De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen: de Universiteit Hasselt en Maastricht University

UNIVERSITEIT VAN DE TOEKOMST

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Use of Zero-Inflated Models for analyses of immunological data

Promotor :
Prof. dr. Geert MOLENBERGHS

Promotor :
Ms. DOROTHÉE MERIC

Aklilu Zemicael Welegebrael

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Use of Zero-Inflated Models for analyses of immunological data

Promotor :
Prof. dr. Geert MOLENBERGHS

Promotor :
Ms. DOROTHÉE MERIC

Aklilu Zemicael Welegebrael

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Interuniversity Institute for Biostatistics and statistical
Bioinformatics (I-BioStat)

Use of Zero-Inflated Models for Analysis of Immunological Data

By

Aklilu Zemicael Welegebrael

Internal Supervisor *Prof.dr.Geert Molenberghs*
External Supervisor *Ms. Dorothee Meric*

Thesis submitted in partial fulfilment of the requirements for the degree of
Master of Science in Biostatistics
September 19, 2011

Certification

This is to certify that this report was written by Aklilu Zemicael Welegebrael under our supervision

Aklilu Zemicael Welegebrael (Student)

…………………………………………………**Date**……………………

…………………………………………………… **Date**…………………………

Professor Dr. Geert Molenberghs **Internal Supervisor**

……………………………………………… **Date**……………………….

Ms. Dorothee Meric **External Supervisor**

Acknowledgments

First of all I would like to thank to God as a grace of him, I could attain this Biostatistics-ICP programme. I would like to express my deepest thanks and respect to my internal supervisor Professor dr. Geert Molenberghs, for his impressing and motivating advice, constructive and instructive ideas. I was really motivated and happy when I am advised by him.

I would like also to extend my deepest thanks to my external supervisor Ms. Dorothee Meric, Human Papilloma Virus (HPV) vaccine coordinator, Glaxo Smith Kline(GSK), for her kind help, continual suggestions, encouragement and comments. I would like to extend my gratitude to the GSK staff members who gave me valuable comments on my work: dr.Fabian Santiago Tibaldi, Mr. Mohamed El Idrissi, and Ms. Yang Feng.

I would like to thank also the VILR-UOS scholarship; it is the basement of mine to be here and attend the Biostatistics–ICP programme. I would like to express my gratitude also to all Censtat staff members.

At last not least my sincere thanks are extended to all my beloved and cooperative families and friends for their continuous encouragement.

September19, 2011

Diepenbeek, Belgium

Contents

Abstract

The immune system is used to recognize and fight foreign agents that invade the body. It detects the pathogen and acts as the first line of defence, clearing the majority of microbial assaults. In the clinical trial study which is conducted by GSK, 213 consented individuals participated and they followed for about two to six visiting measurements after they had taken either of the two Human Papilloma Virus (HPV) vaccines at study start. The main objective of the study was to investigate the statistical method which can be used to compare the B-cell responses over time after the HPV vaccines administered in the presence of zero inflated data and to assess the effect of the two HPV vaccines.

In this study there were excess zeros, about 30.65% of the responses have zero value. The variance of the data was much higher than its mean that is, about 2574 times larger than its mean. Several count models were fitted to select the model which best fits the data, these are: Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB), Poisson mixed, NB mixed, ZIP mixed and ZINB mixed regression models. Each of these models was compared by likelihood ratio test (LR) and the information criteria's and it was found that the ZINB mixed model was the best.

Moreover, in this study it was found that NB, ZINB, NB mixed and ZINB mixed regression models were better fitted the data than Poisson, ZIP, Poisson mixed and ZIP mixed. Some observations were found to be potential outliers; however in this study similar result was found before and after excluding the outlying observations. In both models, treatment (HPV-vaccine) and the interaction between treatment with linear and quadratic time effect were found to be significantly associated with the production of B-cells.

In conclusion the zero-inflated negative mixed models with correlated random intercept fits best the data.

Key words*: B-cells; HPV; Poisson regression model; Negative binomial (NB) regression model; Zero-inflated Poisson (ZIP) regression model; Zero-inflated negative binomial (ZINB)regression model; Poisson mixed regression model; NB mixed regression model; ZIP mixed regression model; ZINB mixed regression model.*

1.0. Introduction

The immune system uses an innate and adaptive immunity to recognize and fight foreign agents that invade the body, such as bacteria, fungi, and viruses. The innate immune system detects the pathogen and acts as the first line of defence, clearing the majority of microbial assaults. It has no specific memory, but is responsible for activating adaptive immunity. The adaptive immune response generates exquisitely specific lethal effect or responses to foreign antigens as well as long-lived cells with memory of the insult. Antibody-mediated humoral immunity such as B lymphocytes (or B-cell) and T lymphocytes (or T-cells), clears virus particles from body fluids and can prevent viral re-infection, while cell-mediated immune responses are essential for the clearance of virus-infected cells and the generation of immune memory [1].

Human Papilloma Virus (HPV) is the name for a group of viruses that includes more than 100 types. Of which more than 40 of them can be transmitted through sexual contact. The types of HPV that infect the genital area are called genital HPV. More than half of sexually active individuals will have the virus at some point in their lives; however most people never know it since HPV often exhibits no symptoms and goes away on its own. Genital HPV is the most common sexually transmitted infection (STI) in the United States. About 20 million Americans aged from 15 to 49 currently have HPV at least half of all sexually active men and women get genital HPV at some time in their lives [2].

Infection with a high-risk type of HPV is considered necessary for the development of cervical cancer, but by itself it is not sufficient to cause cancer because the vast majority of women with HPV infection do not develop cancer. Cervical cancer is the leading cause of cancer mortality among women in developing countries. Approximately 500,000 new cases of cancer are estimated, leading to about 239,000 deaths each year. More than 99% of cervical cancer cases are linked to genital infection with HPV, which is the most common viral infection of the reproductive tract worldwide and infects an estimated 660 million people. While HPV infection resolves spontaneously in the majority of people, it can develop into chronic infection and, in some women, cervical cancer. The disease represents a major health inequity, as 80% of cervical cancer victims live in developing countries. However, developed countries have greatly reduced deaths from cervical cancer through screening programmes that allow early detection and treatment. These programmes are expensive and difficult to implement in low-income (developing) countries. The peak incidence of HPV infection occurs in adolescents and young women, while cervical cancer typically follows 20 to 30 years later [3].

The prevalence of HPV infection is highest in Africa. Among women with HPV infection, compared to HPV-positive women in Europe, HPV-positive women in Africa are relatively less likely to be infected with HPV-16, but they are more likely to be infected with the other types of HPV. Vaccines against HPV infections have the potential to be a more practical and cost–effective way to reduce the incidence of cervical cancer [3].

This paper is organized as follows. In Section 1.1 and Section 1.2, the objective and description of the data will be presented respectively. In Section 2.0, we will explain the methods which will be used to analyse the data. The results will be presented in Section 3.0. At last, the discussion and conclusion will be discussed in Section 4.

1.1. Objective

The objective of this paper is to investigate a statistical method which can be used to compare the B-cell responses over time after administration of two HPV vaccines, in presence of zero inflated data and also to examine the vaccine effect. This will be done through the selection of an appropriate model out of several count data models.

1.2. Data description

This data comes from a GSK clinical trial and it was collected one month after the last vaccine administered and then every six months for the next four years. There are 213 individuals who consented to receive either of the vaccines. The response variable is the number of B-cells produced per million of cells and the covariates are time of measurement and treatment received by the subject. The measurements take place every six months (that is in the first, seventh, twelfth, eighth, twenty fourth, and the thirty sixth months). Note that there was no measurement on month thirty. About 30.65% of the responses have zero values (i.e. no production of B-cells per million cells). Even though the design is a balanced design, the data is an imbalance data due to the missingness. 920 observations were observed out of the 1278 intended observations leading to 358 missed observations.

2.0. Methodology

2.1. Exploratory data analysis

An exploratory data analysis (EDA) was carried out to explore the data. To this effect various data exploration techniques were used such as: individual profile, mean and median evolution plots in order to see how the number of B-cells evolves over time. The variance structure was also used to see how the measurements of the B-cells vary over time. In addition, a scatter plot was used to explore the correlation structure of the measurements at the different time points and to identify the potential outliers in the data.

2.2. Statistical methodology

Even though there are several statistical models, some models may not be appropriate to deal with some specific types of data. Their use is solely depending on the types and nature of the data. In this study, the variable of interest is a count data, which is most often characterized as non-normal distribution. We will discuss statistical methods which can be used to model count data in the next subsections.

2.2.1. Poisson regression model

The standard Poisson distribution is a fundamental distribution to understand regression counts models. It was developed to model discrete count data, since it is easy to interpret in many aspects. According to [4], the apparent simplicity of Poisson comes with two restrictive assumptions. First, the variance and mean of the count variable are assumed to be equal. In reality, however, the variance is usually much larger than the mean. Although Poisson regression models are widely used to handle count data, it may not be well suitable to handle some types of count outcomes such as an over dispersed or under dispersed data. The other restrictive assumption of Poisson models is that occurrences of the event are assumed to be independent of each other.

Poisson regression assumes a Poisson distribution, characterized by a substantial positive skewed with variance equals mean. It tends to fit such data better than the linear regression model. However, if the variance is larger than the mean, it induces deflated standard errors and inflated the standardized normal (i.e. Z-normal) value, resulting in Type I errors and these makes Poisson regression less adequate [5]. Some researchers suggest that, when there is an overdispersion which is not a rise from an excess zeros, it is better to use other models, such as negative binomial which can taker of the overdispersion problem [6].

Let Y be a random variable, which has a Poisson distribution. Its density function is given by:

$$f(Y=y)=\frac{\lambda^{y}e^{-\lambda}}{y!}, \quad y=0,1,2,\dots \quad (1)$$

With $E(Y)=\mathrm{var}(Y)=\lambda$ where y is the realization value of the random variable Y [7]. When there are covariates associated with the parameters, it will be related with covariates by the natural logarithmic link function which leads to Poisson regression model. Suppose: $Y_i \sim Poisson(\lambda_i)$, $\ln(\lambda_i)=X_i^T\beta$, $X_i=(1,x_{i1},x_{i2},.....,x_{ip-1})^T$ is a $p x1$ vector of explanatory variable of the i^{th} observation and $\beta=(\beta_0,...,\beta_{p-1})^T$ is $p x1$ vector of regression parameters. In our case it will be defined as:

$$\ln(\lambda_i)=\beta_0+\beta_1 treat_{ik}+\beta_2 time_i+\beta_3 treat_{ik}time_i+\beta_4 time_i^2+\beta_4 treat_{ik}time_i^2, \quad (2)$$
$$i=1,2,...,920 \quad k=1,2$$

$$treat_i=\begin{cases}1, & if\ the\ i^{th}\ individual\ \ recieved\ \ group\,II\\ 0, & otherwise\end{cases}$$

And $time_i$ is the time point in months at which Y_i is measured, that is $time_i=1,7,12,18,24,36.$ $\beta's$ are the regression coefficients.

2.2.2. Negative binomial regression model

The negative binomial is a conjugate mixture distribution for count data. When the Poisson model assumption fails, negative binomial regression model may fit better, and address the overdispersion problem. However, this is true only if it is not attributed to excess zeros. As we have discussed in section 2.2.1, a severe limitation of the standard Poisson models assumption is, the variance of the data is equal to the mean of the data. Hence, at a fixed mean the variance cannot decrease as additional predictors enter the model.

Like Poisson regression, negative binomial regression model also examines predictive relationships with a count dependent variable. The standard Poisson regression accounts for observed differences among the observations; however negative binomial regression includes a random component that involves unobserved variance among observations. The inclusion of this random component prevents the incorrect Poisson assumption that is all differences among subjects in the dependent variable are equally explained. In an overdispersed data this random component results more accurate standard errors and z-statistics for the regression coefficients than using the standard Poisson regression [5].

