Study of the efficacy of Berenil applied to trypanosomosis's infected cattle using a binary outcome

Gerard Castellà, Mathieu Marauri and Lluis Ramon

Contents

1 Introduction **2**

2 Objective **2**

3 Dataset **2**

4 Outcome categorization **3**

5 Statistical methods **3**

5.1 GEE 3

5.2 GLMM 3

5.3 Transition Models 4

5.4 Missing data 4

6 Results **5**

6.1 Descriptive analysis 5

6.2 GEE 5

6.3 Random effects model 7

7 Missing Data Analysis **8**

8 Conclusion **12**

9 Limitations and Further research **12**

10 Bibliography **12**

1 Introduction

Cattle are an important economic resource and also a major health factor in many countries. Therefore preventing diseases in a herd is vital.

In this study we focus our interest on the parasite of trypanosomosis which can lead to the death of a cow. This disease, transmitted by the tsetse flies, causes an infection characterized by fever, loss of appetite and anemia, which can lead to death depending on different factors.

One medicine, the Berenil, is used to cure the infected cattle. The aim of this research is to determine the efficiency of different doses of Berenil. Finding the most efficient dose, if such a dose exists, is critical when it comes to save both cattle and money. Here lies all the interest of our study.

To determine if a cow is ill or not a binary variable will be studied. This indicator can be 1 for a healthy cow or 0 for a sick one.

2 Objective

The aim of this study is to assess the efficacy of different doses of Berenil in cattle infected with the trypanosomosis parasite.

3 Dataset

A cohort of 10 different cows infected by trypanosomosis parasite was selected for the study. Each Berenil dose (low, medium and high) was administrated three times (time 1, 2 or 3) for each animal. PCV was reported each time as well as the number of calves it had before being infected.

The variables reported for this study are presented below.

- Id: Each cow has its own id. From 1 to 10. *(id)*
- PCV: Binary variable. *(pcv.b)*
- Dose: H High, M Medium L low. *(dose)*
- Time: From 1 to 3. *(time)*
- Number of birth: From 2 to 8. *(nbirth)*

Since the gathering process of the data was unknown, several assumptions were needed.

- Each time PCV is obtained before the treatment. Therefore the effect of the third dose could not be evaluated.
- Dose is assigned randomly in time to the cow.
- For a given cow, the previous treatments (high, medium or low) do not affect the following ones.
- Time intervals are the same and fixed.

4 Outcome categorization

The binary response was related with a cow being healthy or unhealthy. Taking this into consideration, the outcome was categorized using the following criteria:

- Literature review*: A healthy cow is estimated to have a PCV value ranging from 24 to 46.
- Practical Modeling: As the binary response should be modeled in following sections, a suitable one was searched. To this end, a trial and error with a cutoff ranging from 20 to 24 was explored.

The threshold between healthy and unhealthy cow was set at a PCV value of 20. If the PCV value was bigger it was categorized as healthy, if it was lower or equal it was categorized as unhealthy.

Table 1 shows a contingency table for the dichotomised response variable. The number of cows in the healthy category increases in time while the number of unhealthy ones decreases. A Missing value category is also included. One can notice that the number of missing values increases with time. This topic will be further studied in section 7 Missing Data Analysis.

	Time 1	Time 2	Time 3
Unhealthy	28	15	2
Healthy	1	12	18
Missing value	1	3	10

Table 1: Contingence table for the dichotomised response variable (rows) and times.

5 Statistical methods

The methods used in the statistical analyses are detailed in this section. First a Generalized Estimating Equations (GEE) and latter a Generalized Linear Mixed Model. An exploratory data analysis was performed before those regression methods.

5.1 GEE

The response variable was the binary PCV to detect healthy and unhealthy cows. As we wanted to estimate the effect of dose in healthy/unhealthy cows, these two variables were included in the model.

From this initial model, a forward step-wise method was carried for the model selection, including additional covariates or interactions between them. The models were compared QuasiLikelihood ratio test when nested and with QuasiLikelihood Information Criteria (Pan 2001) when non-nested.

