Lecture 23 - Sensitivity, Specificity, and Decisions

Sta102 / BME102

April 25, 2016

Colin Rundel

Bird Keeping and Lung Cancer?

Birdkeeping and Lung Cancer

A 1972 - 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

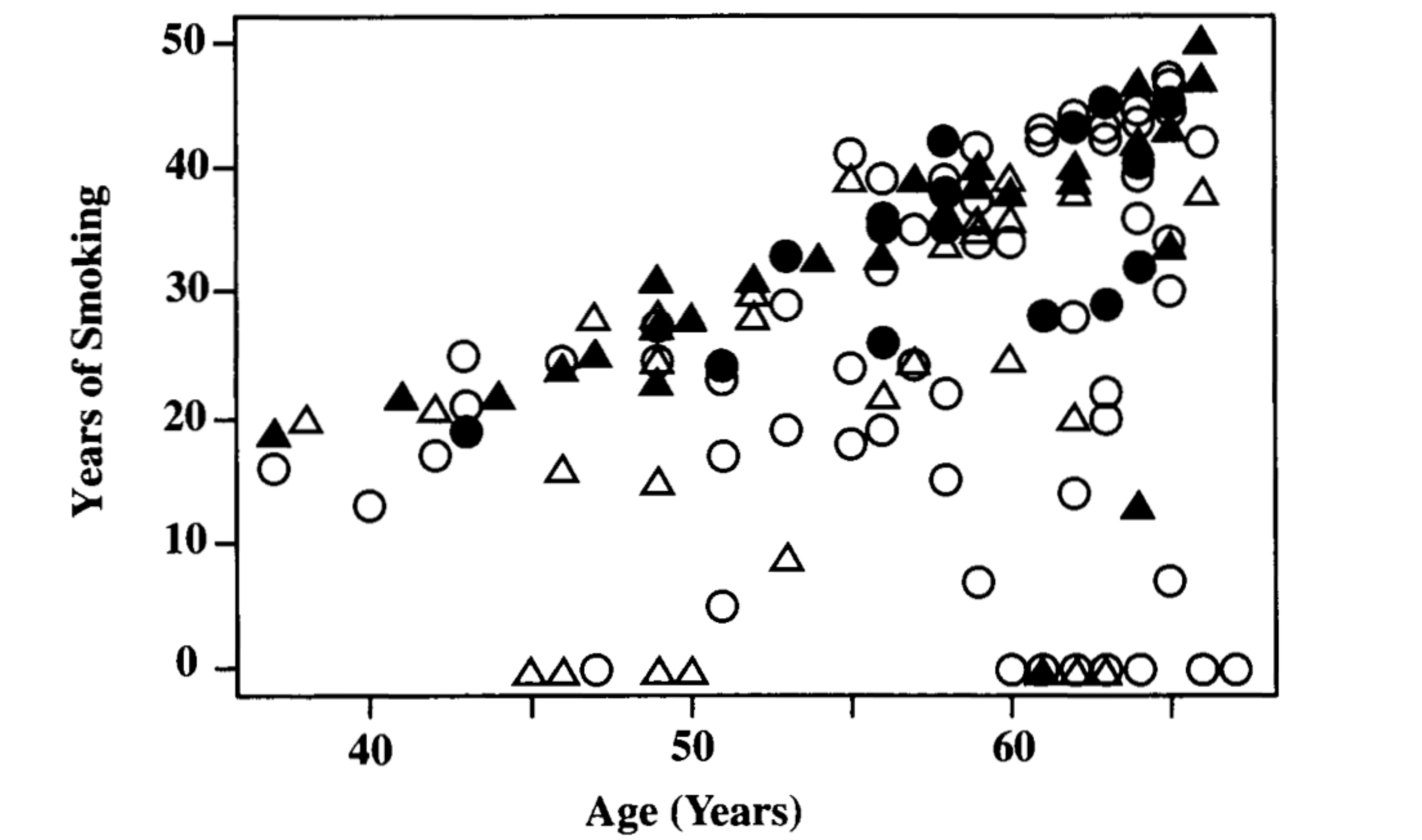
Data

	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37.00	19.00	12.00
2	LungCancer	Male	Low	Bird	41.00	22.00	15.00
3	LungCancer	Male	High	NoBird	43.00	19.00	15.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	NoCancer	Female	Low	NoBird	65.00	7.00	2.00

LC	Whether subject has lung cancer
FM	Sex of subject
SS	Socioeconomic status
BK	Indicator for birdkeeping
AG	Age of subject (years)
YR	Years of smoking prior to diagnosis or examination
CD	Average rate of smoking (cigarettes per day)

Note - NoCancer is the reference response (0 or failure), LungCancer is the contrast response (1 or success).

EDA



	Bird	No Bird
Lung Cancer	▲	●
No Lung Cancer	△	○

Model

```
summary({g=glm(LC ~ ., data=bird, family=binomial)})


##
## Call:
## glm(formula = LC ~ ., family = binomial, data = bird)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5642  -0.8333  -0.4605   0.9808   2.2460
##
## Coefficients:
##             Estimate Std. Error z value Pr(>♦z♦)
## (Intercept) -1.93736    1.80425  -1.074 0.282924
## FMFemale     0.56127    0.53116   1.057 0.290653
## SSHigh       0.10545    0.46885   0.225 0.822050
## BKBird       1.36259    0.41128   3.313 0.000923 ***
## AG          -0.03976    0.03548  -1.120 0.262503
## YR           0.07287    0.02649   2.751 0.005940 **
## CD           0.02602    0.02552   1.019 0.308055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
##
```

$$\underbrace{\log\left(\frac{p_i}{1-p_i}\right)} = b_0 + b_1 X_i \ldots$$

$b_{BK} \; X_{BK}$

$b_{AR} \; X_{AR}$

Model Selection

```
library(MASS)
g2 = stepAIC(g)


## Start:  AIC=168.2
## LC ~ FM + SS + BK + AG + YR + CD
##
##         Df Deviance    AIC
## - SS     1   154.25 166.25
## - CD     1   155.24 167.24
## - FM     1   155.32 167.32
## - AG     1   155.49 167.49
## <none>       154.20 168.20
## - YR     1   163.93 175.93
## - BK     1   165.87 177.87
##
## Step:  AIC=166.25
## LC ~ FM + BK + AG + YR + CD
##
##         Df Deviance    AIC
## - FM     1   155.32 165.32
```