Overdispersion might happen due to some relevant explanatory variables are not included in the model. A mixture model is a flexible way to account for such problem at a fixed setting of the predictors used, given the mean of the distribution of Y is Poisson, but the mean itself varies according to some distributions. Suppose λ_i has a gamma distribution with mean $E(Y_i \mid \lambda_i) = \mu_i$ and variance $\text{var}(Y_i \mid \lambda_i) = \mu_i / k$, $Y_i \mid \lambda_i$ to be a Poisson with conditional mean $E(Y_i \mid \lambda_i) = \lambda_i$. It can be shown that the marginal distribution of Y_i follows a negative binomial distribution with probability density function:

$$f(Y_i = y_i) = \int f(Y_i = y_i \mid \lambda_i) f(\lambda_i)\, d\lambda_i$$

$$= \frac{\gamma(y_i + k)}{\gamma(k)\gamma(y_i + 1)} \left(\frac{k}{k + \mu_i} \right)^k (1 - \frac{k}{k + \mu_i})^{y_i},\ y_i = 0, 1, 2, \ldots \quad (3)$$

With mean, $E(Y_i) = \mu_i$ and variance, $\text{var}(Y_i) = \mu_i (1 + \mu_i k^{-1})$. The index k^{-1} is called the dispersion parameter. As k^{-1} approaches to zero**,** the variance and mean becomes identical**.** Hence the negative binomial distribution will reduce to Poisson distribution. In such cases the data can be modelled easily by Poisson regression model. If $k^{-1} > 0$, the variance will exceed the mean, that is $\text{var}(Y_i) > E(Y_i)$, and the distribution allows for overdispersion [8]. One important characteristic of this distribution is, it accounts naturally for overdispersion. As a result negative binomial regression model has greater flexibility than the highly restrictive Poisson model [9].

Although the negative binomial model can solve an overdispersion problem, it may not be enough flexible to handle when there are excess zeros. In such cases, one can use the zero-inflated Poisson or zero inflated negative binomial model to solve the problem [10].

2.2.3. Poisson and negative binomial mixed regression models

In the case of Poisson, the random parameter can follow Gaussian, gamma, or inverse Gaussian distributions. Gamma is the preferred random distribution to use since it is conjugate to Poisson distribution. This has also an additional feature, which allows an analytic solution of the integral in the likelihood. Other random distributions do not have such favourable features [11]. A model may contain a random intercept and/or slope; however in this study we will deal with the most common, random intercept model. These models are a simple extension of the Poisson and NB models. They include a random intercept in addition to the fixed effect in the Poisson or negative binomial regression model. The effect of adding a random component to the linear predictor is shown in eq. (4). Most of a time the Gaussian or normal distribution is used to characterize the intercept randomness. As a result the Poisson regression mixed model can be given as:

$$\ln(\lambda_{ij}) = X_{ij}^{T}\beta + v_i,\ i = 1, 2, ..., m,\ j = 1, 2, ..., n_i \qquad (4)$$

where $X_{ij} = (1, x_{ij1}, ..., x_{ijp-1})^T$ is a $px1$ matrix of explanatory variable and β is $px1$ vector of regression parameters, m is the number of subjects n_i is the number of measurement for the i^{th} subject and v_i is a random intercept, which is assumed to be normally distributed with mean zero and variance σ_b^2.

In the case of the negative binomial mixed model, the mean, λ_{ij} is expressed in a similar way as eq. (4). The only difference is the response variable is assumed to have a negative binomial distribution [11].

Poisson-normal (or mixed) model can be used to fit longitudinal data. However when there is overdispersion in the data, it may not be enough flexible. In order to include the extra variability which is not taken in to account by the normal-random effect [12] extended this Poisson-normal model to a combined model which includes an overdispersion parameter, which has a gamma distribution. They have also discussed that the combined model (negative binomial mixed model) contributes more to the likelihood than only considering either the Poisson normal (or mixed) model or the negative binomial model.

In our case let $Y_{ij} \sim Poisson(\lambda_{ij})$, then the Poisson regression mixed model can be given as

$$\ln(\lambda_{ij}) = \beta_0 + \beta_1 treat_{ik} + \beta_2 time_{ij} + \beta_3 treat_i time_{ij} + \beta_4 time_{ij}^{\ 2} + \beta_4 treat_i time_{ij}^{\ 2} + v_i, \qquad (5)$$
$$i = 1, 2, ..., 213,\ j = 1, 2, ..., n_i,\ n_i = 1, 2...6, k = 1,2$$

where $time_{ij}$ is the j^{th} time measurement of the i^{th} individual, μ_{ij} is the predictive value of the j^{th} measurement of the i^{th} individual, $\beta' s$ are the regression coefficients and $treat_{ik}$ is the k^{th} treatment assign to the i^{th} individual and v_i is the random intercept as it is defined in eq. (4).

2.2.4. Zero-inflated regression models

There are situations where a major source of overdispersion is a preponderance of zero counts, and the resulting overdispersion cannot be modelled accurately with negative binomial model. In such scenarios, one can use zero-inflated Poisson or zero-inflated negative binomial model to fit the data. The first concept of a zero–inflated distribution originated from the work of [13, 14] who examined the characteristics of mixed Poisson distributions [15].

According to [16], Zero-inflated techniques permit the researcher to answer two questions that pertain to low base rate-dependent variables: (a) what predicts whether or not the event occurs, and (b) if the event occurs, what predicts frequency of occurrence? In other words, two regression equations are created: one predicting whether the count occurs and a second one predicting the occurrence of the count. Moreover, zero-inflated models have a statistical advantage to standard Poisson and negative binomial models in that they model the preponderance of zeros as well as the distribution of positive counts simultaneously [10]. In Section 2.2.4.1 and Section 2.2.4.2 zero-inflated Poisson and zero-inflated negative binomial models will be discussed briefly respectively.

2.2.4.1. Zero-inflated Poisson regression model

In Poisson model, counts are assumed to be generated with mean of λ_i according to the probability function in eq. (1). A characteristic of the Poisson distribution as it present in Section 2.2.1 above, the mean of the distribution is equal to the variance; however when there is an excess zeros, probability of zero in the standard model will be less than the expected. Therefore, in such situation the standard Poison and negative binomial models are not suitable models. In such cases, a ZIP or ZINB models can be used to account the excess zeros. The zero values in the ZIP model can be viewed as comprising two parts. One portion of the zero counts arises from the inflated part of the distribution and the other portion comes from what would be expected given a Poisson distribution with parameter λ.

When there is an excess zeros and high variability in the non-zero outcomes, ZIP models is less adequate than ZINB models. ZINB models will be described briefly in the next section 2.2.4.2.

Suppose Y_i is used to denote a ZIP variate, which is assumed to be generated according to the following probability density function:

$$P(Y_i \mid p_i, \lambda_i) = \begin{cases} p_i + (1-p_i)\exp(-\lambda_i), & if \quad y_i = 0 \\ (1-p_i)\dfrac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}, & if \quad y_i > 0 \end{cases} \quad i = 1,2,....n \qquad (6)$$

where λ_i is the mean of the non-zero outcomes that can be modelled with the associated explanatory covariates using a natural logarithmic link function as:

$$\ln(\lambda_i) = X_i^T\beta \qquad (7)$$

Where $X_i = (1, x_{i1}, x_{i2}, ..., x_{ip-1})^T$ is a *px1* vector of explanatory variable of the i^{th} observation and β is $px1$ vector of regression coefficient parameters. p_i is the probability of an excess zero, which can be estimated by the logistic regression. The associated covariate may be the same with the covariates (X_i\`s) of the count part model is defined as:

$$p_i = \frac{\exp(Z_i^T\gamma)}{1+\exp(Z_i^T\gamma)} \quad i = 1, 2, ...n_i \qquad (8)$$

Where $Z_i = (1, z_{i1}, ..., z_{iq-1})^T$ is a *qx1* vector of explanatory variable for the zero-inflation part model of the i^{th} observation and $\gamma = (1, \gamma_1, ..., \gamma_{q-1})^T$ is $qx1$ vector of regression coefficient parameters.

Unlike the Poisson distribution, which is determined by a single parameter, the ZIP distribution is determined by two parameters, λ_i and p_i The ZIP model is a special case of a two-class finite mixture model with mean and variance $E(Y_i) = (1-p_i)\lambda_i$ and $\mathrm{var}(Y_i) = (1-p_i)(\lambda_i + p_i\lambda_i^2)$ respectively [17].

In our case suppose $Y_i \sim ZIP(\lambda_i, p_i),$ these parameters will be linked with the covariates by the logarithmic and logit function as:.

$$\ln(\lambda_i) = \beta_0 + \beta_1 treat_{ik} + \beta_2 time_i + \beta_3 treat_{ik} time_i + \beta_4 time_i^2 + \beta_4 treat_{ik} time_i^2 \quad .... (9)$$
$$for \quad i = 1, 2, ..., 920$$

$$treat_{ik} = \begin{cases} 1, & if \quad the\, i^{th}\ individual\ recieved\ group\, II \\ 0, & \quad otherwise \end{cases}$$ and p_i is written as :

$$p_{ij} = \frac{\exp(\beta_0 + \beta_1 treat_{ik} + \beta_2 time_i + \beta_3 treat_{ik} time_i + \beta_4 time_i^2 + \beta_4 treat_{ik} time_i^2)}{1+\exp(\beta_0 + \beta_1 treat_{ik} + \beta_2 time_i + \beta_3 treat_{ik} time_i + \beta_4 time_i^2 + \beta_4 treat_{ik} time_i^2)}$$
$$for \quad i = 1,2,....,920 \quad k = 1,2$$

The variables have the same definition as of eq. (2).

2.2.4.2. Zero-inflated negative binomial regression model

Zero-Inflated Negative Binomial (ZINB) regression Model is an extension of the NB regression model that was discussed in section 2.2.2. As the number of zeros in the count distribution is excessive, then the ZIP or ZINB model will be more accurately fit the data than the negative binomial or Poisson model. If overdispersion is not accounted by the ZIP model, then there may be other aspects of the distribution that contribute to overdispersion, in such case the ZINB model is more appropriate [16].

The main difference between ZIP and ZINB model is that the Poisson distribution for the count data is replaced by the negative binomial distribution. The probability function of a ZINB is a simple modification of the ZIP.

Suppose Y_i is used to denote a ZINB variate, which is assumed to be generated according to the following probability function:

$$P(Y_i \mid p_i \lambda_i) = \begin{cases} p_i + \dfrac{(1-p_i)}{(1+\lambda_i/k)^k}, y_i = 0 \\ (1-p_i)\dfrac{\gamma(y_i+k)}{\gamma(k)\gamma(y_i+1)}\left(\dfrac{k}{k+\lambda_i}\right)^k (1-\dfrac{k}{k+\lambda_i})^{y_i}, if \ y_i > 0 \end{cases} \quad i = 1,2,....,n, \qquad (10)$$

Where λ_i is the mean of the non-zero response that can be modelled with the associated explanatory covariates using a natural logarithm link function is defined as eq. (7) and p_i is the probability of an excess zeros, which can be estimated by the logistic regression it is also defined as eq. (8) [18].

The ZINB model is a special case of a two-class finite mixture model like the ZIP model with mean and variance, $E(Y_i) = (1-p_i)\lambda_i$ and $\text{var}(Y_i) = (1-p_i)(\lambda_i + \lambda_i^2/k)$ respectively.

In our case λ_i and p_i are linked with covariates as eq. (9).

2.2.5. Zero- inflated mixed regression models

In healthcare research, count variables with many zeros are quite common in such case a standard Poisson or negative binomial regression models may not be appropriate since it underestimates the zero counts. Moreover, when there is an excess zeros in a cross sectional data we can use the ZIP or ZINB models; however in the case of the cluster and longitudinal data such models are not appropriate instead the zero-inflated mixed (i.e. ZIP mixed or ZINB mixed) models are appropriate. Each of these models will be discussed in section 2.2.5.1 and 2.2.5.2 respectively.

2.2.5.1. Zero-inflated Poisson mixed regression model

A zero-inflated Poisson (ZIP) model was developed by [19] to deal with counts which have extra zeros; however it has limitations for longitudinal and/or clustered count data. Recently, zero-inflated Poisson mixed (ZIP mixed) models have been developed that accommodates both correlated and extra-zero count data [20].

