Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Masterthesis

HPV DNA detection in urine: effect of a first-void urine collection device and time of collection

Dunson Bwese Koge Ejedepang

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Epidemiology & Public Health Methodology

SUPERVISOR :
Mevrouw Robin BRUYNDONCKX

SUPERVISOR :
Dr. Alex VORSTERS

2017
2018

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Masterthesis

HPV DNA detection in urine: effect of a first-void urine collection device and time of collection

Dunson Bwese Koge Ejedepang

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Epidemiology & Public Health Methodology

SUPERVISOR :

Mevrouw Robin BRUYNDONCKX

SUPERVISOR :

Dr. Alex VORSTERS

Dedication

EJEDEPANG BWESE KOGE PETER

Acknowledgements

I would like to express my sincere and immense gratitude to my supervisor BRUYNDONCKX ROBIN for her support and guidance towards the realization of this project.

My appreciation also goes to my external supervisor Dr. VOSTERS ALEX at the university of Antwerpen for all his suggestions towards this project.

My immense appreciation also goes to all the lecturers of the faculty of Science at the university of Hasselt for all the knowledge and skills they transfered to me which was of great help towards the realization of this project.

My appreciation also goes to all all my course mates for being collaborative with a competitive attitude throughout my study at the university of Hasselt.

Finally, i would like to thank Ejedepang Bwese Peter, Rose Epie Mechang, Marino Hauben, Astrid Barthels, Hanne Hauben for their immense support and encouragement to enable me go through the program.

Abstract

Human papillomavirus (HPV) is the most common sexually transmitted infection. HPV is not a sufficient cause, but it is a component cause of cervical cancer and genital warts (Lacy et al ., 2006). HPV can be detected in humans using samples from cervix or urine.
The purpose of this studies is to evaluate the efficiency of a proptotype first-void urine collection device (Colli-Pee) against the standard collection cup and to assess the effect of time of collection on the detection of human and HPV DNA in women. The effect of human DNA on the detection of HPV DNA was also investigated. The dependency of the likelihood of positive HPV and the amount of HPV DNA was also investigated.
The HPV response has excess zeros and is positively skewed. A two-part generalized linear model (GLM) via GEE and random effects model was used to model the excess zeros. For the GLM, a logistic model with logit link was used to model the odds of positive HPV. A log-normal and gamma distribution were considered for the positive response and fitted via GEE using an exchangeable working correlation structure. For the two-part random effects model, a log-normal distribution was assumed for the positive response. This was later extended to generalized gamma distribution and log-skew-normal distribution. For the random effects model, a generalized linear model was used to fit the likelihood of positive HPV and a linear mixed model was used to model the amount of HPV DNA. The human DNA response is positively skewed. A lognormal and gamma model were fitted via GEE using an exchangeable working correlation structure.
The models used actually addressed the objectives of this study. In all models used, Colli-Pee device was significantly correlated with higher amount of human and HPV DNA but there was no significant effect of the period of the day in the detection of of both the amount of human and HPV DNA. Also the random effects model was used to model the 'cross-part' correlation which was not significant though for all models. In this study, a two-par random effects model with a generalized gamma distribution for the positive values was selected as the most parsimonious model for HPV DNA because of it's low AIC and likelihood value. The lognormal model was considered the most parsimonious for human DNA response because of its low QIC value.

Keywords:Two-part model, Log-normal distribution, Gamma distribution, Log-skew-normal distribution, GEE, random effects.

Contents

List of Figures

List of Tables

1 Introduction

Human papillomavirus (HPV) is one of the most common sexually transmitted infection among females (Dunne et al ., 20017). HPV is not a sufficient cause, but it is a component cause of cervical cancer and genital warts (Lacy et al ., 2006). The prevalence of HPV is notably high among young females within the first few years after sexual intercourse with a prevalence rate as high as 40-80%. In most infected persons, the infection clears spontaneously without any clinical signs or symptoms. In a few persons, the infection may become persistent leading to cervical cancer and other genital related cancers (Bosch et al. ,2012).

There are over 170 known HPV types and 40 of these infect the anus and genitals. HPV types are classified based on their ability to cause cervical cancer. "Low risk" HPV (LR-HPV) types are known to cause benign or precancerous lesions in the cervix and genital warts. "High risk" HPV (HR-HPV) types can cause cervical, anal, and other genital cancers.There are about 13 types of HR-HPV that can cause abnormal cells to form on the cervix. These abnormal cells may gradually develop into cervical cancer if not removed. The 13 types of high-risk HPV that are of most concern are 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 68. About 99% of cervical cancers is attributed to HR-HPV types. Approximately 70% of cervical cancers are due to HPV 16 and 18. (Bosch et al., 2012)

HPV can be detected in humans using either cervical or urine sample. Cervical sampling is more sensitive but invasive. The decreased sensitivity of HPV testing in urine samples compared with that of cervical samples may be associated with the low HPV DNA amount in urine. Urine sampling has several advantages over cervical sampling in that it is noninvasive and does not interfere with the natural process of the infection. In contrast to urine sampling, cerivical sampling may lead to increased infection as samples obtained by scraping the epithelium usually create micro-lesions may induce an inflammatory reaction in the presence of viral particles. Urine is easy to collect and can be done by the patient (self-sampling) at home and sent by mail to the laboratory for analysis. Therefore, urine sampling can be performed more frequently. The reasoning behind urine sampling is that debris from exfoliated cells from the cervix can accumulate around the uterus exit and contaminate the first-voided (initial stream) urine. This could contain HPV fragments which could be detected using molecular techniques. The amount of HPV DNA is known to be associated with the amount of human DNA, which is influenced by the amount of debris from exfoliated cells that contaminates the first-void urine (Vorsters et al.,2012).

The purpose of this study is to assess the efficiency of a prototype first-void urine collection device (Colli-Pee) against the standard collection cup and to assess the effect of time of collection on the detection of human and HPV DNA in women. Also we want to know if there is a dependence on the likelihood of positive HPV on HPV viral load. Furthermore, it is a

sub-objective to assess the effect of human DNA on HPV DNA.