Model Selection - Results

```r
summary(g2)
```

```
##
## Call:
## glm(formula = LC ~ BK + YR, family = binomial, data = bird)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6093  -0.8644  -0.5283   0.9479   2.0937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>♦z♦)
## (Intercept) -3.18016    0.63640  -4.997 5.82e-07 ***
## BKBird       1.47555    0.39588   3.727 0.000194 ***
## YR           0.05825    0.01685   3.458 0.000544 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.4756) = 4.37$.

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = \cancel{B_i} \quad b_i$$

$$\log\left(\frac{p_1/1-p_1}{p_2/1-p_2}\right) = b_i$$

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.4756) = 4.37$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0582) = 1.06$.

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.4756) = 4.37$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0582) = 1.06$.

What do these numbers mean in practice?

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are *not* 4x more likely to develop lung cancer than non-bird keepers.

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are *not* 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between *relative risk* and an *odds ratio*.

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

$$OR = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]}$$

Back to the birds - Low Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.05$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 4.37$$

Back to the birds - Low Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.05$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 4.37$$

$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.05}{0.95}}{1 + 4.37 \times \frac{0.05}{0.95}} = 0.187$$

Back to the birds - Low Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.05$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 4.37$$

$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.05}{0.95}}{1 + 4.37 \times \frac{0.05}{0.95}} = 0.187$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.187/0.05 = 3.74$$

