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 π_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 π_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:

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.

$$\text{logit}(P(D_k)) \sim \text{Normal}(\text{logit}(\pi) + \mathbf{x}_k^T \boldsymbol{\beta} + \epsilon_k,\ \delta^2)$$

where π is the population prevalence of disease, $\mathbf{x}_k^T = (1, Age_k, Sex_k)$, and ϵ_k is a random individual effect.

Prior Model

Prevalence:

Fix prevalence at 0.075.

Other parameters:

$$\boldsymbol{\epsilon} \sim Normal(\mathbf{0},\ \text{precison} = 5 \times 10^{-3} * I)$$

There are individual level random effects ϵ_k, $k = 1, ..., K$, and they are assumed to be independent.

OpenBUGS Model 1 Implementation

OpenBUGS Model 1 Disease State Prediction

```
## Set up storage for model results
# pred_df_m1 <- 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_m1$pi.D <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, mean)
# pred_df_m1$SD <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, sd)
# pred_df_m1$LB <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, quantile, probs=0.025)
# pred_df_m1$UB <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, quantile, probs=0.975)
```

Model 2:

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,

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

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.

$$\text{logit}(P(D_k)) \sim \text{Normal}(\text{logit}(\pi) + \mathbf{x}_k^T \boldsymbol{\beta} + \epsilon_k,\ \delta^2)$$

where π is the population prevalence of disease, $\mathbf{x}_k^T = (1, Age_k, Sex_k)$, and ϵ_k is a random individual effect.