2 Methodology

2.1 Data

The dataset consists of 33 participants with self reported HPV infection. They were asked to provide eight first-void (FV) urine samples (four in the morning and four in the evening) over a period of four days. Two FV urine collection methods were alternated ,.ie. the Colli-Pee (TM) device and a collection set with standard urine cup. Human DNA and HPV DNA quantification and detection was performed by real time PCR. The variables used in the study were:

Table 1: Variables used in the study

Variable	Type	Description
NewHPVDNA	semi-continuous	The concentration of HPV DNA
con(hDNA)	semi-continuous	The concentration of human DNA (Response)
ID	continuous	The Subject's ID
Day	continuous	The number of days the subject was followed
Period	Categorical	Period of the day (M=Morning or$E$ =Evening) urine was collected
Device	Categorical	The type of urine collection device (Colli-Pee or Cup)
Type	Categorical	High risk(HR) and Low risk(LR)

2.2 Exploratory Data Analysis

Exploratory analysis was done using box plots and histograms. Also tables were used for summary statistics.

2.3 Statistical Models

The HPV samples are continuous but with a point mass at zero. Several models have been proposed to deal with such response. Firstly, generalized linear model with a gaussian distribution on the transformed outcome after adding a small constant to the zero values . However assuming a parametric distribution such as gaussian cannot account for the excess zeros in the outcome and will inevitably lead to biased inference. Also, performing a log transformation of the outcome in an attempt to normalize the data usually leads to erroneous inference due to the fact that it is heavily skewed and will more often than not result in an asymmetric distribution. Tobin in 1958 and Heckman in 1979 proposed a Tobit and selection model respectively.

These models assume that the dependent outcome follows a censored normal distribution and all zeros are artificial zeros i.e values below detection limit. But in our case, the response is a semicontinuous variable which is different from left-censored or truncated variables in that the zeros are bonafide valid data .
A two-part model for semi-continuous response was proposed by Gragg to model the probability of the zero response and the amount of positive response. An interesting feature of a two-part model as opposed to the Tobit model is that the zeros are assumed to be valid true response and this makes sense for skewed data with a zero point mass which is the case of HPV DNA . (Gragg JG .1971).

Correlated Observations

For each subject, measurements of HPV concentrations were recorded for each urine sample. Since measurements were done on all eight urine samples from the same subject, we expect that the measurements should be correlated within subjects. Treating these measurements as independent samples may lead to biased inference. Two methods were used in this paper to tackle the clustered measurements. These are the two-part generalized linear model (GLM) via Generalized Estimating Equations (GEE) and two-part random effects model.

2.3.1 Two-Part GLM via GEE

Two-part GLMs via GEE is a natural extension of two-part models for cross-sectional data. Two different models were used to describe the ”binary part” and the ”continuous part” of the semicontinuous HPV data, and each part separately accounts for correlation among repeated measures. These two independent models are fitted via GEE.
Generalized Estimating Equation (GEE) was introduced by Liang and Zeger (1986) as a method of parameter estimation for correlated (clustered and repeated) data. It is a common choice for marginal modeling of response if one is interested in the marginal mean parameters (population average) rather than subject-specific estimations. One advantage of GEE is that the estimates of the regression coefficients are consistent even with the misspecification of the variance-covariance structure (Molenberghs and Verbeke, 2005). The correlation between the vector of repeated measurements taken from a given subject Y_i is captured by specifying an association within the subject through a so called working correlation structure.(Molenberghs and Verbeke 2005). The marginal expectation $E(Y_{ij}) = \mu_{ij}$ can be directly modeled through known covariates.
If we consider a random sample of n subjects, Y_{ij} is the response of the ith subject at jth mea-

surement. X_{ij} is a vector of p covariates. The observations within a subject are correlated and observations of different subjects are assumed to be independent. The marginal expectation is estimated by solving the score equation

$$\sum_{i}^{n} D_i^T V_i{-}1(Y_i - \mu_i) = 0$$

with $\mu_i = g^{-1}(\beta^T X_{ij})$; $D_i = D_i(\beta) = \frac{\psi \mu_i(\beta)}{\psi \beta}$; $V_i = U_i^{1/2W(\alpha)}$ the working covariance matrix; $W(\alpha)$ is the working correlation matrix, parameterized by parameter vector α, U_i is a diagonal matrix with diagonal elements $Var(Y_{ij}|X_{ij})$.

Different working correlation structures can be assumed when using GEE such as independent, exchangeable, autoregressive and unstructured. For instance the exchangeable, independent and unstructured working assumptions were used in this study and the empirical based and model based standard errors were compared to validate the choice of the variance-covariance structure. Though the empirical standard errors are roburst to misspecification of the association structure, correctly specifying it might improve efficiency (Molenberghs and Verbeke 2005). In this analysis, the exchangeable working correlation structure was assumed as the empirical and model based estimates were closest compared to the other association structures.

Two-part GLM specification

Let Y_{ij} be a semi-continuous response for subject (cluster) i where i=1,...,n with jth measurements where j=1,...,N. The response variable is represented as Z_{ij}, where ;

$$Z_{ij} = \begin{cases} 0, & \text{if} Y_{ij} = 0 \\ 1, & \text{if} Y_{ij} > 0 \end{cases}$$

The probability that $Y_{ij} > 0$ was assumed to follow a binomial distribution ;

$$logit(Prob(Z_{ij} = 1)) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4}$$

where $x_{i1} = 1$ if subject i provided morning urine sample and zero otherwise, $x_{i2} = 1$ if subject i used Colli-Pee device and 0 otherwise, $x_{i3} = 1$ if subject i was analyzed for HR HPV type and 0 otherwise, x_{i4} is the concentration of human DNA and α_1 to α_4 is the matrix of the model parameters.

Given that $Y_{ij} > 0$, the model assumes that log(Y_{ij}) follows a normal distribution with a constant variance i.e Y_{ij} follows log-normal distribution with constant variance.

$$E[log(Y_{ij}|Y_{ij} > 0)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

where $x_{i1} = 1$ if subject i provided morning urine sample and zero otherwise,$x_{i2} = 1$ if subject i used Colli-Pee device and 0 otherwise,$x_{i3} = 1$ if subject i was analyzed for HR HPV type and 0 otherwise, x_{i4} is the concentration of human DNA, and β_1 to β_4 is the matrix of the model parameters.

