patient also was not undergoing treatment at baseline, female, 35, CD4+ at baseline above 350 cells/mm³ and VL between 500 copies/ml and 5,000 copies/ml (Patient 2). The third patient was undergoing treatment, male, 50 years old, baseline CD4+ below 350 cells/mm³ and VL between 500 copies/ml and 5,000 copies/ml (Patient 3). The fourth patient was undergoing treatment, female, 20, CD4+ at baseline above 350 cells/mm³ and VL less than 500 copies/ml (Patient 4).

The prediction equation is given as:

$$\widehat{log(\mu_i)} = 5.6173 + 0.0325 \ x \ time + 0.3189 \ x \ treatment + 0.0476 \ x \ gender - 0.0009 \ x \ age + 0.3608 \ x \ CD4\_baselina + 0\text{-}0126 \ x \ VL\_dummy1 + 0.0041 \ x \ VL\_dummy2$$

As expected, all four individuals had increasing average numbers of CD4+ cells over time according to their predicted individual profiles (Figure 2). Patient 1 had the worst performance. Interestingly, even though patient 2 was not undergoing treatment at baseline, she performed better than patient 3. This is due to other characteristics of these individuals,

TABLE 3

RELATIVE RISK ESTIMATES FOR ANALYSIS OF LONGITUDINAL CD4+ CELL COUNTS IN HIV-SEROPOSITIVE INDIVIDUALS USING GEE- POISSON AND NEGATIVE BINOMIAL MODELS. SALVADOR-BAHIA. 2002-2012

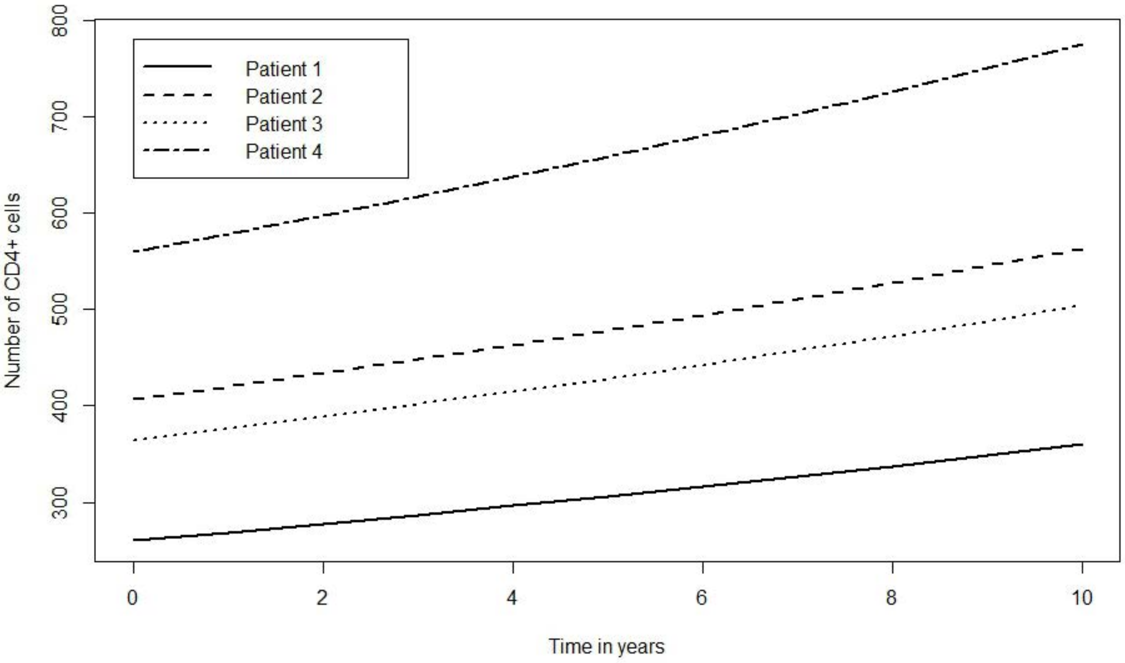
ASSOCIATED FACTORS	GEE-POISSON MODEL		
	INDEPENDENT	AR1	EXCHANGEABLE
	RR (CI 95%)	RR (CI 95%)	RR (CI 95%)
TIME	1.033* (1.033;1.034)	1.031* (1.021;1.036)	1.029* (1.025;1.037)
TREATMENT	1.179* (1.174;1.184)	1.193* (1.026;1.380)	1.482* (1.262;1.746)
GENDER			
MALE	1.000	1.000	1.000
FEMALE	1.076*	1.052	1.078
	(1.074;1.078)	(0.970;1.141)	(0.988;1.177)
AGE IN BASELINE	1.000* (1.000;1.001)	1.000 (0.996;1.005)	1.000 (0.997;1.005)
CATEGORIZED CD4+ COUNT			
< 350 CELLS/MM^3	1.000	1.000	1.000
≥ 350 CELLS/MM^3	1.371*	1.369*	1.324*
	(1.371;1.371)	(1.272;1.480)	(1.210;1.433)
VIRAL LOAD			
VL < 500 COPIES/ML	1.000	1.000	1.000
500≤ VL < 5,000 COPIES/ML	0.927* (0.922;0.928)	0.927 (0.815;1.055)	0.951 (0.824;1.101)
VL≥5,000 COPIES/ML	1.099* (1.003;1.007)	0.992 (0.909;1.086)	1.099 (0.914;1.113)
QIC		44,164,807.9	44,171,363.8
ASSOCIATED FACTORS	GEE-NEGATIVE BINOMIAL MODEL		
	INDEPENDENT	AR1	EXCHANGEABLE
	RR (CI 95%)	RR (CI 95%)	RR (CI 95%)
TIME	1.041* (1.029;1.044)	1.032* (1.024;1.040)	1.031* (1.027;1.039)
TREATMENT	1.169* (1.066;1.279)	1.182* (1.012;1.373)	1.379* (1.172;1.614)
GENDER			
MALE	1.000	1.000	1.000
FEMALE	1.082*	1.048	1.051
	(1.028;1.128)	(0.971;1.144)	(0.965;1.142)
AGE IN BASELINE	1.000 (0.999;1.003)	1.000 (0.996;1.005)	1.000 (0.995;1.003)
CATEGORIZED CD4+ COUNT			
<350 CELLS/MM^3	1.000	1.000	1.000
≥ 350 CELLS/MM^3	1.389* (1.325;1.449)	1.401* (1.295;1.506)	1.427* (1.326;1.552)
VIRAL LOAD			
VL<500 COPIES/ML	1.000	1.000	1.000
500≤ VL < 5,000 COPIES/ML	0.943 (0.867;1.008)	0.939 (0.822;1.065)	0.981 (0.846;1.123)
VL≥5,000 COPIES/ML	1.099 (0.956;1.067)	0.992 (0.912;1.092)	0.999 (0.915;1.102)
QIC		116,849.7	116,829.4

