Perspective on prediction accuracy: Initial rough draft outline

Overview of terminology

"This test has 90% accuracy" is generally an impressive statement. In this perspective, we will unpack what it means by reviewing key concepts that cut across areas like medicine, public health, statistics, and machine learning and set out guidelines for... We note that we use "test" in a very general way - it could in fact be a diagnostic test (like a biopsy), but it could also be a screening test (like a PSA test, a pap smear, or a prenatal test providing the risk of certain genetic abnormalities), or even an algorithm used for basic science purposes (like a method for predicting transcription factor binding sites). In general, a test is meant to detect a certain condition - thus, one can consider the 2x2 table of (presence of a condition) x (result of test): **(Table and definitions are from Wikipedia)** Note that the test being positive means that the result implies the presence of the condition while the test being negative means that it implies the absence of the condition.

	Condition present	Condition absent
Test positive	True positive (TP)	False positive (FP)
Test negative	False negative (FN)	True negative (TN)

Commonly used definitions include:

- Sensitivity - also known as true positive rate (TPR), hit rate, or recall = fraction of all cases with the condition for which the test is positive = $TP/(TP+FN)$.
- Specificity - also known as true negative rate (TNR) = fraction of all cases without the condition for which the test is negative = $TN/(TN+FP)$.
- Accuracy = fraction of all the cases (with and without the condition) for which the test is correct = $(TP+TN)/(TP+FP+FN+TN)$.
- Positive predictive value (PPV), also known as precision = fraction of all cases for which the test is positive who truly have the condition = $TP/(TP+FP)$.
- Negative predictive value (NPV) = fraction of all cases for which the test is negative for whom the condition is truly absent = $TN/(TN+FN)$.

Challenges with quantifying predictions when one class is very small

Difference between sensitivity and specificity and PPV and NPV; how high sens and spec can still lead to very low PPV and NPV

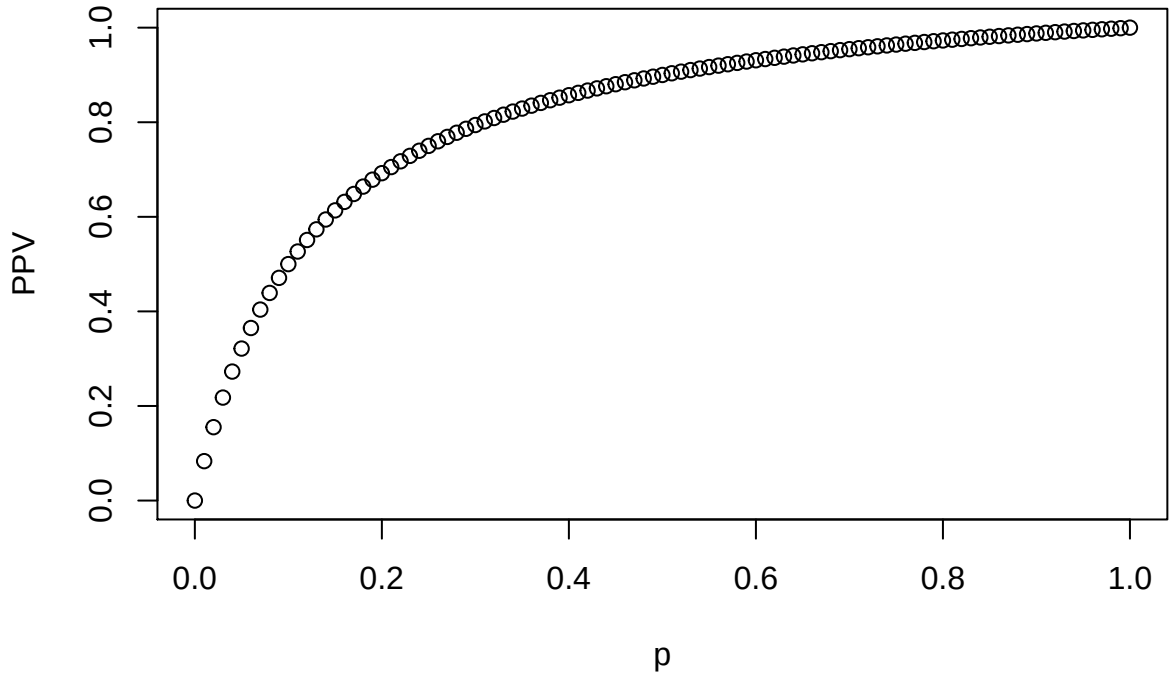
We note that, in many cases, scientists prefer to communicate the sensitivity and specificity. In fact, in the popular ROC plots, sensitivity is plotted against 1-specificity, as detailed in the next section. However, it has often been pointed out that from a clinical or public heath perspective, the more relevant quantities are the PPV and NPV. For instance, take a patient getting tested for a serious disease getting a positive test result back. She may want to know what the chance of having the disease given that she tested positive is - this is exactly the PPV. The sensitivity would instead tell her the chance that someone with the disease tests positive, which is not relevant. One reason why sensitivity and specificity are still more commonly reported compared to PPV and NPV is simply that they are much easier to estimate. In particular, they may be estimated from case-control studies, which take a set of cases (individuals with the condition) and a set of controls (individuals without the condition), then apply the test to each set. **I realized I'm using "cases" in 2 different ways - can adjust terminology** In contrast, the PPV and NPV implicitly require