2.3.2 Gamma Model

In the previous model, the log transformation of the positive values is used part II leading to an assumption that the positive values follow a log-normal distribution with constant variance. However, making such an assumption might be misleading especially in the presence of strong skewness and heteroscedasticity. A gamma distribution was assumed as a variation to this assumption. We assume that the positive values follow a gamma distribution. The gamma model relates the $E(Y_{ij}) = \mu_{ij}$ to the covariates through a log link function;

$$log(E[Y_{ij}]) = \beta_0' + \beta_1' x_{i1} + \beta_2' x_{i2} + \beta_3' x_{i3} + \beta_4' x_{i4}$$

where $x_{i1} = 1$ if subject i provided morning urine sample and zero otherwise, $x_{i2} = 1$ if subject i used Colli-Pee device and 0 otherwise, $x_{i3} = 1$ if subject i was analyzed for HR HPV type and 0 otherwise, x_{i4} is the concentration of human DNA, and β_1' to β_4' is the matrix of the model parameters

2.3.3 Two-part random effects model

2.3.4 Log-normal model

Two-part random effects model proposed by Oslen and Schafer (2001) and Tooze et al (2002), incorporates a random effects in the two-part model to capture the correlation in the response. In part I, the odds of $Y_{ij} > 0$ defined by

$$\pi_{ij} = P(Y_{ij} > 0 | X_{ij}, a_i, b_i)$$

is

$$logit(\pi_{ij}) = \alpha_0' + \alpha_1' x_{i1} + \alpha_2' x_{i2} + \alpha_3' x_{i3} + \alpha_4' x_{i4} + a_i$$

The odds is modeled using a generalized linear mixed model with logit link. This is a subject specific model, thus parameters are conditional upon the random effects. For part II of the model, given that $Y_{ij} > 0$, conditional on the random effects b_i, the model assumes that Y_{ij} follows a log-normal distribution with constant variance.

$$log(Y_{ij} | Y_{ij} > 0, X_{ij}^*) = \beta_0'' + \beta_1'' x_{i1} + \beta_2'' x_{i2} + \beta_3'' x_{i3} + \beta_4'' x_{i4} + b_i + \epsilon_{ij}$$

where $x_{i1} = 1$ if subject i provided morning urine sample and zero otherwise, $x_{i2} = 1$ if subject i used Colli-Pee device and 0 otherwise, $x_{i3} = 1$ if subject i was analyzed for HR HPV type and 0 otherwise, and x_{i4} is the concentration of human DNA. α'_1 to α'_4 and β''_1 toβ''_4 are the fixed effects while a_i and b_i denote the random effects for part I and II respectively. In this two-part random effects model, a generalized linear mixed model was used to model the binary response in Part I and a linear mixed model for the natural log of the positive continuous response in part II. The random effects a_i and b_i are assumed to be normal and correlated

$$v_i = (a_i, b_i)^T \sim N(0, \sum)$$

where

$$\sum = \begin{pmatrix} \sum_{aa} & \sum_{ab} \\ \sum_{ba} & \sum_{bb} \end{pmatrix}$$

The random effects are incorporated to capture the within correlation due to repeated measurements on the same subject, and the "cross-part" correlation between the frequency and the amount of HPV detected. This "cross-part" correlation is of importance and could lead to biased results if ignored. For instance subjects who are more likely to be HPV positive may or may not have a high amount of HPV viral load.
The likelihood contribution of part I model is

$$ll1 = \sum_{j=1}^{n_i} [I(Y_{ij>0}) logit(\pi_{ij}) + log(1 - \pi ij)]$$

The likelihood contribution of part II model assuming $e_{ij} \sim N(0, \sigma_e^2)$ is

$$ll1 = \sum_{j=1}^{n_i} I(Y_{ij>0}[-0.5 log\sigma_e^2 - logY_{ij} - 0.5(\frac{logY_{ij} - \mu_{ij}}{\sigma_e})^2])$$

where $\mu_{ij} = \beta''_0 + \beta''_1 x_{i1} + \beta''_2 x_{i2} + \beta''_3 x_{i3} + \beta''_4 x_{i4} + b_i + \epsilon_{ij}$
The model above assumes that the log of the positive response follows a normal distribution with constant variance as noted above, such assumption might not be entirely true in the presence of strong skewness. Liu et al (2012) generalized this two-part random effects model by proposing other distributional assumption to tackle the skewed nature of the continuous part of the model. They proposed the generalized gamma and log-skew distribution.

2.3.5 Generalized Gamma Distribution

This model assumes that the positive values follow a generalized gamma distribution with parameters k, μ_{ij} and σ_{ij}, where

$$\mu_{ij} = X_{ij}^T \beta + b_i$$

,

$$\sigma_{ij}^2 = exp(X_{ij}^T \delta)$$

where β and δ are regression coefficients for the fixed effects and heteroscedacity.The dependence of σ_{ij} on X_{ij} permits for heteroscedacity.The likelihood is written thus

$$LL_2 = \sum_{j=1} n_i I(Y_{ij} > 0)[(\eta - 0.5)log(\eta) - log\sigma ij - logy_{ij} - log\Gamma(\eta) + \mu ij\sqrt{\eta} - \eta exp(|k|\mu_{ij})]$$

where $\mu_{ij} = sign(k)(logy_{ij} - \mu_{ij})/\sigma_{ij}$
The exponential link function ensures that the estimated scale parameter is positive.

2.3.6 Log-Skew-Normal Distribution

The customary statistical approach of applying a log transformation in setting of right skewness is ad hoc, and may or may not optimally account for distributional characteristics of the data under study. Usually, the log transformation may over-transform the data making the distribution skewed in the opposite direction. In an attempt to remedy this problem (Lui et al.,2010) extended the conventional two-part random effects model by suggesting a log-skew-normal distribution for the positive values. This model assumes that the log of the positive values follows a log-skew-normal distribution with parameters λ, μ_{ij} and σ_{ij},where

$$\mu_{ij} = X_{ij}^T \beta + b_i$$

,

$$\sigma_{ij}^2 = exp(X_{ij}^T \delta)$$

The likelihood is written thus

$$LL_2 = \sum_{j=1} n_i I(Y_{ij} > 0)[-0.5log(\sigma_{ij}^2 + \lambda_{ij}^2) + log\frac{2}{y_{ij}} + log\phi(\frac{logy_{ij} - \mu_{ij}}{\sqrt{\sigma_{ij}^2 + \lambda_{ij}^2}}) + log\Phi(\frac{\lambda}{\sigma_{ij}}\frac{logy_{ij} - \mu_{ij}}{\sqrt{\sigma_{ij}^2 + \lambda_{ij}^2}})]$$

The skew-normal distribution accommodates asymmetry in a more flexible manner, and can model both positively or negatively skewed data (depending on the sign of the skewness parameter) reducing to the normal distribution when the skewness parameter is zero.