Back to the birds - High Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.25$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.25/[1 - 0.25]} = 4.37$$

Back to the birds - High Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.25$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.25/[1 - 0.25]} = 4.37$$

$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.25}{0.75}}{1 + 4.37 \times \frac{0.25}{0.75}} = 0.593$$

Back to the birds - High Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.25$?

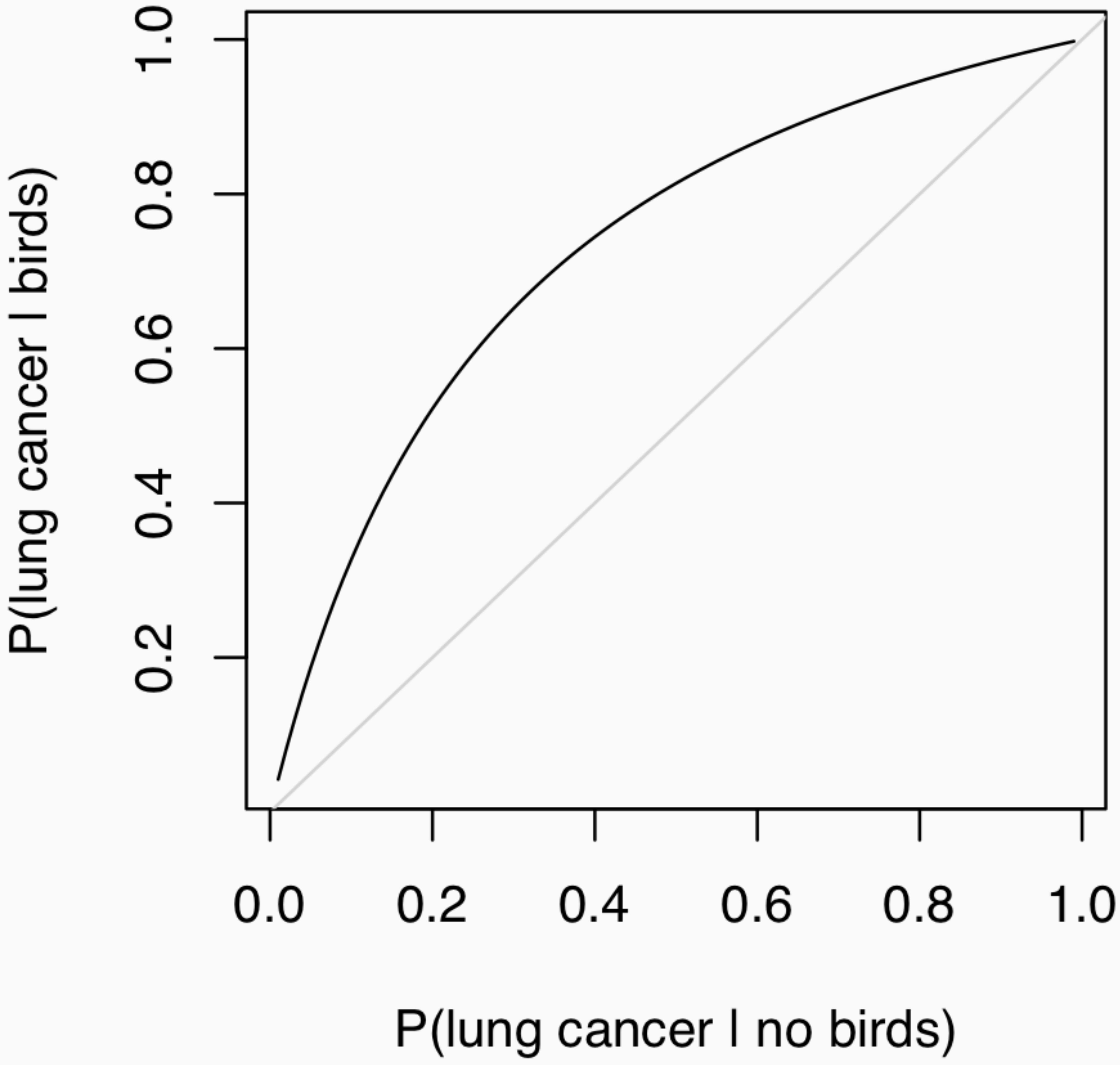
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.25/[1 - 0.25]} = 4.37$$

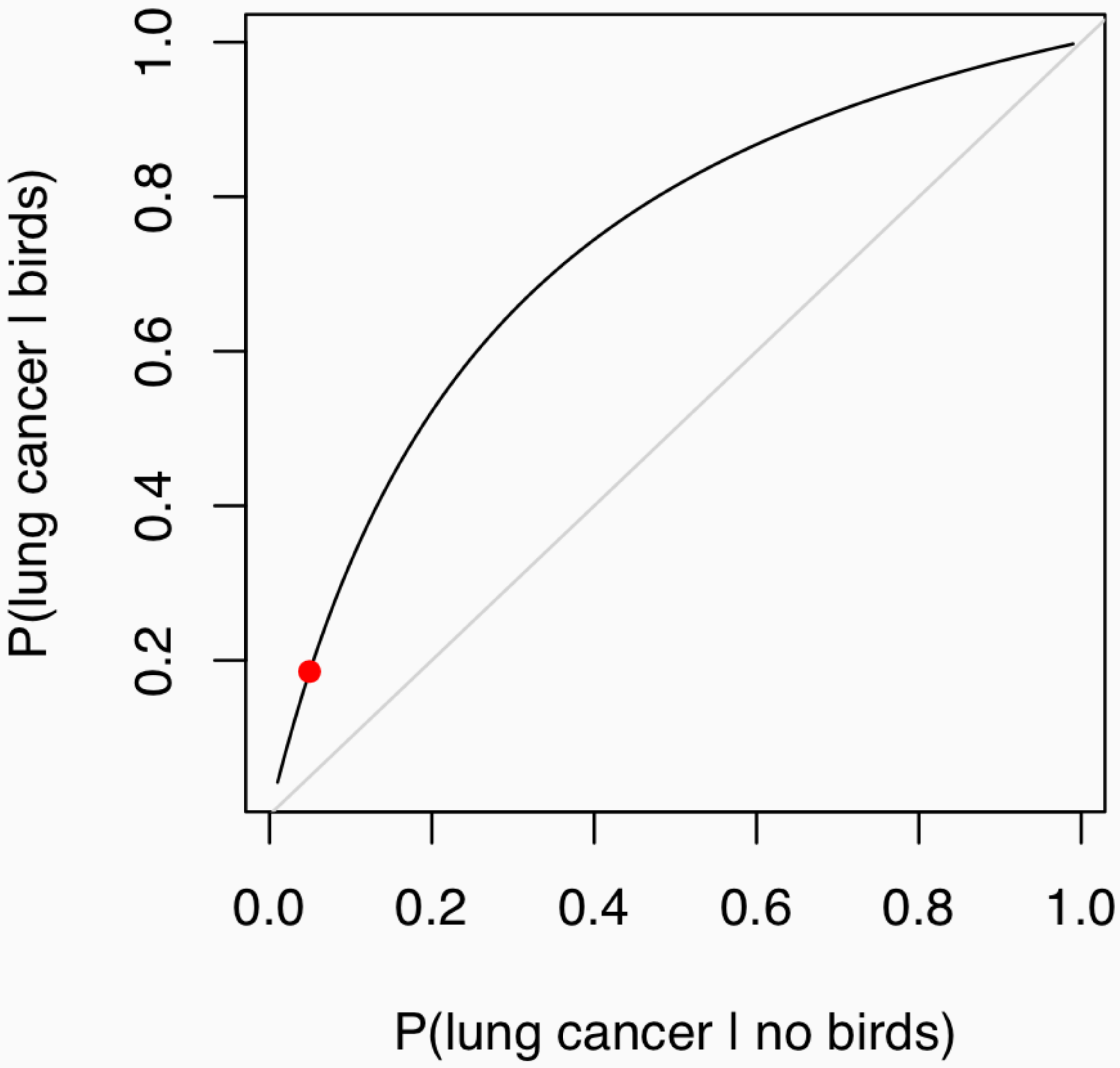
$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.25}{0.75}}{1 + 4.37 \times \frac{0.25}{0.75}} = 0.593$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.593/0.25 = 2.37$$

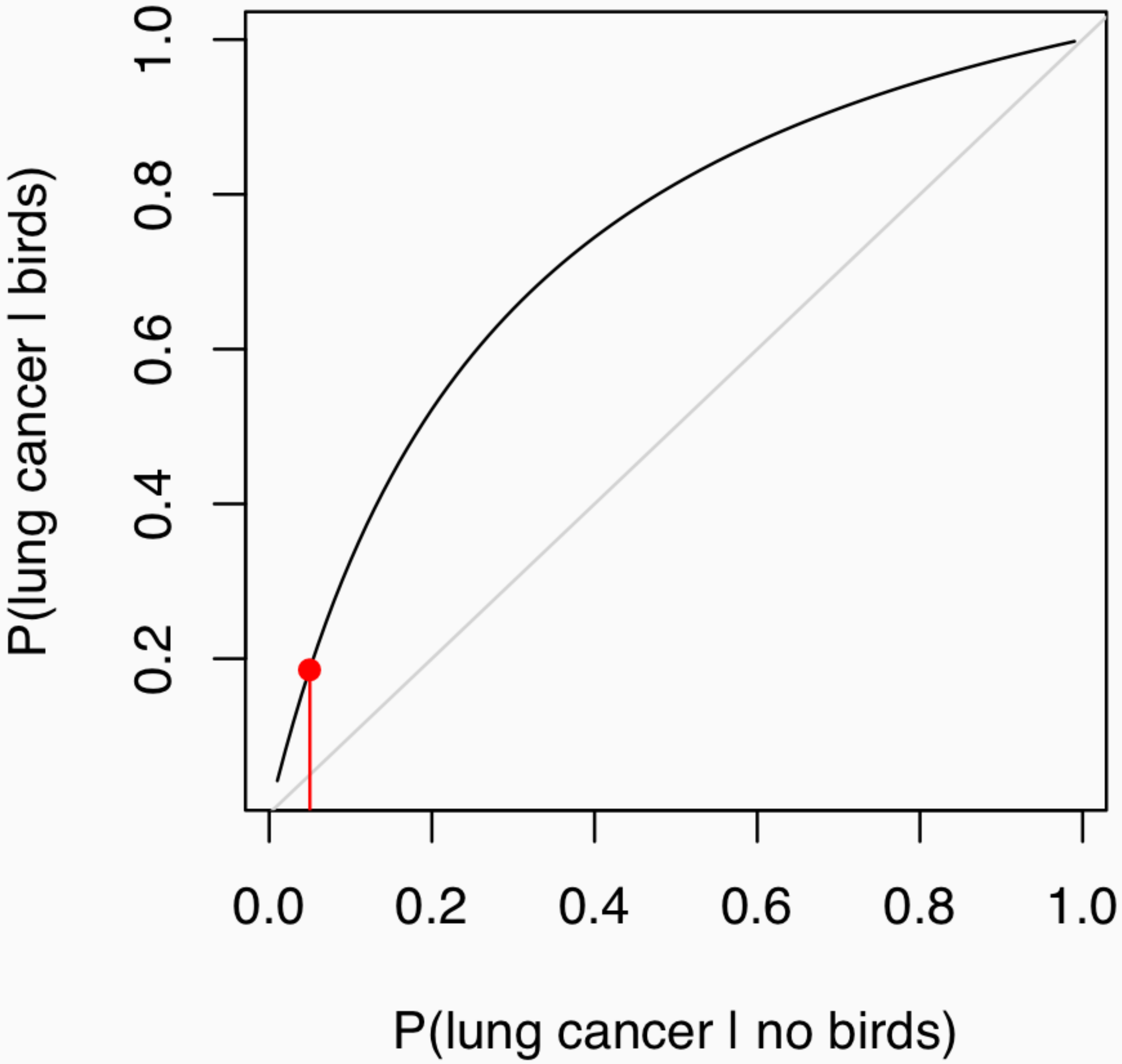
Bird OR Curve



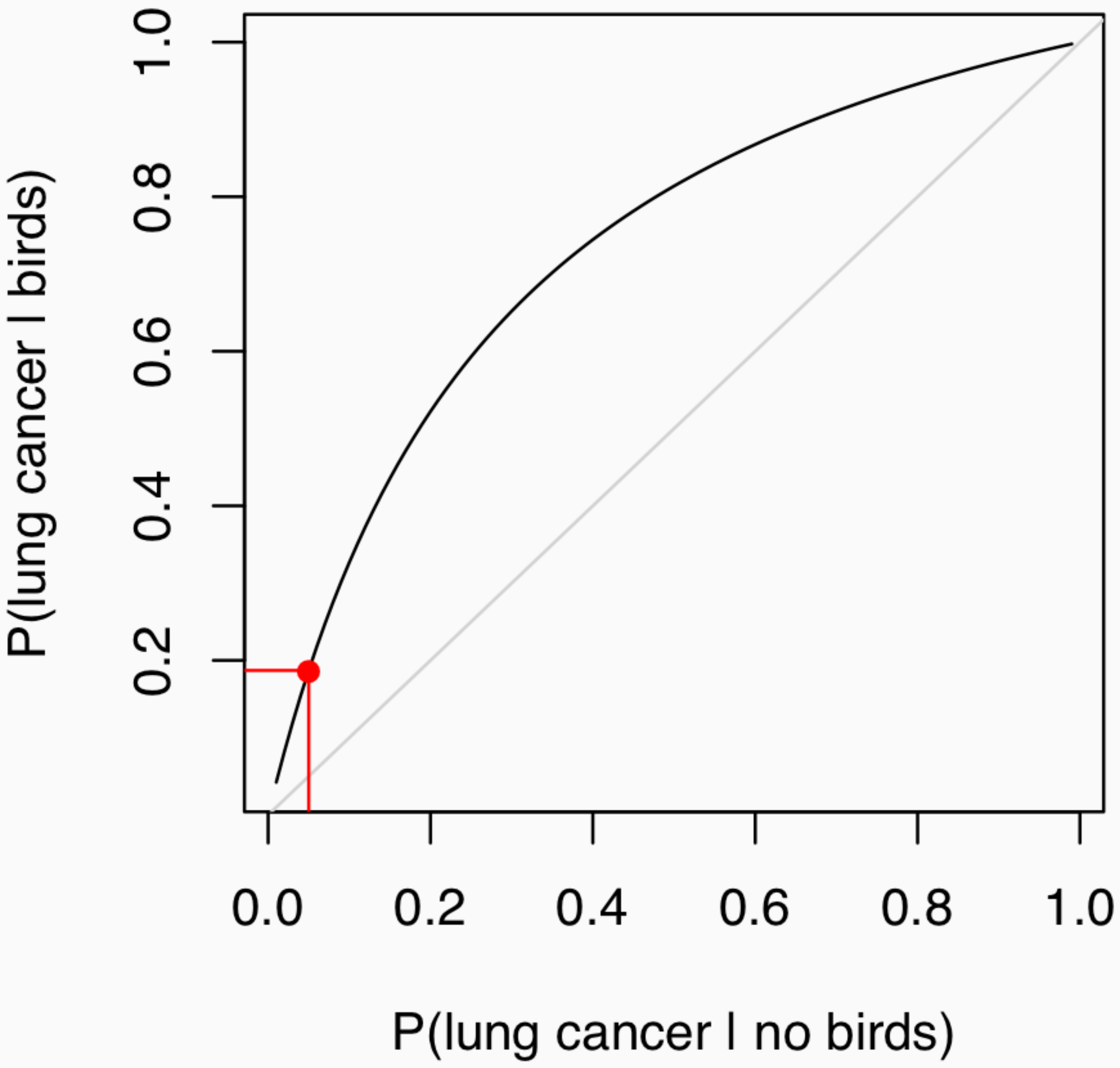
Bird OR Curve



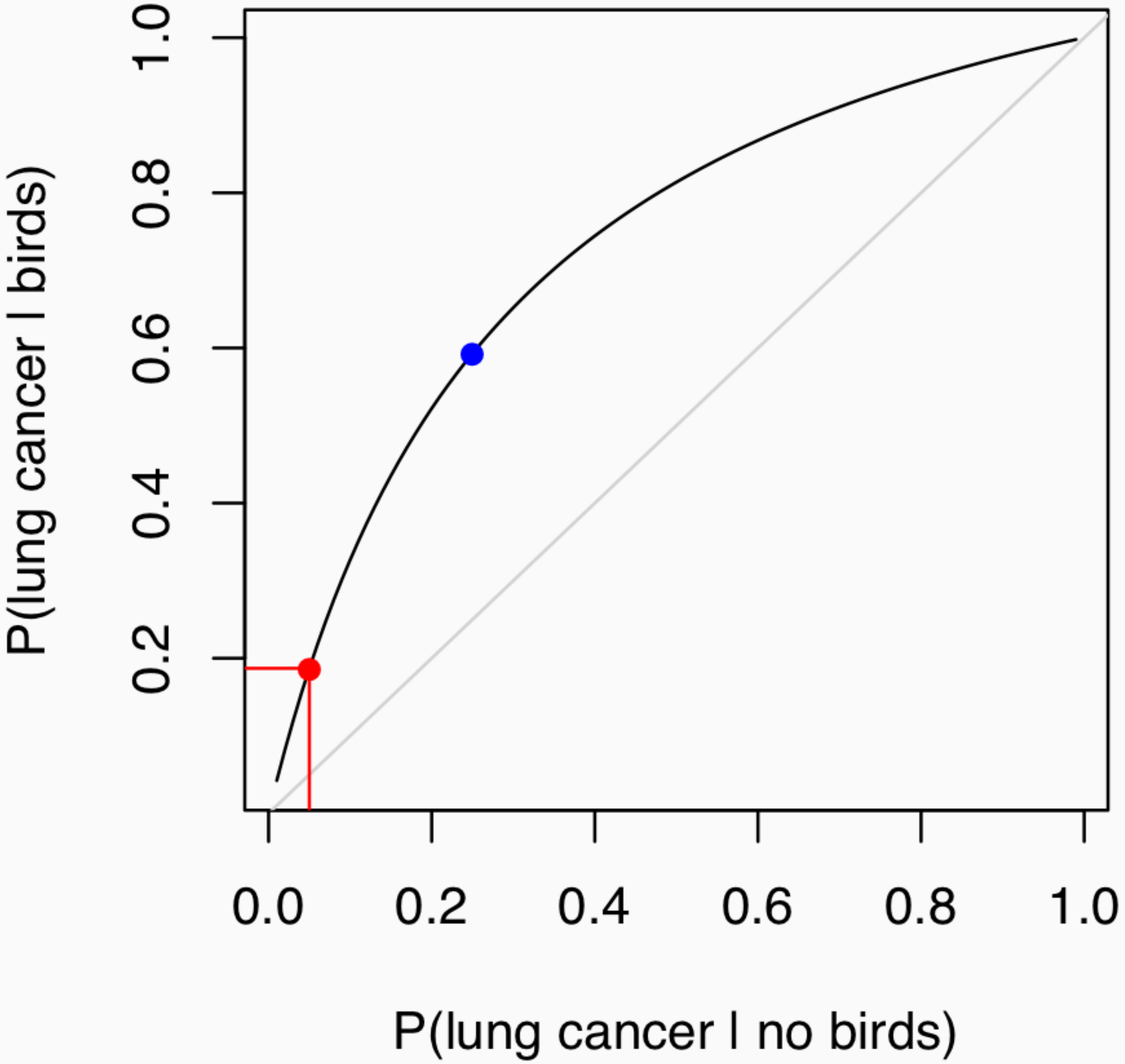
Bird OR Curve



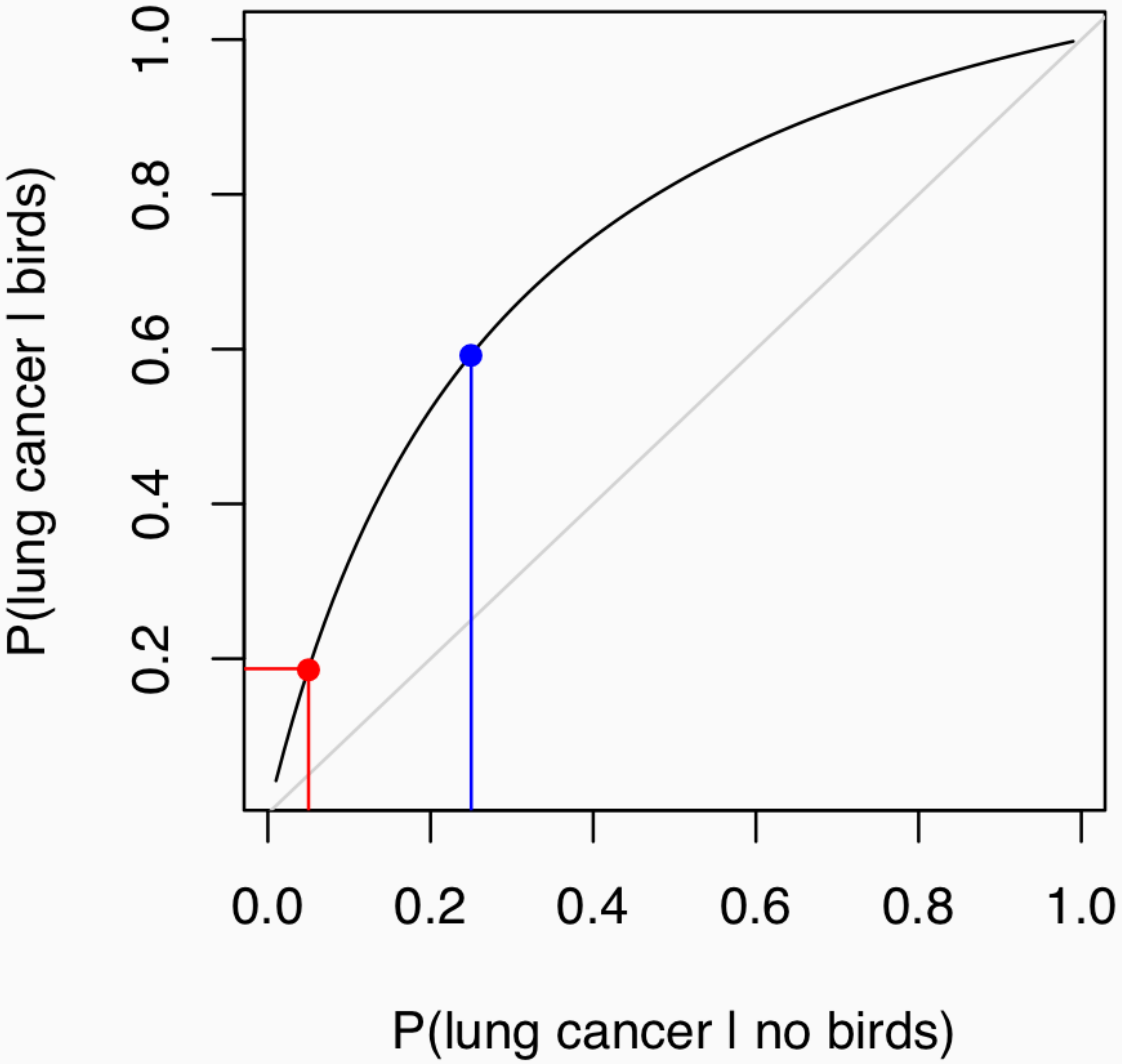
Bird OR Curve



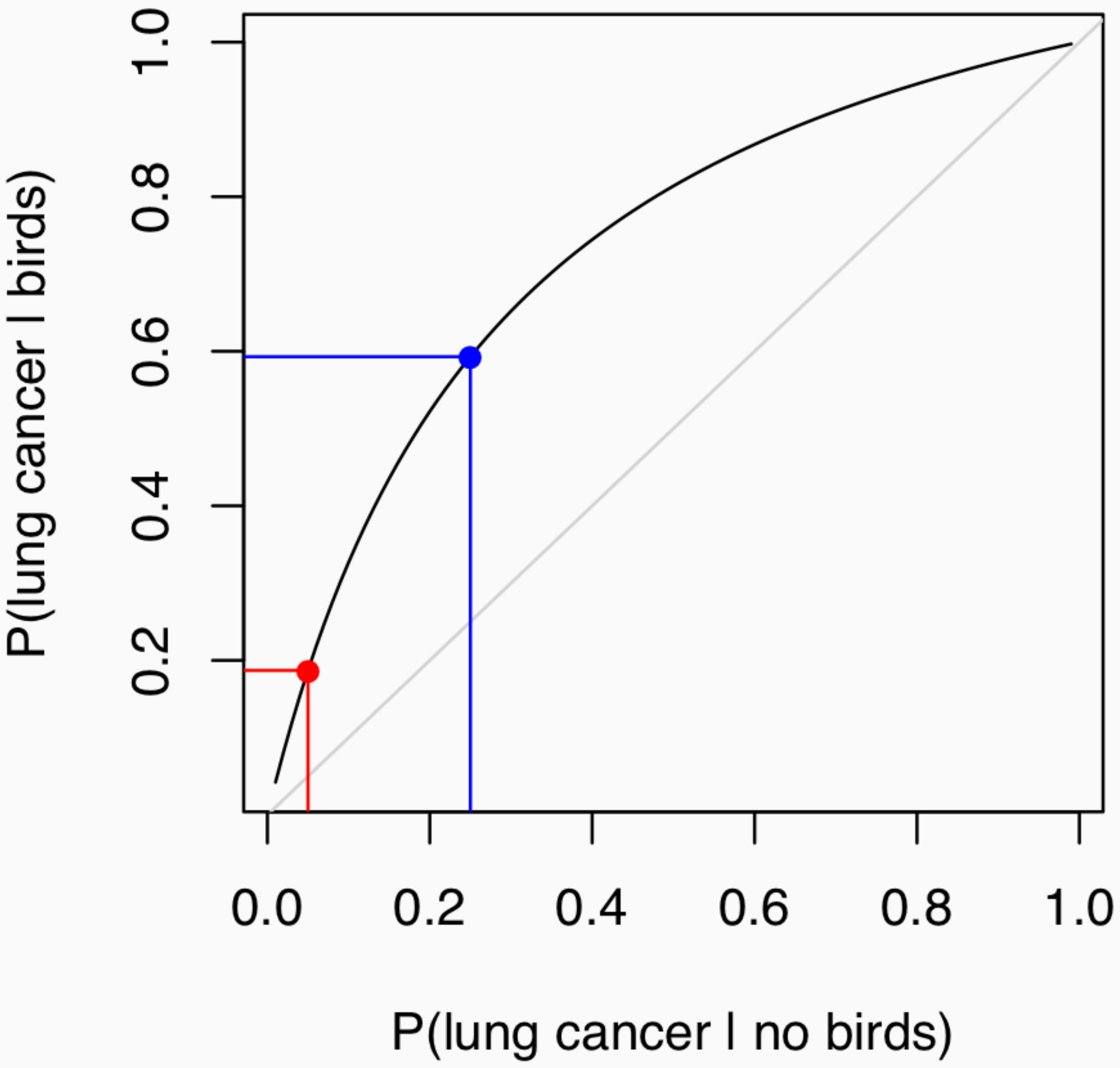
Bird OR Curve



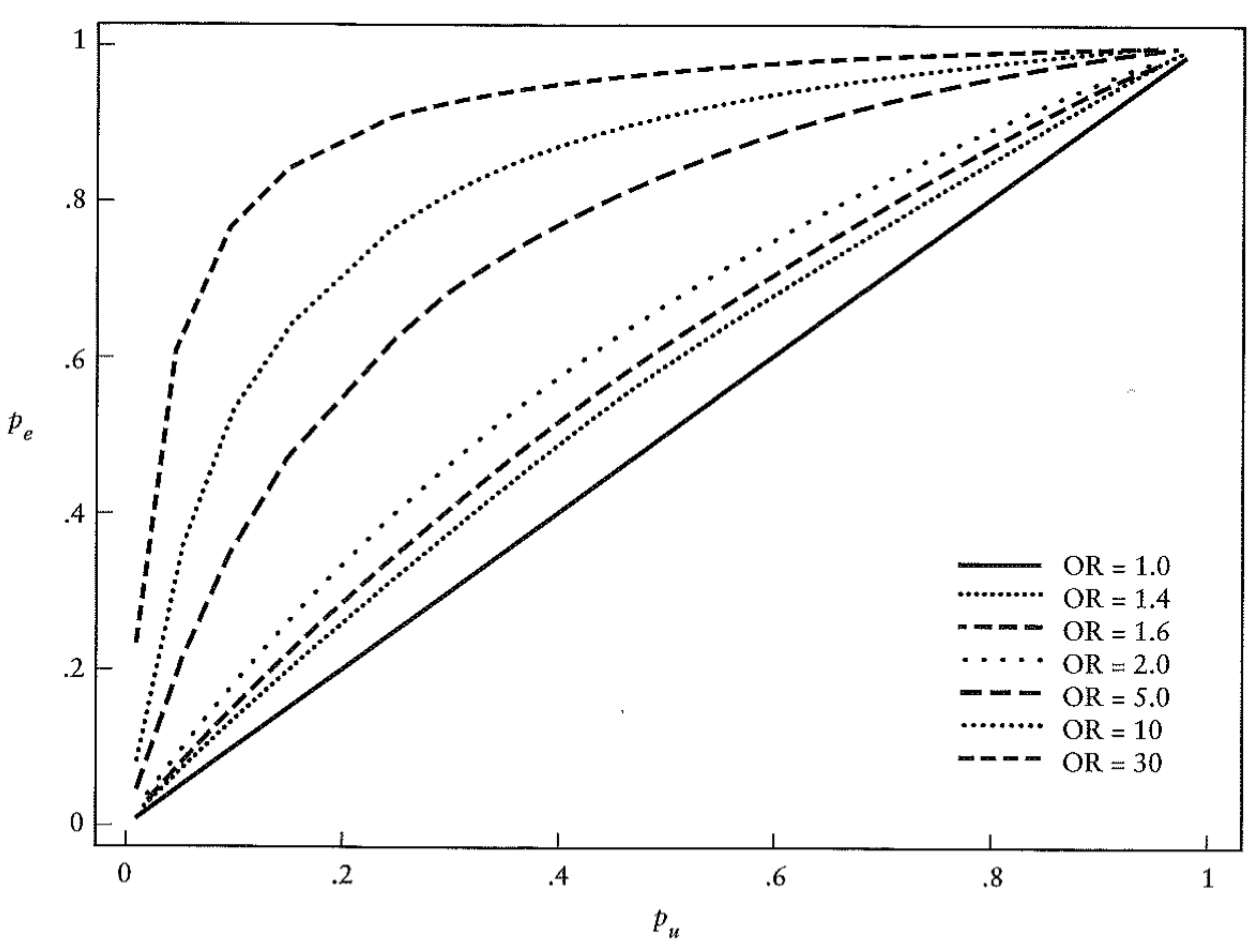
Bird OR Curve



Bird OR Curve



OR Curves



Residuals

Using the logistic regression model we can predict probabilities,

$$\hat{p}_i = \text{logit}^{-1}\left(b_0 + b_1 x_1 + \ldots + b_k x_k\right)$$

Residuals

Using the logistic regression model we can predict probabilities,

$$\hat{p}_i = \text{logit}^{-1}\left(b_0 + b_1 x_1 + \ldots + b_k x_k\right)$$

MLR-like Residual:

$$r_i = y_i - \hat{p}_i$$

Residuals

Using the logistic regression model we can predict probabilities,

$$\hat{p}_i = \text{logit}^{-1}\left(b_0 + b_1 x_1 + \ldots + b_k x_k\right)$$

MLR-like Residual:

$$r_i = y_i - \hat{p}_i$$

$$\underset{b_0 b_1 b_2 \ldots}{\arg\min} \sum r_i^2$$

Deviance Residual:

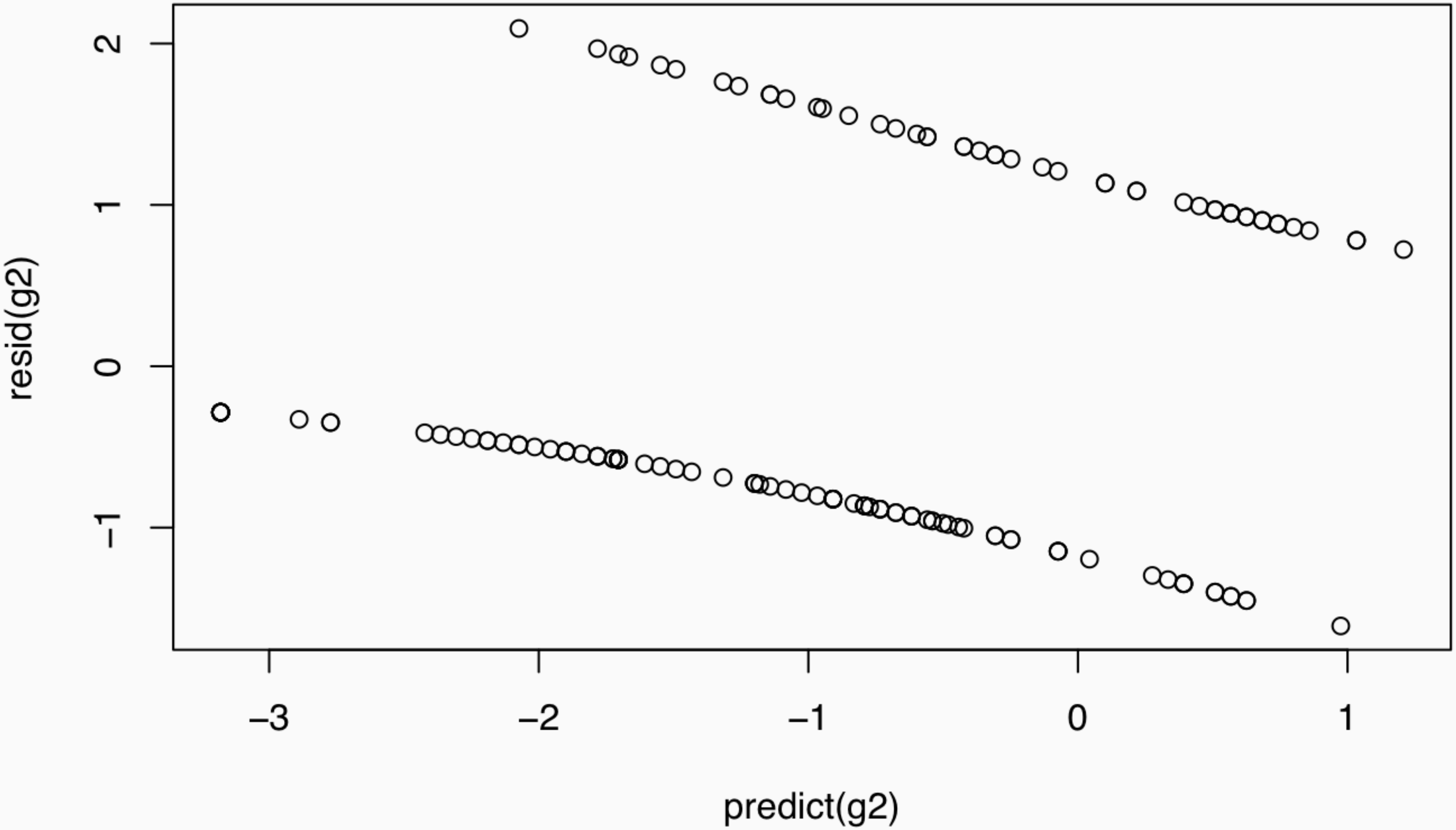
$$r_i = -s_i \sqrt{-2\left(y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right)}$$

where

$$s_i = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = 0 \end{cases}$$

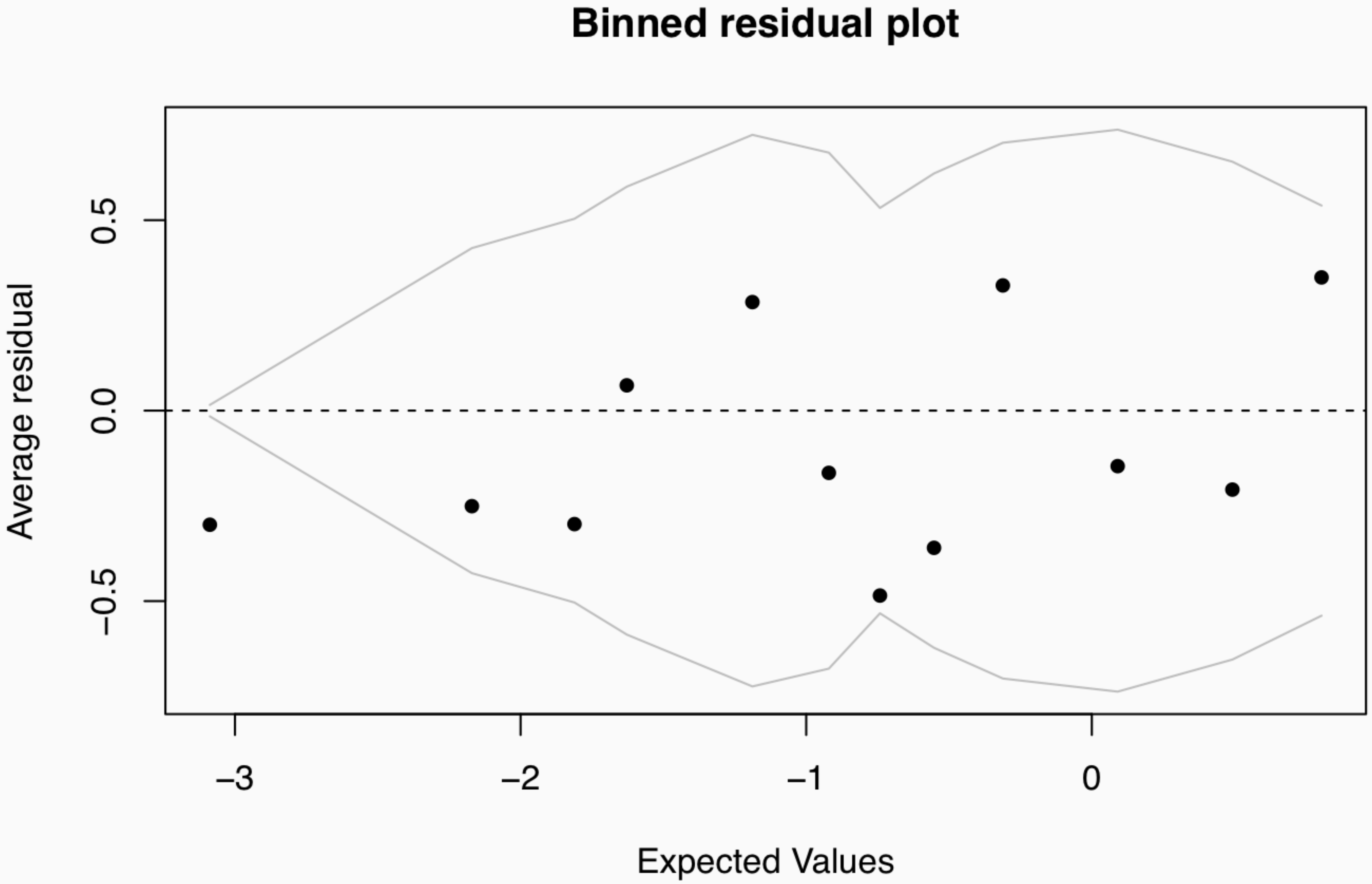
Diagnostics?

