

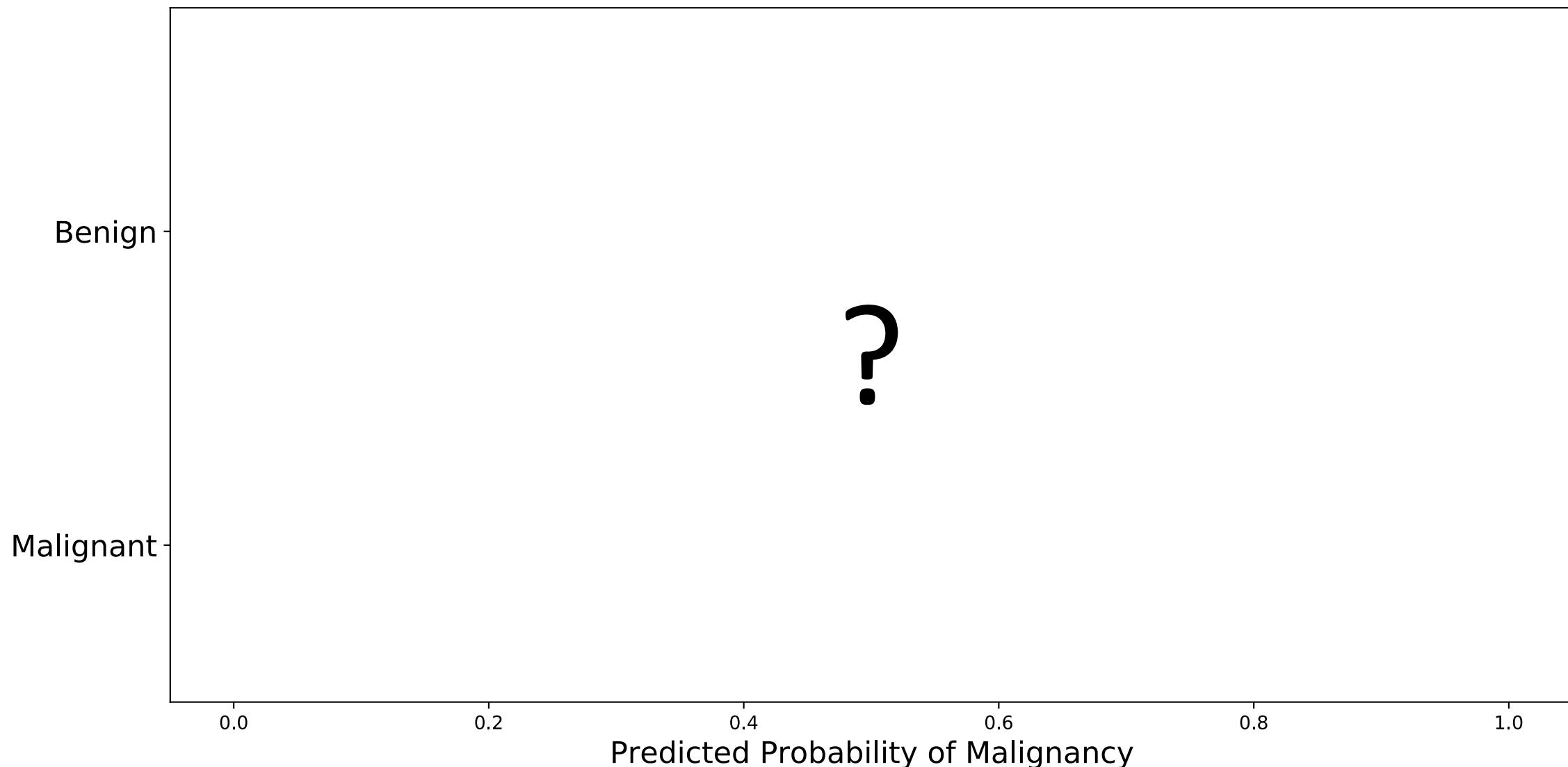
Performance Measures

Matthew Engelhard

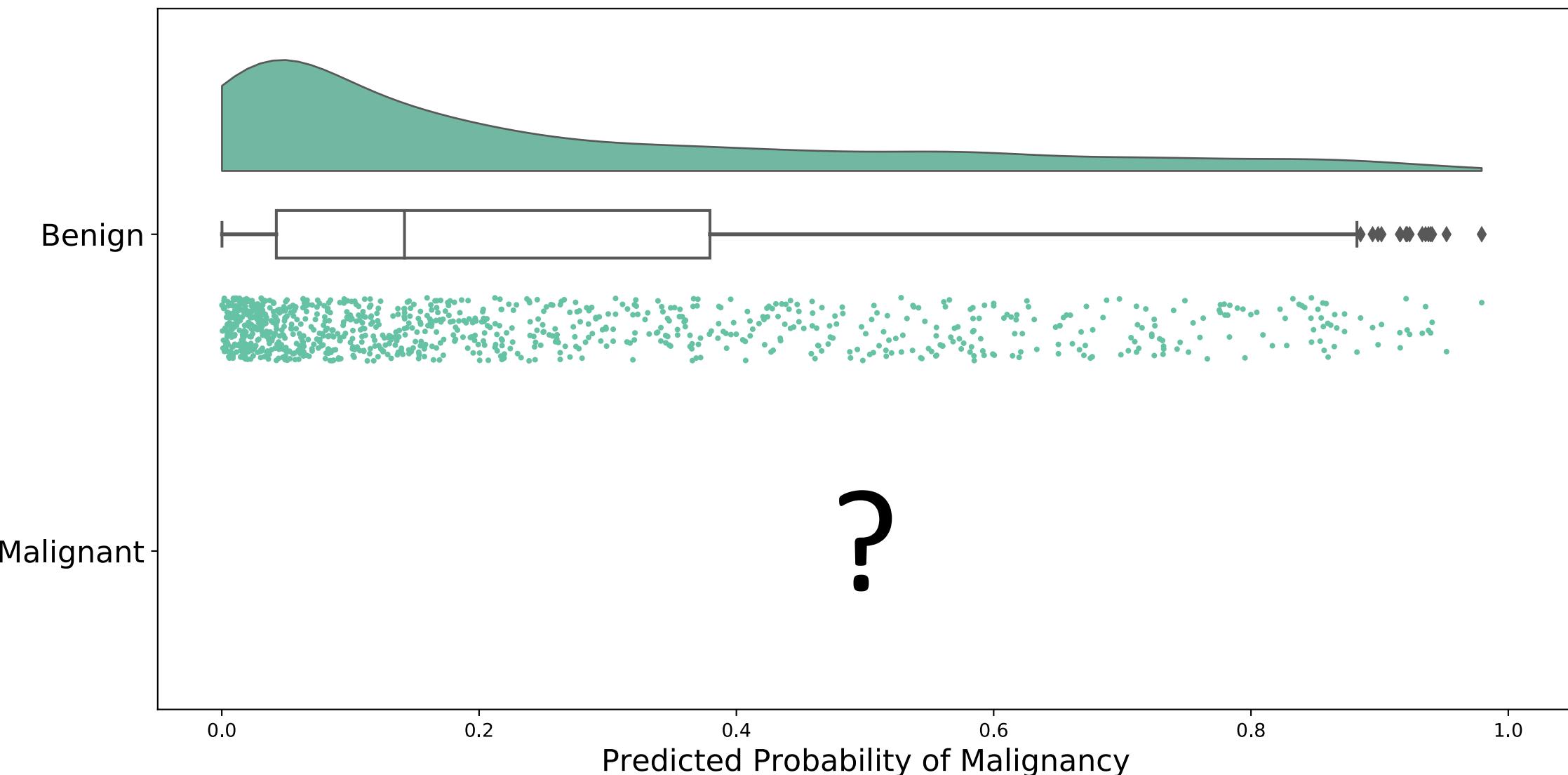
Goals

- Understand and calculate common performance measures for binary classification
- Contextualize performance against that of a *no information* classifier
- Recognize that *good* performance depends on existing alternatives
- Match clinical scenarios to performance measures important in that scenario