2.4 Estimation Procedure

Parameter estimation was performed with GEE using GENMOD for the two-part generalized linear model. For two-part random effects model, estimation was done via maximum-likelihood estimation using PROC NLMIXED in SAS. It is very necessary that this likelihood be approximated in order to yield accurate results. For example, Olsen and Schafer used a sixth-order Laplace approximation, (Olsen and Schafer 2001) whereas (Tooze et al.,2002) made use of the adaptive Gaussian quadrature. Both methods actually yield accurate estimates. In this study, however, adaptive Gaussian quadrature was used as it is much easier to implement.

2.5 Software

SAS 9.4 version was used for statistical analysis. In addition, all statistical tests were done at 5% significance level.

3 Results

3.1 Exploratory Data Analysis (EDA)

3.1.1 Summary statistics

Table 2: Summary statistics the semi-continuous variables by Period and Device

Device	Period	Variable	Mean	S.E	Lower C.L	Upper C.L
Colli-Pee	E	NewHPVDNA	6620985.49	88364732.47	1463314.08	11778656.91
		con_hDNA	15.5066372	19.7205155	14.3555904	16.6576839
	M	NewHPVDNA	15153909.48	48.364950116	-6242583.00	36550401.95
		con_hDNA	15.2631786	23.3964084	13.8914813	16.6348758
Cup	E	NewHPVDNA	1594584.74	24781893.23	162023.29	3027146.18
		con_hDNA	8.9604791	11.4039433	8.2852029	9.6357552
	M	NewHPVDNA	1594584.74	24781893.23	162023.29	3027146.18
		con_hDNA	8.3348750	14.8632364	7.4756792	9.1940708

Summary statistics were used to describe the data set. A histogram was used to explore the distribution of the HPV and human DNA concentration. Table 2 is a summary of the response variables (NewHPVDNA) and con(hDNA) by device and period. The standard deviation of the NewHPVDNA is very high compared to its mean and Colli-Pee seems to be better in the morning than the cup in detecting NewHPVDNA .

Variable	Min	1st Quartile	Median	3rd Quartile	Max
NewHPVDNA	0.000	0.000	0.000	0.000	1.19243E+10
conc(hDNA)	0.000	2.310	5.720	13.240	142.600

Table 3: More than 75% of the HPV responses are zeros, thus it is appropriate to model the zero separately from the positive values. While less than 1% of human DNA was actually zero. The large proportion of zero HPV responses is due to the fact that when a sample has zero HPV, the rest of the 7 samples are also zero for HPV.

Figure 1 and 2 shows a histogram of the distribution of the concentration of human and HPV DNA. It can be seen from the plots that the mass of the values is around zero for HPV DNA concentration and the nonzero values are positively skewed. This motivated the use of a two part model for such a distribution. Figure 2 shows the responses for human DNA are positively skewed.

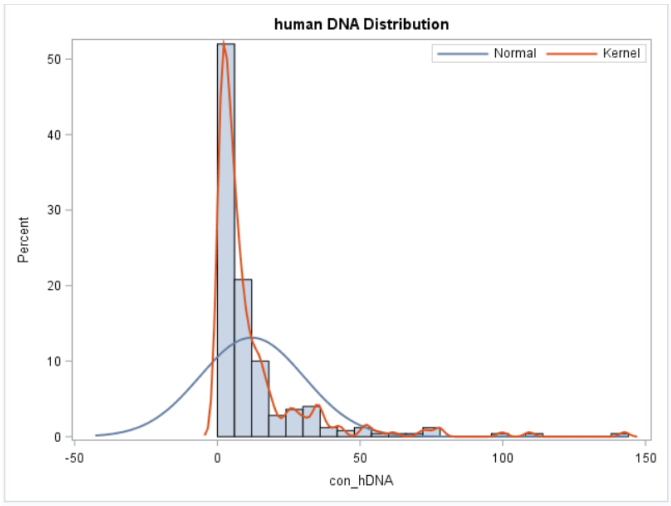


Figure 1: Histogram of the concentration human DNA

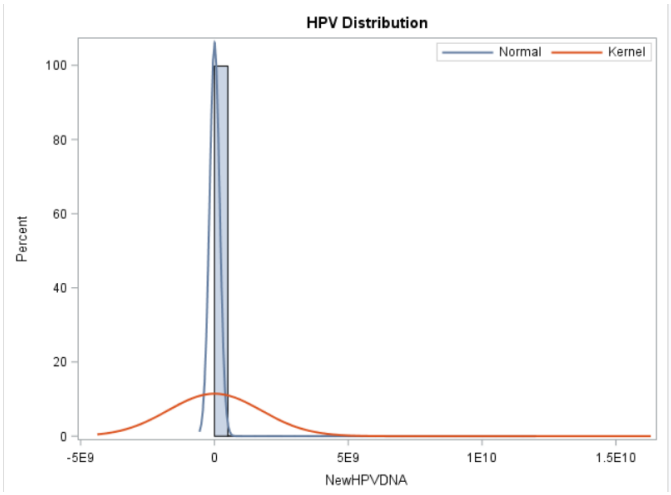


Figure 2: Histogram of the concentration of HPV DNA.