```r
plot(predict(g2),resid(g2))
```



Diagnostics - Binning

```r
library(arm)
binnedplot(predict(g2),resid(g2))
```



Sensitivity and Specificity

(An old) Example - *House*

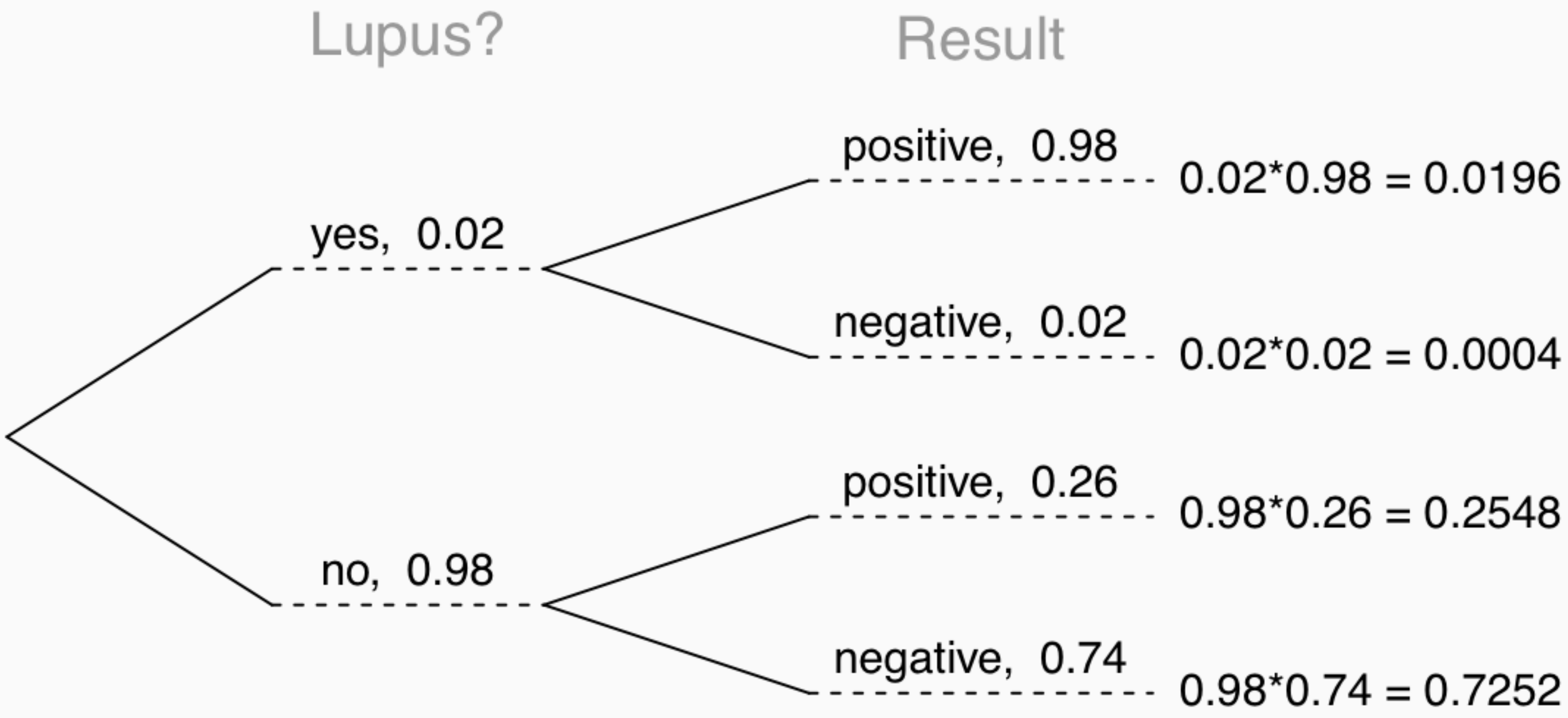
If you've ever watched the TV show *House* on Fox, you know that Dr. House regularly states, "It's never lupus."

Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease.

The test for lupus is very accurate if the person actually has lupus, however is very inaccurate if the person does not. More specifically, the test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease.

Is Dr. House correct even if someone tests positive for Lupus?

(An old) Example - *House*



$$
\begin{aligned}
P(\text{Lupus}|+) &= \frac{P(+,\text{Lupus})}{P(+,\text{Lupus}) + P(+,\text{No Lupus})} \\
&= \frac{0.0196}{0.0196 + 0.2548} = 0.0714
\end{aligned}
$$

Testing for lupus

It turns out that testing for Lupus is actually quite complicated, a diagnosis usually relies on the outcome of multiple tests, often including: a complete blood count, an erythrocyte sedimentation rate, a kidney and liver assessment, a urinalysis, and or an antinuclear antibody (ANA) test.

It is important to think about what is involved in each of these tests (e.g. deciding if complete blood count is high or low) and how each of the individual tests and related decisions plays a role in the overall decision of diagnosing a patient with lupus.

Testing for lupus (cont.)

At some level we can view a diagnosis as a binary decision (lupus or no lupus) that involves the complex integration of various explanatory variables.

The example does not give us any information about how a diagnosis is made, but what it does give us is just as important - the *sensitivity* and the *specificity* of the test(s). These values are critical for our understanding of what a positive or negative test result actually means.

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

