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

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 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

TODO(Lluis): Add model information. TODO(All): Include pros/cons with our data. Decide if our data could be modeled using this kinds of models.

6 Results

6.1 GEE

TODO(Mathieu): Revise text.

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. Neither it did to include the interaction between time and dose.

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 X shows the coefficient 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 dose covariate. Even medium dose was not significant in the model, we kept it in the model because High dose was significant. As part of a further research 8 is let the models with Low and Medium dose categories merged as a single category.

TODO(Gerard): Create table with coefficients and alpha

```
##
## Call:
## geeglm(formula = pcv.b ~ dose + time, family = binomial, data = cows.com,
##     id = idDose, corstr = "ar1", scale.fix = TRUE)
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  -8.5234  1.7975 22.484 2.12e-06 ***
## doseM         1.1988  1.2221  0.962   0.3266
## doseH         3.3817  1.1873  8.112   0.0044 **
## time          3.2046  0.6289 25.966 3.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale is fixed.
##
## Correlation: Structure = ar1  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.1707 0.04126
## Number of clusters:   29   Maximum cluster size: 3
```

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

TODO(Gerard): Create table with classification

```
##
##      0  1
##   0 41  5
##   1  4 26
```

6.2 Random effects model

6.2.1 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. The first thing to know is what e 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 2 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 2: 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:

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

6.2.2 Interpretation of the selected model.

In table 3 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	1563726860268608474348 8.00

Table 3: 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.

Extremely high (or low) values were obtained because there are only twice as much observations than parameters to estimates. Besides in our dataset one can notice that the zeros are mainly at time 1 and for the low dosage and the one at time 3 for the high dosage.

To conclude one could say that 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.

7 Missing Data Analysis

TODO(Gerard): Missing exploration, in process

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 1 and Table 4 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 values in low and medium dose than in high dose, and also as time increases.

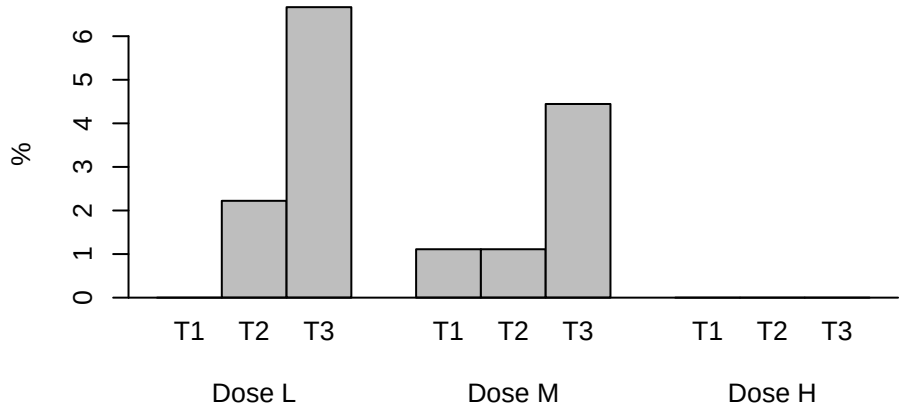

Figure 1: Percentage of missing values for each time and dose with respect to the whole sample.

For all this, one could think that missing data are produced because of the death or drop out of the cow for any reason. This makes sense also when looking at Figure ????: the cow with ID 5, for example, 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, this explains the missing values for the medium dose. Following this rationale, cows number 1, 2 and 4 would have died at the time 3 of dose L (their sequence would end with L, since there are no more missing data); cows number 3, 8 and 10 before time 3 of the medium dose (ending with M dose) and cows with ID 6 and 9 before time 2 of the L dose.

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 4: Number of missing data and available data in the outcome.

The point in all this is that, if we think that the doses have different effects in the cows, then, the sequence of doses will affect the number of missing values for a cow, introducing a huge source of bias in the data. 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 missing values are produced because of the death or that the cow is dropped out from the study due to having too many health issues. With this assumption, the cows with worse prognostic would not have been analysed in the previous sections. Also, the missing data could be understood like if the cow was not healthy, making the

Therefore, the pattern for missing values in this data (maybe more than one) is, probably, missing not at random (MNAR).

8 Limitations and Further research

- Jacknife estimators for small sample better than sandwich estimator

  Jackknife variance estimators are preferable to the sandwich estimator in case of a small number of clusters. (Højsgaard, Halekoh, Yan 2006)

- Compare a new factor L and M dose against H dose
- Perfect Separation Problem with gee

  It apeared when using a 22 PCV value threshold for healthy or unhealthy cow.

- Compare GEE model with GLMM. Coefficients and SE.

9 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