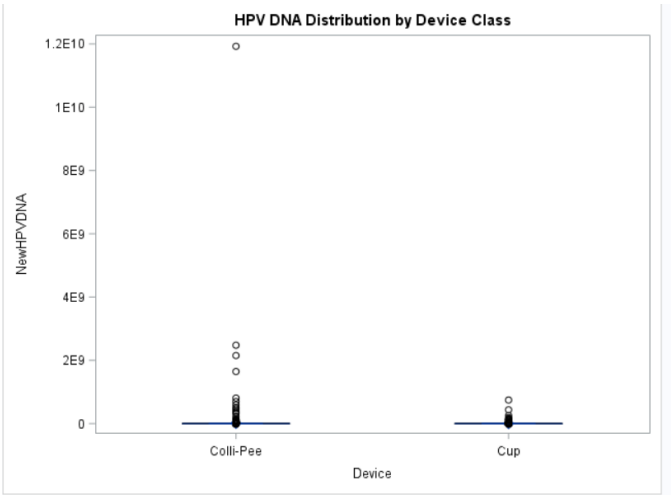


Figure 3: Boxplot of the concentration HPV DNA by device type

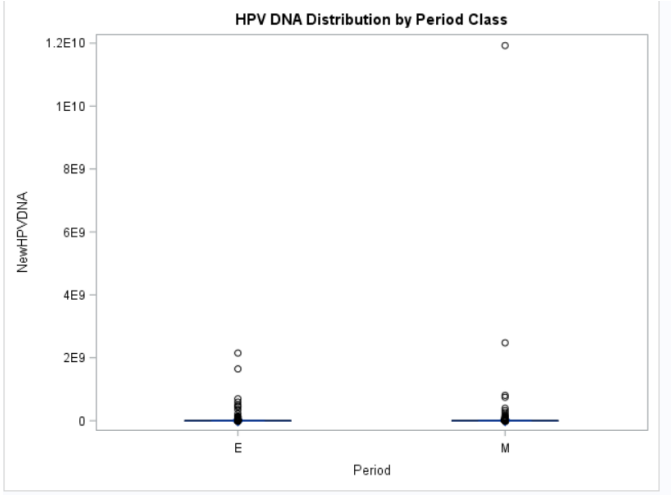


Figure 4: Boxplot of the concentration of HPV DNA by period type

Figure 3 is a boxplot of the concentration of HPV by the device used. It can be seen that Colli-Pee seems to be a better device as compared to the standard cup but there seems to be no difference in the collection period in the detection of HPV. Also there seems to be an inlier (outlier within the subgroup). This was investigated by analyzing the data with and without the outlier but the result was the same.

3.2 Results for GLM via GEE

The results of the logistic, log-normal and gamma model are shown in table 4 below.

Variable	Logistic GEE (S.E) with logit link	log-normal (S.E) with identity link	Gamma GEE(S.E) with log link
Intercept	-2.922(0.282)***	12.298(0.791)***	15.300(0.302)***
Period=('E')	-0.060(0.066)	0.078(0.269)	-0.283(0.278)
Device=('Colli-Pee')	-0.024(0.078)	1.100(0.331)***	1.507(0.517)***
Type=('HR')	0.296(0.350)***	0.672(0.640)	1.807(0.340)***
conc(hDNA)	0.012(0.003)***	0.006(0.013)	0.044(0.004)***

Table 4: GEE Parameter Estimates for the human DNA : where *=significant effect.

Assuming that the log transformed positive values follow a normal distribution, we confirm that Colli-Pee urine collection device is associated with lower odds of positive HPV though the effect wasn't significant but significantly correlated with higher amount of HPV DNA with a mean estimate of exp(0.088)=1.092 and a p-value of 0.0009. HR-HPV was significantly associated with higher odds of positive HPV and was correlated with higher amount of HPV DNA though the effect was not significant. The concentration of human DNA was significantly associated with higher odds of positive HPV p=0.003, but was correlated with higher amount of HPV DNA though the effect was not significant (p=0.658).

Assuming that the positive values follow a gamma distribution, we again confirm that Colli-Pee is significantly correlated with higher amount of HPV DNA with an estimated value of 1.507 with $p = 0.003$. Also there was no significant effect of the period of the day in the detection of HPV. The concentration of human DNA was significantly correlated with higher amount of HPV DNA (p<.0001). The log-normal model had a lower QIC (322) than the gamma model (968) and the residual plot in figure 8 in the appendix shows that the log-normal model seems a better fit to the data compared to the gamma. Thus the log-normal model was preferred to the gamma model.

3.3 Results of two-part random effects model

Variable	Log-skewed (S.E) -Normal model	log -Normal(S.E) Model	Generalized Gamma (S.E) Model
Part I (binary outcome α)			
Intercept	-3.582(0.311)**	-3.586(0.213)**	-3.591(0.314)**
Period=('E')	-0.057(0.124)	-0.057(0.124)	-0.056(0.124)
Device=('Colli-Pee')	-0.040(0.129)	-0.040(0.129)	-0.045(0.129)
Type=('HR')	0.313(0.124)**	0.313(0.124)**	0.313(0.124)**
conc(hDNA)	0.014(0.004)	0.014(0.004)	0.015(0.004)***
Part II (continuous outcome β)			
Intercept	12.077(0.932)**	11.998(0.992)**	13.419(0.927)**
Period=('E')	0.059(0.292)	0.073(0.291)	0.075(0.215)
Device=('Colli-Pee')	1.104(0.309)**	1.080(0.307)**	0.976(0.230)**
Type=('HR')	0.690(0.421)	0.677(0.411)	0.326(0.349)
conc(hDNA)		0.008(0.011)	0.012(0.009)
Heteroscedacity (δ)			
Intercept	1.842(0.191)**		1.029(0.306)**
Period=('E')	0.050(0.182)		-0.112(0.214)
Device=('Colli-Pee')	-0.071(0.196)		0.123(0.235)
Type=('HR')	-0.069(0.190)		-0.016(0.219)
conc(hDNA)	0.0008(0.003)		-0.003(0.004)
Variance (σ^2)			
σ_1^2	2.014(0.772)**	2.034(0.784)**	2.050(0.790)**
σ_2^2	10.104(3.207) **	1(3.113)**	8.889(3.203)**
σ_{12}	0.347(1.828)	0.239(1.910)	1.228(1.728)
λ	-4.014(0.211)***		
k			1.467(0.260)**
AIC	11518	11509	11466
-2Loglikelihood	11480	11481	11428

Table 5: Random effects model for all three distributional assumptions where **=significant effects, σ_1^2 and σ_2^2 are the variance of part I and II random intercept. λ is the skew parameter. The estimates for all three models are very similar except for the heteroscedacity estimates.

The result of all three models is shown in table 5 above. The most appropriate model was selected base on loglikelihood and AIC . The model with the smallest loglikelihood or AIC value was considered the most parsimoniuous model.

