Conditional Probability: Estimating the Probability of Disease

BIOS 6611

CU Anschutz

Week 5

Background and Example Refreshers

The Probability of Disease

Suppose a patient arrives to have a diagnostic test administered. Our prior probability that the patient has disease is the prevalence of the disease, $P(D)$. The result of the test should alter this to $P(D|T)$ if the test is positive or to $P(D|\bar{T})$ if the test is negative.

The effectiveness of diagnostic tests is described by sensitivity, $P(T|D)$, and specificity, $P(\bar{T}|\bar{D})$. *Recall, these do not depend on the prevalence of the disease* and are usually estimated from studies with a large number of persons with and without the disease.

Bayes' theorem shows how to relate them to predictive values using the actual population prevalence.

Bayes' Theorem Refresher

We will apply Bayes' Theorem to calculate the *posterior probability* of an event based on some prior probability by utilizing conditional probabilities.

The theorem shows how to take prior probabilities (e.g., assumed prevalence of disease), incorporate new information (e.g., diagnostic test results), and obtain revised (posterior) probabilities (e.g., predictive values).

In our context of diagnostic testing we have two random variables: D_i represents i mutually exclusive and exhaustive disease states ($i = 1, ..., k$) and T represents a positive test or presence of a symptom. Bayes' theorem states:

$$P(D_i|T) = \frac{P(T \cap D_i)}{P(T)} = \frac{P(T|D_i)P(D_i)}{\sum_{i=1}^{k} P(T|D_i)P(D_i)}$$

Example Study Results

We have enrolled 50 participants with CHD and 50 without CHD (i.e., we have fixed the GS numbers in our study) and want to evaluate the treadmill test as a less invasive and cheaper test to administer over the angiogram in a population where $P(\text{CHD}) = 0.2$:

Treadmill Test	**Angiogram (GS)**		**Total**
	Positive (D)	**Negative** ($\bar{D}$)	
Positive (T)	40	5	45
Negative ($\bar{T}$)	10	45	55
Total	50	50	100

Predictive Values

Predictive Values

The *predictive values* of a test depend on sensitivity, specificity, and the underlying prevalence (i.e., prior probability) of having the disease.

Through Bayes' rule we can relate sensitivity, specificity, and the underlying prevalence of disease to the positive and negative predictive values (PPV and NPV) using the actual population prevalence.

Note: PPV and NPV are ***very*** dependent on prevalence. Using study prevalence is not appropriate when calculating their values, unless it's from a large prospective study from which prevalence can be well estimated.

Positive Predictive Value

Positive Predictive Value (PPV): the probability of disease given a positive test result:

$$
\begin{aligned}
\text{PPV} &= P(D|T) \\
&= \frac{P(D \cap T)}{P(T)} \\
&= \frac{P(T|D)P(D)}{P(T \cap D) + P(T \cap \bar{D})} \\
&= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} \\
&= \frac{\text{(sensitivity)(prior probability)}}{\text{(sensitivity)(prior prob)} + \text{(1-specificity)(1-prior prob)}}
\end{aligned}
$$

PPV Example

Using our previous results that the sensitivity of the treadmill test is 0.8, the specificity is 0.9, and $P(\text{CHD}) = 0.2$, our PPV is:

$$PPV = P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} =$$

Interpretation: If the treadmill test is positive, there is a ____% chance that an individual in the population with 20% prevalence of CHD actually has CHD.

Negative Predictive Value

Negative Predictive Value (NPV): the probability of not having a disease given a negative test result:

$$
\begin{aligned}
\text{NPV} &= P(\bar{D}|\bar{T}) \\
&= \frac{P(\bar{D} \cap \bar{T})}{P(\bar{T})} \\
&= \frac{P(\bar{T}|\bar{D})P(\bar{D})}{P(\bar{T} \cap D) + P(\bar{T} \cap \bar{D})} \\
&= \frac{P(\bar{T}|\bar{D})P(\bar{D})}{P(\bar{T}|D)P(D) + P(\bar{T}|\bar{D})P(\bar{D})} \\
&= \frac{\text{(specificity)(1-prior probability)}}{\text{(1-sensitivity)(prior prob)} + \text{(specificity)(1-prior prob)}}
\end{aligned}
$$

