# Sensitivity/Specificity Analyses - Canine Leishmaniosis

*Marie Ozanne*

*January 30, 2019*

## Exploratory Analyses

```
##           DPP
## PCR        Negative Positive
##   Negative      743       16
##   Positive        1       11
```

## Models

Angela's paper (Toepp et al., 2019, https://doi.org/10.1371/journal.pntd.0007058) uses logistic regression, with age, sex, and variables that have to do with diagnostic tests as explanatory variables. They are something like this:

**Model A 1:** $logit(\pi_k) = \beta_0 + \beta_1 Age_k + \beta_2 Sex_k + \beta_3 Y_k$, where $Y_k$ is diagnostically positive (as defined in Model 1 below), but for the mom and $\pi_k$ is the probability of disease for individual $k$

**Model A 2:** $logit(\pi_k) = \beta_0 + \beta_1 Age_k + \beta_2 Sex_k + \beta_3 T_{1k} + \beta_4 T_{2k}$, where $T_{jk}$ is the result for Test $j$ (as defined in Model 1 below), but for the mom and $\pi_k$ is the probability of disease for individual $k$

Note, these models were fit with a log link function, presumably so that relative risks could be recovered?

We plan to evaluate similar models for our data and to then incorporate sensitivity and specificity of the tests into these models. Then we will compare the model performance to that of other methods. Hopefully we will see an improvement/some details that we miss when we do not include the sensitivity and specificity for the tests.

In all these models, we will assume that the observations are independent.

### Model 1:

The data outcome we are using is "diagnostically positive", meaning that an individual tests positive on at least one diagnostic test. This is what we have used in our other papers and seems to be popular in the literature (add some references to this). In this model, we assume that the two diagnostic tests are independent, and that there is some imprecision in the test results, so we include sensitivity and specificity for each test in the model.

#### Data Model

$$Y_k | T_{1k}, T_{2k} \sim Bernoulli\left(P(T_{1k} = 1) \cup P(T_{2k} = 1)\right)$$

where $P(T_{1k} = 1) \cup P(T_{2k} = 1) = P(T_{1k} = 1) + P(T_{2k} = 1) - P(T_{1k} = 1) \times P(T_{2k} = 1)$ since we are assuming that the test outcomes are independent.

For the probability of a positive test result for individual $k$ on test $j$,

$$\begin{aligned} P(T_{jk}=1) &= P(T_{jk}=1 \cap D_k=1) + P(T_{jk}=1 \cap D_k=0)\\ &= P(T_{jk}=1|D_k=1)P(D_k=1) + P(T_{jk}=1|D_k=0)P(D_k=0)\\ &= \underbrace{P(T_{jk}=1|D_k=1)}_{\text{Sensitivity}} P(D_k=1) + \underbrace{[1-P(T_{jk}=0|D_k=0)]}_{1-\text{Specificity}} P(D_k=0) \end{aligned}$$

### Process Model

Now we need a model for the probability of disease for individual $k$ that depends on disease prevalence, and some individual level factors.

$$\text{logit}(P(D_k)) = \text{logit}(\pi) + \mathbf{x}_k^T \boldsymbol{\beta}$$

where $\pi$ is the population prevalence of disease; $\mathbf{x}_k^T = (Age_k, Sex_k)$.