$$P(\text{Test } + \mid \text{Conditon } +) = P(+ \mid \text{lupus}) = 0.98$$

Specificity - measures a tests ability to identify negative results.

$$P(\text{Test } - \mid \text{Condition } -) = P(- \mid \text{no lupus}) = 0.74$$

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

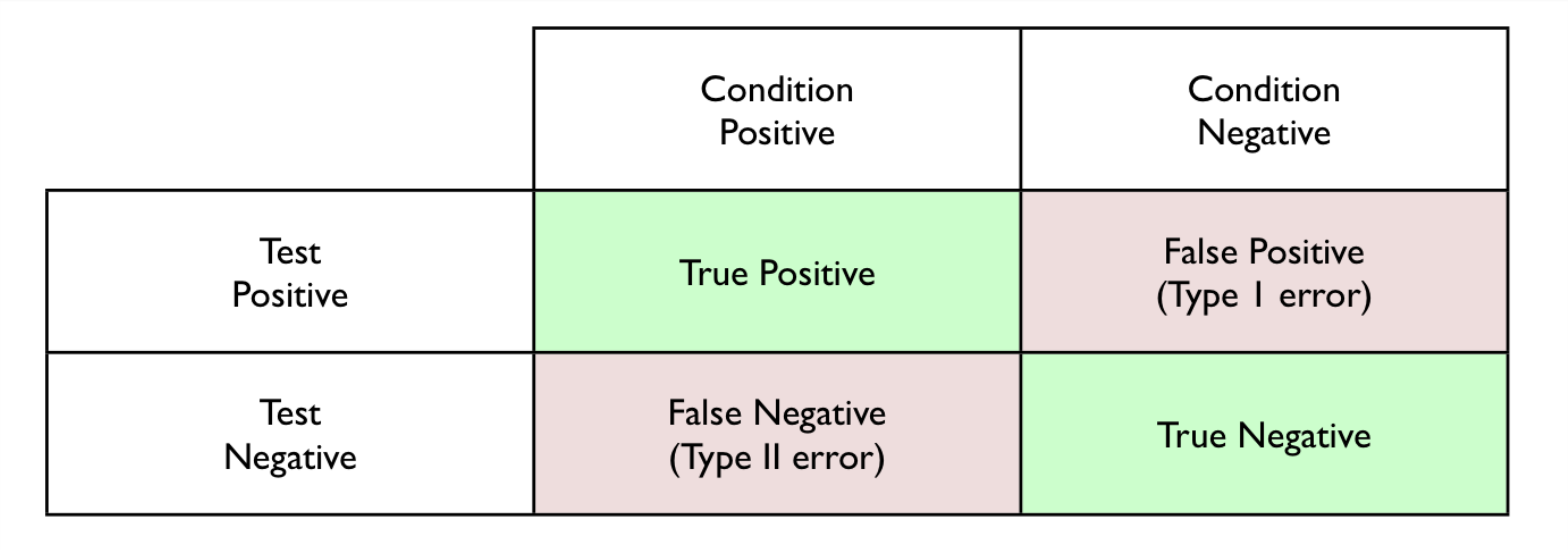
$$P(\text{Test } + \mid \text{Conditon } +) = P(+ \mid \text{lupus}) = 0.98$$

Specificity - measures a tests ability to identify negative results.

$$P(\text{Test } - \mid \text{Condition } -) = P(- \mid \text{no lupus}) = 0.74$$

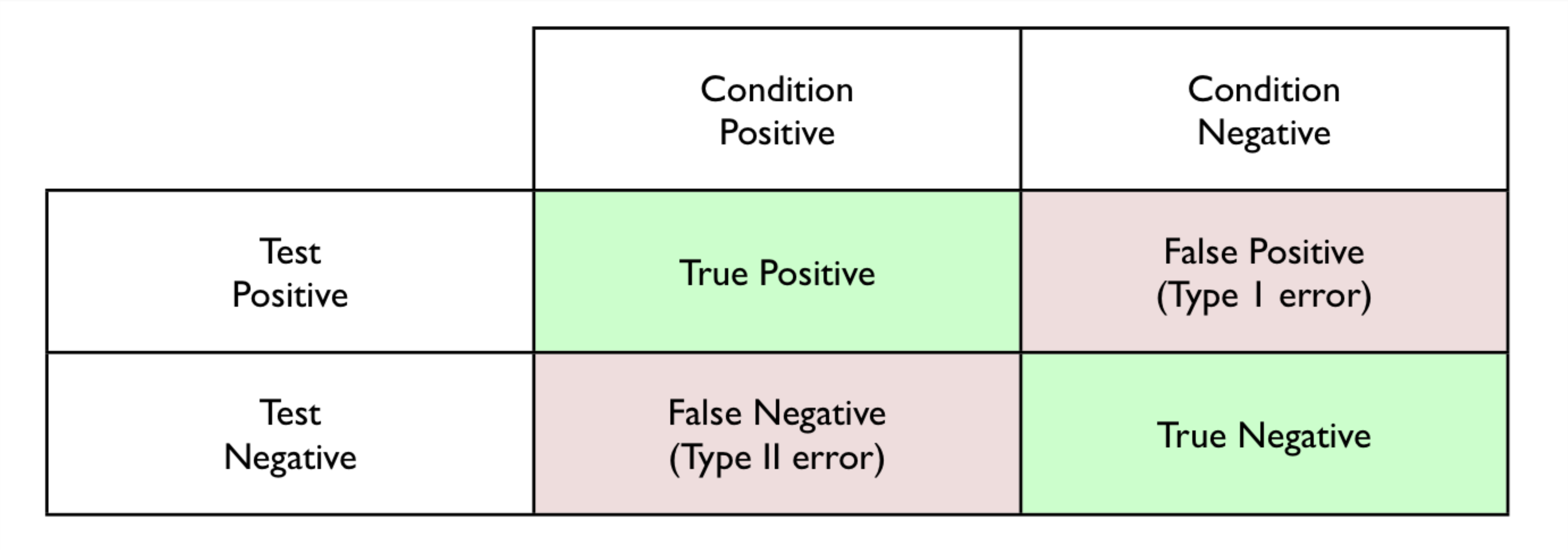
It is illustrative to think about the extreme cases - what is the sensitivity and specificity of a test that always returns a positive result?

Sensitivity and Specificity (cont.)



	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

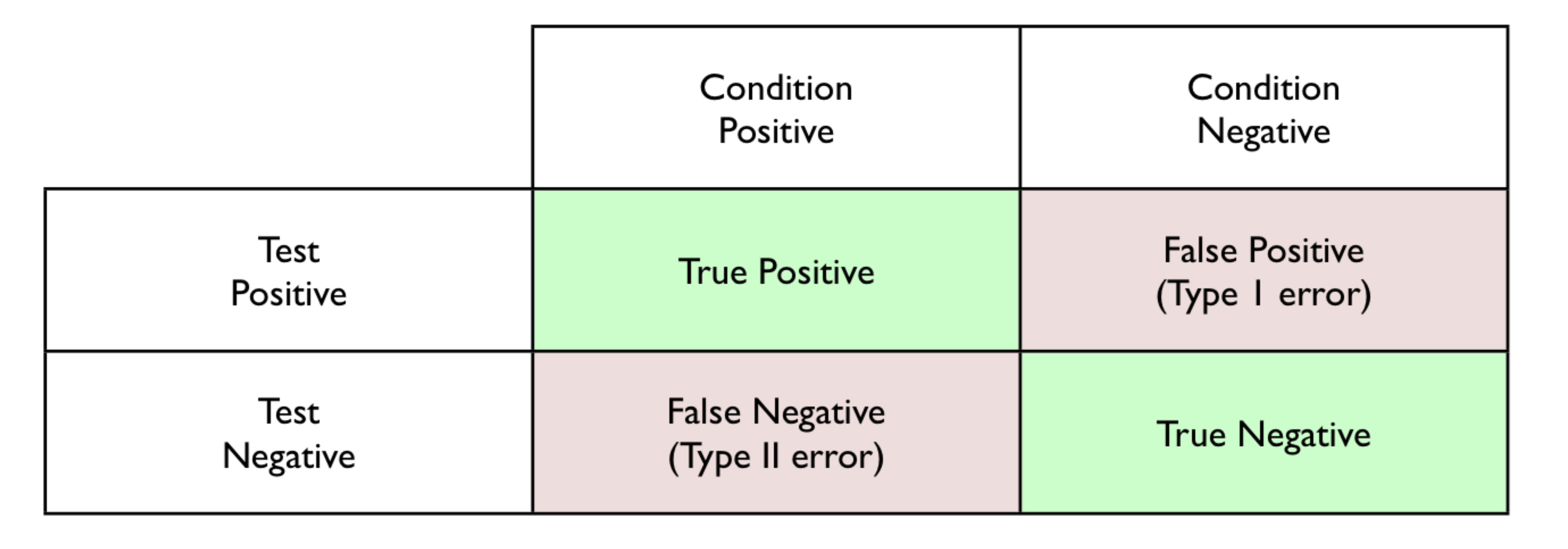
Sensitivity and Specificity (cont.)



	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +)$$

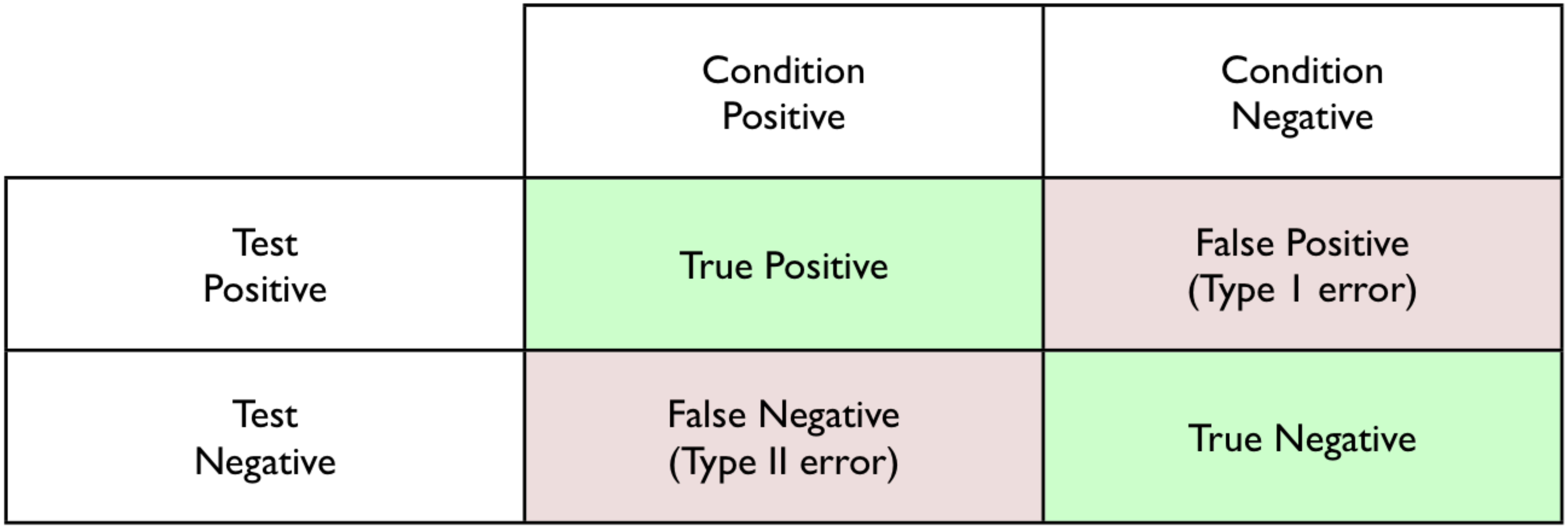
Sensitivity and Specificity (cont.)



	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

Sensitivity and Specificity (cont.)

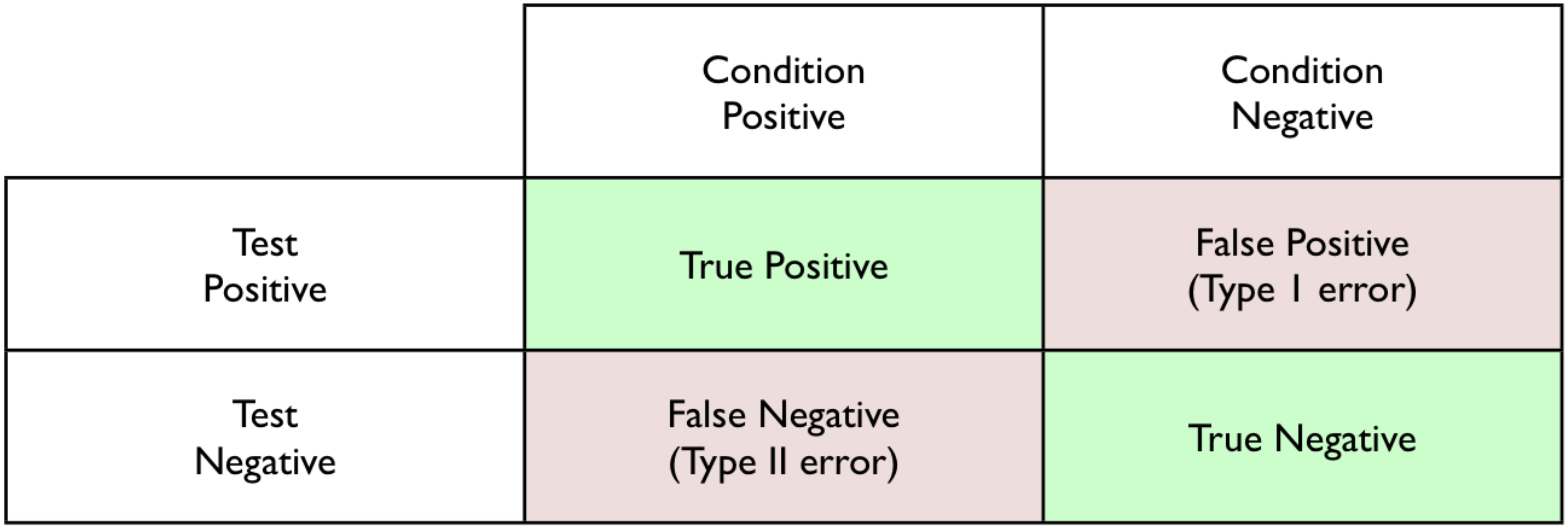


	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -)$$

Sensitivity and Specificity (cont.)

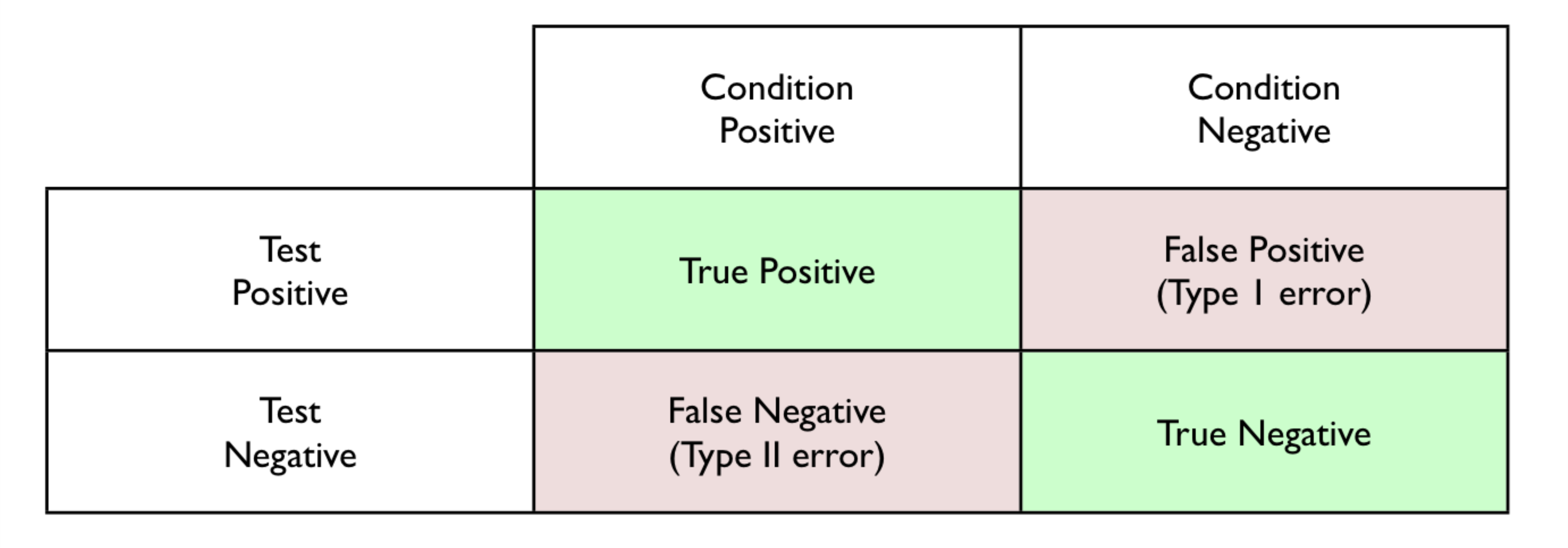


	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

Sensitivity and Specificity (cont.)



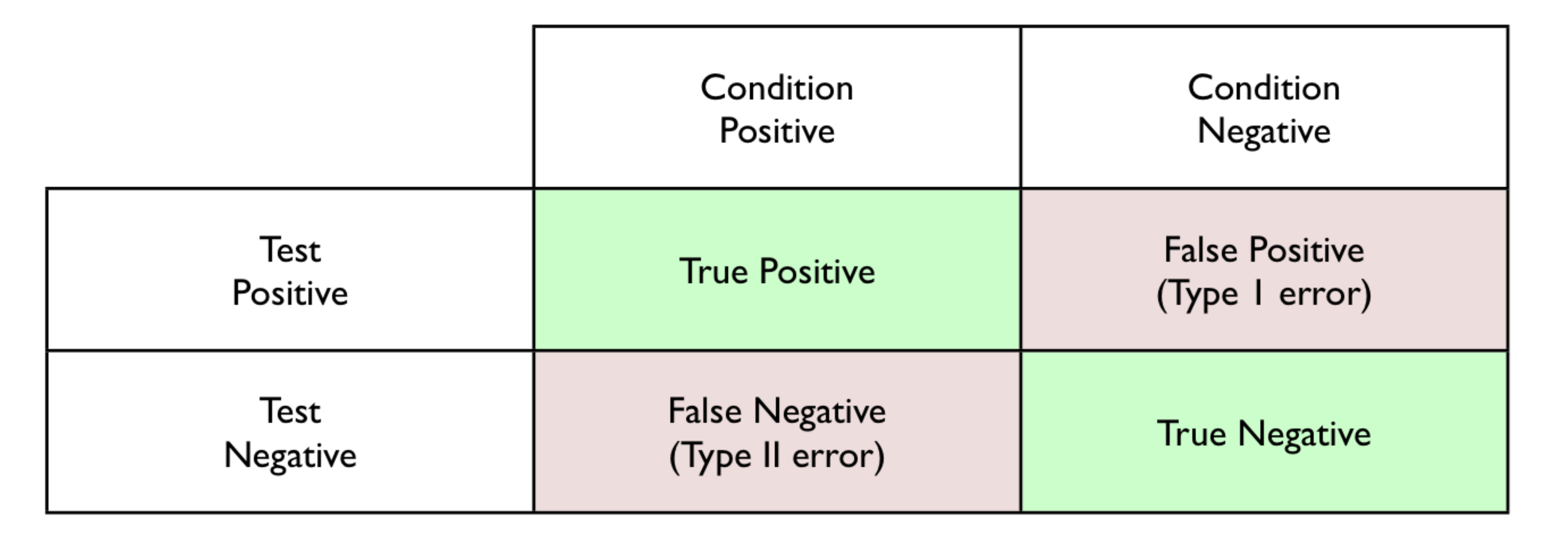
	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$Sensitivity = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$Specificity = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$False\ negative\ rate\ (\beta) = P(\text{Test } - \mid \text{Condition } +)$$

Sensitivity and Specificity (cont.)



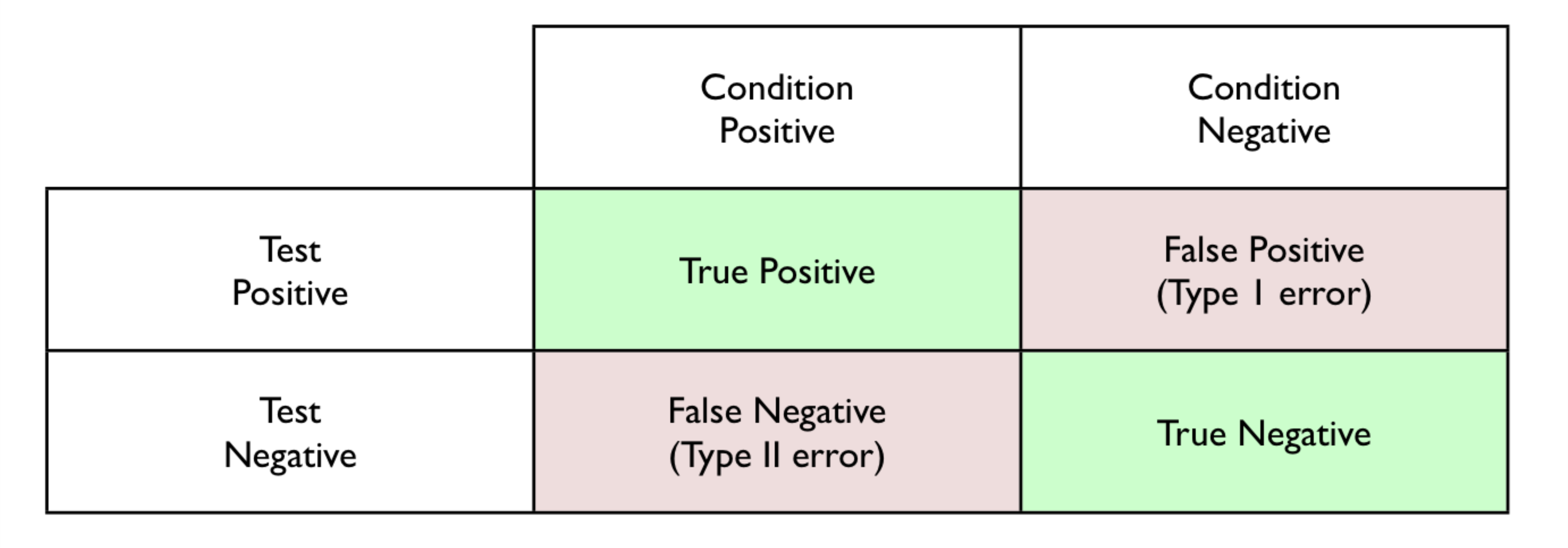
	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

Sensitivity and Specificity (cont.)