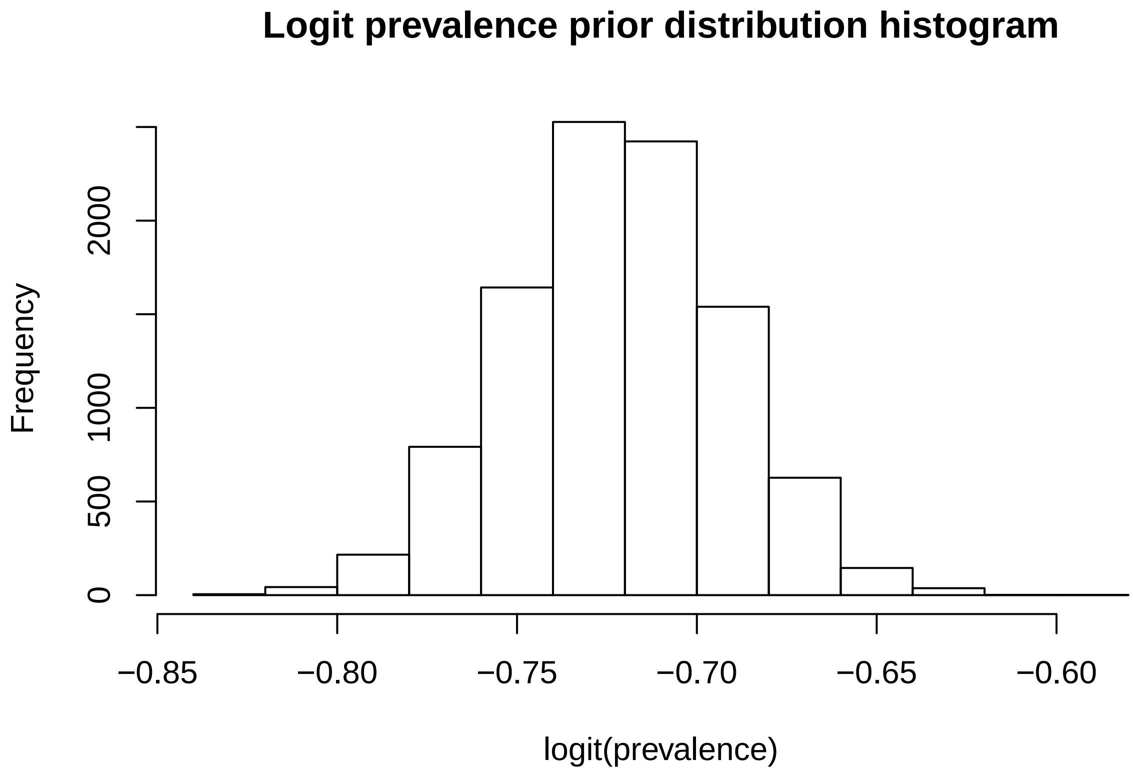
Prior Model

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)")
```



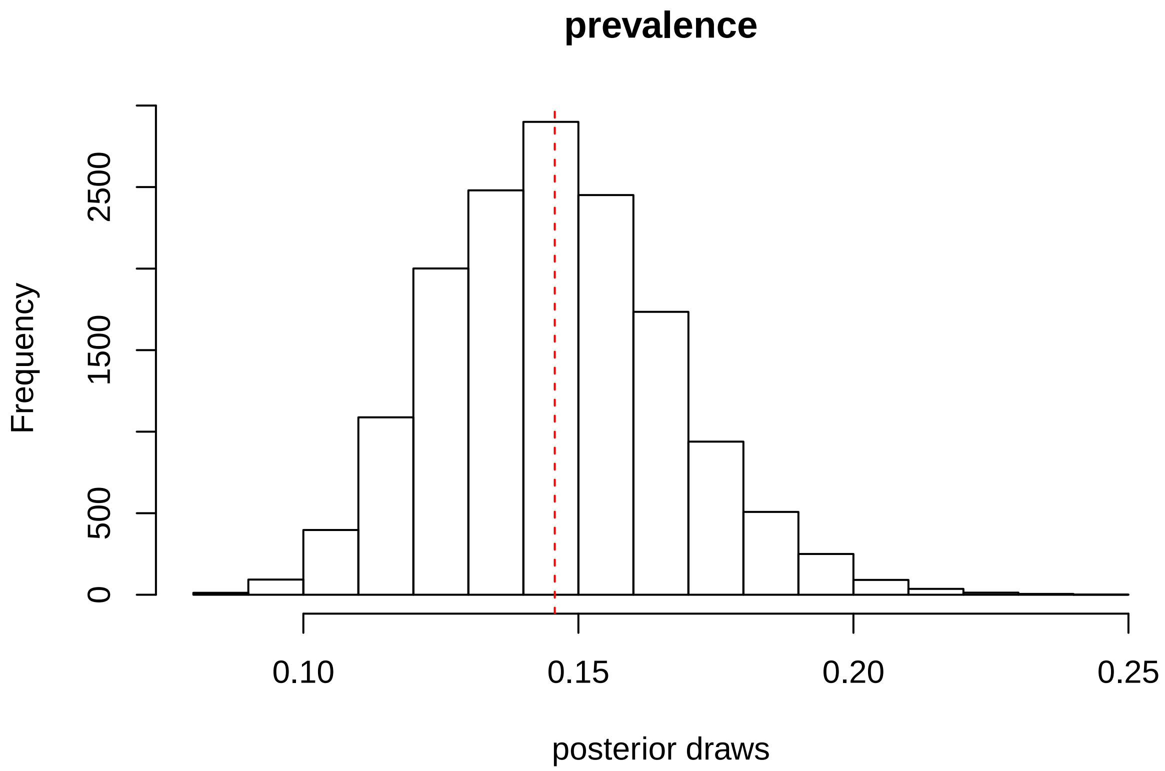
Other parameters:

$$\boldsymbol{\epsilon} \sim Normal(\mathbf{0},\ \text{precison} = 5 \times 10^{-3} * I)$$

There are individual level random effects ϵ_k, $k = 1, ..., K$, and they are assumed to be independent.

OpenBUGS Model 2 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")
```



```
## 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 0.1457 and the 95% credible interval is: (0.1076, 0.1908).

OpenBUGS Model 2 Disease State Prediction

```
## 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$LB <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, quantile, probs=0.025)
pred_df_m2$UB <- apply(model2_df[,grep("pi.D", names(model2_df))], 2, quantile, probs=0.975)
```