A classification table with predicted values and original data was created to asses how the model performed. Sensitivity and specificity are calculated for the final model.

5.2 GLMM

In order to take into account the specific effect due to the cow itself we had to introduce random effects in our model. To do so we performed a Generalized Linear Mixed Model.

We start by having fixed effects on the covariates selected in the model obtained with the GEE model. A grouping factor is needed to add random effects. The dataset gives us two different grouping factor: the *idDose* and the *dose* by *id*. The choice will be made in the section 6.3.1 Selection of the model.

A similar process that the one used in the GEE to select the covariates will be performed to select the random effects. The models will be compared using a QuasiLikelihood ratio test if they are nested or the Akaike Information Criterion if they are not.

5.3 Transition Models

http://faculty.washington.edu/yanez/b540/lectures/lectureWk082010-2x2.pdf

For the transition model, one models the conditional distribution of the responses, Y_{ij}, on covariates, x_{ij}, and past responses, $Y_{i1}, Y_{i2}, \ldots, Y_{i,j-1}$

- Let $H_{ij} = Y_{i1}, Y_{i2}, \ldots, Y_{i,j-1}$ = history of past responses.
- The conditional mean of the transitional model is $\mu_{ij}^c = E[Y_{ij}|H_{ij}, x_{ij}]$

We consider transition models where the conditional mean satisfies the equation

$$g(\mu_{ij}^c) = {x_{ij}}'\beta + \sum_{r=1}^{s} f_r(H_{ij}, \alpha)$$

for suitable functions f_r and parameters α.

Past responses (or functions thereof) are treated as additional explanatory variables.

- The present is affected by the past through the sum of s terms.
- These models nicely characterize change for longitudinal categorical data.
- The likelihood factorization means that standard GLM software can be used to fit the models.
- The coefficients for the previous responses summarize the strength of the longitudinal dependence (compared to our other models)

In our case, a possible transition model would be,

$$logit[P(PCV_{ij}|x_{ij}, PCV_{i,j-1}] = \alpha_0 + \alpha_1 doseMedium + \alpha_1 doseHigh + \alpha_3 PCV_{i,j-1}$$

We consider that this model could be good and worth a try. Eventhough the serie is quite short with just 3 elements.

5.4 Missing data

Also, a missing data analysis was performed to give consistence to the previous results. Fisher test were carried out to assess the association between the fact that an observation is missing and the time or the dose. Also a sensitivity analysis was done in order to better assess the dependence between the missings and the covariates as well as the PCV. Some assumptions had to be made to give some sense and better understand the data we were given.

6 Results

6.1 Descriptive analysis

Knowing the general behavior of the data is a starting point to every other analysis. Figure 1 shows the value of the binary outcome for all the cows in the dataset.