A ZIP mixed is an extension of ZIP by taking the clustering effect into account. Its corresponding density function is given by:

$$P(Y_{ij} = y_{ij} \mid p_{ij,}\lambda_{ij}) = \begin{cases} p_{ij} + (1-p_{ij})\exp(-\lambda_{ij}), y_{ij} = 0 \\ (1-p_{ij})\dfrac{\lambda_{ij}^{\ y_{ij}}\exp(-\lambda_{ij})}{y_{ij}!}, if y_{ij} > 0 \end{cases} \quad for\ i = 1,2,.....m, \qquad (11)$$

$$j = 1,2,...n_i$$

where m is the number of individuals included in the study and n_i is the number of measurements of the i[th] individual [20]. The mean and variance of the ZIP random variable are given by $E(Y_{ij}) = (1-p_{ij})\lambda_{ij}$ and $\mathrm{var}(Y_{ij}) = (1-p_{ij})\lambda_{ij}(1+p_{ij}\lambda_{ij}\ )$

In the regression setting, both λ_{ij} and p_{ij} parameters are related to covariate vectors X_{ij} and Z_{ij} as follow:

$$\begin{aligned} \log(\lambda_{ij}) &= X_{ij}^{\ T}\beta + v_i \\ p_{ij} &= \frac{\exp(Z_{ij}^{\ T}\gamma + u_i)}{1+\exp(Z_{ij}^{\ T}\gamma + u_i)} \quad for\ \ i = 1,2,.....m,\ j = 1,..n_i \end{aligned} \qquad (12)$$

Where $X_{ij} = (1, x_{ij1},...., x_{ijp-1})^T$ *and* $Z_{ij} = (1, z_{ij1},...., z_{ijq-1})^T$ *are* $px1$ *and* $qx1$ vectors of known covariates for the Poisson and logistic parts, respectively, from the i[th] individual. $\beta_{ij} = (\beta_0, \beta_1,...., \beta_{p-1})^T$ *and* $\gamma_{ij} = (\gamma_0,..., \gamma_{q-1})^T$ *are* $px1$ *and* $qx1$ are Poisson and logistic regression parameter vectors associated with covariates X_{ij} and Z_{ij} Here, p_{ij} is considered to be a mixing parameter for the mixture of a binary and a Poisson process. υi and νi are random effects and are assumed to be normally distributed, υi ~ $(0, \sigma_u^2)$ and νi ~ (0, $\sigma^2{}_v$). In our case λ_{ij} and p_{ij} are linked with covariates as:

$$\ln(\lambda_{ij}) = \beta_0 + \beta_1 treat_{ik} + \beta_2 time_{ij} + \beta_3 treat_{ik} time_{ij} + \beta_4 time_{ij}{}^2 + \beta_4 treat_{ik} time_{ij}{}^2 + v_i, \quad .(13)$$
$$for\ i = 1,2,...,213,\ j = 1,2,...n_{i,}\ n_i = 1,2,...,6\ and\ k = 1,2$$

$$treat_{ik} = \begin{cases} 1, if\ the\ i^{th}\ individual\ recieved\ group\ II \\ 0, otherwise \end{cases}$$ and p_{ij} is written as

$$p_{ij} = \frac{\exp(\beta_0 + \beta_1 treat_{ik} + \beta_2 time_{ij} + \beta_3 treat_{ik} time_{ij} + \beta_4 time_{ij}{}^2 + \beta_4 treat_{ik} time_{ij}{}^2 + u_i)}{1+\exp(\beta_0 + \beta_1 treat_{ik} + \beta_2 time_{ij} + \beta_3 treat_{ik} time_{ij} + \beta_4 time_{ij}{}^2 + \beta_4 treat_{ik} time_{ij}{}^2 + u_i)}$$
$$for \quad i = 1, 2,..., 213,\ j = 1,\ 2,...n_{i,}\ n_i = 1,2,...,6\ and\ k = 1,2$$

where v_i and u_i is the random intercepts of the Poisson and binomial models respectively, $time_{ij}$ is the j^{th} time measurement of the i^{th} individual, μ_{ij} is the predictive value of the j^{th} measurement of the i^{th} individual, $\beta`s$ are the regression coefficients and $treat_{ik}$ is the k^{th} treatment assign to the i^{th} individual.

2.2.5.2. Zero-inflated negative binomial mixed regression model

Zero-inflated negative binomial mixed model is an extension of the ZINB model. Unlike to the ZINB model it takes into account the clustering effect. If the overdispersion problem is not only arising from excess zeros, a ZINB mixed model is used to overcome the problem , which is attributed to the non-zero count data in a clustered or longitudinal data . In such cases the count variable Y_{ij} follows a ZINB distribution, which is given by:

$$P(Y_{ij} \mid p_{ij}, \mu_{ij}) = \begin{cases} p_{ij} + \dfrac{(1-p_{ij})}{(1+\lambda_{ij}/k)^k}, y_{ij} = 0 \\ (1-p_i)\dfrac{\gamma(y_{ij}+k)}{\gamma(k)\gamma(y_{ij}+1)}\left(\dfrac{k}{k+\lambda_{ij}}\right)^k (1-\dfrac{k}{k+\mu_{ij}})^{y_{ij}}, if\ y_{ij} > 0 \end{cases} \quad for\ i = 1,2,...,m$$
$$and \quad j = 1,2,...n_i \qquad (14)$$

where m and n_i are the same as they defined in eq.(11) and k^{-1} is an overdispersion parameter [21]. Both λ_{ij} and p_{ij} parameters are related to covariate vectors X_{ij} and Z_{ij} in a similar fashion as eq. (12) model.

2.2.6. Goodness of fit

2.2.6.1. Likelihood ratio test

The maximum likelihood estimation method is used to assess the adequacy of any two or more than two nested models by using the likelihood ratio test. it compares the maximum likelihood under the alternative hypothesis with the null hypothesis. For instance, the null hypothesis can be the overdispersion parameter is equal to zero (i.e. the Poisson distribution can be fitted well the data) and the alternative hypothesis can be the data would be better fitted by the Negative binomial regression (i.e. the overdispersion parameter is different from zero). The likelihood ratio test is defined as: $X^2 = -2(l - l_0)$

where l and l_0 are the log likelihood of models under the alternative and null hypothesis respectively. This has a chi-square distribution. As a result this test of statistics will be compare with the tabulated chi-square with a degree of freedom, the difference between the degree of freedom of the model under null hypothesis and the alternative hypothesis respectively. This method is not appropriate for models which are not nested one on the other, in such situation; we will use another method such as the Akakie information criteria (AIC) and Bayesian information criteria (BIC) [22].

In this study a likelihood ratio was used to compare the Poisson with the negative binomial and zero-inflated Poisson with zero-inflated negative binomial since Poisson is nested on negative binomial and zero-inflated Poisson is nested in zero-inflated negative binomial; However this will not be used to compare Poisson or negative binomial with the zero inflated Poisson and negative binomial as long as these models are not nested one on the other.

2.2.6.2. Information criteria

If there are several models to be compared in order to select the best model which fits the data instead of using the likelihood ratio test, it can be easily select by using the Akakie information criteria (AIC) and Bayesian information criteria (BIC).

2.2.6.3. Akakie information criteria (AIC)

AIC is the most common means of identifying the model which fits well by comparing two or more than two models. It is trying to balance the goodness of fit against the complexity of the model It is similar as of the coefficient of multiple determination (R^2); however, it penalized by the number of parameter included in the model (i.e. the complexity of the model). Unlike theR^2, the good model is the one which has the minimum AIC value. It is given by the following formula

$$AIC = -2l + 2k$$

Where l are the log likelihood of a model that will compare with the other models and k is the number of parameter in the model including the intercept [22].

Unlike the Akakie information criteria the Bayesian information matrix (BIC) takes in to account the size of the data under considered. It is given by:

$$BIC = -2l + k \log(n)$$

where l are the log likelihood of a model that will compare with the other models, n is the sample size of the data and k is the number of parameters in the model including the intercept.

2.2.7. Software

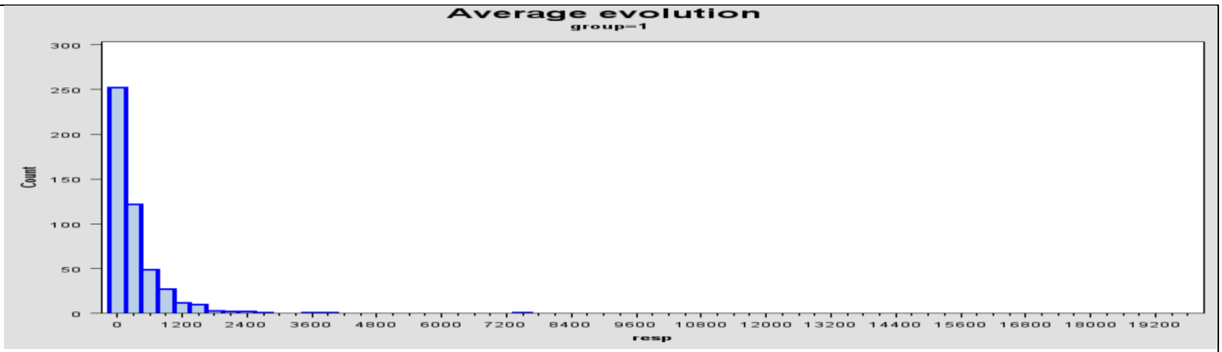
In this study, SAS (version 9.2) and R (2.10) were used to analyse the data. In addition all hypotheses were tested at 0.05 level of significance.

3.0. Results

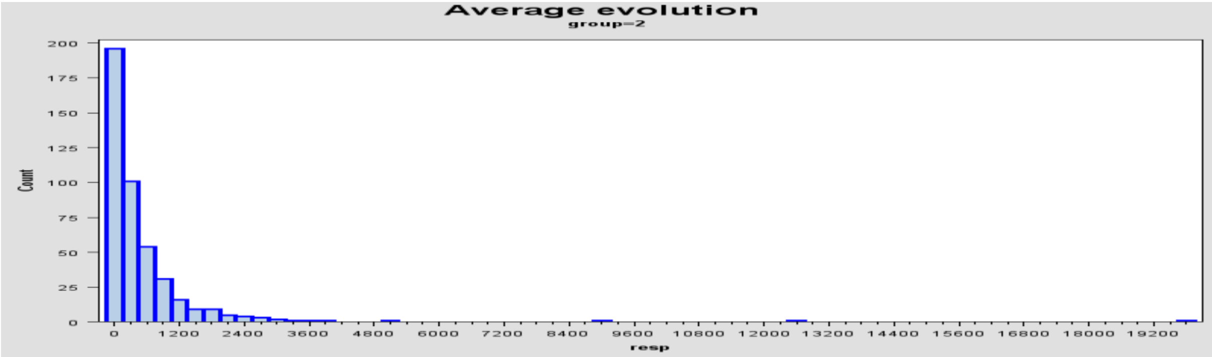
3.1. Exploratory data analysis

To have an insight on the data, an exploratory data analysis was conducted. In this study the mean of B-cells produced per million cells was 420.396, which is much smaller than the variance, 1035753.998. This indicates that there is an over dispersion. In such case the standard Poisson regression model is not an appropriate model to fit the data. In addition the median of the data was 161.50, which is smaller than the mean.

Figure 1 presents the distribution of the number of B-cells produced per million cells in each group. Since there is large number of zero outcomes, the histograms are highly picked at the very beginning (about the zero values) in both groups. However large observations (i.e. large number of B-cells) are less frequently observed. This leads to have a positively (or right) skewed distribution in each group. This could be fitted better by count data models which takes into account excess zeros like zero-inflated models.



Panel A:Histogram of the number of B-cells in group I



Panel B:Histogram of the number of B-cells in group II

Figure 1: Histogram of the B-cells found per million of cell by group

The mean profile B-cell produced over time (months) of the two treatment groups, group I and group II are presented in Figure 2. It indicates that there is a higher treatment effect in group II than in group I. Especially at month 7, the treatment in group I and group II produces high number of B-cells as compared to other time point measurements. In general this Figure indicates that the average production of B-cells by the two group vaccines is increasing until the first time measurement (7th month) and then starts to decline.