3.3.1 Generalized Gamma Distribution

Assuming that the positive values in part II follow a generalized gamma distribution, the Colli-Pee device is significantly correlated with higher amount HPV DNA and thus significantly better than the standard cup (p=0.0002) but was associated with lower odds of positive HPV though the effect was not significant (p=0.749). We also found out that morning samples were associated with lower odds of positive HPV though the effect was not significant (p=0.651) but contain higher amount of HPV than evening samples though the effect still was not significant (p=0.730). The concentration of human DNA was associated with higher odds of positive HPV DNA (p=0.0008) and was correlated with higher amount of HPV DNA though the effect was not significant (p=0.196). There is a positive cross-part correlation but the effect is not significant. This implies there is no significant dependence of the odds of positive HPV on the amount of HPV.
The shape parameter was significant ($p < 0.0001$) against all nested models of the generalized family such as log-normal, weibul, and gamma confirming that none provides a good fit to the data as generalized gamma distribution.

3.3.2 Log-skew normal Distribution

Assuming that the log-transformed values follow a log-skew normal distribution, we confirmed that the log-transformed values are left-skewed, indicating an overcorrection of the skewness by log transformation as the parameter estimate is -4.014, with a significant skewness ($p < 0.0001$). The results are identical to the generalized gamma model.

3.3.3 Log-normal model

Given that the positive HPV values follow a log-normal distribution, the concentration of human DNA was associated with higher odds positive HPV DNA (0.0009) but was not significantly associated with the amount . We also find that Colli-Pee was associated with lower odds of positive HPV though the effect was insignificant but it was correlated with higher amounts of HPV (p=0.001). There was no significant association of the period of day with the odds of positive HPV and the amount of HPV detected. The random intercept was significant in both parts suggesting that a random intercept model is required. There was also a positive

cross part correlation (0.239) though it was not significant (p=0.761). This means there was no dependence between the frequency of HPV testing and the amount of HPV detected.

3.4 Model Comparison

The models were compared using the loglikelihood and AIC. It can be seen that the generalized gamma model has the lowest -2loglikelihood value (11428) and AIC (11466) suggesting that it somewhat has a comparative better fit to the data than the other models. Also the model assumes that the bivariate random effects are normally distributed with a constant variance. A quantile-quantile normal plot shows the random effects are jointly normally distributed as shown in figure 5 in appendix.

A generalized gamma model was refitted with only the significant covariates.We find that human DNA is associated with higher odds of positive HPV (p=0.0007). Colli-Pee was significantly better in predicting positive HPV than the standard cup $p < .0001$. The shape parameter is significant against all nested models. This means the generalized gamma distribution is significantly better compared to all nested models (gamma, lognormal and weibull which are special cases of generalized gamma model).

Variable	**Parameter**	**Estimate(S.E)**	**Pr**$>|t|$
Part I (binary outcome α)			
Intercept	α_0	-3.443(0.301)	<.0001
Type=('HR')	α_1	0.311(0.124)	0.017
conc(hDNA)	α_2	0.014(0.004)	0.0007
Part II (continuous outcome β)			
Intercept	β_0	13.997(0.901)	<.0001
Device=('Colli-Pee')	β_1	1.130(0.211)	<.0001
Heteroscedacity (δ)			
Intercept	δ_0	0.987(0.188)	<.0001
Variance (σ^2)			
variance a	σ_1^2	2.109(0.815)	0.015
variance b	σ_2^2	8.279(2.758)	0.005
Shape parameter	k	1.477(0.253)	<.0001

Table 6: Parameter estimates for generalized gamma model σ_1^2 and σ_2^2 are the variance of part I and part II random intercept.

3.5 Results for Human DNA

The distribution of human DNA is skewed positively. In this study, we use a one-part generalized linear model via generalized estimating equations. A one-part GLM fit via GEE actually

treats the observed human DNA as realization of a single process. The GLM makes use of a link function and therefore avoids the need to transform the data before modeling:

$$g(E(Y_{ij})) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

where Y_{ij} is the concentration of human DNA for individual i for j measurement,$g(.)$ is the link function relating the outcome to the linear predictor and x_{i1}=1 if period is morning and o otherwise,x_{i2}=1 if device is Colli-Pee and 0 otherwise. In this study a lognormal model with a log link was fitted to render the response normally distributed with a constant variance. Also a conventional alternative to the log-normal distribution i.e the gamma distribution with a log link was fitted. The QIC of both models were compared. Both models produced similar results

Variable	**log-normal (S.E) with logit link**	$P > \|t\|$	**Gamma GEE(S.E) with log link**	$P > \|t\|$
Intercept	2.158(0.210)	<.0001	2.133(0.225)	<.0001
Period=('E')	0.015(0.085)	0.856	0.039(0.073)	0.856
Device=('Colli-Pee')	0.577(0.145)	<.0001	0.582(0.155)	0.0002
QIC	4669.0		14266.5	

Table 7: GEE Parameter Estimates for the human DNA.

but the log-normal model had a better fit QIC=4669.0. Assuming that the concentration of human DNA follows a log-normal distribution, we find Colli-Pee significantly correlated with higher amount of human DNA with a mean estimate of exp(0.577)=1.78 and (p<.0001). There was no significant difference between morning and evening period in the detection of human DNA (p=0.856).

4 Discussion and Conclusion

In this study, several models were used to assess the efficiency of Colli-Pee against the standard cup in the detection of HPV virus. We started by considering two marginal models to account for the clustered semi-continuous data. Firstly a log-normal model was fitted and then a gamma model was considered to tackle the issue of non-constant variance. Both models were fitted via GEE. These marginal models offered us the possibility of performing marginal inference on the data. A drawback of this model is that it fits the models separately and there is no connection between both parts. Therefore a two-part random effects model was considered were the positive HPV responses were assumed to follow a log-normal distribution with constant variance and the random effects of both parts were assumed to be jointly normal and correlated. The advantage of this model is that it captures the 'cross part correlation' between both models which is very important in health studies such as this. The two-part random effects model was then extended to other distribution forms so as to account for the skewness and non-constant variance. The generalized gamma distribution and log-skew-normal distribution were considered. All three models yielded similar results but the generalized gamma model had an even better fit to the data as it had lower loglikelihood and AIC value.

Another area of interest is in relaxing the normality assumption of the random effects. (Liu and Yu,. 2008) proposed the use of Clayton copula to handle bivariate normal distribution. Also, a probability integral transformation method was proposed by (Nelson et al.,2006) for estimation in models with non-normal random effects.