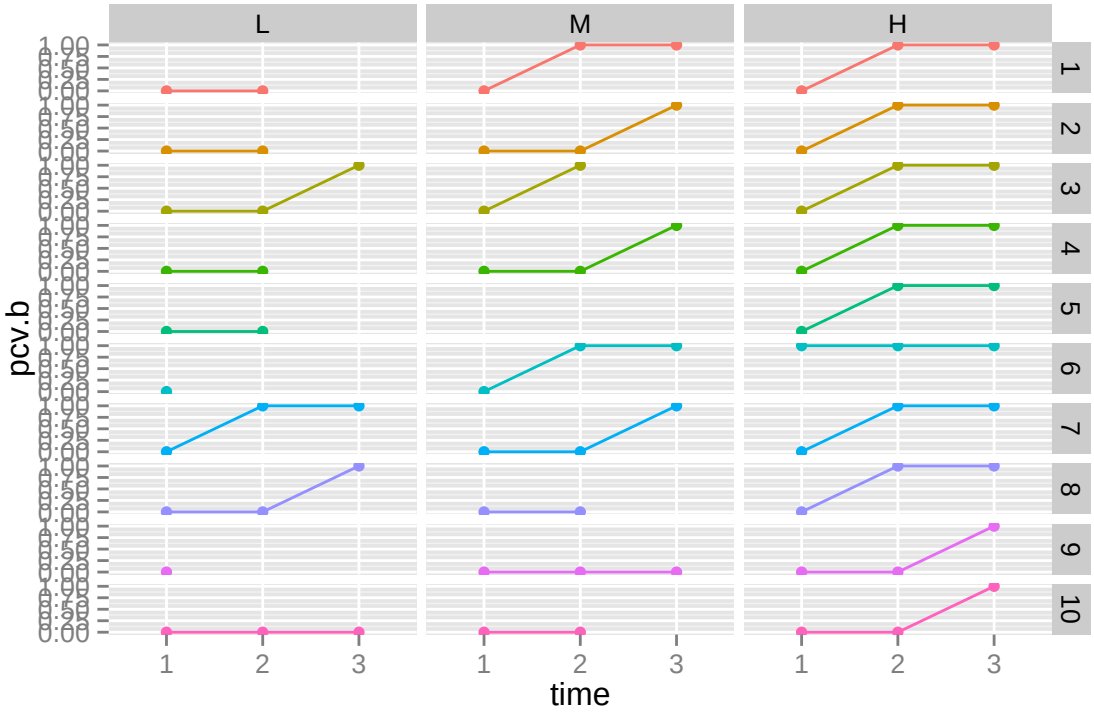

Figure 1: Representation of the data.

It clearly appears that cows which are treated with the low dose are unhealthy at time 1 and 2 (except for cow 7). The second major observation is that at time 3 cows that are treated with the high dose are healthy. This is still true for almost every cows at time 2.

The effect of the treatment can be seen when a given cow goes from unhealthy to healthy. Such an evolution appears 18 times. That is to say that among the 30 "different" cows (identified by the variable *idDose*) 60% have a positive reaction to the treatment. The healing of the cows is done after the first injection of Berenil in 47% of the cases.

6.2 GEE

As explained in the section 5.1, a starting model for PCV was fitted with dose. Following a step-wise approach, time was included in the model. The number of births and its interaction with time or dose did not improved the model. Adding the interaction between *time* and *dose* did not improve the model either.

The final model was:

$$logit(p) = \beta_0 + \beta_1 time + \beta_2 doseMedium + \beta_3 doseHigh$$

Several working correlation matrices were used and an AR1($\alpha = 0.171$) was chosen because of its smaller QIC.

Table 2 shows the coefficients estimate for the final model. The odds ratio of having a high dosage is $exp(3.382)$ times of having low dosage adjusted for time covariate. An increase of one unit in time increases the odds ratio by $exp(3.205)$ times adjusted for the *dose* covariate. Even if medium dose was not significant in the model, we kept it in the model because High dose was significant. The merging of the medium dose and the low dose into one covariate to be compared to the high dose is something that could be done in the section 9 Further research.

	Estimate	Std.err	p-value
(Intercept)	-8.52	1.80	0.00
doseM	1.20	1.22	0.33
doseH	3.38	1.19	0.00
time	3.20	0.63	0.00

Table 2: Coefficients, SD and p-values from the GEE model. Alpha estimation (SD) = 0.171 (0.0413).

Table 3 shows the classification table for the chosen model. The selected model has a sensitivity of 0.911 and a specificity of 0.839.

	Original	data
Predicted	41	5
	4	26

Table 3: Crosstable with the predicted data with the model and the original data.

6.3 Random effects model

6.3.1 Selection of the model

The GEE model does not take into account the particularity of each cow. In order to do so we performed a GLMM model. As explained in the section 5.2 the first thing to know is what will be the exact grouping factor. Our dataset provides us with two different ways of grouping our data. We have the idDose variable and the id then the dose which create a hierarchical model with two levels of random effects. In the last case we have a random effect for the cow among all the cows and a random effect for the dose among a given cow. We used the second case because the first one erases the subject specific effect and does as if we have 29 cows instead of only 10.