NPV Example

Using our previous results that the sensitivity of the treadmill test is 0.8, the specificity is 0.9, and $P(\text{CHD}) = 0.2$, our PPV is:

$$NPV = P(\bar{D}|\bar{T}) = \frac{P(\bar{T}|\bar{D})P(\bar{D})}{P(\bar{T}|D)P(D) + P(\bar{T}|\bar{D})P(\bar{D})} =$$

Interpretation: If the treadmill test is negative, there is a ____% chance that an individual in the population with 20% prevalence of CHD actually does not have CHD.

Other Summaries

Likelihood Ratios

We may wish to answer the question, *is the treadmill test a good test?*

Likelihood ratios (aka Bayes factors) are one useful metric. They are defined as the positive likelihood ratio (LR+) and negative likelihood ratio (LR-):

$$\text{LR+} = \frac{P(T|D)}{P(T|\bar{D})} = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{\text{TPR}}{\text{FPR}}$$

$$\text{LR-} = \frac{P(\bar{T}|D)}{P(\bar{T}|\bar{D})} = \frac{1 - \text{sensitivity}}{\text{specificity}} = \frac{\text{FNR}}{\text{TNR}}$$

LR+ is the number of true positive results per false positive result. Large ratios are desirable and it serves as a prevalence-free measure of the strength of a positive test.

LR- is the number of false negative results per true negative result. Small ratios are desirable and it serves as a prevalence-free measure of the strength of a negative test.

LR+ Example

Using our previous results that the sensitivity of the treadmill test is 0.8, the specificity is 0.9, and $P(\text{CHD}) = 0.2$, the positive likelihood ratio for the treadmill test is:

$$\text{LR+} = \frac{\text{sensitivity}}{1 - \text{specificity}} =$$

Interpretation: There are ____ true positives per one false positive treadmill test result.

LR- Example

Using our previous results that the sensitivity of the treadmill test is 0.8, the specificity is 0.9, and $P(\text{CHD}) = 0.2$, the negative likelihood ratio for the treadmill test is:

$$\text{LR-} = \frac{1 - \text{sensitivity}}{\text{specificity}} =$$

Interpretation: There are ____ false negatives per one true negative treadmill test result.

Alternative Interpretation: There are ____ false negatives per ____ true negative treadmill test results.

Posterior Odds - I

Using the likelihood ratio of a test we are also able to take advantage of Bayes' theorem to calculate the **posterior odds** of disease (or no disease).

Let $O(D)$ represent the prior odds of having the disease. From our prior probability (i.e., prevalence) of having the disease, $P(D)$, we can calculate the prior odds:

$$O(D) = \frac{P(D)}{1 - P(D)}$$

This can also be solved to calculate the probability if we were given the odds:

$$P(D) = \frac{O(D)}{1 + O(D)}$$

Similarly, the prior odds for $\bar{D}$ are calculated as

$$O(\bar{D}) = \frac{P(\bar{D})}{1 - P(\bar{D})} = \frac{1 - P(D)}{P(D)}$$

Posterior Odds - II

In general, the calculation of the posterior in Bayesian analysis is (posterior)=(prior)(likelihood).