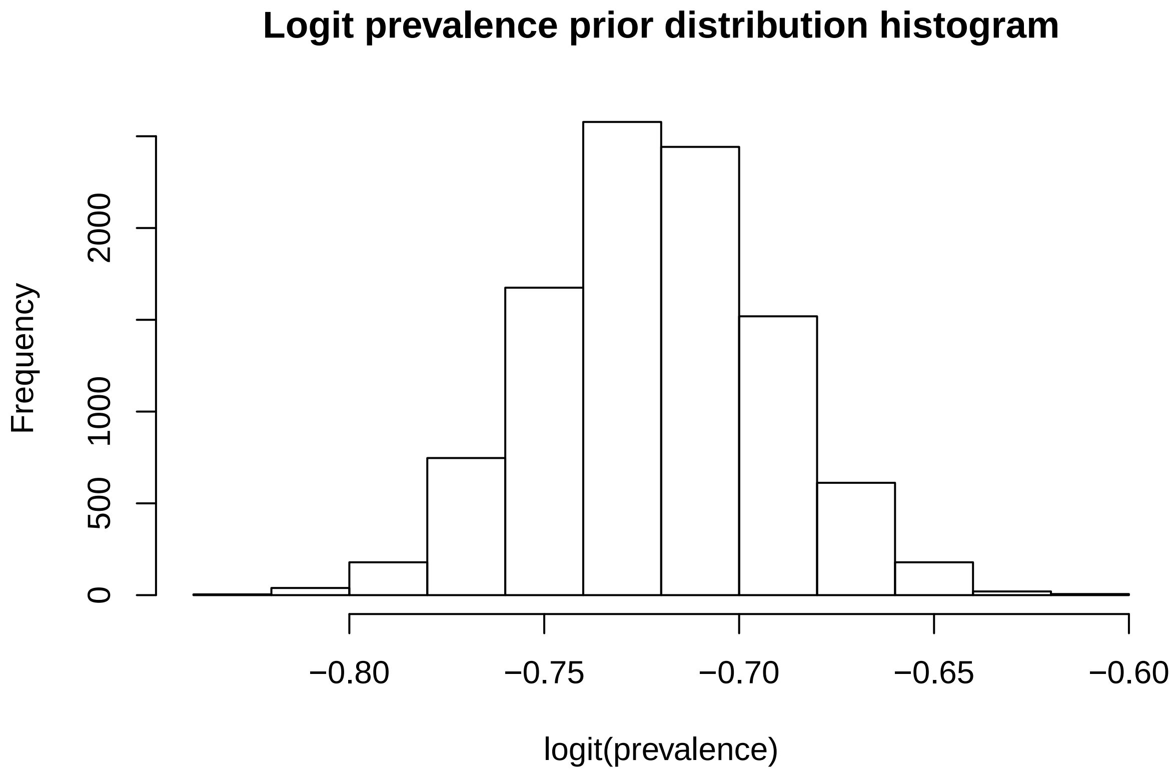
### Prior Model

**Prevalence:**

$$logit(\pi) \sim Normal(\mu_\pi, \sigma_\pi^2)$$

We have a range for the prevalence of (0.05, 0.10). This corresponds to a range of (-0.7497, -0.6972) on the logit scale.

```r
hist(rnorm(10000, mean = log(0.075)/(1-log(0.075)), sd = 0.03),
     main="Logit prevalence prior distribution histogram",
     xlab="logit(prevalence)")
```