	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

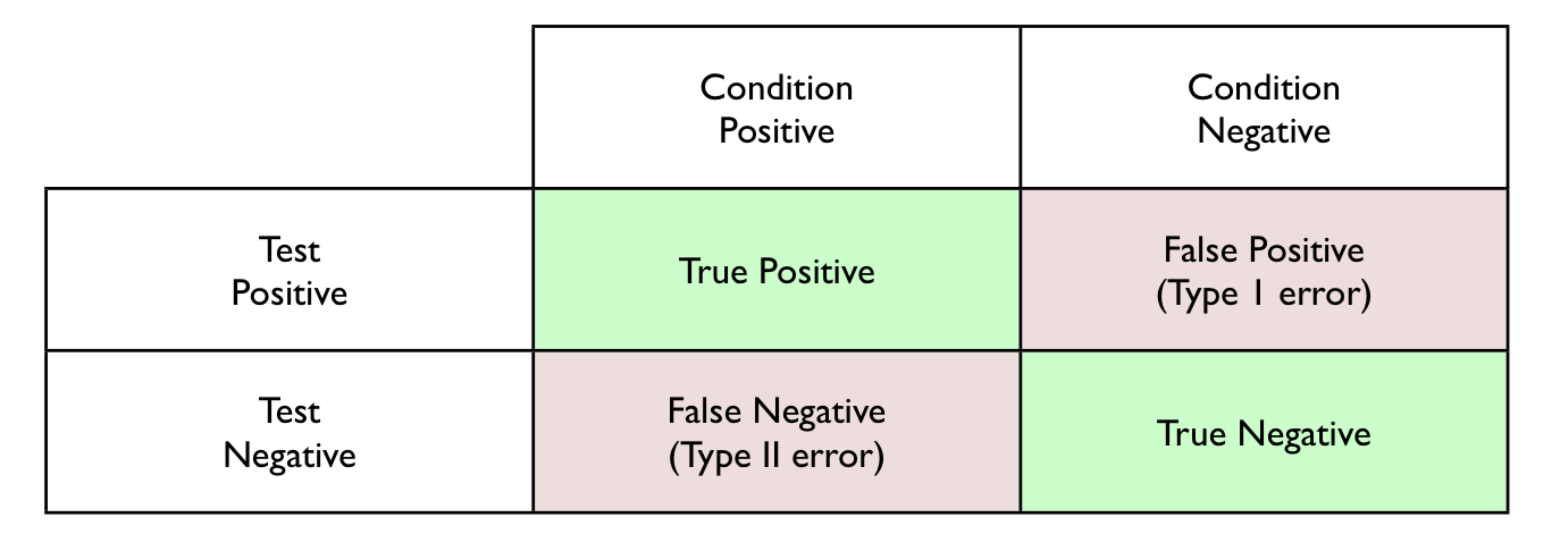
$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

$$\textit{False positive rate } (\alpha) = P(\text{Test } + \mid \text{Condition } -)$$

Sensitivity and Specificity (cont.)



	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

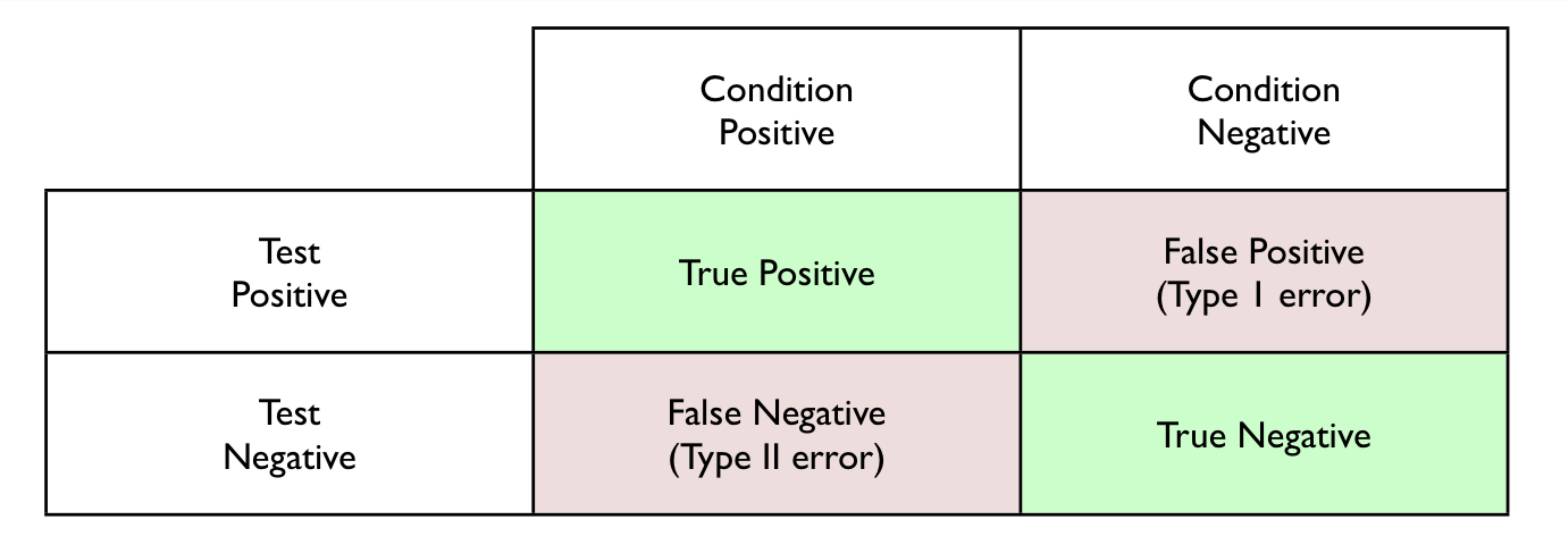
$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP+FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP+TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP+FN)$$

$$\textit{False positive rate } (\alpha) = P(\text{Test } + \mid \text{Condition } -) = FP/(FP+TN)$$

Sensitivity and Specificity (cont.)



	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

$$\textit{False positive rate } (\alpha) = P(\text{Test } + \mid \text{Condition } -) = FP/(FP + TN)$$

$$\textit{Sensitivity} = 1 - \textit{False negative rate} = \text{Power}$$

$$\textit{Specificity} = 1 - \textit{False positive rate}$$

So what?

Clearly it is important to know the Sensitivity and Specificity of a test (and or the false positive and false negative rates). Along with the incidence of the disease, e.g. $P(\text{lupus})$, these values are necessary to calculate important quantities like $P(\text{lupus}|+)$.

Additionally, our foray into power analysis after the first midterm should also give you an idea about the trade offs that are inherent in minimizing false positive and false negative rates (increasing power required either increasing α or n).

How do we use this information when we are trying to come up with a decision?

ROC curves

Back to Spam

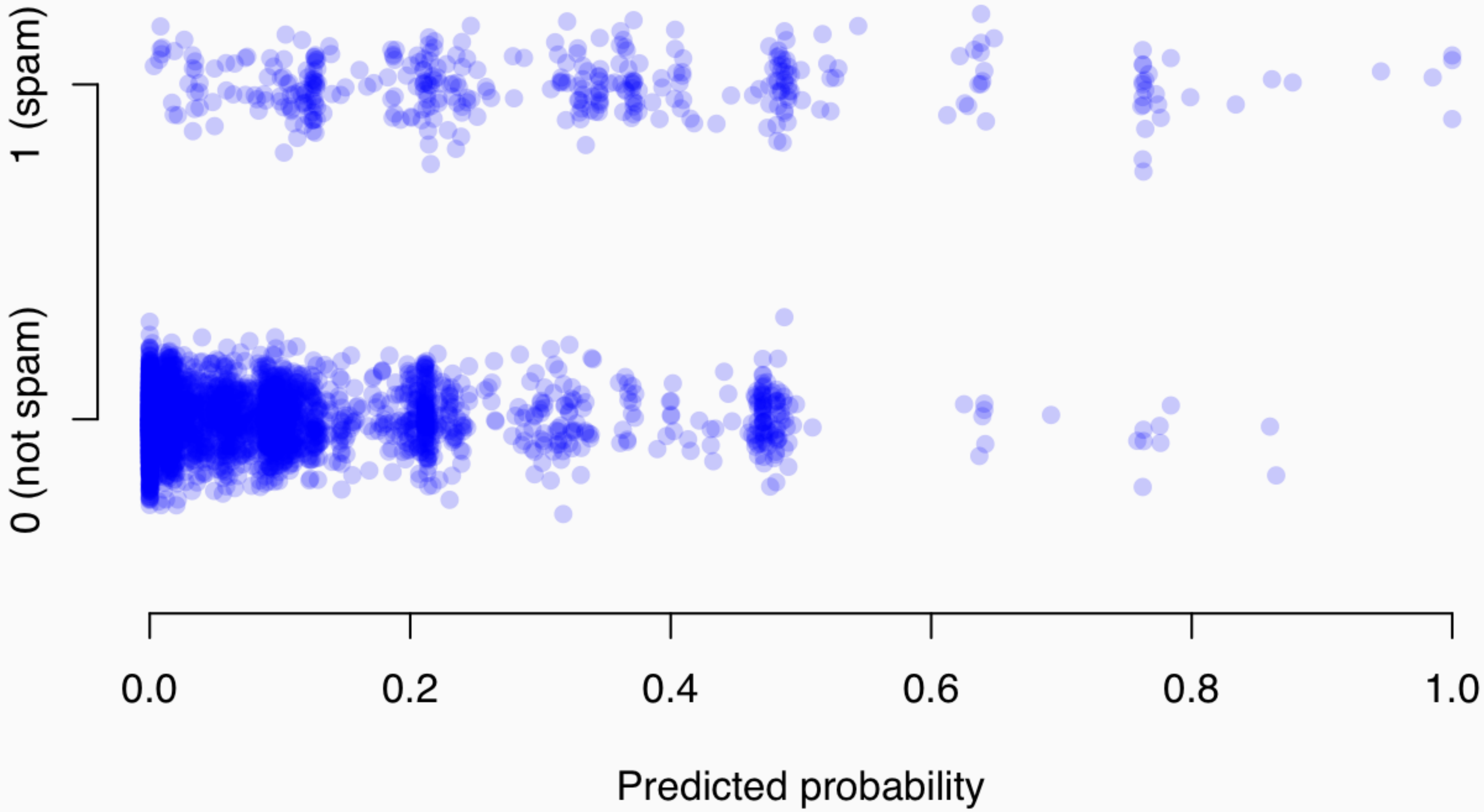
we will now examine a data set of emails where we are interested in identifying spam messages. We will examine several different logistic regression models, however these models only predict the probability an incoming message is spam. If we were designing a spam filter this would only be half of the battle, we also need to design a decision rule about which emails get flagged as spam (e.g. what probability should we use as out cutoff?)