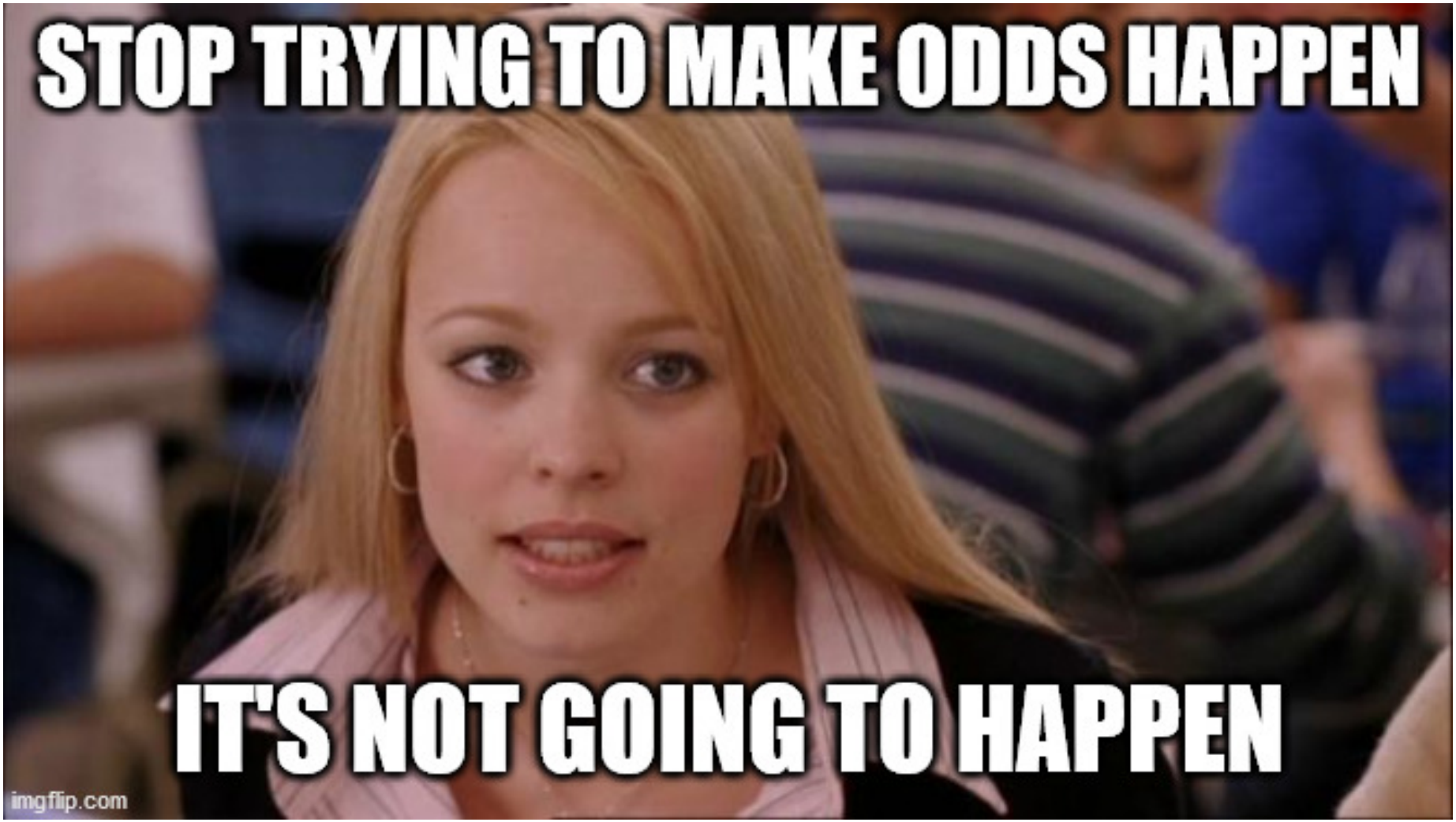
Therefore, in our context, the **posterior odds** of having (or not having) the disease are:

$$\text{Posterior odds of } D = \text{Prior odds of } D \times \text{LR+} = \frac{P(D)}{1-P(D)} \times \text{LR+}$$

$$\text{Posterior odds of } \bar{D} = \text{Prior odds of } \bar{D} \times (\text{LR-})^{-1} = \frac{1-P(D)}{P(D)} \times \frac{1}{\text{LR-}}$$

The interpretation of the posterior odds of D $[\bar{D}]$ are that, after observing the test results for an individual, we are X:1 in favor of D $[\bar{D}]$ (i.e., in favor of having [not having] the disease), where X is our posterior odds.

Posterior Probability



Posterior Probability

Odds range from 0 to ∞ and represent a ratio of outcomes.

