Receiver Operating Characteristics and the Area Under the Curve

- Receiver operating characteristic (ROC) curves tell us how good a predictive model, e.g., a logistic regression model, is at differentiating between two possible outcomes.
- The area under the ROC curve can be used as a measure of model performance.
- The higher the area under the curve (AUC) value, the better the model's performance.
- There is no standard interpretation for AUC values, so there is no universal "good" value.

by ALEXANDER THORPE, GARSTON LIANG, & QUENTIN F. GRONAU

1 Evaluating Models

Predictive models are useful for understanding what factors may contribute to good or bad outcomes for patients. For example, spinal cord stimulation (SCS) may have different chances of success for patients with long-term chronic pain compared to those who have only experienced pain for a short time. Let's consider a simple example study that examines this relationship with the research question *"does duration of chronic pain predict pain relief after spinal cord stimulation?"*. The researchers could use a logistic regression to find that patients with short-term chronic pain are more likely to find relief after SCS than patients with long-term pain.

But how do our researchers know they can have confidence in their model's predictions? To answer this question, they would need to evaluate the model's performance by considering its *sensitivity* and *specificity*. Sensitivity is a model's ability to detect true positives without committing Type II errors—saying "no" when the answer is really "yes". Specificity is a model's ability to detect true negatives without committing Type I errors—saying "yes" when the answer is really "no". If a model is not sensitive, it will miss true positives, and if it is not specific, it will sound false alarms. We can assess both the sensitivity and specificity of a model using *receiver operating characteristic* curves.

2 Receiver Operating Characteristics

Receiver Operating Characteristic (ROC) curves are a way of assessing how sensitive a predictive model is—how often it says "yes" when the answer is really "yes"—over a range of specificity

levels—how often it says “yes” no matter what the real answer is. We can think of this like a filter, that starts closed and moves to being completely open. The graphs in Figure 1 show what effect these extreme positions have on model performance. On the left side of each figure, the model says “no” to everything, no matter what evidence it is presented with. Under these conditions, the model never gives a false positive but it also never detects true positives. At the other extreme, the model always says “yes”. This means it always detects true positives, but it also gives false positives every time the answer was really “no”. The right proportion of “yes” answers must lie somewhere in between these extremes, but we don’t know where. We can check every possible proportion by sweeping our filter from “always no” to “always yes” and graphing the rate of true positives against false positives at every point. If the model is no better than chance, then the true positives and false positives will increase at the same rate, signified by the left panel of Figure 1, and the red dotted line in both panels. However, if the model is sensitive to true positives while still ruling out true negatives, the line on our graph will curve up towards the top-left of the figure. The better the model performs, the further above the dotted line it will rise. The right panel shows a better-than-chance model, where the ROC curve is above chance level for at least some of its length.

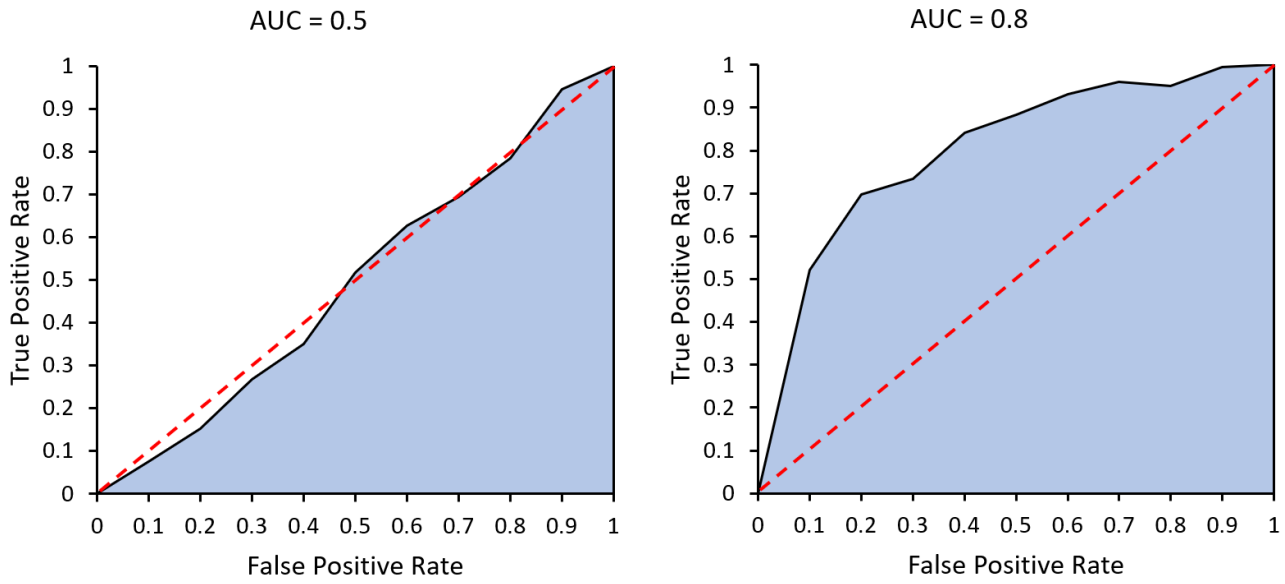

Figure 1: Receiver Operating Characteristic (ROC) curves. The red dotted line represents chance-level performance. The chance-level ROC (left) has an area under the curve (AUC) of 0.5. The better-than-chance ROC (right) has an AUC of 0.8; greater area under the curve indicates better performance.

3 Area Under the Curve

We can boil the model’s performance down to a single value by calculating the area under the ROC curve. This value simply represents the proportion of the graph that lies below the curve, depicted as the blue area in Figure 1, so it is always expressed as a unit-less number between 0 and 1. However, chance-level performance runs straight from bottom-left to top-right, so the

lowest meaningful value is 0.5. Anything less than 0.5 would indicate the model's predictions are worse than just guessing! The greater the area under the curve (AUC) value, the better the model's performance. These values can be used to compare the performance of multiple models. The curves in Figure 1 are labelled with their respective AUC values.

4 Interpreting Area Under the Curve

So we have a number to report, but what does it mean? Because it is a proportion of "correct" vs "incorrect" classifications, we can interpret it in terms of probability when talking about a single case. A model with an AUC of 0.8 has an 80% chance of correctly distinguishing between the two outcomes. Going back to our example study, in which pain duration predicted success of spinal cord stimulation, we could take two randomly sampled patients, one whose treatment brought them pain relief, and one whose treatment did not. If the researchers' model produced an AUC of 0.8, it would successfully rate the first patient's probability of treatment success as being higher than the second patient's 80% of the time.

As mentioned before, a bigger AUC is better, but there is no standard classification of AUCs. In other words, what counts as a "good" AUC depends on the context in which it is used. The researchers in our example study may be satisfied with 80% predictive accuracy, considering the difficulty of predicting the success of treatment. However, a doctor using a diagnostic tool to detect a communicable disease may require a higher standard of performance, due to the high cost of a false negative result. The most commonly accepted interpretations from diagnostic medicine literature rate anything below 0.7 as "poor", 0.7 to 0.8 as "acceptable", 0.8 to 0.9 as "good", and 0.9 and above as "excellent". However, these interpretations vary, and should be used as a loose guide only.

5 Further Reading

Below are some open-source articles that discuss ROC curves and their interpretation in greater detail, with a focus on their use in medical research.

- de Hond, A. A. H., Steyerberg, E. W., & van Calster, B. (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet*, *4*(12), E853-E855.
- Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315-1316.
- Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *155*(5), 654-657.