the prevalence of the condition in the general population, which by definition cannot be estimated from case-control studies. This can be seen directly by applying Bayes' rule:

$$\begin{aligned} PPV &= P(\text{Condition present}|\text{Positive test}) = \frac{P(\text{Positive test}|\text{Condition present})P(\text{Condition present})}{P(\text{Positive test})} \\ &= \frac{P(\text{Positive test}|\text{Condition present})P(\text{Condition present})}{P(\text{Positive test}|\text{Condition present})P(\text{Condition present}) + P(\text{Positive test}|\text{Condition absent})P(\text{Condition absent}} \\ &= \frac{TPR \times p}{TPR \times p + (1 - TNR) \times (1 - p)}, \end{aligned}$$

where p =prevalence of the condition.

Consider a test with TPR=0.9 and TNR=0.9 and look at the PPV as a function of p:



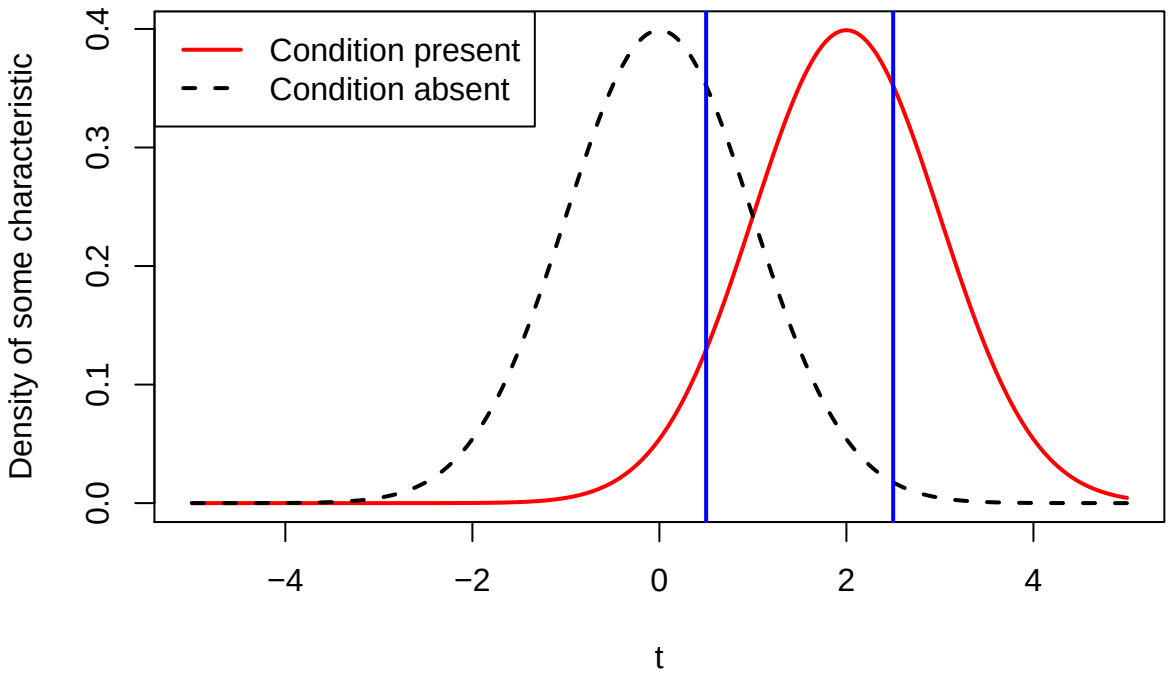
Note that PPV increases as a function of p and for low values of p, the PPV can also be quite small - for example, for $p = 0.01$, $PPV = 0.08$, which means that a positive result would mean only an 8% chance of actually having the condition! In contrast, for $p = 0.1$, $PPV = 0.5$, and for $p = 0.2$, $PPV = 0.69$. **Cite some work on biomarkers, maybe some work by Wacholder on PPV, NPV etc. I think Frank Harrell has argued that sensitivity and specificity should never be studied or reported, since they're just misleading and focus on the wrong thing.** If a test is used in clinical practice, this is a very important consideration and is one reason why, for example, screening tests are not applied to the general population, but rather a population that is already *enriched for some risk factor*, which in effect increases the prevalence of the disease: For example, only performing colonoscopies in individuals only 50, only considering certain prenatal tests in women over 35 etc.