Using the generalized gamma model, we estimated that Colli-Pee device was more efficient in HPV detection than the standard cup. Also, the period of the day was not significant i.e there was no significant difference between morning and evening urine sample in the detection of HPV. The concentration of human DNA was associated with higher odds of positive HPV but was not correlated with the amount of HPV. The study also confirms no dependency of the likelihood of HPV detection on the amount of HPV detected. A lognormal distribution and gamma distribution were assumed for the concentration of human DNA. The lognormal model had a lower QIC and therefore has a better fit compared to the gamma model. The Colli-Pee device was significantly associated with higher amount of human DNA compared to the standard cup. The period of the day was not associated with the concentration of human DNA.

The generalized gamma model is most appropriate for HPV response because it accounts for clustering, heteroscedacity which results from unstabilized variance after log transformation and also models cross-part correlation. The log-normal GLM via GEE was most appropriate for the skewed positive human DNA.

5 References

[1] Dunne EF, Unger ER, Sternberg M, McQuillan G, Swan DC, et al. (2007) Prevalence of HPV infection among females in the United States. JAMA 297: 813–819.

[2] Lacey CJ, Lowndes CM, Shah KV (2006) Chapter 4: Burden and management of non-cancerous HPV-related conditions: HPV-6/11 disease. Vaccine 24 Suppl 3S3/35–41.

[3] Manhart LE, Holmes KK, Koutsky LA, et al.(2006). Human papillomavirus infection among sexually activeyoung women in the United States: implications for developing a vaccination strategy. Sex Transm Dis. 33:502-508.

[4] Walboomers JM, Jacobs MV, Manos MM, et al.(1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide.J Pathol. 189:12-19

[5] Bosch FX, de Sanjose S.(2003). Chapter 1: human papillomavirus and cervical cancer-burden and assessment of causality. J Natl Cancer Inst Monogr. (31):3-13.

[6] Bosch FX, Tsu V, Vorsters A, Van Damme P, Kane MA.(2012). Reframing Cervical Cancer Prevention. Expanding the Field Towards Prevention of Human Papillomavirus Infections and Related Diseases. Vaccine. 30:F1-F11.

[7] Moscicki AB, Shiboski S, Broering J, et al.(1998). The natural history of human papillomavirus infection as measured by repeated DNA testing in adolescent and young women. J Pediatr. 132:277-284

[8] Franco EL, Villa LL, Sobrinho JP, et al.(1999). Epidemiology of acquisition and clearance of cervical human papillomavirus infection in women from a high-risk area for cervical cancer.J Infect Dis. 180:1415-1423. [9] 9.0 Liu l, Strwderman RL, Johnson BA, Q'Quigley JM.(2012). Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. Stat Methods Meds Res.

[10] Cragg JG (1971): Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica, 39:829-844.

[11] Olsen M and Schafer J.(2001). A two-part random effects model for semicontinuous longitudinal data. J Am Stat Assoc 96: 730–745.

[12] Molenberghs, G and Verbeke, G. (2005).*Models for discrete longitudinal data*. Springer Series in Statistics. Springer, NewYork.

[13] Liang, K.Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73:13–22.

[14] Liu L and Yu Z.(2008). A likelihood reformulation method in non-normal random effects models. Stat Med ; 27: 3105–3124.

[15] Nelson K, Lipsitz S, Fitzmaurice G, et al.(2006). Use of the probability integral transformation to fit nonlinear mixedeffects models with nonnormal random effects. J Comput Graph Stat ; 15: 39–57.

[16] Vorsters A., Micalessi I., Bilcke J., Ieven M., Bogers J., van Damme P.(2012). Detection of human papillomavirus DNA in urine: A review of the literature. Eur. J. Clin. Microbiol. 31:627–640. doi: 10.1007/s10096-011-1358-z.

[17] Heckman, J. (1979). Sample Selection Bias as a Specification Error. Econometrica, Vol. 47, pp. 153-161.

[18] Tobin J. (1958). Estimation of Relationships for Limited Dependent Variables. Econometrica, Vol. 26, pp. 24-36.

[19] Tooze J, Grunwald G and Jones R.(2002). Analysis of repeated measures data with clumping at zero. Stat Meth Med Res 11: 341–355.

[20] Liu L, Strawderman R, Cowen M, et al.(2010). A flexible two-part random effects model for correlated medical costs. J Health Econom 2010 29: 110–123.

6 Appendix

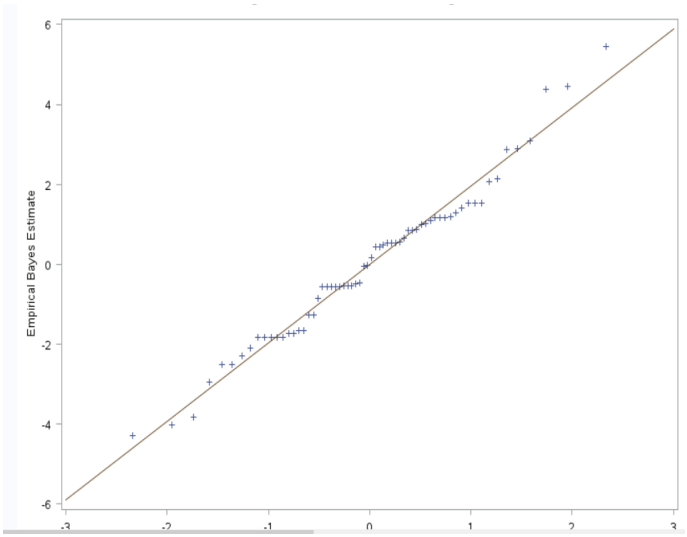


Figure 5: The qqplot of random effects suggest that the random effects are normally distributed. The kolmorogorov-Smirnov confirms the distribution is normally with a ($p > 0.15$)