First we tried to have random intercept on both the intercept and the time but the model did not converge so we decided to take out the random effect on the intercept. This decision was also motivated by the fact that at the beginning all the cows are sick and we are not interested in the differences between them before the treatment but in the differences in their response to the treatment. The aim of this study is to assess the efficiency of the treatment so it is of a crucial importance to add random effects for the intercept. The random effects are only used for the time covariate. It is impossible to have random effects on the dose since we would have too many parameters to estimate regarding the number of observations.

Now that we know where to put random effects we can start building the model with the fixed effects. We start with the covariates used in the GEE model that is to say with fixed effects on the dose and on the time. We have the following model:

$$PCV.b_i = \beta_0 + (\beta_1 + b_{1i})time + \beta_2 doseMedium + \beta_3 doseHigh + \epsilon_i$$

Then we tried to add a fixed effect on *nbirth* and on the interaction between *time* and *dose*. Table 4 gives the results of the tests we performed to know which model was the best.

	df	AIC	p-value
$\beta_0 + +(\beta_1 + b_{1i})time + \beta_2 doseMedium + \beta_3 doseHigh$	6	36.592	
$\beta_0 + +(\beta_1 + b_{1i})time + \beta_2 doseMedium + \beta_3 doseHigh + \beta_4 nbirth$	7	38.335	0.6123
$\beta_0 + (\beta_1 + b_{1i})time + \beta_2 doseMedium + \beta_3 doseHigh + \beta_4 dose * time$	8	38.998	0.4506

Table 4: Anova table comparing the different models with the one with fixed effects for time and dose.

As a result we kept the initial model. Adding random effects on *nbirth* does not make sense since we do not have a fixed effect on *nbirth*. Besides even when trying to add a fixed effect and a random effect on it, the model has a hard time estimating the parameters. Therefore our final model is:

$$PCV.b_i = \beta_0 + (\beta_1 + b_{1i})time + \beta_2 doseMedium + \beta_3 doseHigh + \epsilon_i$$

6.3.2 Interpretation of the selected model.

In table 5 are presented the values of the odds ratio for the fixed effects of the model.

	inf	est	sup
(Intercept)	0.00	0.00	0.00
doseM	0.15	6982.82	323921533.94
doseH	411020.60	25495859855125468.00	1581523827264867511306062042.00
time	199935.59	55914635647086.74	15637268602686084743488.00

Table 5: Odds ratio with their confidence interval.

It appears that an increase of one unit in time increases the odds ratio by 5.59e+13 which means that it is 5.59e+13 times more likely to be healthy than unhealthy when the time increases. The random effects on the time add variability to the effect of time depending on the id and on the dose. it means that cows have a different response to the treatment as regard to the time. The use of the high dose has an equivalent high impact on the probability to be healthy.

The more the cow receives the treatment (e.g. as time increases) the more likely it is that it will be healthy. The same relation goes for the choice of the high dosage as compared to the low dosage. Cows react in different ways to the treatment as shown by the random effects on time but the overall is still positive.

6.3.3 Perfect separation

Perfect separation occurs in logistic regression when a covariate or a combination of them almost completely separates the value of the outcome variable. In this case the likelihood does not have a maximum for the corresponding variable and so the estimates became very large. It frequently occurs with small sample. A way to detect it is when the model has very big coefficients and standard errors. It is fully described in (Heinze 2002) 10.

Our model has the previous conditions, so we consider that this model it is not appropriate and it is suggested to switch for a model that takes into account this issue. Several alternative models are described in section Further research 9.

7 Missing Data Analysis

In this section, the patterns of missing data will be discussed and analyzed, in order to assess its influence in the analyses performed in previous sections.

Figure 2 and Table 6 show the missing data distribution over time and for each dose. Clearly, the fact that an observation is missing is associated with the dose and the time (fisher tests p-values= 0.005 and 0.007, respectively). Actually, there are more missing data in low and medium dose than in high dose, and also as time increases. Missing values in the PCV are not produced completely at random.