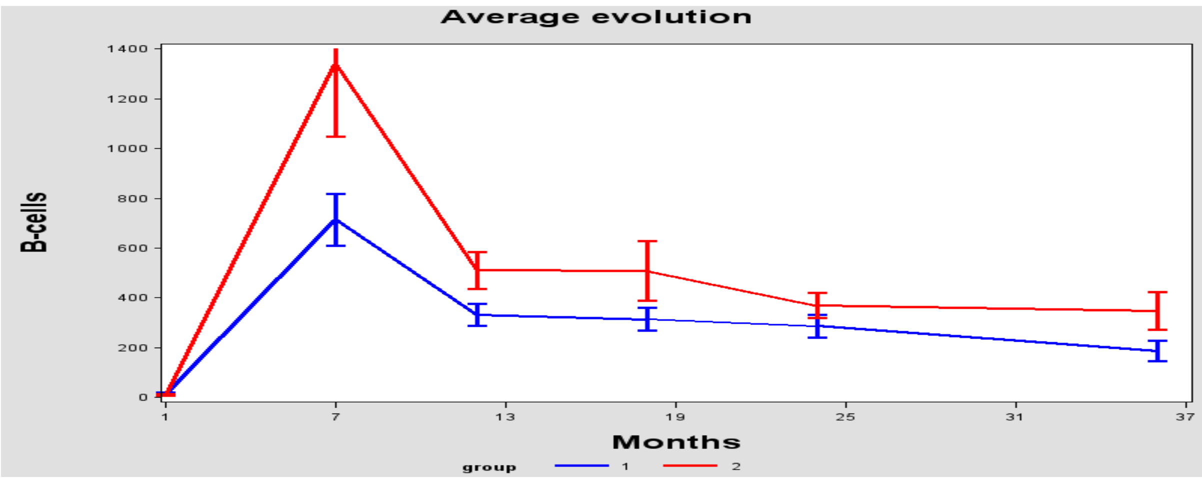


Figure 2: mean profile of B-cells over time with respect to the two group I and group II

As there is a high variability in the data, the median would be better in explaining the measure of central tendency of the data than the mean since mean is highly affected by extreme observations. The median function of the measurement across the different time points of the two vaccine groups is shown in Figure 3. The median measurement at each time point is much smaller than the corresponding mean measurement in Figure 2.

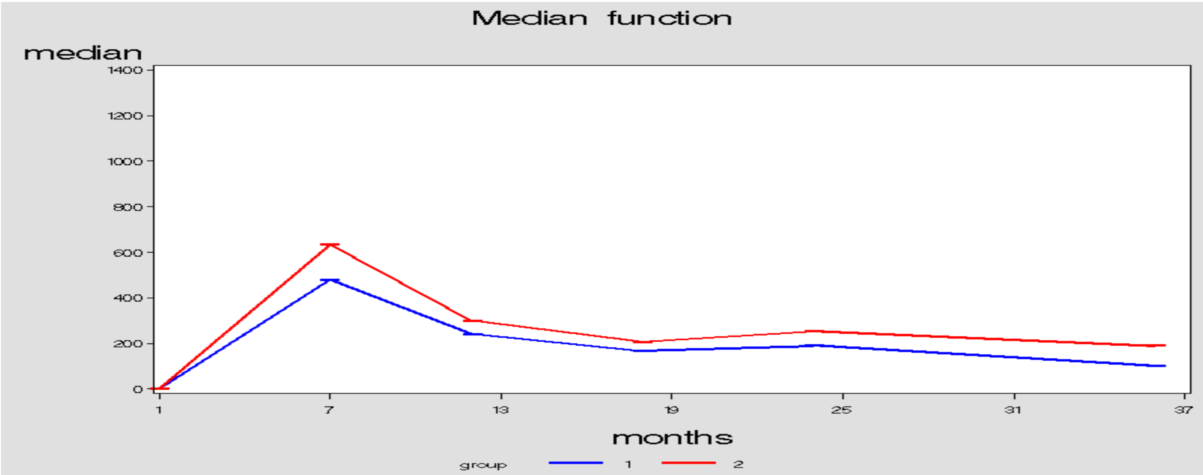


Figure3: median structure of the B-cells over time with respect to the two groups: group I and group II

The individual profile plot present in Figure 10A (Appendix) shows that there is a substantial between and within variability of the B-cells production. Thus, it is better to consider models which take into account the heterogeneity nature of the data.

To assess the variability across the different time points, a variance structure was used. As is shown in Figure 4, there is a high variability of producing B-cells over time. Especially in the seventh month, it picks up and then starts to decline and then it starts to rise up again. Then after it remains constant over the rest time measurements (in the 24th and 36th month measurement). It suggests that as the random intercept model may not be enough. We should include also random slopes. The variance functions for the two treatment (vaccine) groups are also summarized in Figure 1A (Appendix). The plot shows as there is high variability over time, particularly individuals who received a group II vaccine.

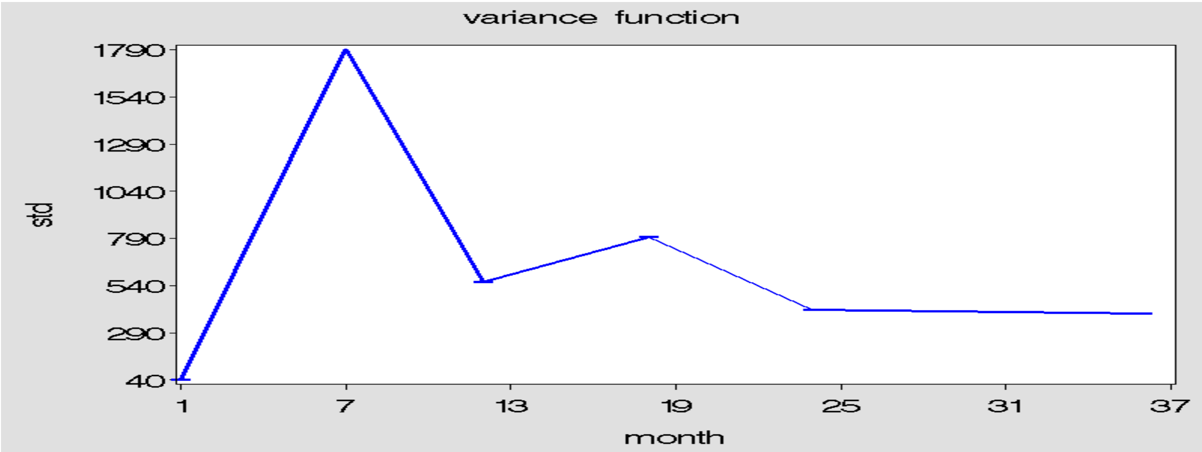


Figure 4: Variance structure of the data

The scatter plot matrix of the number of B-cells observed at each time point is shown in Figure 2A (Appendix). It shows that as there are potential outlying observations in the data. In addition, the histogram in the diagonal of the scatter plot matrix shows that the response variable is not normally distributed rather it is positively (or right) skewed. We have also shown the mean evolution and variance structure for the non-zero outcomes Figure 3A (Appendix) in order to assess the variability in the non-zero outcomes. It shows that the variance at each time point is substantially higher than the mean. This gives us a clue as there is an overdispersion in the non-zero value of the response variable.

3.1. Statistical data analysis

The variable of interest in this study was the number of B-cells produced per million cells. Such data can be well fitted by the count models rather than the linear regression model. In this study we have considered different possible count data models. Likelihood ratio test (LR), Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to compare the candidate models to identify the most parsimonious model.

3.1.1. Model comparison

In order to select an appropriate model which fits the data well, eight different models were considered namely: the standard Poisson, negative binomial, Poisson mixed, negative binomial mixed, zero-inflated Poisson, zero-inflated negative binomial, zero-inflated Poisson mixed and negative binomial mixed models.

3.1.1.2. Fixed effect models

Table 1 presents the parameter estimates with their corresponding standard error of the Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial regression models with their corresponding deviance (-2*l*), AIC and BIC values. The overdispersion parameter (k^{-1}) is significantly different from zero in both NB and ZINB regression models. Hence there is an overdispersion problem in the data. As a result of this the standard error of the standard Poisson regression model is smaller than the standard error of the other models. Especially when compared to the NB regression model, the standard error of the standard Poisson regression model is very small. Thus, the z-statistic will be inflated consequently the covariates may wrongly interpreted.

As one can be seen from this Table 1, all covariates included in the standard Poisson model such as: treatment, linear time and the interaction between treatment with linear and quadratic time effect are significantly associated with the B-cells production; however in the case of the NB model only linear and quadratic time effects are significantly associated. This is due to the fact that the standard Poisson regression model does not take in to account overdispersion; however it can be handled by the NB model if it is not arise from an excess of zero observations. ZIP and ZINB regression models were better fitted than Poisson and NB respectively based on their corresponding AIC as well as BIC. The parameter estimate of time main effect is positive in Poisson and negative binomial regression models; however it is negative in the zero-inflated regression model.

A likelihood ratio test was used to compare the nested models, standard Poisson and ZIP regression models with NB and ZINB respectively. It was found that NB and ZINB regression models were well fitted the data than the standard Poisson and ZIP respectively since their LR, $X^2(1) = 774198$ and $X^2(1) = 492957$ both were highly significant (p-value<0.0001). This also supported by the information criteria's (Table 1). The overdispersion parameter (k^{-1}) in the ZINB regression model is significantly different from zero since there is a high variability in the non-zero outcomes. In such scenario, it would be better to use the model which takes into account the excess zeros and high variability due to non-zero outcomes. The zero-inflated negative binomial (ZINB) regression model was found to be the most parsimonious model which fits the data better than the other possible candidate models. Since it has the smallest AIC (10277) as well as BIC (10339) values as presented in Table 1.

Table 1: Parameter estimates of Poisson, NB, ZIP and ZINB regression models

Parameters	***Poisson model estimates(s.e.)***	***NB model estimates(s.e.)***	***ZIP model estimates(s.e.)***	***ZINB model estimates(s.e.)***
Poisson/negative binomial part				
intercept	5.1908(0.0070)*	4.4402(0.3469)	6.3821(0.00818)*	6.4693(0.2241)*
treat	0.6828(0.0088)*	0.7992(0.5255)	1.1737(0.01056)*	1.3133 (0.3283)*
time	0.1119(0.0009)*	0.1998(0.0473)*	-0.0011(0.00106)*	-0.0145 (0.02700)
Treat*time	-0.0206(0.0012)*	-0.03774(0.0714)	-0.0914(0.00137)*	-0.1089 (0.03884)*
Time2	-0.0036(0.00003)*	-0.0052(0.0012)*	-0.0007(0.00003)	-0.0003(0.00067)
Treat*tim2	0.0004(0.00004)*	0.0009(0.0017)	0.00214(0.00004)*	0.0025(0.00096)*
k^{-1}	0	4.8143(0.2195)*	0	1.0944(0.05845)*
Logistic(inflated) part				
intercept	-------------	------------	1.0052(0.22420)*	1.0235(0.2258)*
treat	-----------	-------------	0.7028(0.35460)*	0.7080(0.3588)*
time	------------	------------	-0.2647(0.03247)*	-0.2694(0.0330)*
Treat*time	------------	-------------	-0.1769(0.05398)*	-0.1810(0.0557)*
Time2	------------	-------------	0.00628(0.00088)*	0.0064(0.00089)*
Treat*tim2	-----------	-------------	0.00437(0.00142)*	0.0045(0.00146)*
	$-2l$=774198 AIC =774210 BIC=774239	$-2l$=10921 AIC = 10935 BIC= 10968	$-2l$=503208 AIC =503232 BIC =503290	$-2l$=10251 AIC=10277 BIC=10339

3.1.1.3. Mixed effect models

Table 2 summarizes the parameter estimates and their corresponding standard errors of the Poisson mixed, negative binomial mixed, zero-inflated Poisson mixed and zero-inflated negative binomial mixed regression models. Each model contains one random component in each of the count part (Poisson or negative binomial part). The overdispersion parameter in the NB mixed and ZINB mixed regression models were significantly different from zero; however the magnitude of the overdispersion was higher in the NB mixed regression model than ZINB mixed.

The AIC and BIC values of the models presented in Table 2 was smaller as compared to the corresponding models in Table 1. Hence including a random component increases the fitness of the models. Especially the AIC and BIC value of the Poisson mixed and ZIP mixed regression models were greatly reduced; however, in the case of the NB mixed and ZINB mixed regression models reduced relatively small. From this we can say that keeping the longitudinal nature of the data in the NB and ZINB regression models may not be a serious issue; however in the case of the Poisson and ZIP is a crucial thing.

From Table 2 the Poisson mixed regression model has a very small standard error than NB mixed. This will result a high probability of committing Type I error (wrongly rejecting the null hypothesis when it is true). Thus, some covariates will wrongly interpret as they have an association with the response variable in fact they are not. The negative binomial mixed and zero-inflated negative binomial mixed regression models were better fitted the data than the standard Poisson mixed and zero-inflated Poisson mixed based on their corresponding AIC and BIC values. This was further confirmed by the likelihood ratio test ($p<0.0001$). The AIC (10160) as well as BIC (10207) values of the ZINB mixed regression model was the smallest as compared with other models presented in Table 2. Hence this model fits better than the other models.