*p-value < 5%

Abbreviations: RR, relative risk; CI, confidence interval; QIC, Quasi-likelihood Information Criteria.

FIGURE 2

PREDICTION OF FOUR INDIVIDUAL PROFILES FOR CD4 COUNT FOR A PERIOD OF 10 YEARS ACCORDING TO GEE-NB MODEL



especially for their CD4+ counts at baseline.

It is worth noting that the interpretation of the results from marginal and conditional models differs. Although both approaches model the average number of CD4+ cells, the marginal model has a population-average interpretation while the multilevel model, being conditional on random effects, provides an individual interpretation. Therefore, the results from these models should not be directly compared.

An important step in fitting a regression model is the verification of possible departures from the assumptions of the model. This diagnostic analysis is usually performed through residual analysis. However, due to the complexity of our data structure (unbalanced and unevenly spaced) there were no diagnostic methods available in R or Stata. Thus, their implementation represents a challenge for future work.

DISCUSSION

Analysis of longitudinal data using conventional regression models is inadequate as they fail to consider the dependence between observations over time. Longitudinal data may also present additional complexities in its structure, which may occur due to the imbalance and/or the fact that they are unevenly spaced, or owing to missing data. It is up to the data analyst to conduct a thorough exploratory analysis to evaluate the data structure and choose the statistical model that best suits it. For the analysis of count data in particular, the two most widely used statistical models are the Poisson and NB models. The choice depends on characteristics inherent to the data, for example the NB model is appropriate when overdispersion is suspected [20,21]. The parameter estimates based on NB are not very different from those based on the Poisson model. However, the Poisson regression underestimates the standard errors when overdispersion is present, leading to inappropriate inference. A simple way to choose between these two models is to compare them based on some criteria, such as AIC, CAIC or QIC, depending on the adopted modeling strategy. Another way is to estimate the scale parameter from NB and to test the null hypothesis that it is equal to zero.