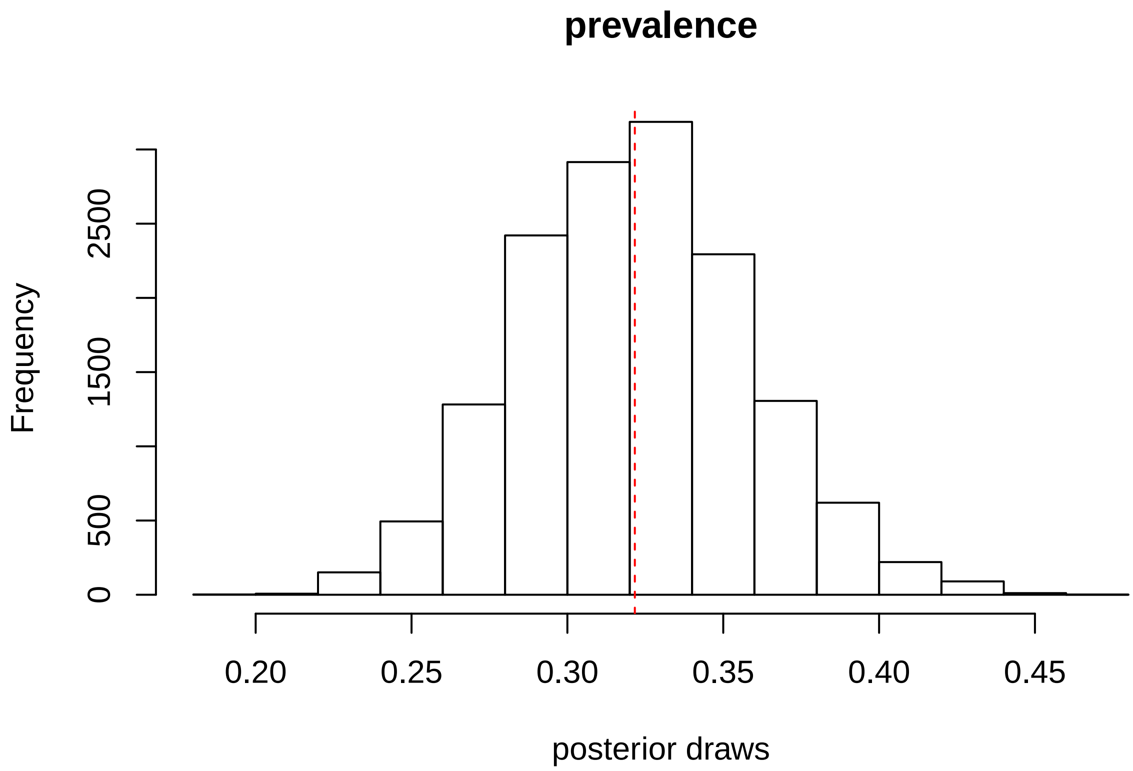
**Other parameters:**

$$\boldsymbol{\beta} \sim Normal(\boldsymbol{\mu}_\beta, \Sigma_\beta)$$

We will assume that the regression coefficients are independent, so $\Sigma_\beta$ is a diagonal matrix.

### OpenBUGS Model 1 Implementation

```r
## Graphial summaries of posterior distribtuions
hist(exp(model2_df$lpi)/(1+exp(model2_df$lpi)), main="prevalence", xlab="posterior draws")
abline(v=mean(exp(model2_df$lpi)/(1+exp(model2_df$lpi))), lty="dashed", col="red")
```



```r
## Numeric summaries of posterior distributions
#boxplot(model2_df[,!(names(model2_df) %in% c("deviance"))])
```

Removing the age and sex parameters made the estimate for prevalence make a lot more sense. Is there are good reason for this? Now the mean prevalence is 0.3216 and the 95% credible interval is: ( 0.2499, 0.3974 ).

### OpenBUGS Model 1 Disease State Prediction

```r
## Set up storage for model results
pred_df_m2 <- data.frame(obs=1:nind,
                         pi.D=rep(NA,nind), ## average estimate
                         SD=rep(NA,nind),
                         LB=rep(NA,nind), ## 2.5th percentile
                         UB=rep(NA,nind), ## 97.5th percentile
                         model_assignment=rep(NA,nind),
                         Clinical_status=ss_data2$ClinicalStatus,
                         Diagnostic_status=ss_data2$Diagnostically_positive)

## Calculate probabilities of compartment membership for each posterior draw
pred_df_m2$pi.D <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, mean)
pred_df_m2$SD <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, sd)
pred_df_m2$LB <- apply(model2_df[,grep("pi.D", names(model2_df))], 2,
                       quantile, probs=0.025)
```

```r
pred_df_m2$UB <- apply(model2_df[,grep("pi.D", names(model2_df))], 2,
                        quantile, probs=0.975)


summary(pred_df_m2)
```

```
##       obs              pi.D                  SD
##  Min.   :  1.0   Min.   :1.200e-07   Min.   :1.410e-06
##  1st Qu.:193.5   1st Qu.:2.870e-06   1st Qu.:2.558e-05
##  Median :386.0   Median :1.119e-05   Median :1.052e-04
##  Mean   :386.0   Mean   :7.825e-04   Mean   :8.943e-04
##  3rd Qu.:578.5   3rd Qu.:2.427e-05   3rd Qu.:1.823e-04
##  Max.   :771.0   Max.   :7.924e-02   Max.   :4.308e-02
##        LB                  UB            model_assignment Clinical_status
##  Min.   :0.000e+00   Min.   :1.300e-07   Mode:logical     A:  0
##  1st Qu.:0.000e+00   1st Qu.:1.228e-05   NA's:771         N:736
##  Median :0.000e+00   Median :6.166e-05                    S: 35
##  Mean   :7.189e-05   Mean   :3.074e-03
##  3rd Qu.:0.000e+00   3rd Qu.:2.167e-04
##  Max.   :1.974e-02   Max.   :1.635e-01
##  Diagnostic_status
##  Negative:743
##  Positive: 28
##
##
##
##
```

```r
## Apply a cut off of point estimate of 0.5; if pi.D > 0.5, classify as S (symptomatic), otherwise as N
## Summarize in a table (clinical status versus diagnostic status)
table(pred_df_m2[pred_df_m2$pi.D > 0.5,]$Clinical_status,
      pred_df_m2[pred_df_m2$pi.D > 0.5,]$Diagnostic_status)
```

```
##
##     Negative Positive
##   A        0        0
##   N        0        0
##   S        0        0
```