Table 2: Parameter estimates of Poisson mixed, NB mixed, ZIP mixed and ZINB mixed regression models

Parameters	***Poisson mixed model estimates (s.e.) with one random int.***	***NB mixed model estimates (s.e.) with one random int.***	***ZIP mixed model estimates (s.e.) with one random int.***	***ZINB mixed model estimates (s.e.) with one random int.***
Poisson/negative binomial part				
intercept	4.5937(0.1301)*	4.3416(0.3509)*	6.2301(0.0908)*	6.3472(0.2089)*
treat	0.6350(0.1867)*	0.6826(0.5312)	1.1187(0.1295)*	1.0001(0.3050)*
time	0.1164(0.0010)*	0.2078(0.0472)*	-0.0250(0.0012)*	-0.0222(0.0240)
Treat*time	-0.0166(0.0013)*	-0.0300(0.0711)	-0.1038(0.0016)*	-0.093(0.0346)*
Time2	-0.0037 (0.00003)*	-0.0054(0.0012)*	-0.0003(0.00003)*	-0.0001(0.0006)
Treat*tim2	0.0005(0.00004)*	0.0008(0.0017)	0.0025(0.00004)*	0.0023(0.0009)*
k^{-1}	0	4.6846(0.2247)*	0	0.7068(0.0436)*
$\boldsymbol{var(v_i)}$	1.8409(0.1919)*	0.1249(0.0827)	0.8628(0.08538)*	0.4543(0.0749)*
Logistic(inflated) part				
intercept	-------------	------------	1.0050(0.2243)*	1.0317(0.2252)*
treat	-----------	------------	0.7035(0.3546)*	0.6415(0.3541)
time	------------	-------------	-0.2670(0.0325)*	-0.2684(0.0327)*
Treat*time	------------	------------	-0.1714(0.0539)*	-0.1699(0.0542)*
Time2	------------	-------------	0.0064(0.0009)*	0.0064(0.0009)*
Treat*tim2	-----------	-------------	0.0042(0.0014)*	0.0042(0.0014)*
	$-2l$=389061 AIC =389075 BIC=389098	$-2l$=10917 AIC =10933 BIC=10960	$-2l$=188152 AIC=188178 BIC=188222	$-2l$=10132 AIC= 10160 BIC=10207

The parameter estimates and their corresponding standard errors of the ZIP and ZINB mixed regression models with one and two random components, and their corresponding AIC and BIC values is presented in Table 3. The overdispersion parameter estimates in both of the ZINB mixed regression models were significantly different from zero. Hence there was a high variability in the non-zero outcomes. As a consequence of this the standard errors of both ZIP mixed regression models were much smaller than those of the ZINB mixed. This leads to the inflation of type I error. All covariates included in both logistic (inflated) and

Poisson part of any ZIP mixed models were significantly associated with the production of B-cells; however the linear and quadratic time effects were not significantly associated with the production of B-cells in the negative binomial part of both ZINB mixed models.

The AIC and BIC values of the ZIP mixed and ZINB mixed regression models present in Table 3 are smaller than those of ZIP and ZINB mixed present in Table 2 as well as ZIP and ZINB models present in Table 1 respectively. Hence as the random components were added to each part of the zero-inflated models, the fitness of the models was improved. As we can see from Table 3, both of the ZINB mixed regression models have a smaller AIC as well as BIC value as compared to those of ZIP mixed. Therefore the ZINB mixed regression models which have two random intercepts were appeared to fit the data better than the ZIP mixed models. This was also supported by the likelihood ratio test (P<0.0001). The ZINB mixed models with correlated random components have the smallest AIC (10123) as well as BIC (10177) values as compared to all other possible models considered. This could be considered as the parsimonious model.

Table 3: Parameter estimates of ZIP mixed and ZINB mixed models with two random intercepts

Parameters	***ZIP mixed model estimates with uncorrelated random int.***	***ZINB mixed model estimates (s.e.) with uncorrelated random int.***	***ZIP mixed model estimates with correlated random int.***	***ZINB mixed model estimates with correlated random int.***
		Poisson/negative binomial part		
intercept	6.2136(0.0906)*	6.3345(0.2086*	6.229(0.0913)*	6.2220(0.2086)*
treat	1.1496(0.1292)*	1.0182(0.3048)*	1.0871(0.1305)*	1.1121(0.3048)*
time	-0.0250(0.0012)*	-0.0208(0.0240)	-0.0250(0.0012)	-0.0172(0.0238)
Treat*time	-0.1038(0.0016)*	-.0949(0.0345)*	-0.1037(0.0016)*	-0.098(0.0344)*
Time2	-0.0003(0.00003)*	-0.0002(0.0006)	-.0003(0.00003)*	-0.0003(0.0006)
Treat*time2	0.0025(0.00004)*	0.0023(0.0009)*	0.0025(0.00004)*	0.0024(0.0009)*
k^{-1}	0	0.7064(0.0435)*	0	0.6989(0.0428)*
$\boldsymbol{var(v_i)}$	0.8584(0.0845)*	0.4540(0.0748)*	0.8770(0.0878)*	0.5012(0.0809)*
		Logistic(inflated) part		
intercept	1.1002(0.2556)*	1.1637(0.2622)*	1.0459(0.2500)*	1.1097(0.2597)*
treat	0.7545(0.3936)*	0.7465(0.4020)	0.7412(0.3828)*	0.8283(0.4002)*
time	-0.2974(0.0366)*	-0.3079(0.0376)*	-.2912(0.0359)*	-0.302(0.0372)*
Treat*time	-0.1902(0.0576)*	-0.1924(0.0589)*	-.1805(0.0562)*	-0.204(0.0586)*
Time2	0.0070(0.0010)*	0.0073(0.0010)*	0.0069(0.0010)*	0.0071(0.0010)*
Treat*time2	0.0048(0.0015)*	0.0048(0.0016)*	0.0045 (0.0015)*	0.0051(0.0015)*
$\boldsymbol{var(u_i)}$	0.6377(0.2306)*	0.7478(0.2617)*	0.5466(0.2046)*	0.7537(0.2602)*
$\boldsymbol{Cov(v_i, u_i)}$	-----	------	-0.5778(0.1099)*	-0.546(0.1196)*
	-2l =188135 AIC = 188163 BIC = 188210	-2l =10115 AIC= 10145 BIC=10195	-2l =188106 AIC=188136 BIC=188136	-2l =10091 AIC= 10123 BIC= 10177

3.1.2. ***Zero-inflated negative binomial mixed regression model***

Table 4 summarizes the parameter estimates and their corresponding standard errors of the zero-inflated negative binomial mixed (ZINB mixed) regression model. This model is a mixture of two regression models one is for the logistic (inflated) part and the other is for the count data (negative binomial part). As is shown in Table 4 treatment, linear time, quadratic time and the interaction between treatment with linear and quadratic time effect were significantly associated with the production of the B-cells in the logistic (inflated) part; however in the negative binomial part the linear and quadratic time effects were not significant. In addition the overdispersion parameter is significantly different from zero (p-value<.0001) this confirm that there was an overdispersion with an excess zeros in the data.

Table 4: Parameter estimates of the zero-inflated negative binomial mixed regression models

parameters	***Parameter estimates***	***Standard error***	***Pr > |t|***
Negative binomial part			
intercept	6.22200	0.20860	<.0001
treat	1.11210	0.30480	0.0003
time	-0.017160	0.02384	0.4724
Treat*time	-0.09828	0.03435	0.0046
Time2	-0.00026	0.00059	0.6590
Treat*time2	0.002374	0.00085	0.0056
k^{-1}	0.69890	0.04277	<.0001
$\boldsymbol{var(v_i)}$	0.50120	0.08086	<.0001
Logistic(inflated) part			
intercept	1.10970	0.25970	<.0001
treat	0.82830	0.40020	0.0397
time	-0.30210	0.03720	<.0001
Treat*time	-0.20380	0.05857	0.0006
Time2	0.00711	0.00100	<.0001
Treat*time2	0.00510	0.00150	0.00110
$\boldsymbol{var(u_i)}$	0.75370	0.26020	<.0001
$\boldsymbol{Cov(v_i, u_i)}$	-0.54570	0.11960	0.0042

The number of B-cells produced was positively associated with the treatment and interaction between treatment with linear and quadratic time effect; however it was negatively associated with treatment by time interaction in the negative binomial part of the model. Thus the main effect is interpreted by taking into account the interaction effect. Thus there was different production of B-cells at the different time measurements. Group II vaccine was superior than group I at all time measurements for instance, at the first measurement time (at one month), the group II vaccine was producing 2.7156 times higher than the one produced by group I. A similar interpretation can be drawn at all other time points. The random intercepts in the logistic and negative binomial part of the model were negatively correlated. Thus as the one increases the other decreases and vice-versa. In contrast the negative binomial part, all covariates were significantly associated with the B-cell production in the logistic part. Here also the main effect interpretation is given by taking in to account the interaction effect.

Therefore the final model which fits best the data can be written as follows:

$$P(Y_{ij} \mid p_{ij}, \lambda_{ij}) = \begin{cases} p_{ij} + \dfrac{(1-p_{ij})}{(1+\lambda_{ij}/k)^{k}}, & y_{ij} = 0 \\ (1-p_i)\dfrac{\gamma(y_{ij}+k)}{\gamma(k)\gamma(y_{ij}+1)}\left(\dfrac{k}{k+\lambda_{ij}}\right)^{k}(1-\dfrac{k}{k+\lambda_{ij}})^{y_{ij}}, & if \quad y_{ij} > 0 \end{cases} \quad for\ i = 1,2,\dots,m$$

$$and\ j = 1,2,\dots n_i$$

$$\ln(\lambda_{ij}) = \beta_0 + \beta_1 treat_{ik} + \beta_2 time_{ij} + \beta_3 treat_{ik} time_{ij} + \beta_4 time_{ij}{}^2 + \beta_4 treat_{ik} time_{ij}{}^2 + v_i,$$
$$\text{logit}\,(p_{ij}) = \beta_0 + \beta_1 treat_{ik} + \beta_2 time_{ij} + \beta_3 treat_{ik} time_{ij} + \beta_4 time_{ij}{}^2 + \beta_4 treat_{ik} time_{ij}{}^2 + u_i,$$
$$fori = 1,2,\dots,213,\ j = 1,2,\dots n_{i,} n_i = 1,2,..6\, and\ k = 1,2$$

$$treat_{ik} = \begin{cases} 1, ifthei^{\,th} individualrecievedagroupII \\ 0, otherwise \end{cases}$$

Thus the final estimated model was:

$$\ln(\lambda_{ij}) = 6.2220 + 1.1121 treat_{ik} - 0.01716 time_{ij} - 0.0983 treat_{ik} time_{ij} - 0.0003 time_{ij}{}^2$$
$$+ 0.0024 treat_{ik} time_{ij}{}^2,$$
$$\log it(p_{ij}) = 1.10970 + 0.82830 treat_{ik} \text{ - } 0.30210 time_{ij} \text{ - } 0.20380 treat_{ik} time_{ij} + 0.00711 time_{ij}{}^2$$
$$+ 0.00510 treat_{ik} time_{ij}{}^2$$
$$fori = 1,2,\dots,213,\ j = 1,2,\dots n_{i,} n_i = 1,2,..6 andk = 1,2$$

$$treat_{ik} = \begin{cases} 1, if\ \ the\ i^{th}\, individual\, recieved\ a\, group\, II \\ 0, otherwise \end{cases}$$

3.2. Model diagnostics

To check the predicting power and to identify the potential outlying observations of the data different diagnostic plots were used, namely: Plot of the observed versus predicted, the average evolution of the observed and predicted number of B-cells produced over time of measurement, the normal q-q plot and the plot of Pearson residual against the predicted values. In addition, the scatter plot of the two random components was used to check the potential outlying observations. Figure 6 shows that the mean profile of the observed and predicted values of the final model (ZINB mixed model) which was relatively well fitted as compared to other models such as zero-inflated poison mixed (ZIP mixed) is presented in Figure 5A (Appendix) and negative binomial mixed (NB mixed) is presented in Figure 6A (Appendix).