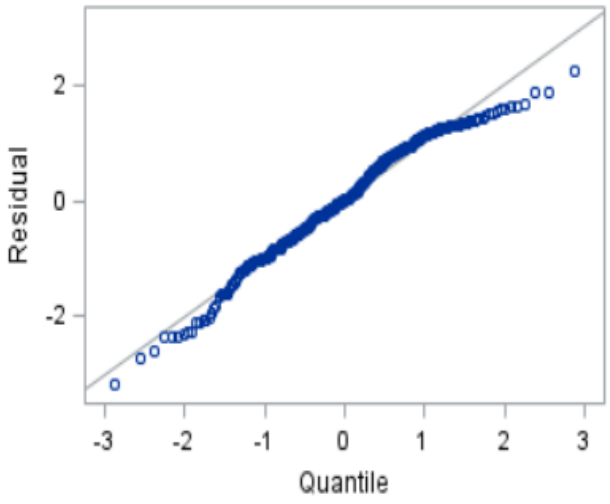
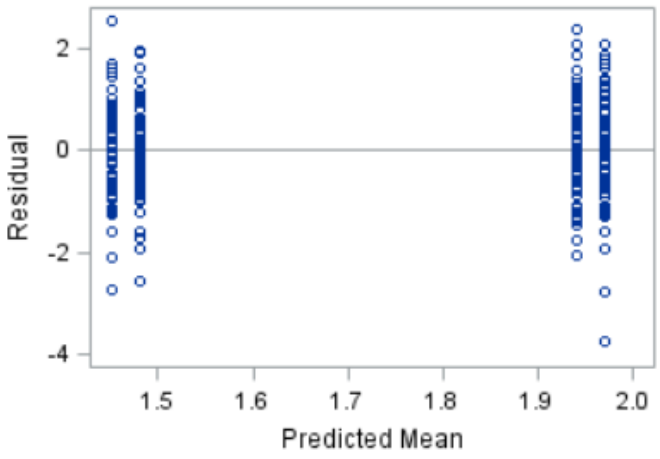


Figure 6: The normal quantile-quantile plot of the log-transformed positve values shows that the transformed values are not normally distributed but rather left skewed.The kolmorogorov-Smirnov test ($p < 0.001$) confirms the transformed values are not normally distributed



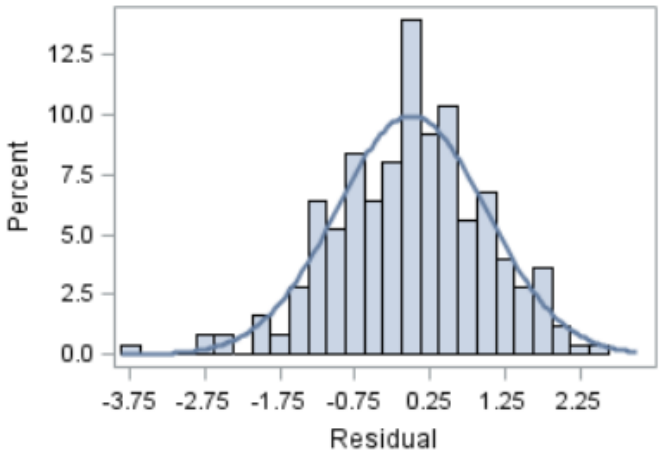


Figure 7: The plots the log transformed human DNA concentrations looks approximately normal with a heavy left tail.There is no pattern in the residual plot which shows a good fit

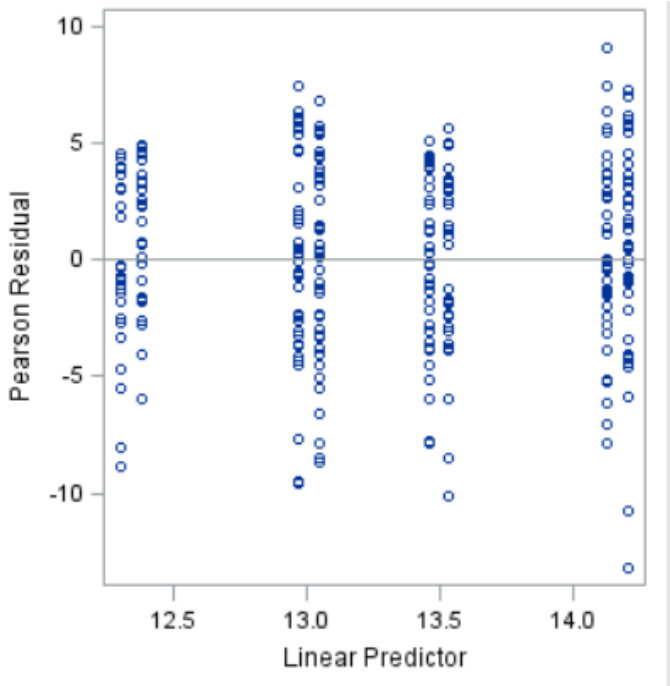


Figure 8: Pearson reesiduals of log-normal model

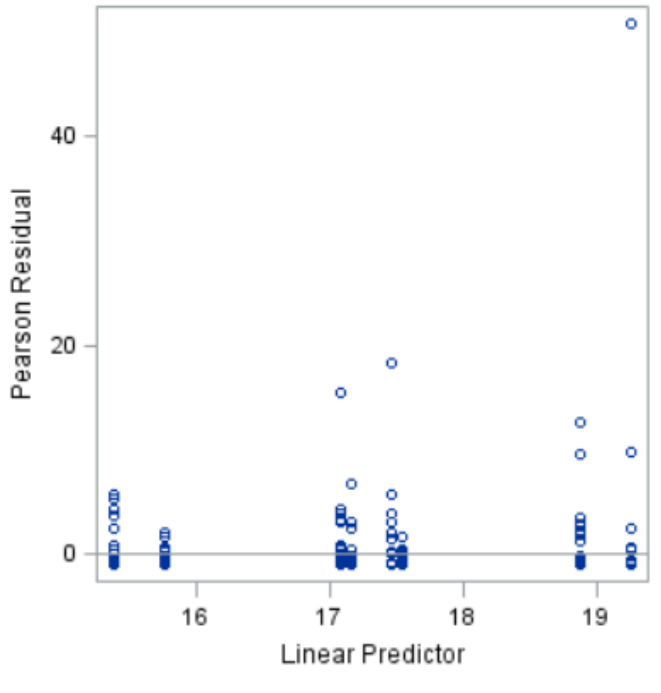


Figure 9: Pearson residuals of Gamma model

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
HPV DNA detection in urine: effect of a first-void urine collection device and time of collection

Richting: **Master of Statistics-Epidemiology & Public Health Methodology**
Jaar: **2018**

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,

Ejedepang, Dunson Bwese Koge

Datum: **22/01/2018**