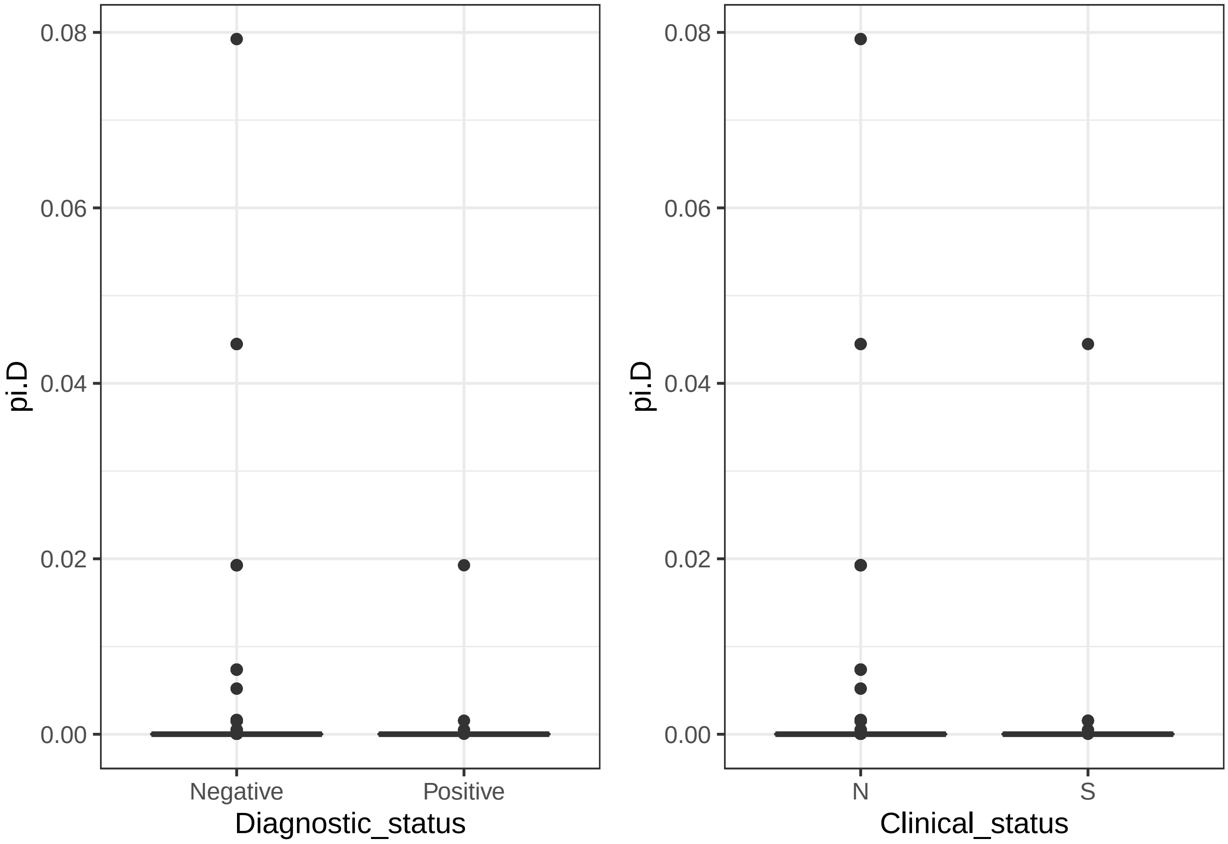
```r
## Print summary table of clinical status versus diagnostic status from the original data
table(pred_df_m2$Clinical_status, pred_df_m2$Diagnostic_status)
```

```
##
##     Negative Positive
##   A        0        0
##   N      728        8
##   S       15       20
```

```r
## boxplots
p1 <- (ggplot(data=pred_df_m2, aes(x=Diagnostic_status, y=pi.D))
       + geom_boxplot()
```

```
        + theme_bw())
p2 <- (ggplot(data=pred_df_m2, aes(x=Clinical_status, y=pi.D))
        + geom_boxplot()
        + theme_bw())

ggarrange(p1,p2, nrow=1)
```



From the first table, we can see that we identify all of the diagnostically positive individuals as having disease. We supplied diagnostic status, so the model is perfectly recovering the diagnostic status, but missing all those that are diagnostically negative but are symptomatic based on clinical status (15 - see second table). It seems like the sensitivity and specificity pieces are not making a difference right now..