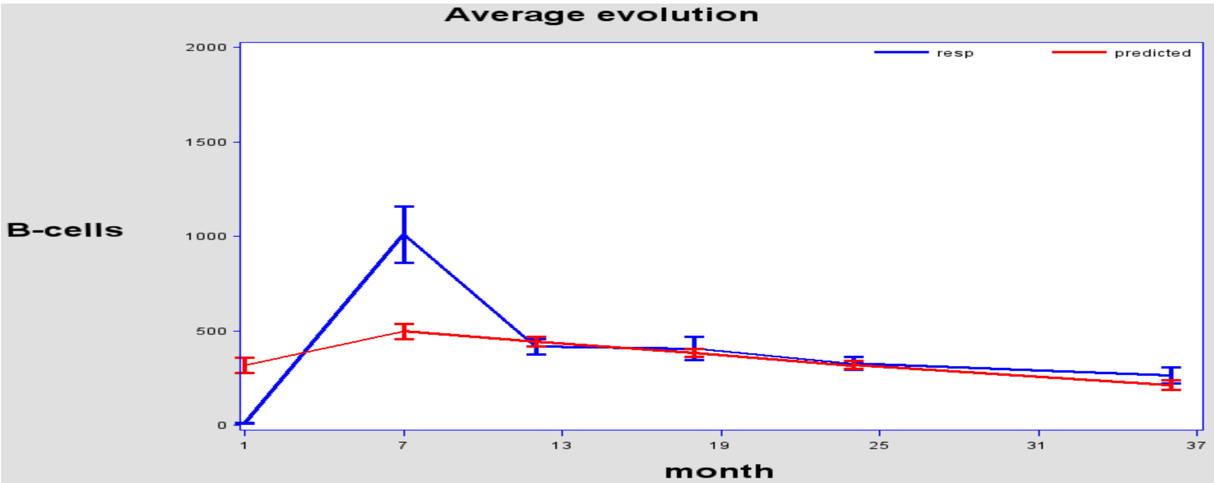


Figure6: The average plot of the predicted (red) and observed response (blue) of the ZINB mixed model

There is a much difference between the predicted and observed values of the B-cells produced of the NB mixed model. Thus, the NB mixed model was not fit the data very well that may be due to the presence of excess zeros in the data. As one can see, there is a visible difference in the first and second measurement time points (i.e. the 1st and 7th month) in almost all possible fitted models. ZINB mixed regression model was selected as best model; however, there seems a potential outlying observations as we can identify from the scatter plot matrix in Figure 2A (appendix), plot of Pearson versus predicted values presented in Figure 8A (appendix), plot of the two random intercepts presented in pane lA of Figure 9A (Appendix) and plot of observed versus predicted values of the ZINB mixed regression model in panel A of Figure11A (appendix). To end this problem 16 possible potential outliers was discarded and the data was fitted again. The zero-inflated negative binomial mixed regression model was also found as a better model fits the data with AIC (9727.1) and BIC

(9780.9). The parameter estimates of ZINB mixed model before and after excluding the outlying observation are summarized in Table 5.

Table 5: Parameter estimates of the ZINB mixed models with and without outlier

parameters	***ZINB mixed model without Outlier estimates (s. e.)***	***ZINB mixed model of the full data estimates (s. e.)***
	Negative binomial part	
intercept	5.9555(0.1982)*	6.2220(0.2086)*
treat	1.0375(0.2931)*	1.1121(0.3048)*
time	0.01014(0.0230)	-0.01716(0.0238)
Treat*time	-0.1057(0.0333)*	-0.0983(0.0344)*
Time2	-0.0008(0.0006)	-0.0003(0.0006)
Treat*time2	0.0027(0.0008)*	0.0024(0.0009)*
k	0.6630(0.0419)*	0.6989(0.0428)*
$\boldsymbol{var(v_i)}$	0.3328(0.06563)*	0.5012(0.0809)*
	Logistic (inflated) part	
intercept	1.1305(0.2609)*	1.1097(0.2597)*
treat	0.8000(0.4010)*	0.8283(0.4002)*
time	-0.3021(0.0373)*	-0.3021(0.0372)*
Treat*time	-0.1932(0.05831)*	-0.2038(0.0586)*
Time2	0.00707(0.0010)*	0.0071(0.0010)*
Treat*time2	0.00484(0.0015)*	0.0051(0.0015)*
$\boldsymbol{var(u_i)}$	0.7470(0.2625)*	0.7537(0.2602)*
$\boldsymbol{cov(v_i, u_i)}$	-0.3975(0.1055)*	-0.5457(0.1196)*

Table 5 shows the parameter estimates of the models before and after excluding the outlying observations are not much apart they are almost similar. In addition, all covariates significant in one model are also significant on the other. Thus In this study the ZINB mixed regression model is robust to outliers.

The ZINB mixed model estimate obtained after excluding the outliers is given by:

$$\ln(\lambda_{ij}) = 5.9555 + 1.0375 treat_{ik} + 0.01014 time_{ij} - 0.3021 treat_{ik} time_{ij} - 0.00080 time_{ij}{}^2$$
$$+ 0.002675 treat_{ik} time_{ij}{}^2,$$
$$\log it(p_{ij}) = 1.1305 + 0.8000 treat_{ik} - 0.3021 time_{ij} - 0.1932 treat_{ik} time_{ij} + 0.007086 time_{ij}{}^2$$
$$+ 0.004839 treat_{ik} time_{ij}{}^2$$
$$for\, i = 1,2,...,213,\ j = 1,2,...n_{i,} n_i = 1,2,..6 \ and\ k = 1,2$$

$$treat_{ik} = \begin{cases} 1, \ if \ \ the\ i^{th}\ individual\ recieved\ a\ group\ II \\ 0, otherwise \end{cases}$$

The Plot of the observed versus predicted values based on the two estimated models, (i.e. obtained before and after excluding the outlier) are presented in Figure 11A (appendix). The points were relatively lying on the straight line after excluding the outlier. This also supported by the mean evolution plot in Figure 4A (Appendix) that is the observed and predicted mean evolution after excluding the outliers is more closer than before excluding the outlier. The q-q plot of Pearson error of the two conditions is plotted in Figure 7A (Appendix). It shows that the variability is smaller after the outliers are excluded, the error varies from -2 to 4 in the full data; however it varies from -0.5 to 2 after excluded the outliers.

4.0. Discussion and conclusion

In this clinical trial study 213 consented individuals were participated. Individuals who participated in the study followed from two to six measurement times after the individual has taken either of the two HPV vaccines. The main objective of the study was to investigate a statistical methodology to compare the B-cell responses per million of cells over time after the two HPV vaccines has been administered and to assess the two HPV vaccines effect in producing the B-cells, in the presence of zero inflated data. From the exploratory data we could identify that as there is an excess zeros and high variability in the non-zero B-cells produced values. The mean of the B-cells produced was much lower than the variance. This might occur due to an excess of zeros and high variability of the non-zero outcomes. Since the number of zero outcomes was about 30.65% of the observed data and a high variability in the non-zero observations was also identified from the variance function.

The data had about 358 missing observations. The missing mechanism was treated as missing at random (MAR). Under the likelihood or Bayesian approach this missing mechanism is ignorable [23]. In this study the data was analysed by a likelihood approach using the SAS procedure NLMIXED.

The best model was selected from different possible models namely: Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, Poisson mixed, negative binomial mixed, zero-inflated Poisson mixed and zero-inflated negative binomial mixed model with one and two random intercepts. The comparison was conducted by using likelihood ratio test (LR), Akakie information criteria (AIC) and Bayesian information criteria (BIC). Likelihood ratio test (LR) was used to compare any two nested model such as Poisson with Poisson mixed, negative binomial, negative binomial mixed model, and zero-inflated Poisson with zero-inflated Poisson (ZIP), zero-inflated Poisson mixed (ZIP mixed), zero-inflated negative binomial and zero-inflated negative binomial mixed models; however any two non-nested models was compared by either AIC or BIC.

Of these the zero-inflated negative binomial regression mixed model with two random intercepts was selected as the best model. When the random components were introduced to each of the models, the goodness of fit was improved. Since the data has excess zeros the standard Poisson and negative binomial regression models were not appropriate this is due to the fact that the number of zeros in the data were beyond the model could predict. Consequently the standard error of the parameter estimates of the standard Poisson model was too small as compared with models that take in to account the variability in the data such as the Poisson mixed negative binomial, zero-inflated and other models. As a result, all covariates were significantly associated with B-cell production. This would lead to a wrong conclusion due to the inflation value of the Z-statistic; however in the case of negative binomial the standard error of the parameter estimates were too big as compared to the standard Poisson model.

In this study, it was found that NB, ZINB, NB mixed and ZINB mixed regression models were better fitted the data than ZIP and ZIP. This may be due to the high variability of the B-cells productions. ZIP mixed regression model with two random intercepts was better fitted the data than the standard Poisson and ZIP regression models. Furthermore zero-inflated negative binomial mixed regression model with two correlated random intercepts, one is for the logistic (inflated) part and the other is for the negative binomial part of the model was found to be the best. The data was also fitted again after removing the potential outlying observations in order to study their effect on the model that would be selected. In addition, to examine their impact on the parameter estimates of the model. In this case also ZINB mixed regression model with two correlated random intercepts was selected as the best.

The parameter estimates of the final model before and after excluding the outlying observations were close to each other. Thus in this study the zero-inflated negative binomial mixed (ZINB mixed) regression model was robust to the outlying observation. The different diagnostic tools such as the plot of the predicted Vs observed of the B-cells, Pearson Vs predicted value of the B-cells, endorsed that the new model fits best the data. In both final models, treatment (HPV-vaccine) and the interaction between treatment with linear and quadratic time effects were found to be significantly associated with the production of the B-cells. Hence the two HPV-vaccine effects are different in different time points.

In conclusion the zero-inflated negative binomial mixed model with two correlated random intercepts was better fitted with data which is characterized by excess zeros and high variability in the non-zero outcome.

From this study we can recommend that as this study is a small study, the result may not be generalizable, that is its external validity may not be valid. So that it would be better to examine in a large data set.

5.0. References

[1] .Stanley, M. (2005). Immune responses to human popilloma virus. *Science direct*, Elsevier.

[2] http://www.womenshealth.gov/publications/our-publications/fact-sheet/human-papillomavirus.cfm accessed on August 08,2011.

[3] World Health Report (2005). Report of the Consultation on Human Papillomavirus Vaccines Geneva, World Health Organization.

[4] Sturman, M. C. (1999). Multiple approaches to analyzing count data in studies of individual differences: The propensity for type I errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement*, **59**:414–430.

[5] Elhai , J. D., Calhoun, P.S., and Ford, J.D.(2008).Statistical procedures for analyzing mental health services data. *J. Psychiatry Research*, Elsevier, **160**: 129-136.

[6] Bonate, L. P. , Sung, C. and Richards, S.(2009).Conditional modelling of antibody titters using a zero-inflated Poisson random effects model: application to Fabrazyme® , . *J. Pharmacokinet Pharmacodyn*, **36**:443–459.

[7] Cameron AC. and Trivedi PK (1999). *Regression analysis of count data*. Cambridge University Press, Cambridge.

[8] Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley & Sons.

[9] *www2.sas.com/proceedings/sugi26/p247-26.pdf, accessed on july 12, 2011.*

[10] Karazsia,B.T. ,and Van Dulmen, M.H.(2008). Regression Models for Count Data: Illustrations using Longitudinal Predictors of Childhood Injury*. *Journal of Paediatric Psychology*, Kent State University, Belgium, **33**(10):1076-1084.

[11] Hilbe, M.J. (2007). Negative Binomial Regression. Second ed., Cambridge University Press, New York.

[12] Molenberghs,G.,Verbeke,G., Dem´etrio, G.B. C. and Vieira,A.(2007). A Family of Generalized Linear Models for Repeated Measures With Normal and Conjugate Random Effects.*J. Lifetime Data Anal.* **13**(4):513-31.

[13] Rider, P. R.(1961). Estimating the Parameters of Mixed Poisson, Binomial and Weibull Distributions by Method of Moments. *Bulletin de l'Institut International de Statistiques* 38, Part 2.

[14] Cohen, A. C. (1963). Estimation in Mixtures of Discrete Distributions. In Proceedings of the International Symposium on Discrete Distributions, Montreal, Quebec.

[15] Lord, D., Washington, S. P. and Ivan, J. N. (2005) .Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *J. Accident Analysis and Prevention*, Elsevier, USA **37**:35-46.