```
summary(pred_df_m2)
```

```
##       obs             pi.D               SD               LB           
##  Min.   :  1.0   Min.   :0.05816   Min.   :0.1844   Min.   :8.080e-16  
##  1st Qu.:193.5   1st Qu.:0.06705   1st Qu.:0.2036   1st Qu.:3.575e-15  
##  Median :386.0   Median :0.06993   Median :0.2091   Median :5.424e-15  
##  Mean   :386.0   Mean   :0.09044   Mean   :0.2173   Mean   :1.835e-13  
##  3rd Qu.:578.5   3rd Qu.:0.07300   3rd Qu.:0.2150   3rd Qu.:8.122e-15  
##  Max.   :771.0   Max.   :0.65411   Max.   :0.4507   Max.   :1.251e-11  
##        UB         model_assignment Clinical_status Diagnostic_status
##  Min.   :0.8763   Mode:logical     A:  0           Negative:743     
##  1st Qu.:0.9894   NA's:771         N:736           Positive: 28     
##  Median :0.9962                    S: 35                            
##  Mean   :0.9912                                                     
##  3rd Qu.:0.9988                                                     
##  Max.   :1.0000                                                     
```

```
## 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_
```

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

```
## 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
```

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..

Model 3:

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. We include Sex and Age as explanatory variables, which we didn't do in Model 2.

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.

$$\text{logit}(P(D_k)) \sim \text{Normal}(\text{logit}(\pi) + \mathbf{x}_k^T\boldsymbol{\beta} + \epsilon_k,\ \delta^2)$$

where π is the population prevalence of disease, $\mathbf{x}_k^T = (1, Age_k, Sex_k)$, and ϵ_k is a random individual effect.

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.

Other parameters:

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

We will assume that the regression coefficients are independent, so Σ_β is a diagonal matrix.

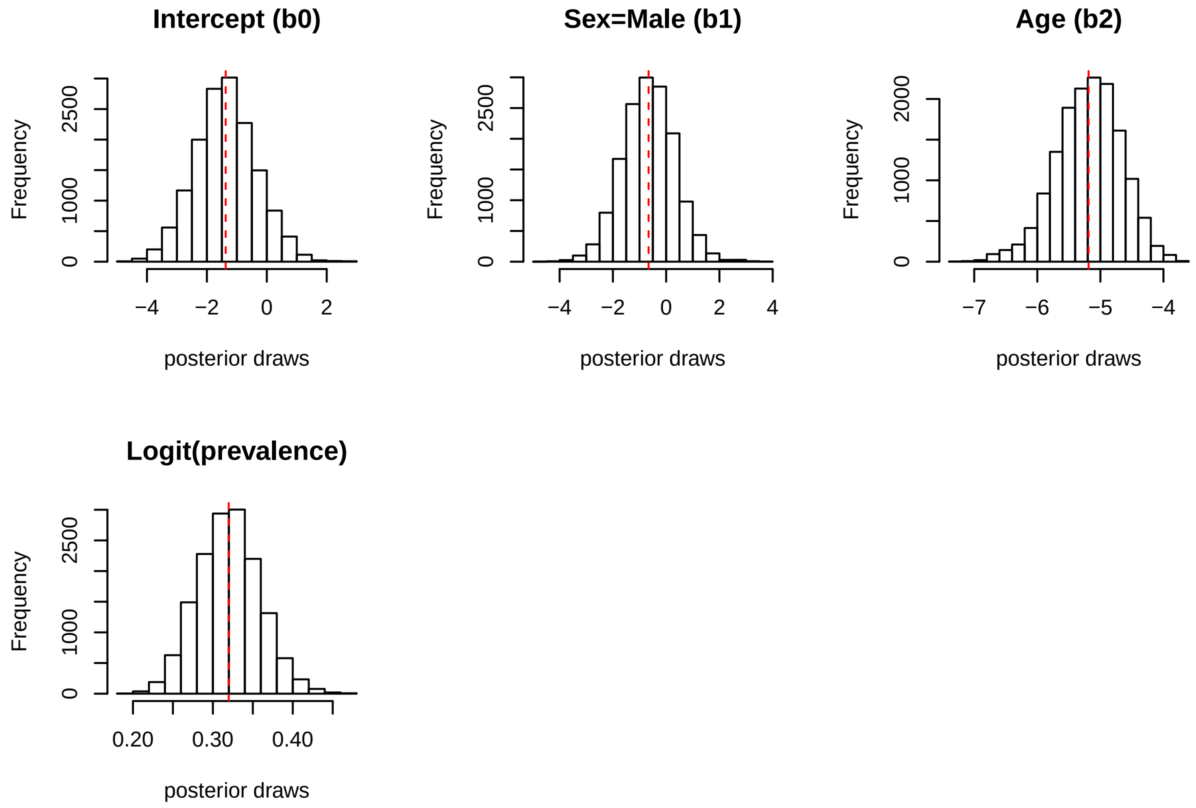
$$\boldsymbol{\epsilon} \sim Normal(\mathbf{0},\ \text{precison} = 5 \times 10^{-3} * I)$$

There are individual level random effects ϵ_k, $k = 1, ..., K$, and they are assumed to be independent.

OpenBUGS Model 3 Implementation