Back to Spam

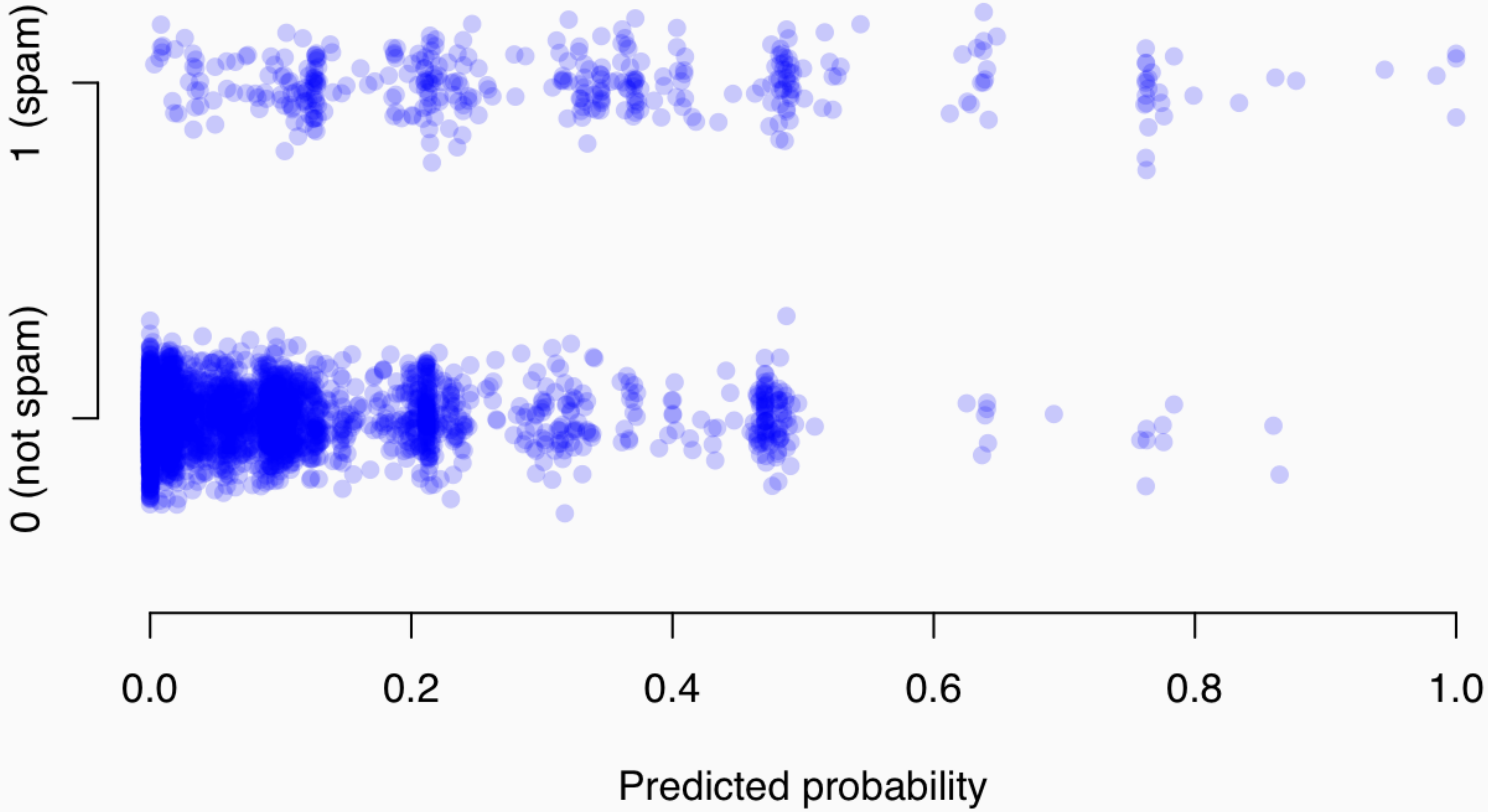
we will now examine a data set of emails where we are interested in identifying spam messages. We will examine several different logistic regression models, however these models only predict the probability an incoming message is spam. If we were designing a spam filter this would only be half of the battle, we also need to design a decision rule about which emails get flagged as spam (e.g. what probability should we use as out cutoff?)

While not the only possible solution, we will consider a simple approach where we choose a single threshold probability and any email that exceeds that probability is flagged as spam.

Picking a threshold

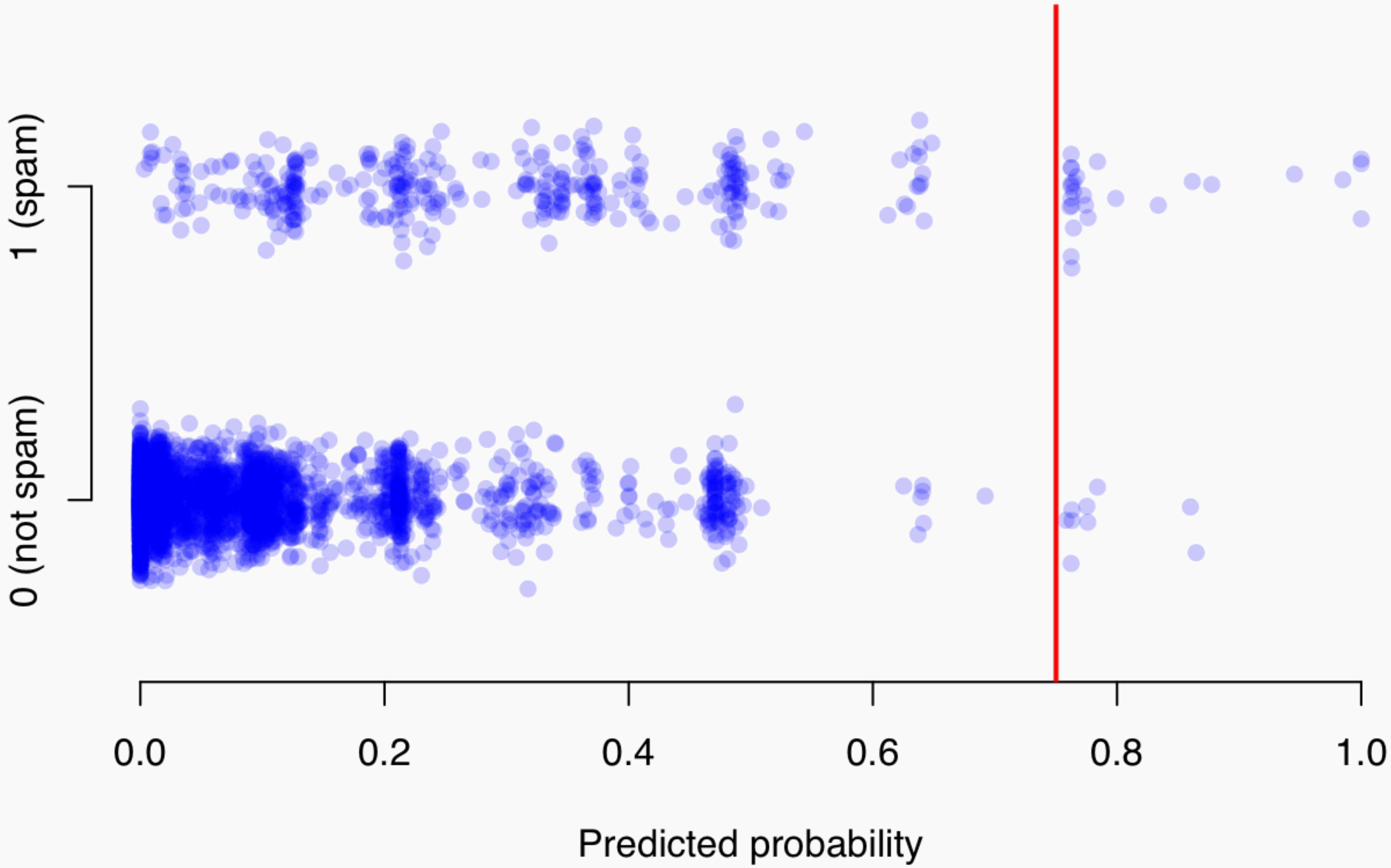


Picking a threshold



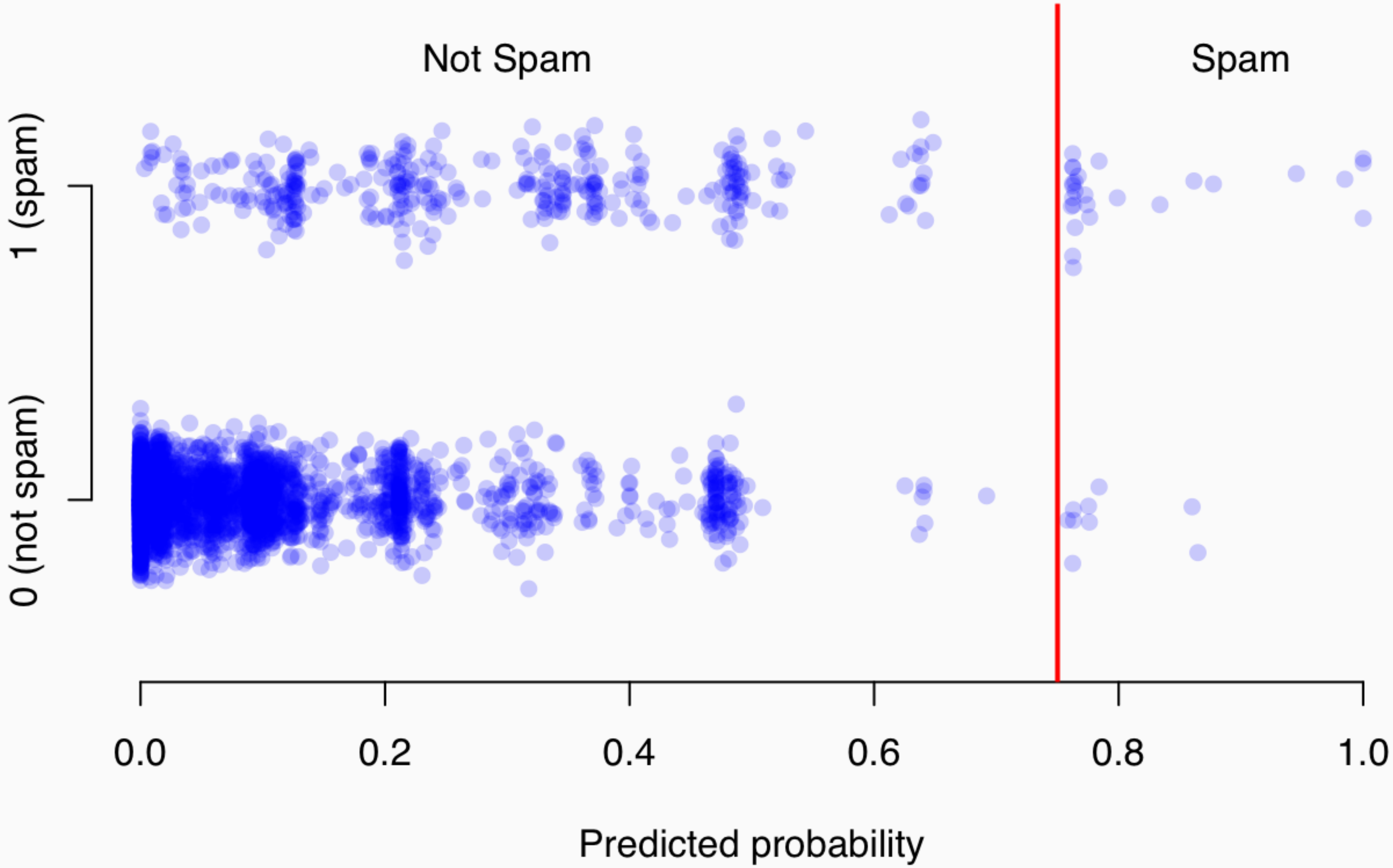
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



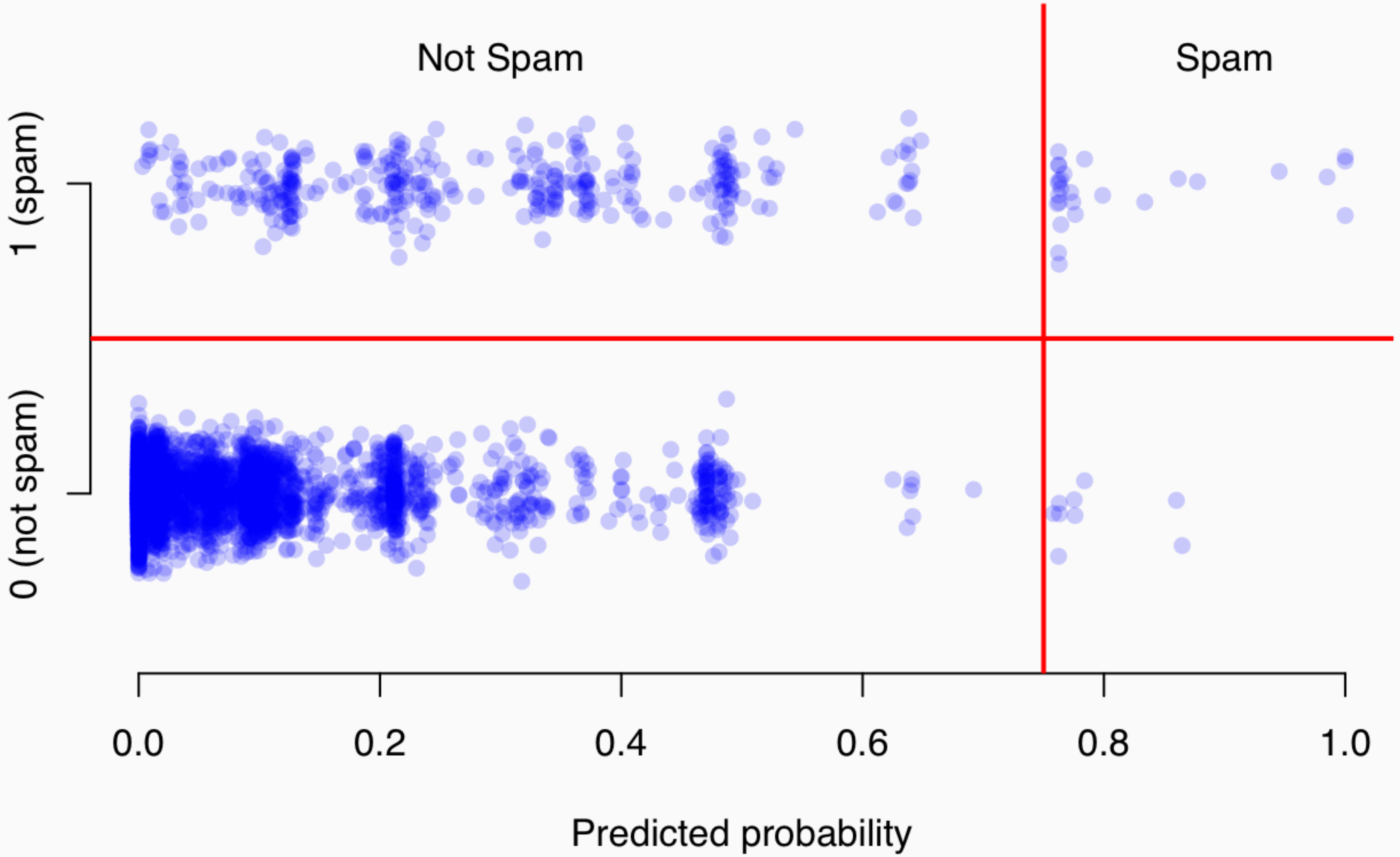
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



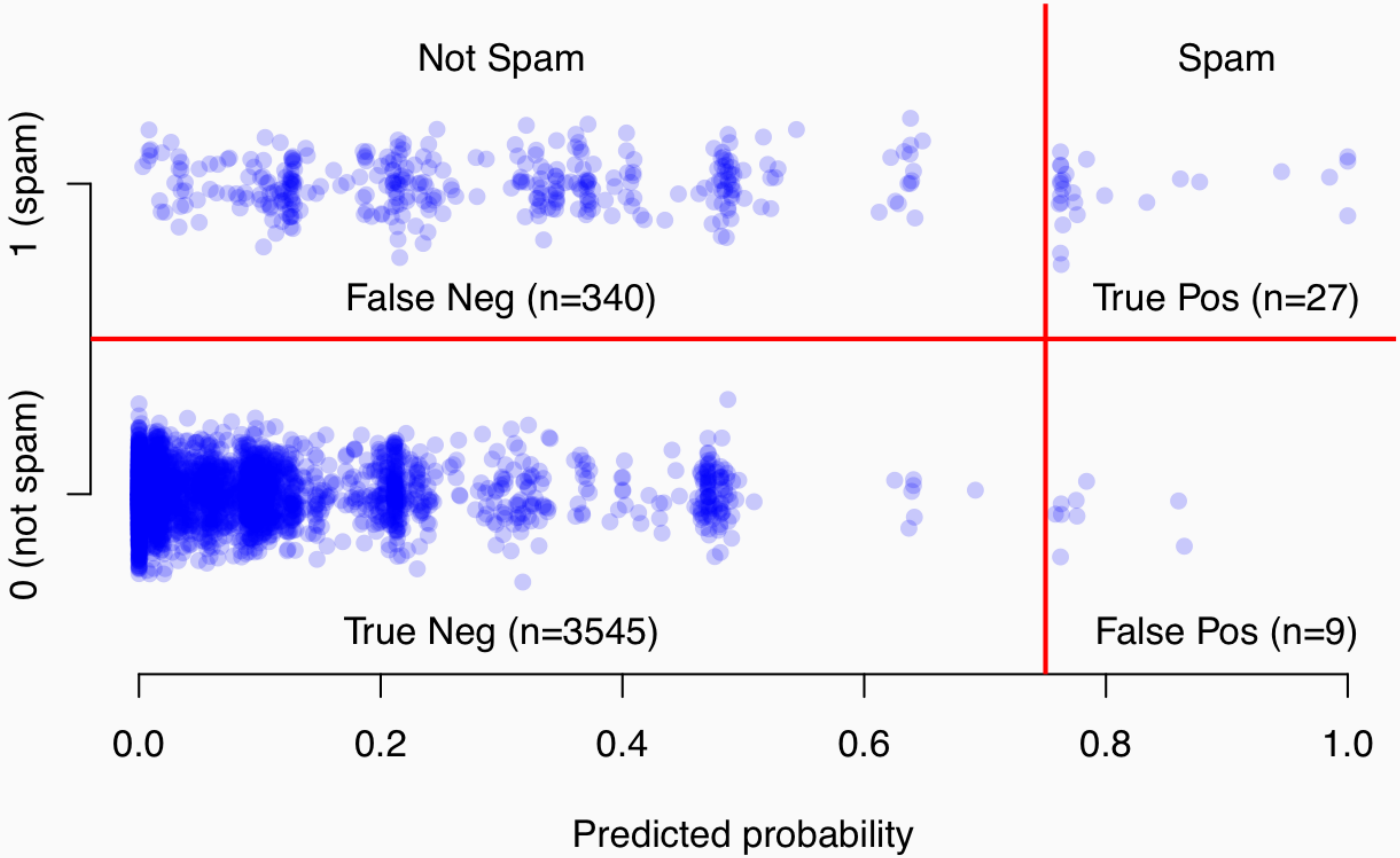
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



Lets see what happens if we pick our threshold to be *0.75*.

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

$$\begin{aligned} FN &= 340 & TP &= 27 \\ TN &= 3545 & FP &= 9 \end{aligned}$$

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$

$$TN = 3545 \qquad FP = 9$$

What are the sensitivity and specificity for this particular decision rule?

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

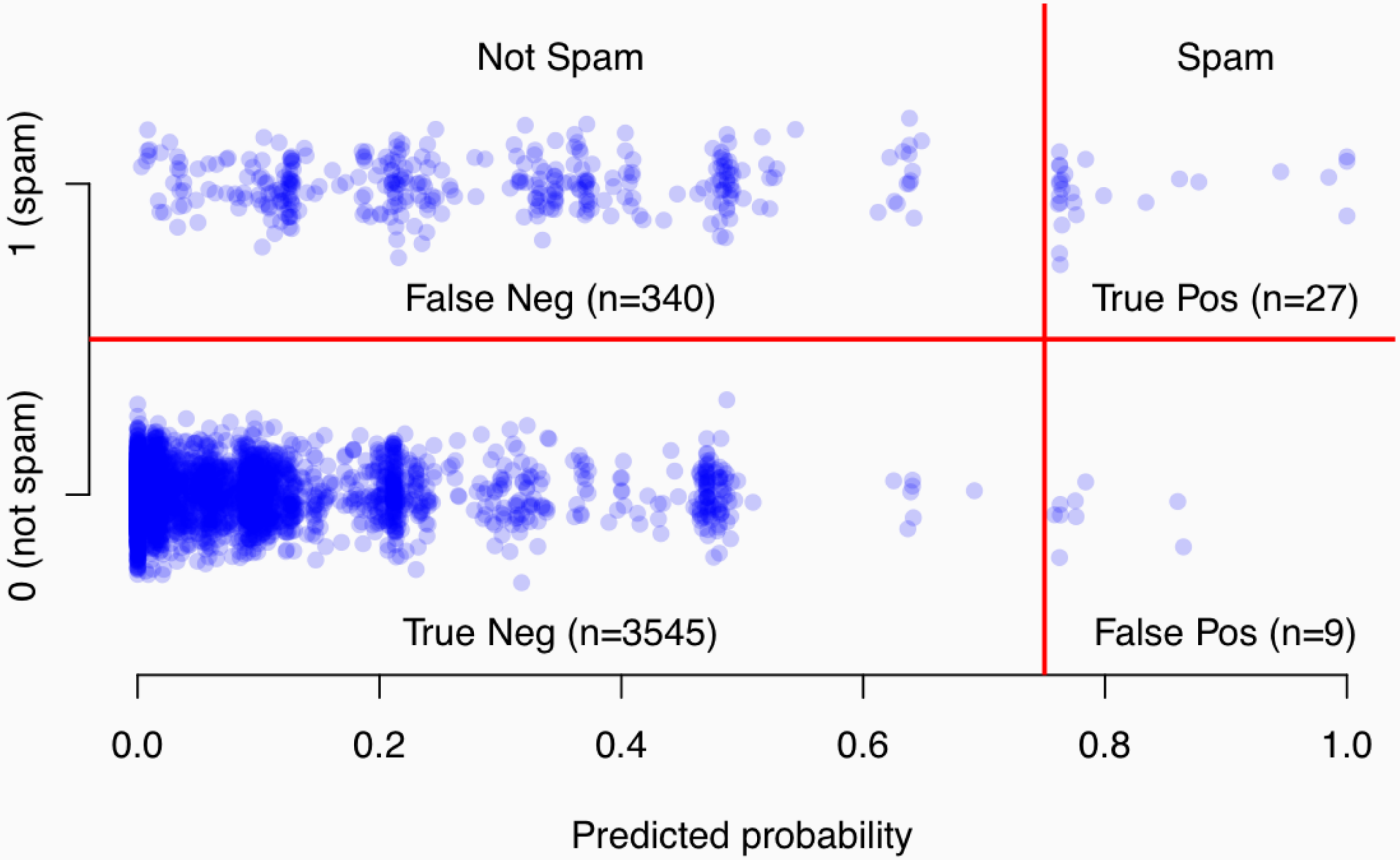
$$\begin{aligned} FN &= 340 \qquad & TP &= 27 \\ TN &= 3545 \qquad & FP &= 9 \end{aligned}$$

What are the sensitivity and specificity for this particular decision rule?

$$Sensitivity = TP/(TP + FN) = 27/(27 + 340) = 0.073$$

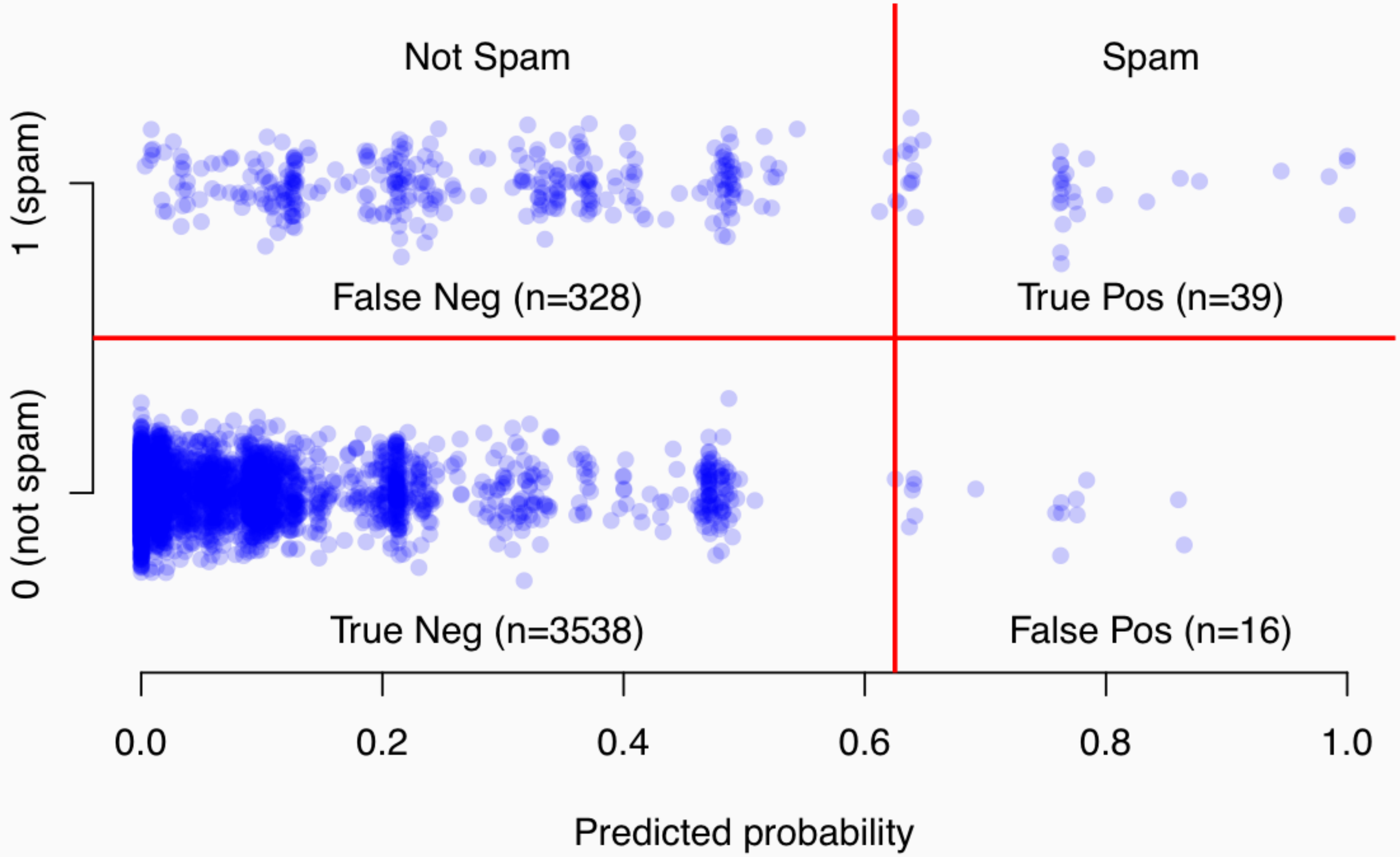