Dose	Time	Missing	Available
L	1	0	10
L	2	2	8
L	3	6	4
M	1	1	9
M	2	1	9
M	3	4	6
H	1	0	10
H	2	0	10
H	3	0	10

Table 6: Number of missings and available data in the outcome.

For all this, it is pretty clear that there is some systematic mechanism that produces the missing data. Probably, given that 8 of the 9 cows that have some missing data have the first missing after being unhealthy, it would be logic to understand that the missings are produced because of the death or that the cow is dropped out from the study due to having too many health issues. The fact that assuming a sequence of

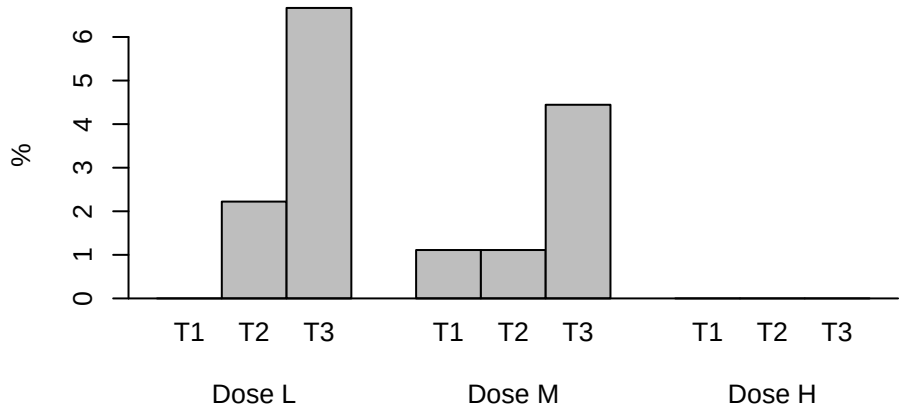

Figure 2: Percentage of missings for each time and dose with respect to the whole sample.

doses for every cow, once there's a missing, no more values are reported anymore. For instance, looking at Figure 3, the cow with ID 5, would have died (or left the study) before taking the time 3 low dose measure, and if one assumes that its sequence of doses was H-L-M (the sequence is not provided), this explains the missings for the medium dose. So under these assumptions, the missing data are monotone pattern of missing data. In this sense, our sample could be easily biased, because the missings are not equally distributed among the data. Therefore, the pattern for missings in this data is MAR or MNAR, depending on the mechanism that is producing the missings.

Since all the analyses performed previously were done with the complete cases (missing PCV's were omitted), the results obtained could be far from the reality (or not). The complete cases analysis can only deal with MCAR type of missing data. Therefore, several approaches will be used to analyse the data taking this into account.

The following table shows the coefficients of the logistic regression performed with the covariates time, dose and PCV (as binary factor) to predict the missings. The model was:

$$logit(p) = \beta_0 + \beta_1 PCV_{healthy} + \beta_2 doseMedium + \beta_3 doseHigh$$

All the coefficients have a lot of variation depending on the values of the coefficient of the PCV (Table 7). The missings then, depend a lot on the covariates and also on the values of PCV.

Also a sensivity analysis was carried out using GEE, pointing to the same direction. The results can be seen in Table 8