[16] Long, J. S., and Freese, J. (2006). *Regression models for categorical dependent variables using stata* (2nd, ed.)., College Station, TX: Stata Press.

[17] Liu, H. and Powers, A. D. (2007). Growth Curve Models for Zero-Inflated Count Data: An Application to Smoking Behaviour, *J. structural equation modelling*, **14**(2), 247–279.

[18] Zuur, F.A., Ieno, N. E, Walker, N. J., Saveliev, A. A. and Smith,M.G. (2009). *Mixed effects Models and Extensions in Ecology with R.* Springer, New York, USA.

[19] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to detects in manufacturing, *Technometrics*, **34**:1-14.

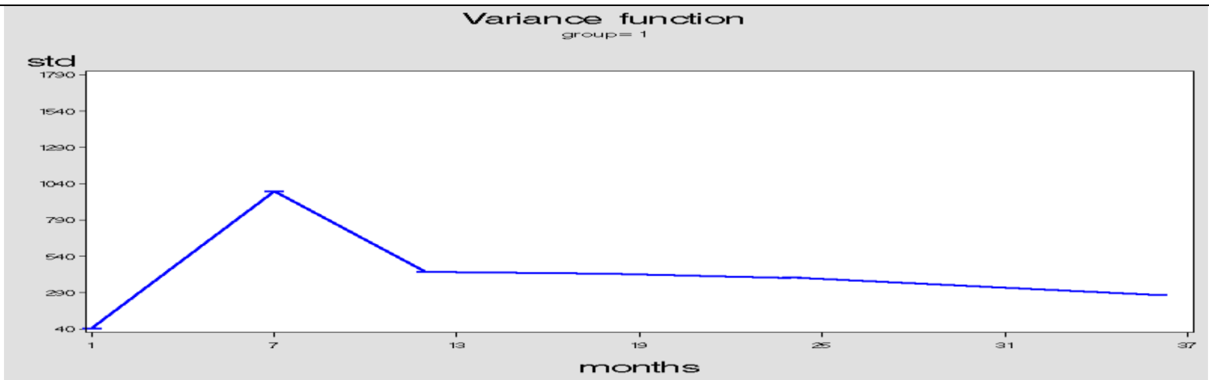
[20] Hur, K. (1998). A random-effects zero-inflated Poisson regression model for clustered extra-zero. Unpublished PHD thesis.

[21] Xiang, L., Lee, H. A., Yau, W.K. and McLachlan. J. G. (2007). A score test for overdispersion in zero-inflated Poisson, *J. Statist. Med.*, **26**:1608–1622.

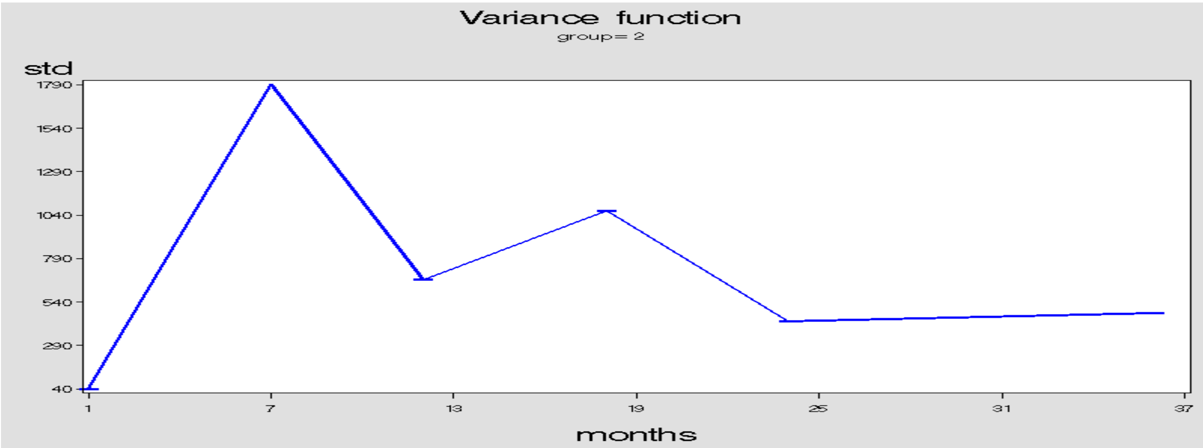
[22] Ismail, N. and Jemain, A. A. (2007). Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. *Casualty Actuarial Society Forum*,103-158.

[23] Molenberghs, G. And Verbeke, G. (2005). *Models for Discrete longitudinal Data.* Springer series in statistics, New York.

6.0. Appendix A



Panel A Variance function of the response for group I vaccine



Panel B Variance function of the response for group II vaccine

Figure 1A: Variance structure of group I (panel A) and group II (panel B)

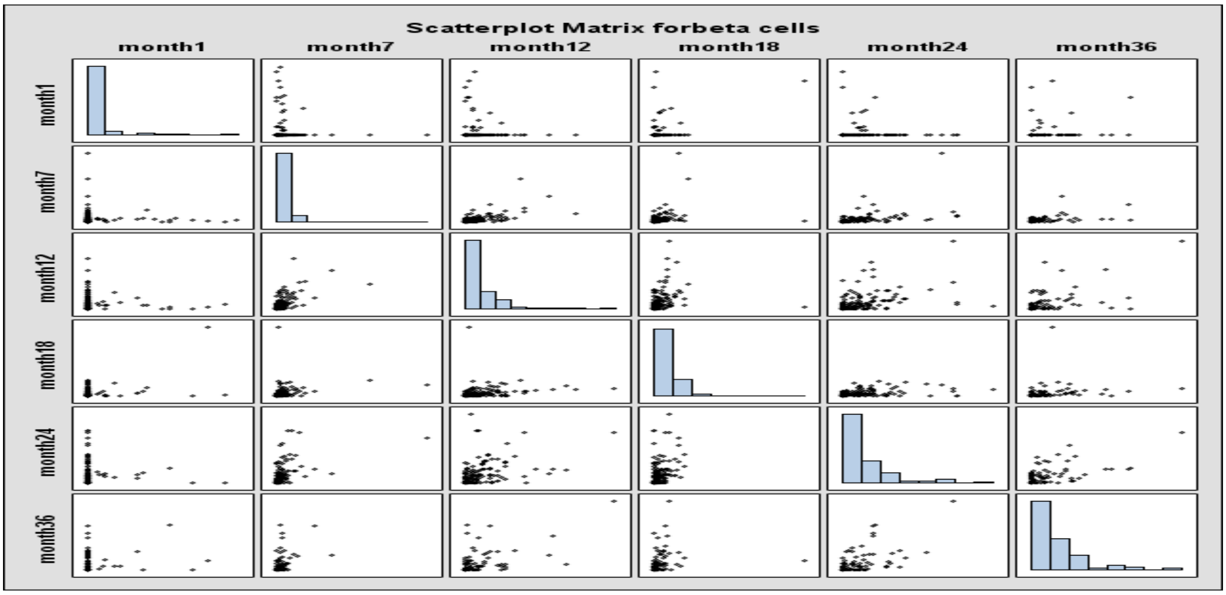


Figure2A: scatter plot matrix of the number of beta cell produced at each measurement time

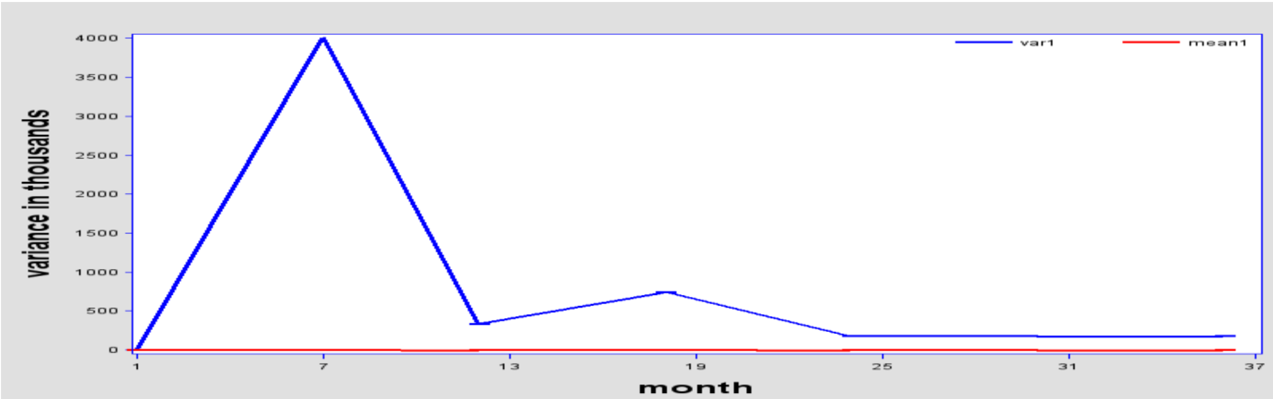
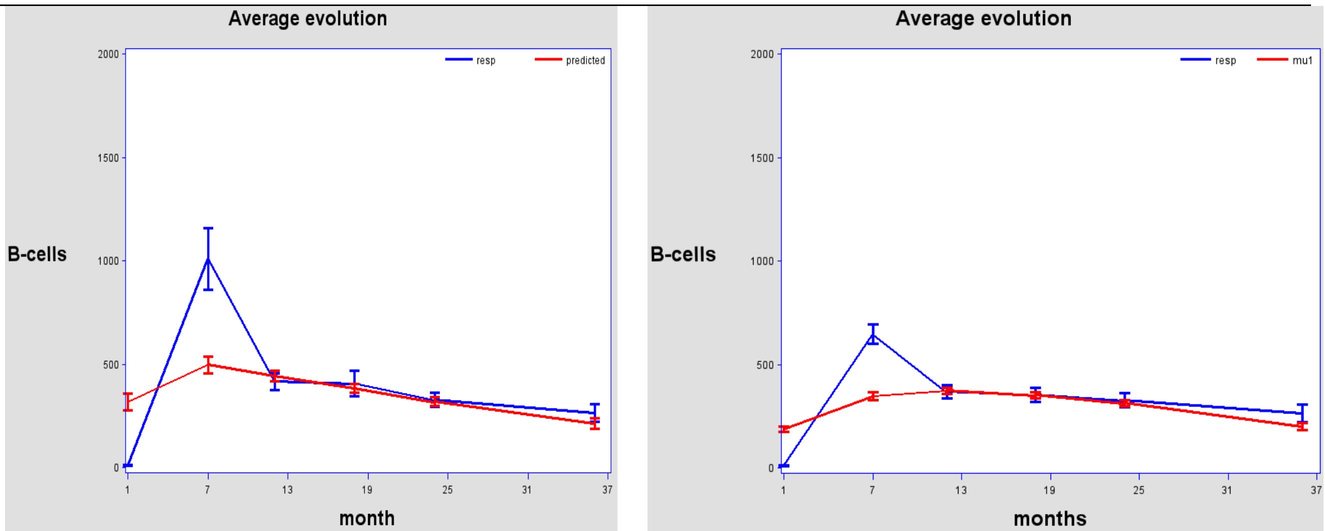


Figure 3A: Plot of the mean and variance function over time



Panel A: Before excluding an outlier

Panel B: After excluding an outlier

Figure 4A: Mean evolution of the number of B-cells produced per million cell before and after excluding outlier