The choice of the modeling strategy

depends on the purpose of the study, especially because the results from these models can lead to different interpretations. The GEE estimates the regression coefficients as in a cross-sectional study, modelling the population-average response and, separately, it models the correlation between two observations from the same individual at two different points in time. The multilevel model, on the other hand, deals with the regression coefficients and the intra-subject correlation simultaneously in a single equation, in which the response is modeled as a function of the covariates and random effects[14]. These methods are complex and some of them are still under development. To date we know of no implementation of regression diagnostic methods for very complex data structures as the described in our application. It should also be noted that the distribution associated to the random effects varies according to the statistical software.

Despite the limitations and methodological complexity, the use of LS with counts as responses is important in epidemiological studies, as is the case of our application. Other authors have monitored and evaluated the natural history of HIV using repeated measurements of CD4+, which were analyzed by using the multilevel linear model. To consider this type of modeling, an alternative embodiment is to use the CD4+ percentage as the outcome rather than a count [22] or using a transformation to the cell count (for example, square root of the number of cells) [23] so that the assumptions of normality and homoscedasticity of the errors are fulfilled. We implemented these strategies on our data but we found evidence of violation of the model assumptions. An additional limitation of using linear mixed models for longitudinal count data is that they do not enable the estimation of measures of association such as the relative risk. In addition to allowing the specification of a dependency structure between multiple measurements on the same individual, one of the advantages of the models for longitudinal count data described in this article is the possibility of including both fixed covariates, such as race or sex, and time-dependent covariates, such as type of treatment regimen or number of infections in the last quarter.

The methods described in this work enable the description of the impact of several factors on lymphocyte CD4+ counts in HIV-seropositive patients using all available information. We believe that this type of analysis can be useful to address several important issues in public health as well as help in monitoring patients and checking the effectiveness of their treatments.

ACKNOWLEDGEMENTS: *This work was supported in part by Coordenação de Aperfeiçoamento de Pessoal do Nível Superior (CAPES)/Ministry of Education-Brazil (scholarship to D.B.T.), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grant 478556/2010-1 to L.D.A and grant 470810/2013-0 to R.O.) and Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE) (grant APQ-0682-1.02/14). We are grateful to the Department of STD, AIDS and Viral Hepatitis of the Brazilian Health Ministry for allowing the use of their anonymous data, and particularly to Silvano B. Oliveira for his assistance during the project.*

CONFLICT OF INTEREST: *none declared.*

APPENDIX: COMPUTATIONAL SYNTAX

The **R** software refers to a language and an integrated development environment for statistical calculations and graphics, being freely distributed and available at www.r-project.org[21]. **Stata** software is a statistical program that was developed in C and released in 1985[22]. The most current version of **Stata** is 14. There are available versions of **R** and **Stata** for Windows, Macintosh, Linux and Unix.

Software R

To fit the multilevel Poisson models one can used the **lme4** library through the **glmer** function. Particularly to fit the random intercept Poisson multilevel model it is necessary to consider the argument (1 | id). The syntax is:

glmer(CD4 ~ time + treat + gender + age + CD4_baseline + factor (VL_baseline) + (1|id), data=database, family=poisson)

On the other hand, for fitting the multilevel model with two random effects, being associated to the intercept and time variable, it is necessary to use the argument (time | id). In this case, the syntax is:

glmer(CD4 ~ time + treat + gender + age + CD4_baseline + factor (VL_baseline)+ (time|id), data= database, family=poisson)

The Negative Binomial multilevel models are implemented using the **glmmADMB** library through the function **glmmadmb**. The syntax is similar to the previous one, substituting the name of the function and the argument concerning the distribution to:

family="nbinom".

For the random intercept Poisson and NB models implemented in R, b_{0i} follows a Normal distribution. For multilevel models with two random effects the joint distribution of random effects b_{0i} and b_{1i} is considered to be bivariate normal in the software **R**.

To fit GEE using the Poisson distribution, the **geepack** library can be used along with **geeglm** function. The syntax considering the autoregressive (AR1) correlation structure is:

geeglm(CD4 ~ time + treat + gender + age + CD4_baseline + factor (VL_baseline), data = database, family = poisson, id = id, corstr = "ar1")

For other correlation structures one should only change the argument *corstr* for "*exchangeable*" or "*unstructured*". To date this function does not support the fit of NB models.