```
## Graphial summaries of posterior distribtuions
par(mfrow=c(2,3))
hist(model3_df$b0, main="Intercept (b0)", xlab="posterior draws")
abline(v=mean(model3_df$b0), lty="dashed", col="red")
hist(model3_df$b1, main="Sex=Male (b1)", xlab="posterior draws")
abline(v=mean(model3_df$b1), lty="dashed", col="red")
hist(model3_df$b2, main="Age (b2)", xlab="posterior draws")
abline(v=mean(model3_df$b2), lty="dashed", col="red")
hist(exp(model3_df$lpi)/(1+exp(model3_df$lpi)), main="Logit(prevalence)", xlab="posterior draws")
abline(v=mean(exp(model3_df$lpi)/(1+exp(model3_df$lpi))), lty="dashed", col="red")

## Numeric summaries of posterior distributions
#boxplot(model3_df[,!(names(model3_df) %in% c("deviance"))])
```



On this run, the mean prevalence is 0.32, which is outside the range that we have from the literature: (0.05, 0.10). Why is this so high? Should there be an intercept in this model? If we add in the estimate for the intercept and claim that this is the mean estimate for the disease prevalence (which I don't think makes any sense), the resulting estimate is: 0.142, which still seems to be too high, but not nearly as bad.

Options:

(1) Simplify the model by fixing prevalence, and see what we get.
(2) Change the prior on prevalence? Make it less vague?

OpenBUGS Model 3 Disease State Prediction

```
## Set up storage for model results
pred_df_m3 <- 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_m3$pi.D <- apply(model3_df[,grep("pi.D", names(model3_df))], 2, mean)
pred_df_m3$SD <- apply(model3_df[,grep("pi.D", names(model3_df))], 2, sd)
pred_df_m3$LB <- apply(model3_df[,grep("pi.D", names(model3_df))], 2, quantile, probs=0.025)
pred_df_m3$UB <- apply(model3_df[,grep("pi.D", names(model3_df))], 2, quantile, probs=0.975)

summary(pred_df_m3)
```

```
##       obs             pi.D                 SD
##  Min.   :  1.0   Min.   :0.0000000   Min.   :0.000000
##  1st Qu.:193.5   1st Qu.:0.0000092   1st Qu.:0.000709
##  Median :386.0   Median :0.0001630   Median :0.006477
##  Mean   :386.0   Mean   :0.0024453   Mean   :0.016204
##  3rd Qu.:578.5   3rd Qu.:0.0006376   3rd Qu.:0.015983
##  Max.   :771.0   Max.   :0.4489169   Max.   :0.462152
##        LB                   UB          model_assignment Clinical_status
##  Min.   :0.000e+00   Min.   :0.0000000   Mode:logical     A:  0
##  1st Qu.:0.000e+00   1st Qu.:0.0000000   NA's:771         N:736
##  Median :0.000e+00   Median :0.0000001                    S: 35
##  Mean   :4.377e-19   Mean   :0.0142204
##  3rd Qu.:0.000e+00   3rd Qu.:0.0000123
##  Max.   :1.158e-16   Max.   :1.0000000
##  Diagnostic_status
##  Negative:743
##  Positive: 28
##
##
##
##
```

```
## 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_m3[pred_df_m3$pi.D > 0.5,]$Clinical_status, pred_df_m3[pred_df_m3$pi.D > 0.5,]$Diagnostic_
```

```
##
##     Negative Positive
##   A        0        0
##   N        0        0
##   S        0        0
## Print summary table of clinical status versus diagnostic status from the original data
table(pred_df_m3$Clinical_status, pred_df_m3$Diagnostic_status)
```

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