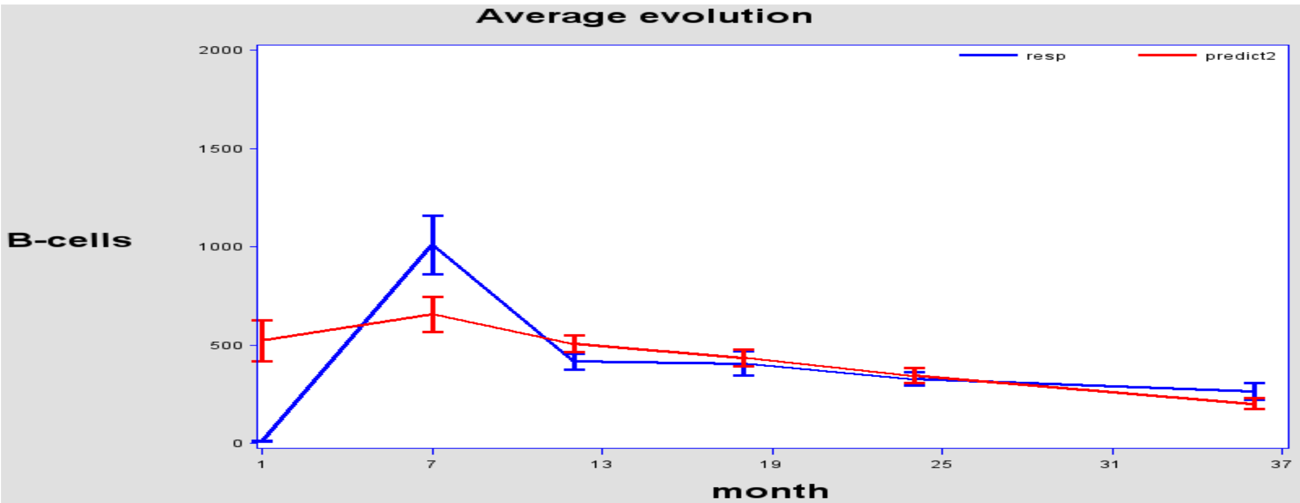


Figure5A: mean evolution plot of predicted and observed B-cells of the ZIP mixed model

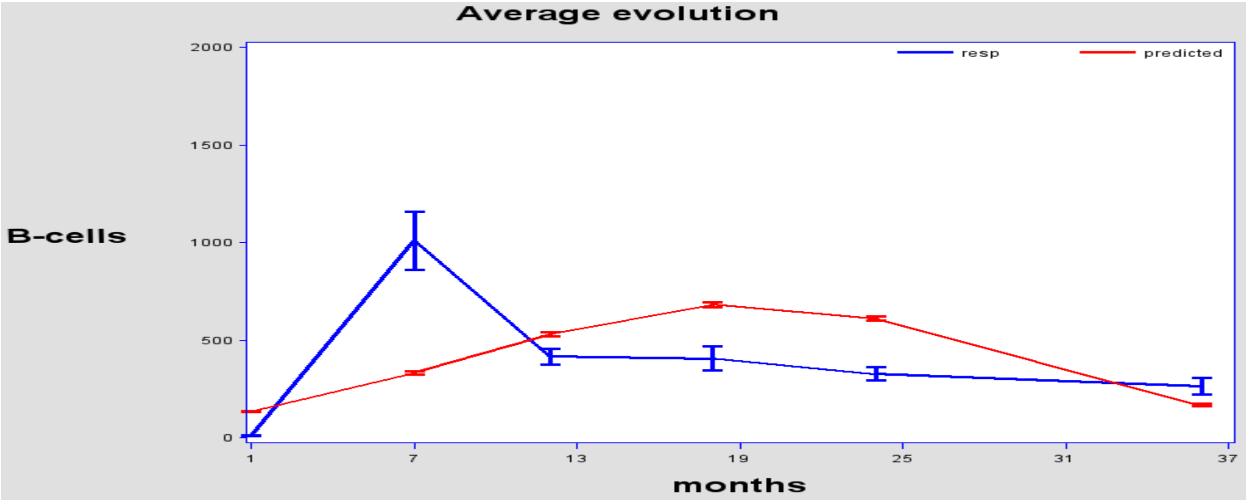
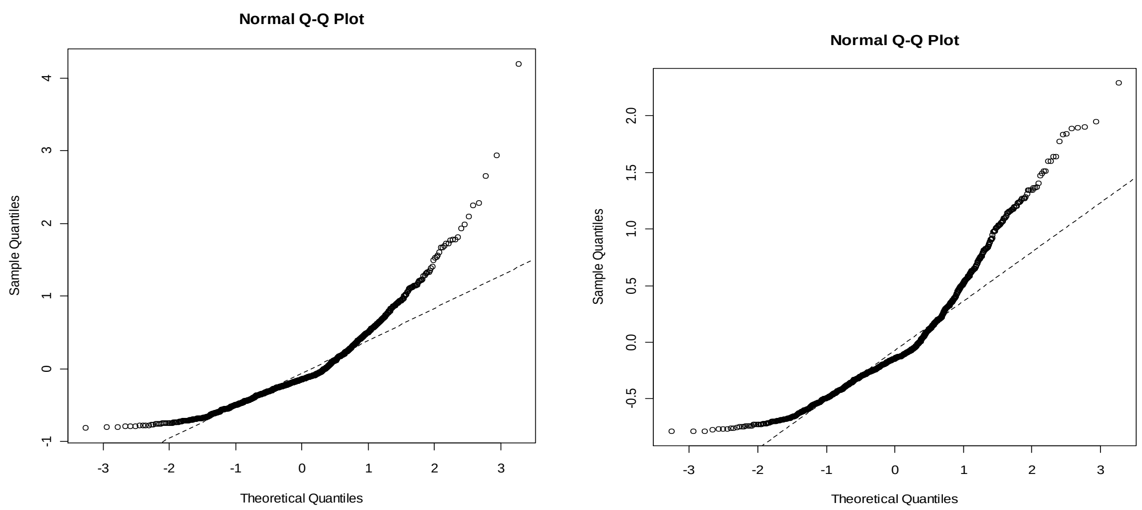


Figure 6A: Mean evolution plot of observed Vs predicted value of the negative binomial mixed model



Panel A: Q-Q plot of the full data *Panel B: Q-Q plot after excluding the outlier*

Figure7A: Q-Q-plot before (panel A) and after excluding the outlier (panel B)

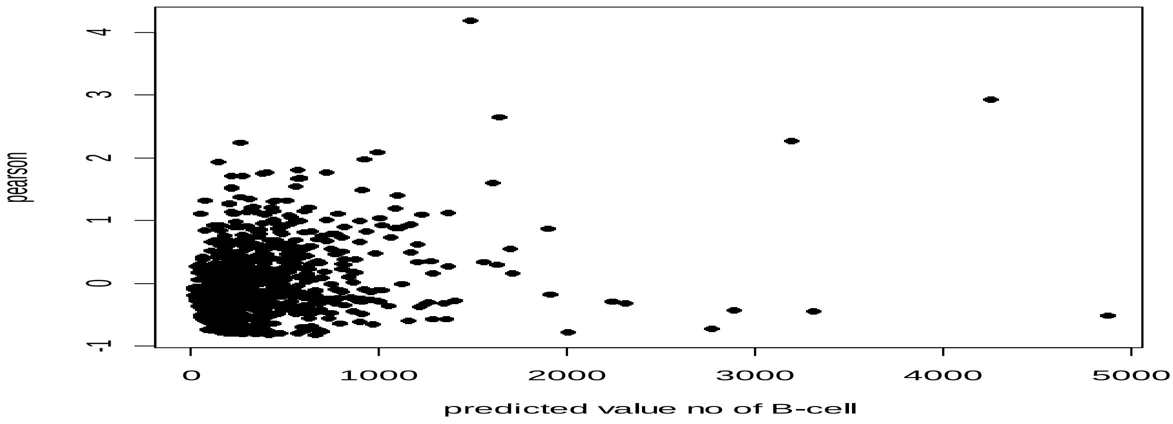
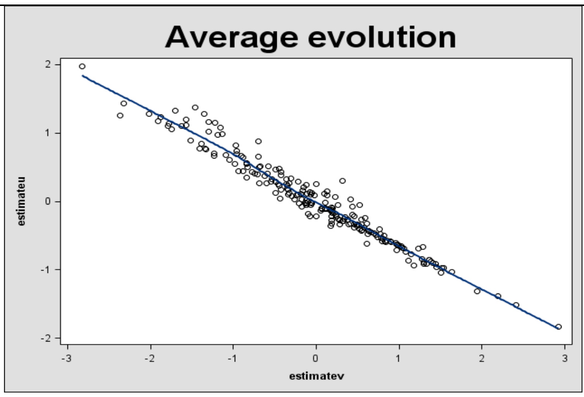
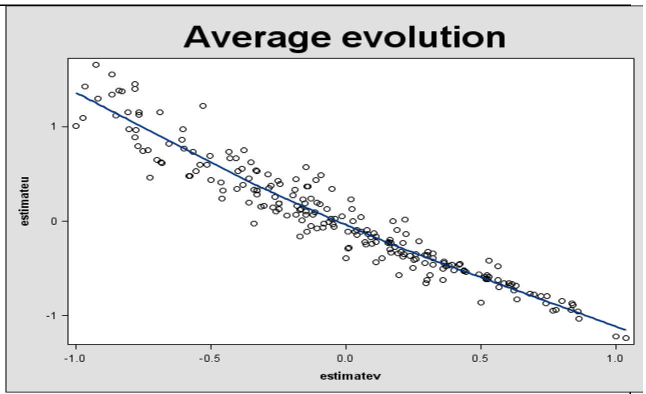


Figure8A: plot of Pearson residuals versus predicted value of the full data



Panel A: Plot of the random intercepts before excluding the outlier



Panel B: Plot of the random intercepts after excluding an outlier

Figure9A: Plot of the random intercepts *before (panel A) and after excluding the outlier (panel B)*

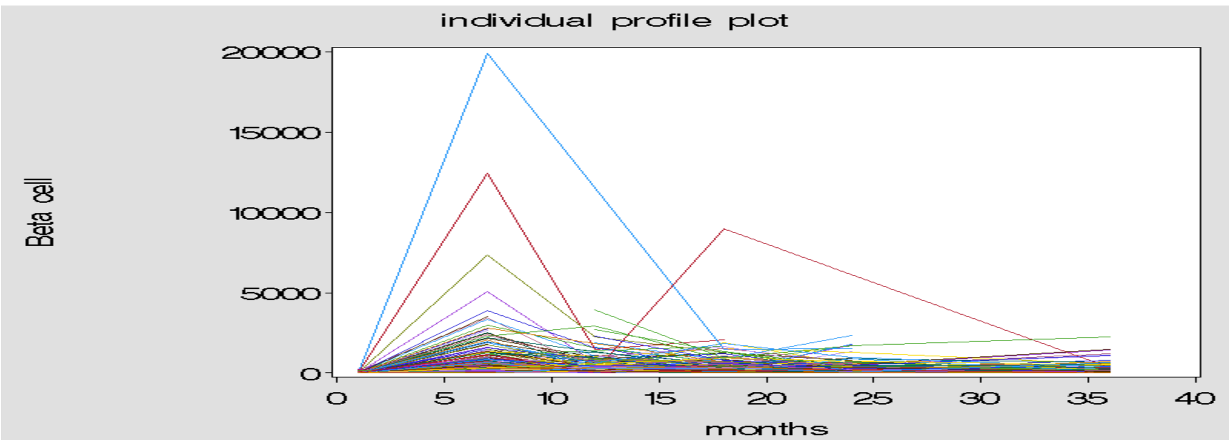
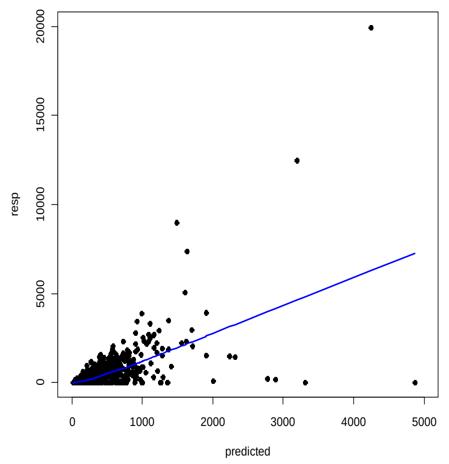
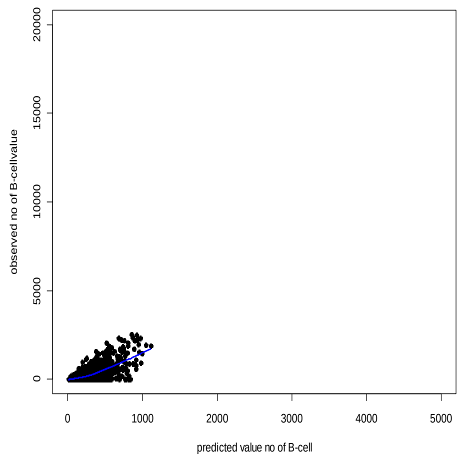


Figure 10A: individual profile plot



Panel A: Before excluding an outlier



Panel B: after excluding an outlier

Figure 11A: The plot predicted versus observed values

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Use of Zero-Inflated Models for analyses of immunological data

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Welegebrael, Aklilu Zemicael

Datum: **12/09/2011**