How high accuracy can be misleading

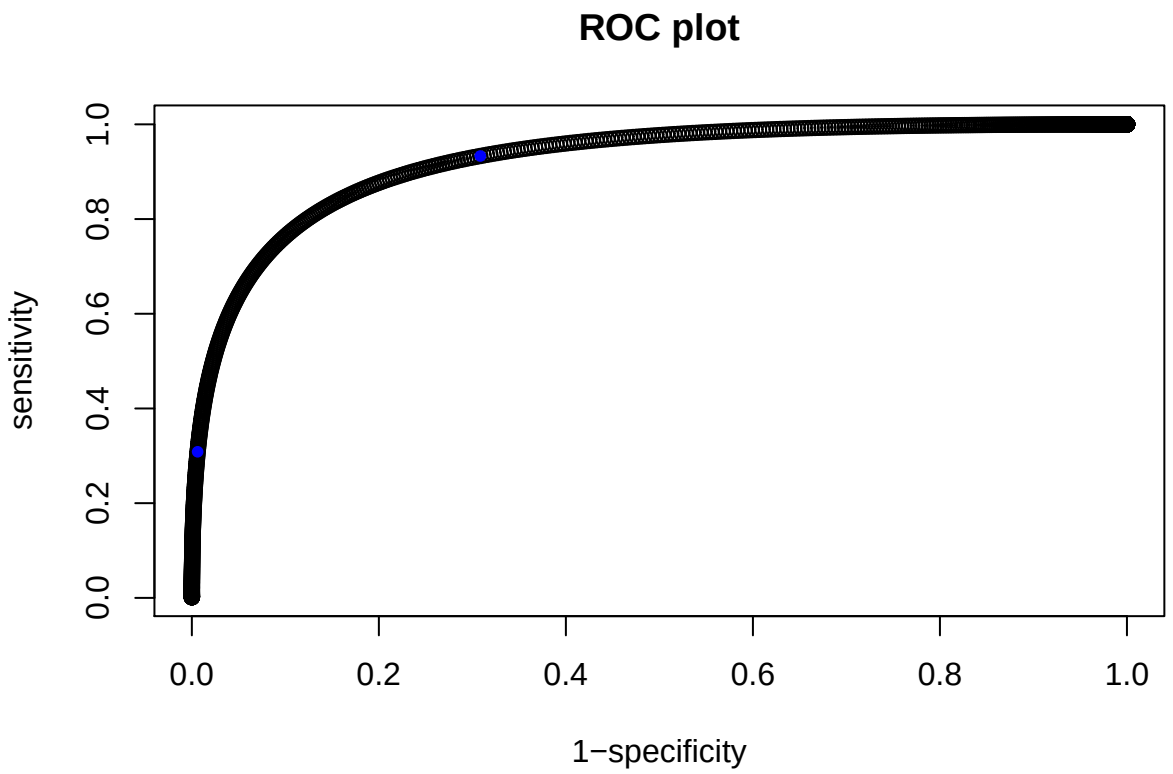
This can affect other measures, such as the accuracy. For example, if the class with the condition is very small - say, having a prevalence $p = 0.1$, then simply setting the test to always be negative will result in an accuracy of 90%. However, this will not help at all in distinguishing between the two classes. This is another reason why using the PPV and NPV may be more useful in practice.

Sensitivity and specificity tradeoff; ROC curves

In general, there is a trade-off between sensitivity and specificity, since a test can often be thought of as being "positive" if it is above some threshold and "negative" if it is below the threshold. Consider the situation where the distribution of a certain characteristic is from a N(0,1) if the condition is absent and N(2,1) if the condition is present. Thus, the two densities look like this:



Note that this plot does not hold any information for the prevalence of the condition, which can also be seen as the probability that any case has the condition, i.e. has the characteristic from the N(2,1) distribution. The blue vertical lines represent two thresholds (t=0.5 and t=2.5), which values of the characteristic above the threshold representing a positive test. For test 1 (t = 0.5), the sensitivity = 0.93 (area to the right of the threshold on the N(2,1) curve), while the specificity = 0.69 (area to the left of the threshold on the N(0,1) curve). As the threshold increases, the sensitivity decreases (0.31 for test 2, t=2.5) and the specificity increases (0.99 for test 2). An extreme example is of setting the test to always be positive - in that case, the sensitivity is 100%, while the specificity is 0% - or always to be negative - sensitivity = 0%, specificity = 100%. This can be seen in the classical ROC plots, where sensitivity is plotted against 1-specificity. Each point on the ROC plot represents these two values at a different threshold, with the (0,0) and (1,1) points representing the scenario where the test is always negative, respectively where it is always positive. The plot below highlights the two tests discussed above:



Thus, with the change in threshold, one may only move on this ROC curve, but other tests that use different methods will result in other curve. In general, the better the test is, the closer it is to the (0,1) corner. A simple way of quantifying this is by using the "area under the ROC curve" (AUC, auROC), which is simply the integral of sensitivity with respect to 1-specificity. Better tests will have values closer to 1. In particular, the identity line on an ROC plot is equivalent to sensivity + specificity = 1; this means that the two densities above actually overlap and there is no discrimination offered by the test, with the resulting AUC = 0.5. As before though, we caution that this type of analysis completely ignores the prevalence of the condition and that in the case of a low prevalence, focusing on the sensitivity and specificity only may result in overenthusiasm.