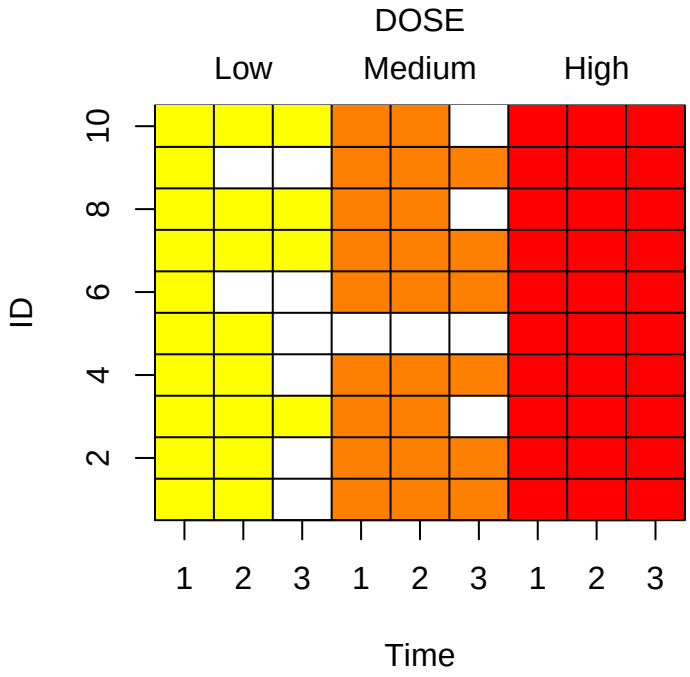

Figure 3: Missings distribution with reordered doses for each cow (columns). Missing data is plotted in white, low, medium and high doses are plotted in yellow, orange and red, respectively.

Beta1	doseM	doseH
-6	1.731 (0.55-5.45)	7.723 (2.575-23.167)
-5	1.58 (0.608-4.109)	5.493 (2.199-13.721)
-4	1.442 (0.671-3.097)	3.907 (1.879-8.127)
-3	1.316 (0.742-2.335)	2.779 (1.605-4.813)
-2	1.201 (0.819-1.76)	1.977 (1.371-2.851)
-1	1.096 (0.905-1.327)	1.406 (1.171-1.688)
0	0.936 (0.792-1.105)	0.766 (0.648-0.905)
1	0.913 (0.754-1.105)	0.711 (0.592-0.854)
2	0.833 (0.568-1.22)	0.506 (0.351-0.73)
3	0.76 (0.428-1.348)	0.36 (0.208-0.623)
4	0.694 (0.323-1.49)	0.256 (0.123-0.532)
5	0.633 (0.243-1.646)	0.182 (0.073-0.455)
6	0.578 (0.183-1.818)	0.129 (0.043-0.388)

Table 7: Odds ratio and 95 percent confidence interval, given the coefficient for the PCV.

Beta1	doseH	time
-2	0.159 (0.476)	0.686 (0.463)
-1.9	0.153 (0.476)	0.652 (0.463)
-1.8	0.146 (0.476)	0.618 (0.463)
-1.7	0.139 (0.476)	0.583 (0.463)
-1.6	0.133 (0.476)	0.549 (0.463)
-1.5	0.125 (0.476)	0.514 (0.463)
-1.4	0.118 (0.476)	0.48 (0.463)
-1.3	0.11 (0.476)	0.445 (0.463)
-1.2	0.103 (0.476)	0.411 (0.463)
-1.1	0.095 (0.476)	0.376 (0.463)
-1	0.087 (0.476)	0.342 (0.463)
-0.9	0.079 (0.476)	0.308 (0.463)
-0.8	0.07 (0.476)	0.273 (0.463)
-0.7	0.062 (0.476)	0.239 (0.463)
-0.6	0.053 (0.476)	0.205 (0.463)
-0.5	0.045 (0.476)	0.171 (0.463)
-0.4	0.036 (0.476)	0.136 (0.463)
-0.3	0.027 (0.476)	0.102 (0.463)
-0.2	0.018 (0.476)	0.068 (0.463)
0.1	0.009 (0.476)	0.034 (0.463)
0	0 (0.476)	0 (0.463)
0.1	-0.009 (0.476)	-0.034 (0.463)
0.2	-0.018 (0.476)	-0.068 (0.463)
0.3	-0.028 (0.476)	-0.102 (0.463)
0.4	-0.037 (0.476)	-0.136 (0.463)
0.5	-0.047 (0.476)	-0.171 (0.463)
0.6	-0.057 (0.476)	-0.205 (0.463)
0.7	-0.066 (0.476)	-0.239 (0.463)
0.8	-0.076 (0.476)	-0.274 (0.463)
0.9	-0.086 (0.476)	-0.308 (0.463)
1	-0.096 (0.476)	-0.343 (0.463)
1.1	-0.107 (0.476)	-0.378 (0.463)
1.2	-0.117 (0.476)	-0.413 (0.463)
1.3	-0.128 (0.476)	-0.448 (0.463)
1.4	-0.139 (0.476)	-0.483 (0.463)
1.5	-0.15 (0.476)	-0.518 (0.463)
1.6	-0.161 (0.476)	-0.554 (0.463)
1.7	-0.172 (0.476)	-0.59 (0.463)
1.8	-0.184 (0.476)	-0.626 (0.463)
1.9	-0.195 (0.476)	-0.662 (0.463)
2	-0.207 (0.476)	-0.698 (0.463)