Software Stata

It is possible to fit the multilevel Poisson and NB models using **xtpoisson** and **xtnbreg** commands, respectively. The default distribution for the random intercept for multilevel Poisson model is the Gamma distribution, but one can alter that to normal distribution using the **normal** argument. The syntax for fitting the corresponding models is:

xi: xtpoisson CD4 time treat gender age CD4_baseline i.VL_baseline, nolog re i(id)

xi: xtpoisson CD4 time treat gender age CD4_baseline i.VL_baseline, nolog normal re i(id)

For multilevel NB models, the default distribution of the random intercept is Beta. The syntax is given by:

xi: xtnbreg CD4 time treat gender age CD4_baseline i.VL_baseline, nolog re i(id)

Regarding the GEE approach, the functions used to fit the NB and Poisson models are, respectively, **xtgee** and **xtnbreg**. The syntax using AR1 correlation structure is:

xi: xtgee CD4 time treat gender age CD4_baseline i.VL_baseline, nolog fam(poisson) corr(ar1) i(id) *xi: xtnbreg CD4 time treat gender age CD4_baseline i.VL_baseline, nolog pa corr(ar1) i(id)*

Other options for *corr* argument are *exch*, *ind* and *unstr*. The QIC can be calculated after installation of the function QIC.do. The syntax associated with the GEE- AR1 Poisson model, for example, is:

qic CD4 time treat gender age CD4_baseline i.VL_baseline, nolog eform fam(poisson) corr(ar1) i(id)

References.

[1] Casals M, Girabent-Farrés M, Carrasco JL. Methodological Quality and Reporting of Generalized Linear Mixed Models in Clinical Medicine (2000–2012): A Systematic Review. PLoS ONE. 2014;9(11):e112653.

[2] Christofides NJ, Jewkes RK, Dunkle KL, et al.,. Perpetration of physical and sexual abuse and subsequent fathering of pregnancies among a cohort of young South African men: a longitudinal study. BMC Public Health, 2014;14: 947.

[3] Faires MC, Pearl DL, Ciccotelli WA, Berke O, Reid-Smith RJ, Weese JS. Detection of Clostridium difficile infection clusters, using the temporal scan statistic, in a community hospital in southern Ontario, Canada, 2006–2011. BMC Infectious Diseases. 2014;14: 254.

[4] Diggle, PJ, Heagerty, P, Liang, KY., Zeger, S. Analysis of Longitudinal Data. 2nd Ed., New York: Oxford university Press; 2002.

[5] Burton P, Gurrin L, Sly P. Tutorial in Biostatistics. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Statistics in Medicine. 1998;17;1261-91.

[6] Goldstein H, Browne W, Rasbash J. Tutorial in Biostatistics. Multilevel modelling of medical data. Statistics in Medicine, 2002;21:3291-315.

[7] Merlo J, Chaix B, Ohlsson H, et al., A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. Journal Epidemiology Community Health. 2006; 60:290-7.

[8] Twisk, JWR. Longitudinal data analysis. A comparison between generalized estimation and random coefficient analysis. European Journal of Epidemiology, 2004;19:769-76.

[9] Bouwmeester W, Twisk, JWR, Kappen, TH, Wilton A van KW, Moons KGM, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Medical Research Methodology. 2013;13:19.

[10] UNAIDS. Global Report. UNAIDS report on the global AIDS epidemic; 2013.

[11] Liang K., Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13-22.

[12] Hilbe JM. Negative binomial regression. 2nd ed. New York: Cambrigde University Press; 2011.

[13] Hedeker D, Gibbons RD. Longitudinal data analysis. Ed., New Jersey: John Wiley & Sons; 2006.

[14] Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Longitudinal data analysis. Chapman & Hall/CRC; 2009.

[15] Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987;52:345-70.

[16] Pan W. Akaike's information criterion in generalized estimating equations. Biometrics. 2001;57:120-25.

[17] Cui J. Information Criterion and computation for GEE model selection, with an application to a longitudinal study of skin cancer. Biometrics: Methods, Applications and Analysis. 2010;167-75.

[18] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2015.

[19] Stata StataCorp. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP, 2007.

[20] Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34:1-14.

[21] Heilbron D. Zero-altered and other regression models for count data with added zeros. Biometrical Journal. 1994;36:531-47.

[22] Paintsil D, Ghebremichael M, Sostena RS, Andiman WA. Absolute CD4+ T-Lymphocyte Count as a Surrogate Marker of Pediatric HIV Disease Progression. NIH Public Access. 2008;27:629-35.

[23] Reda AA., Biadgilign S, Deribew A, Gebre B, Deribe K. Predictors of Change in CD4 Lymphocyte Count and Weight among HIV Infected Patients on Anti-Retroviral Treatment in Ethiopia: A Retrospective Longitudinal Study. PLoS ONE. 2013;8(4): e58595.

*