Probabilities range from 0 to 1 and, depending on your target audience, may be more easily understood.

Leveraging the relationship between odds and probabilities from the previous slides, we can calculate the **posterior probability** of an outcome given our observed test response and interpret it like we do other probabilities:

$$\text{Posterior probability of } D = \frac{\text{Posterior odds of } D}{1 + \text{Posterior odds of } D}$$

$$\text{Posterior probability of } \bar{D} = \frac{\text{Posterior odds of } \bar{D}}{1 + \text{Posterior odds of } \bar{D}}$$

Posterior Odds and Probability of D Example

Using our previous results that the LR+ is 8 and $P(\text{CHD}) = 0.2$, the posterior probability of having CHD given a positive treadmill test is:

$$\text{Prior odds of } D = \frac{P(D)}{1 - P(D)} =$$

$$\text{Posterior odds of } D = \text{Prior odds of } D \times \text{LR+} =$$

$$\text{Posterior probability } D = \frac{\text{Posterior odds of } D}{1 + \text{Posterior odds of } D} =$$

Interpretation: After observing a positive TT, our odds in favor of CHD are ____ :1, or the (posterior) probability of having the disease is ____% (a change from our prevalence of 20%).

Posterior Odds and Probability of $\bar{D}$ Example

Using our previous results that the LR- is $\frac{2}{9}$ and $P(\text{CHD}) = 0.2$, the posterior probability of not having CHD given a negative treadmill test is:

$$\text{Prior odds of } \bar{D} = \frac{1 - P(D)}{P(D)} =$$

$$\text{Posterior odds of } \bar{D} = \text{Prior odds of } \bar{D} \times \frac{1}{\text{LR-}} =$$

$$\text{Posterior probability } \bar{D} = \frac{\text{Posterior odds of } \bar{D}}{1 + \text{Posterior odds of } \bar{D}} =$$

Interpretation: After observing a negative TT, our odds in favor of not having CHD are ____ :1, or the (posterior) probability of not having the disease is ____% (a change from our prior probability of 80% not having CHD).

A PPV/NPV and PP Easter Egg

If we look back at the start of the lecture we may remember that our PPV was 67% and our NPV was 94.7%. . . *which is the same as our posterior probabilities!!*

It turns out if we are interested in the posterior probabilities we can just look to the predictive values in our context of a diagnostic test that can be summarized with a 2x2 table. A nice shortcut to take instead of calculating the likelihood ratios and odds, then transforming back to the probability scale.

Formula Summary

For our confusion matrix we can calculate the following summaries:

$$\text{Sens} = P(T|D) = \frac{P(T \cap D)}{P(D)} = \frac{a}{a+c}$$

$$\text{Spec} = P(\bar{T}|\bar{D}) = \frac{P(\bar{T} \cap \bar{D})}{P(\bar{D})} = \frac{d}{b+d}$$

$$\text{FNR} = P(\bar{T}|D) = \frac{P(\bar{T} \cap D)}{P(D)} = \frac{c}{a+c}$$

$$\text{FPR} = P(T|\bar{D}) = \frac{P(T \cap \bar{D})}{P(\bar{D})} = \frac{b}{b+d}$$

$$\text{LR+} = \frac{P(T|D)}{P(T|\bar{D})} = \frac{\text{Se}}{1-\text{Sp}} = \frac{a(b+d)}{b(a+c)}$$

$$\text{LR-} = \frac{P(\bar{T}|D)}{P(\bar{T}|\bar{D})} = \frac{1-\text{Se}}{\text{Sp}} = \frac{c(b+d)}{d(a+c)}$$

Test	**GS/Disease Status**	
	Positive (D)	**Negative** $(\bar{D})$
Positive (T)	a	b
Negative $(\bar{T})$	c	d

If we believe the study prevalence is a good estimate for the population prevalence:

$$\text{PPV} = P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{a}{a+b}$$

$$\text{NPV} = P(\bar{D}|\bar{T}) = \frac{P(\bar{D} \cap \bar{T})}{P(\bar{T})} = \frac{d}{c+d}$$