Table 8: Regression coefficients and SD given the coefficient for the PCV.

8 Conclusion

The main objective of this study was to assess the efficiency of the different kinds of doses used to cure cows infected with the trypanosomosis parasite. As expected, time and dose have a positive effect on the cow's health. However the effect of the dose varies with the kind of doses used. No differences were found when comparing medium dose with low dose, using different models. However, high dose was found to be significantly better than the low dose.

For the missing analysis part, assuming a particular sequence of doses for the cows (were not given), the missings have a monotonic pattern: once there is the first missing, no more values are reported until the end of the study. A possible explanation for this, could be that they die during the treatment due to the disease or some other cause. The fact that all almost all the missings happened after reporting a low PCV (unhealthy) also supports this idea, even though the causes for this may be others, too. Also the missings were found to be associated with the time and the dose, so the missing data are not completely at random for sure. The sensitivity analyses also suggest this idea. If the cows truly are dropouts due to deaths, the missings will clearly be not at random (since the lower the PCV easier to be missing or death).

9 Limitations and Further research

The relative low number of observations in the dataset makes difficult to generalize the results of the study. A lot of assumptions were needed. Besides issues arose when we analyzed the data. In order to verify whether or not those assumptions can be validated we would have needed more information on the dataset.

In order to further analyse the data some improvements or ideas are possible. Some of them are listed below.

In GEE modeling for small sample, Jacknife estimators are better than sandwich estimator as explained in (Højsgaard, 2006).

As Medium dose seems to be not significant for our models, a joint Low and Medium dose factor can be created and compared against High dose.

In this study we dealt at several occasions with the Perfect Separation Problem. It appeared when a 22 PCV value threshold was used for healthy or unhealthy cow. The GEE models with time and dose interaction with this model resulted to be perfectly separated. Then we changed to 20 to avoid this problem. Unfortunately it latter appeared with the GLMM modeling. Several options such as Bayesian Modeling with a flat prior seems to be an option to try in the future. It is implemented at (Kosmidis, 2013) or at (Hadfield, 2010).

Transition models are explained in section 5.3 and a better exploration could be interesting.

10 Bibliography

TODO(Gerard): Normalize citations

- Pan 2001, Biometrics, Akaike's Information Criterion in Generalized Estimating Equations.
- Højsgaard, S., Halekoh, U. & Yan J. (2006) The R Package geepack for Generalized Estimating Equations Journal of Statistical Software, 15, 2, pp1–11

- Hardin & Hilbe 2002, Generalized Estimating Equations
- Kamil Barton (2015). MuMIn: Multi-Model Inference. R package version 1.13.4. http://CRAN.R-project.org/package=MuMIn
- A solution to the problem of separation in logistic regression. http://www.ncbi.nlm.nih.gov/pubmed/12210625
- complete separation/ http://www.ats.ucla.edu/stat/mult_pkg/faq/general/complete_separation_logit_models.htm
- Ioannis Kosmidis (2013). brglm: Bias reduction in binomial-response Generalized Linear Models. http://www.ucl.ac.uk/~ucakiko/software.html
- Jarrod D Hadfield (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software, 33(2), 1-22. URL http://www.jstatsoft.org/v33/i02/.