$$Specificity = TN/(FP + TN) = 3545/(9 + 3545) = 0.997$$

Trying other thresholds



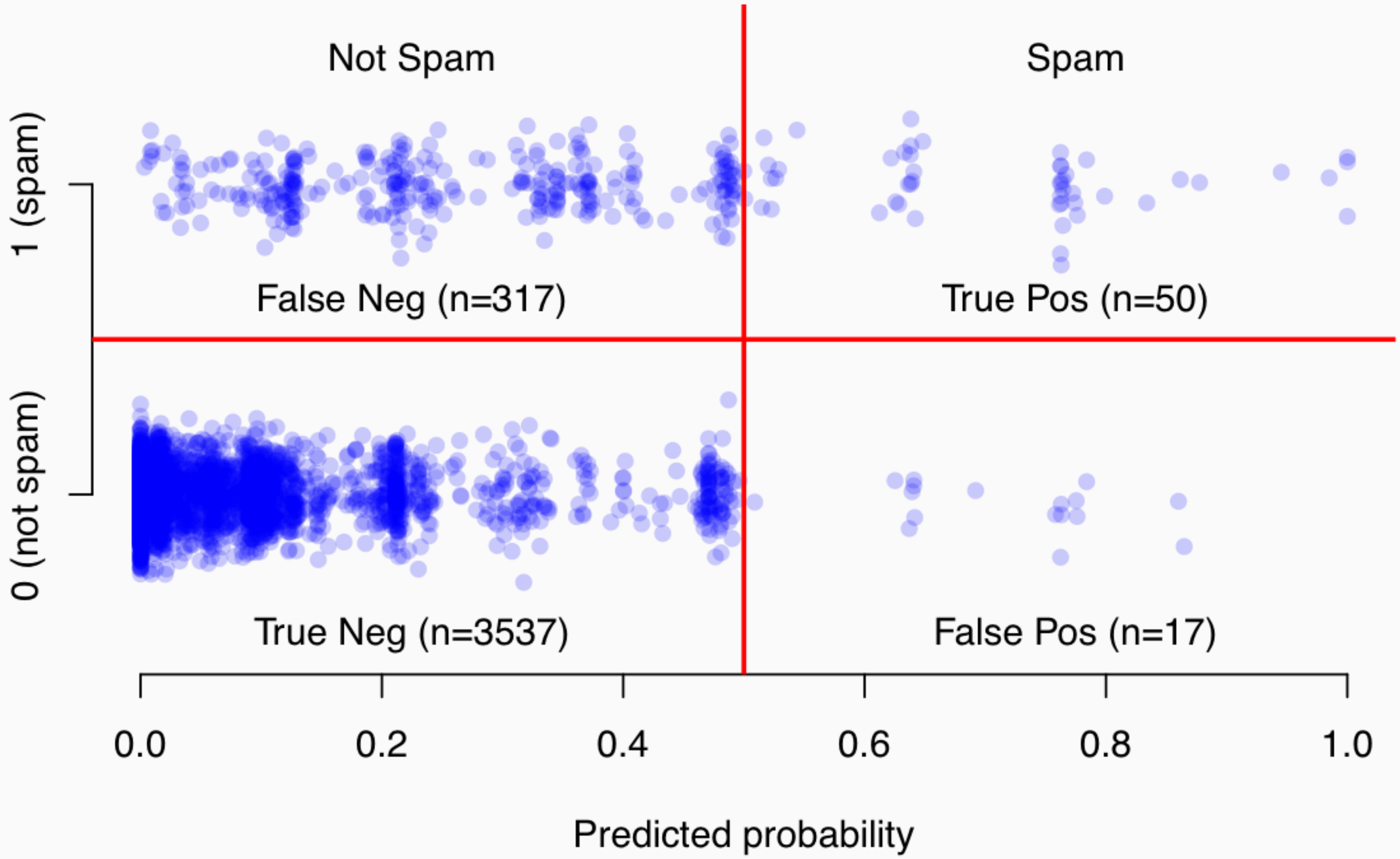
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074				
Specificity	0.997				

Trying other thresholds



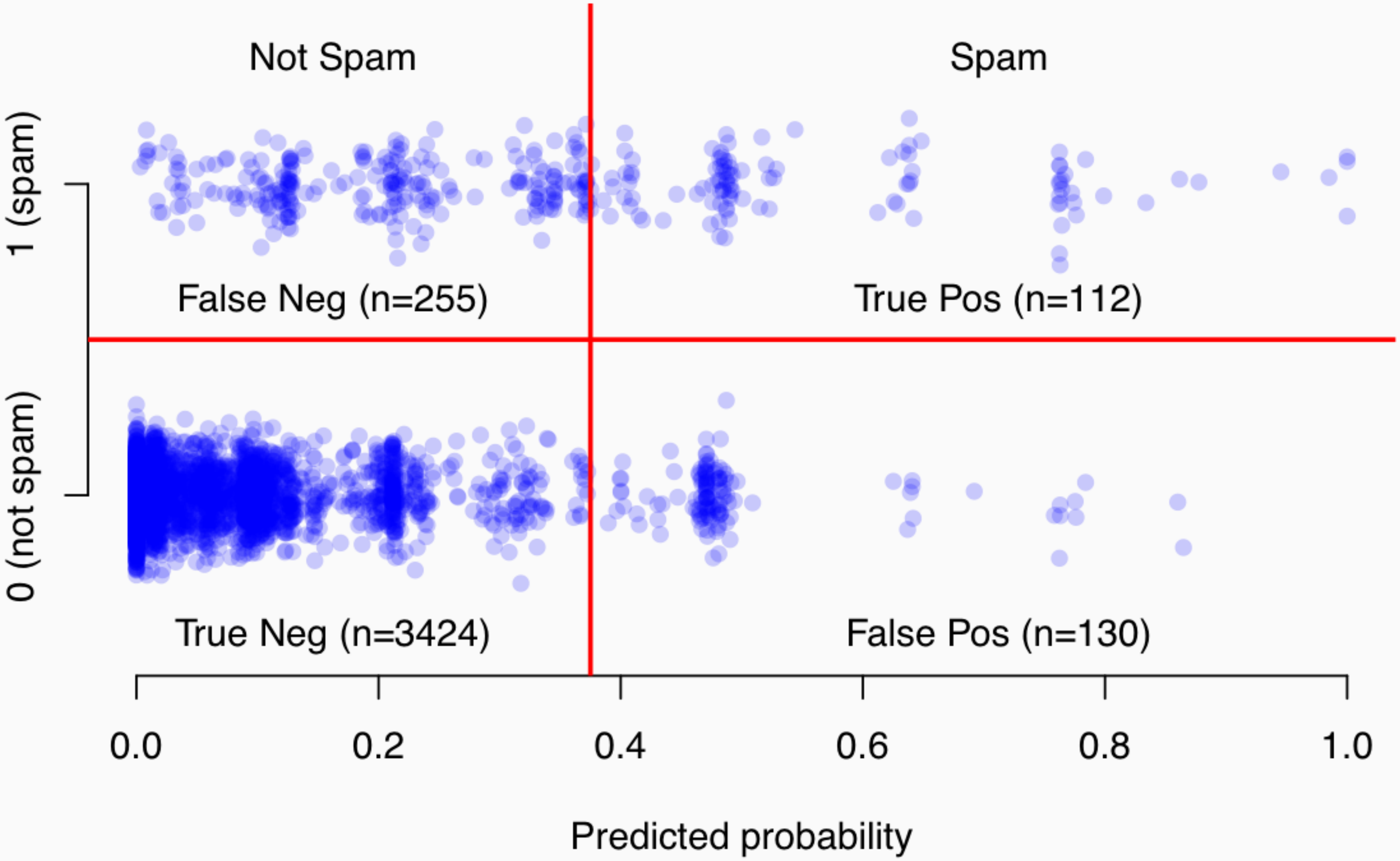
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



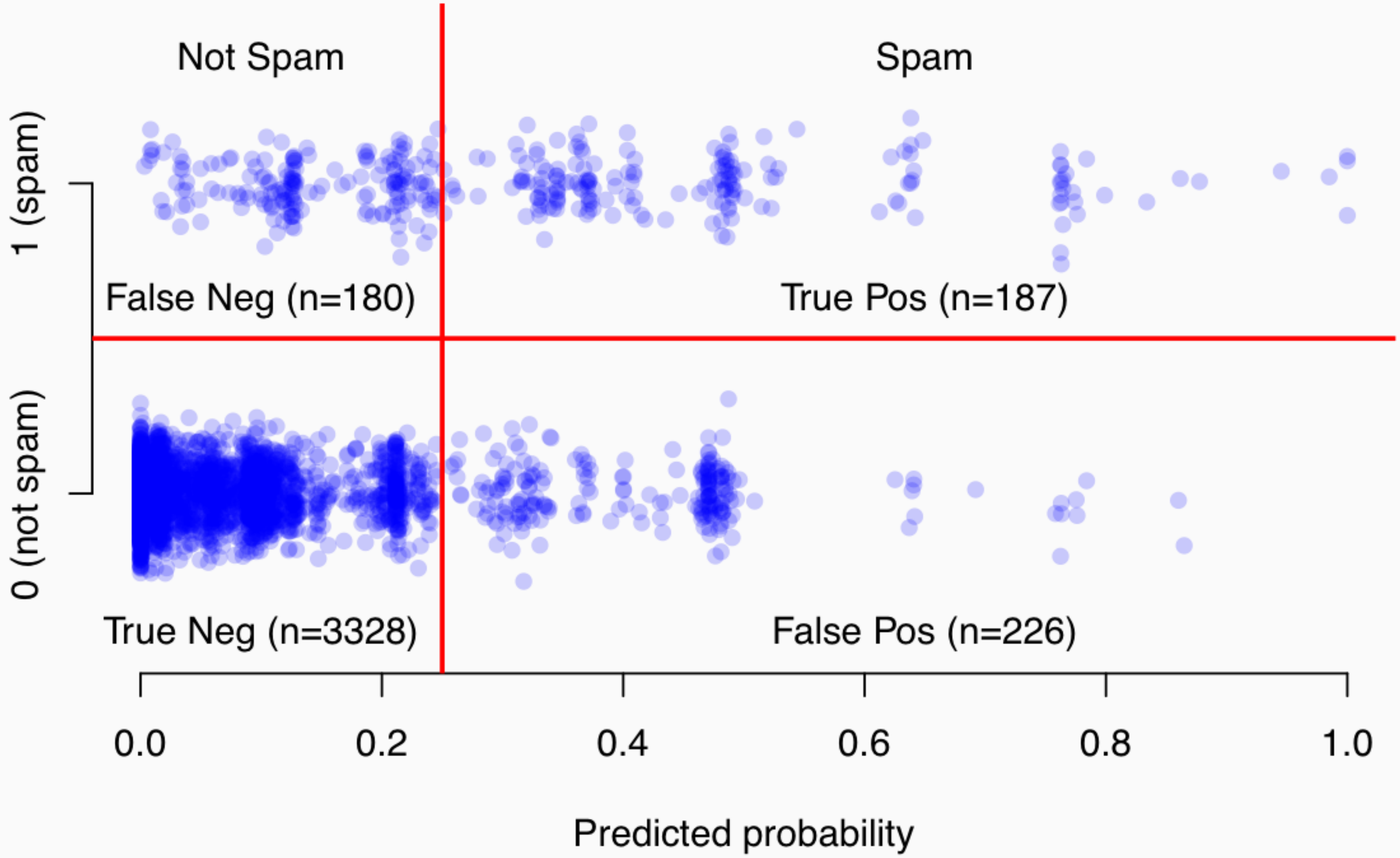
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	
Specificity	0.997	0.995	0.995	0.963	

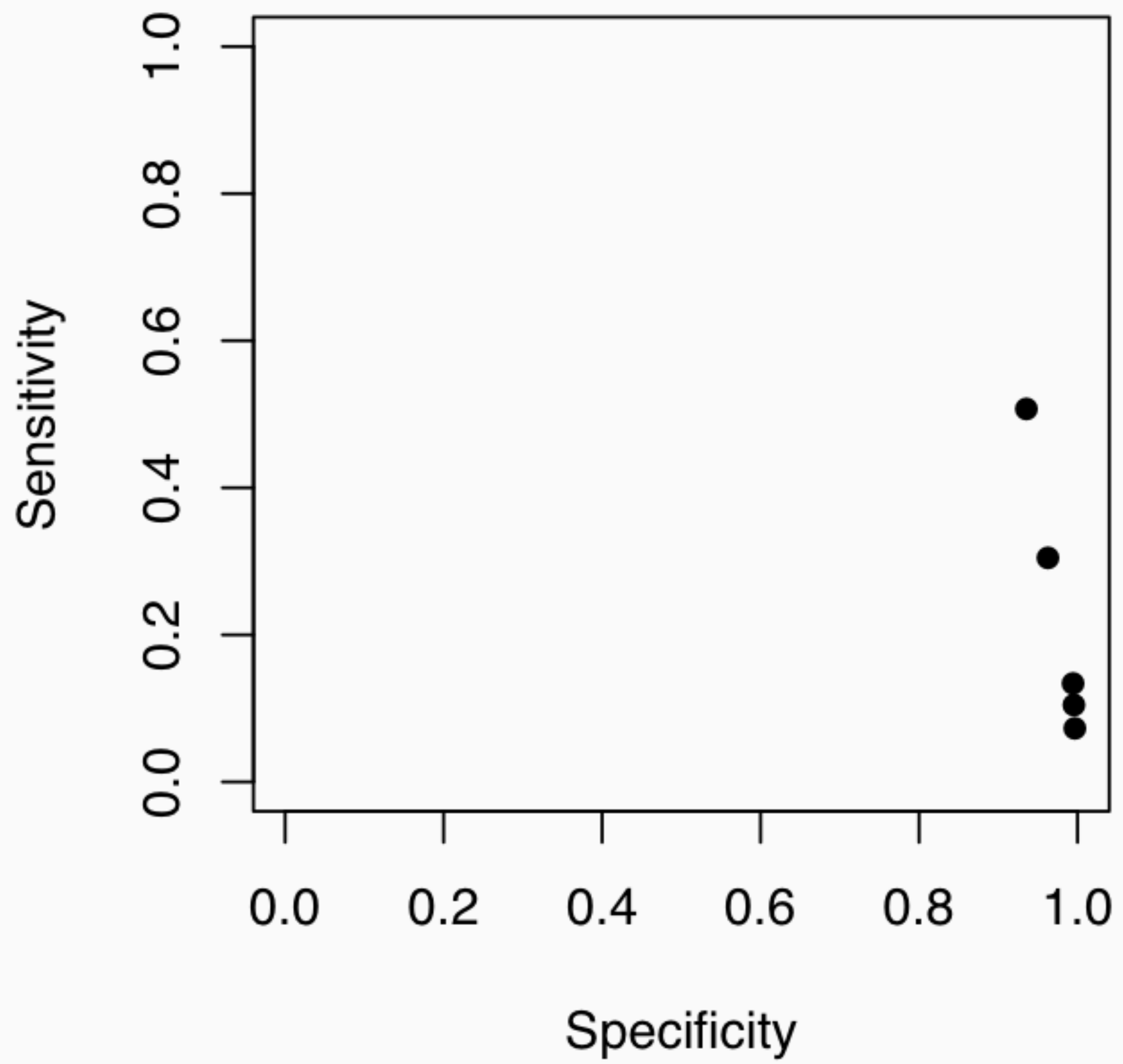
Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

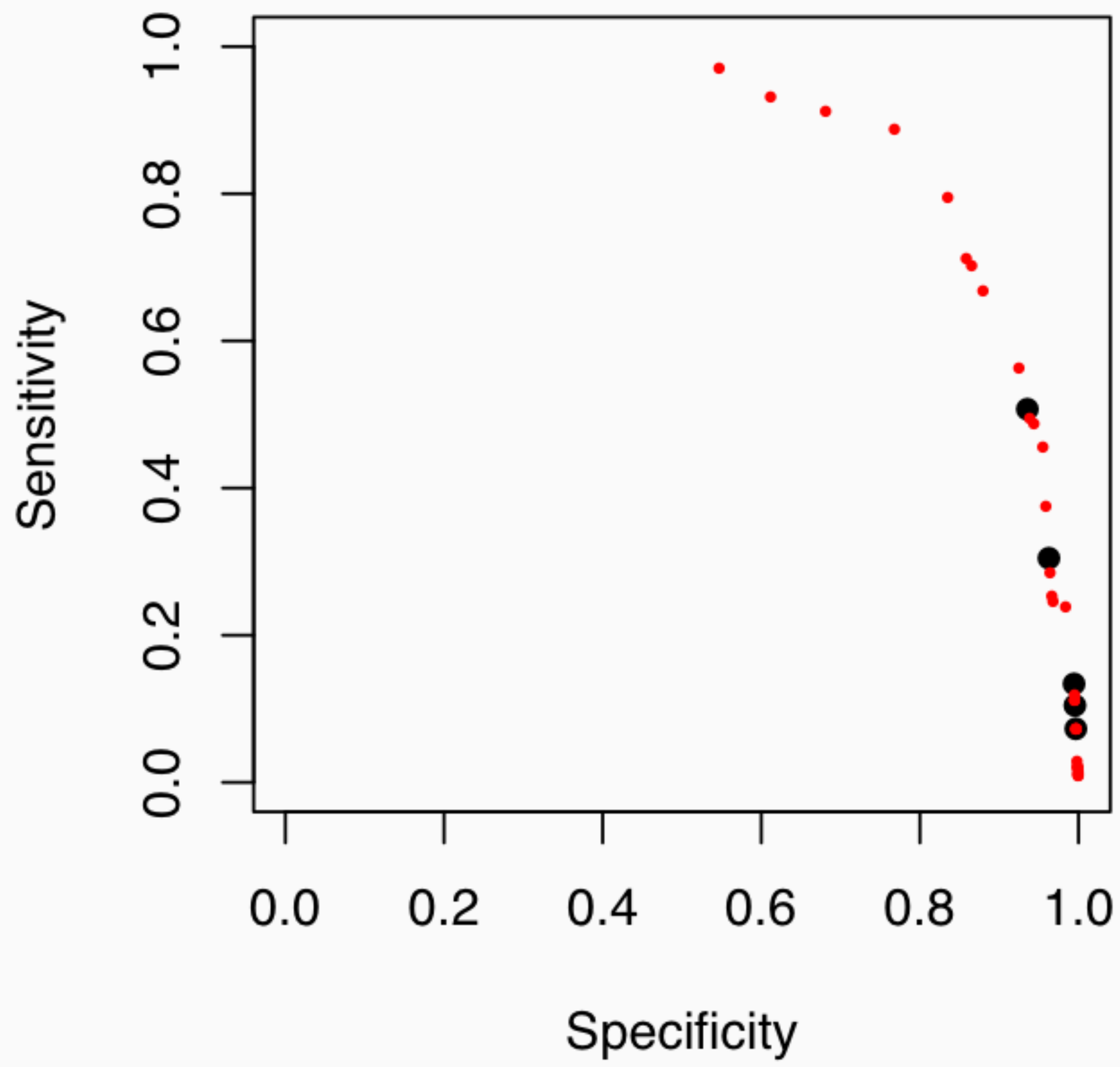
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



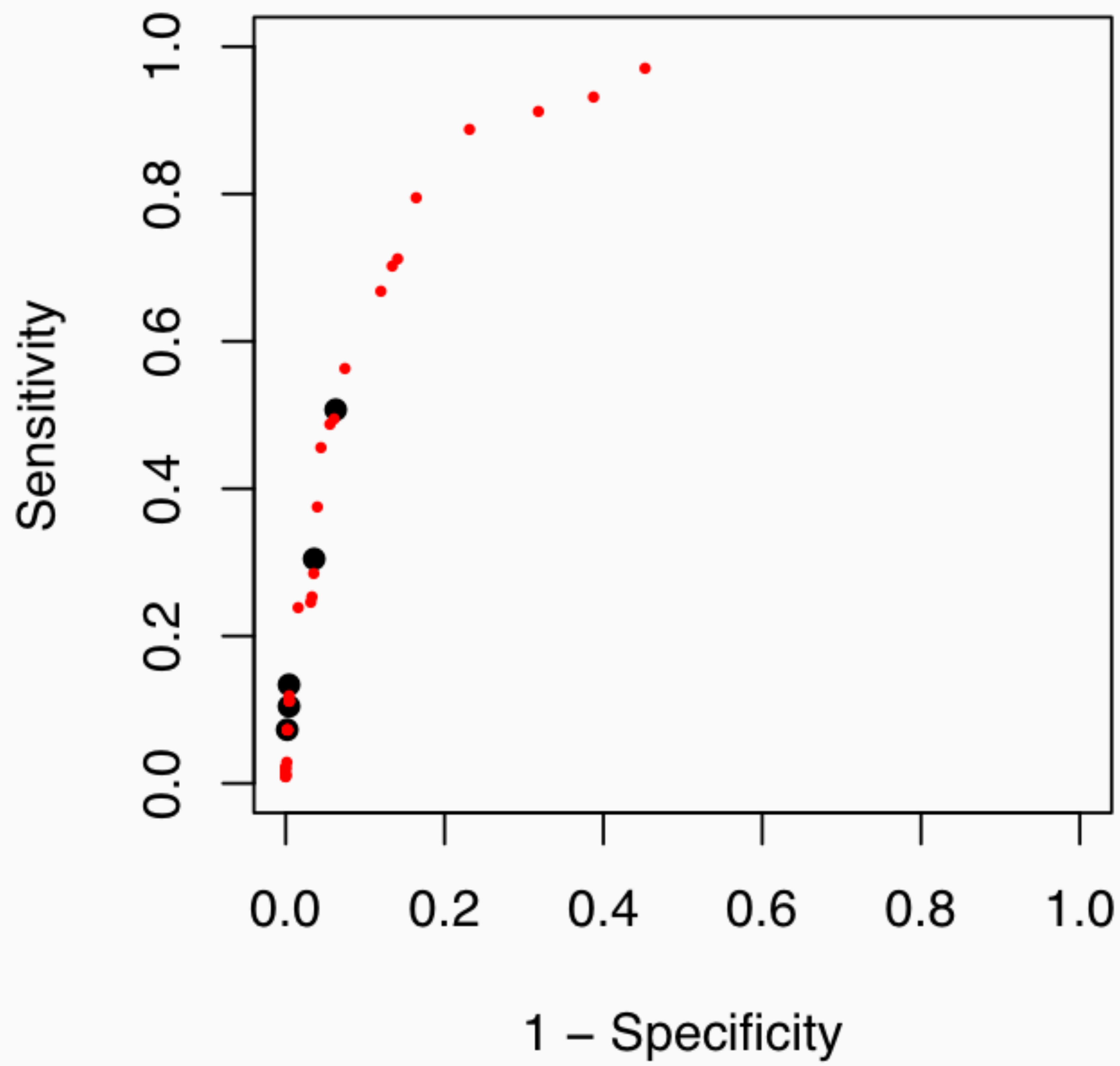
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

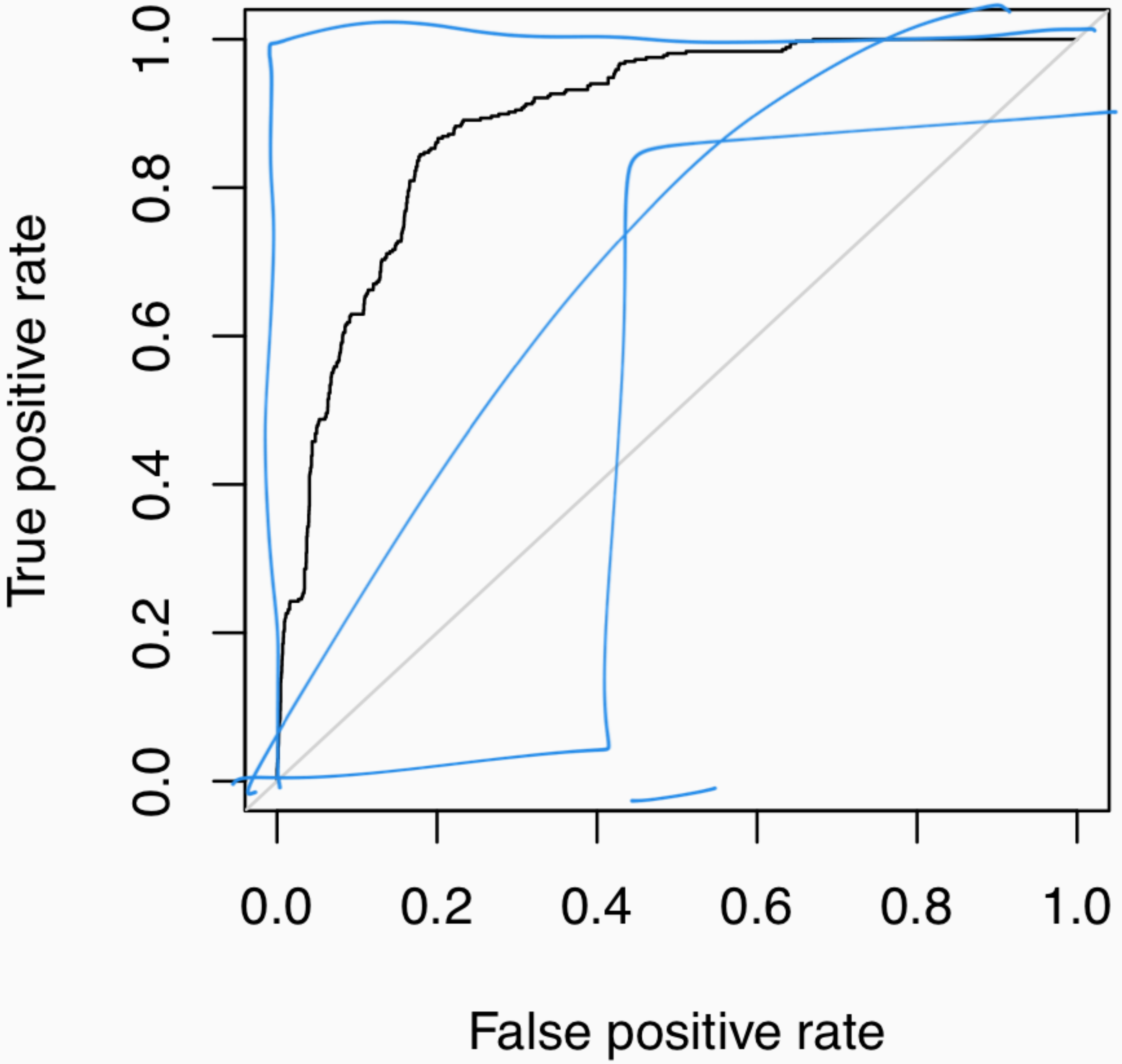


Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



Receiver operating characteristic (ROC) curve



Receiver operating characteristic (ROC) curve (cont.)

Why do we care about ROC curves?

- Shows the trade off in sensitivity and specificity for all possible thresholds.
- Straight forward to compare performance vs. chance.
- Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.

Refining the Spam model

```
refined = glm(spam ~ to_multiple+cc+image+attach+winner
                    +password+line_breaks+format+re_subj
                    +urgent_subj+exclaim_mess,
              data=email, family=binomial)
summary(refined)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7594	0.1177	-14.94	0.0000
to_multipleyes	-2.7368	0.3156	-8.67	0.0000
ccyes	-0.5358	0.3143	-1.71	0.0882
imageyes	-1.8585	0.7701	-2.41	0.0158
attachyes	1.2002	0.2391	5.02	0.0000
winneryes	2.0433	0.3528	5.79	0.0000
passwordyes	-1.5618	0.5354	-2.92	0.0035
line_breaks	-0.0031	0.0005	-6.33	0.0000
formatPlain	1.0130	0.1380	7.34	0.0000
re_subjyes	-2.9935	0.3778	-7.92	0.0000
urgent_subjyes	3.8830	1.0054	3.86	0.0001
exclaim_mess	0.0093	0.0016	5.71	0.0000

Comparing models