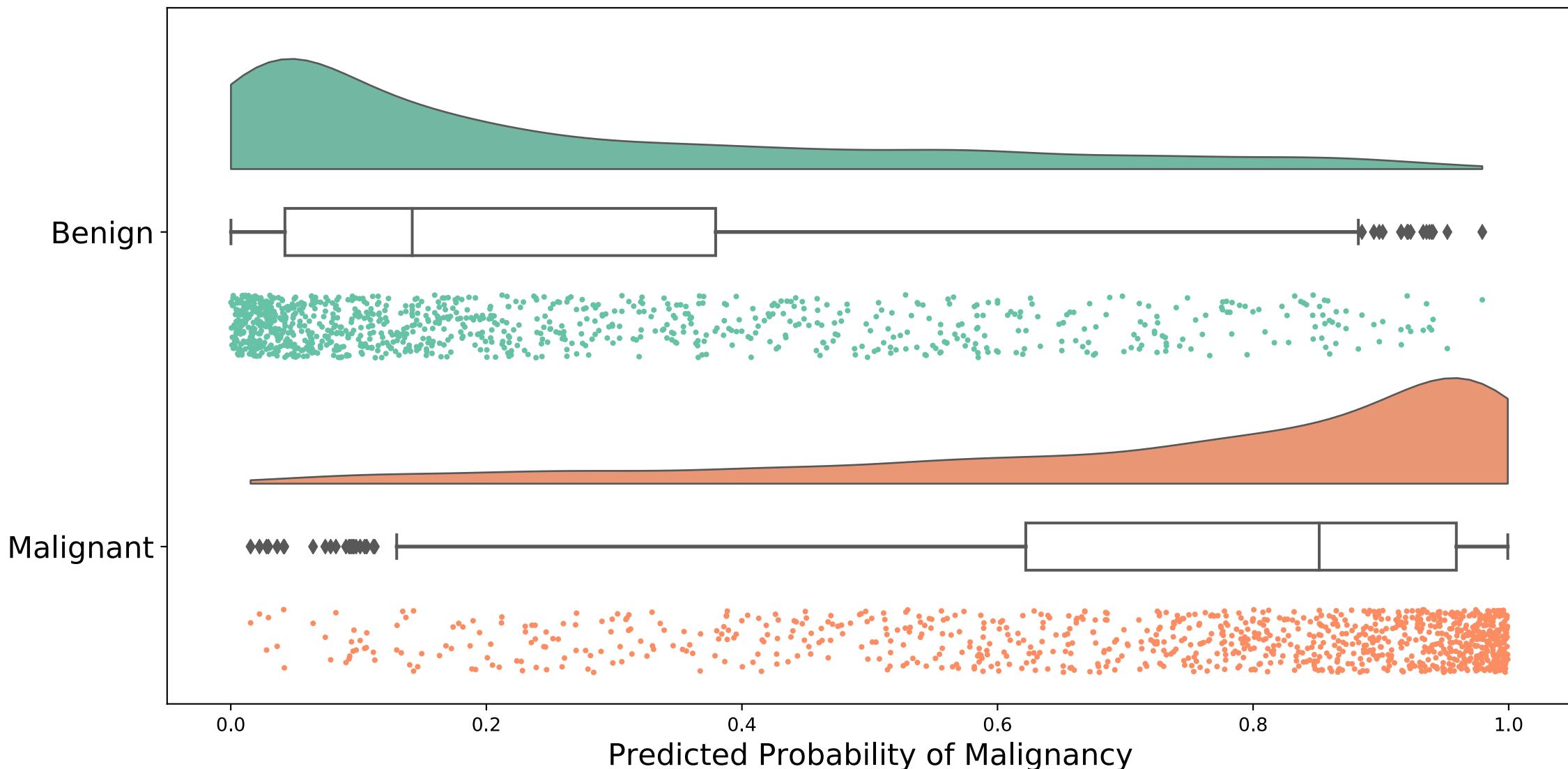
Back to cancer prediction. Suppose our features are highly informative. What might our model's predictions look like?

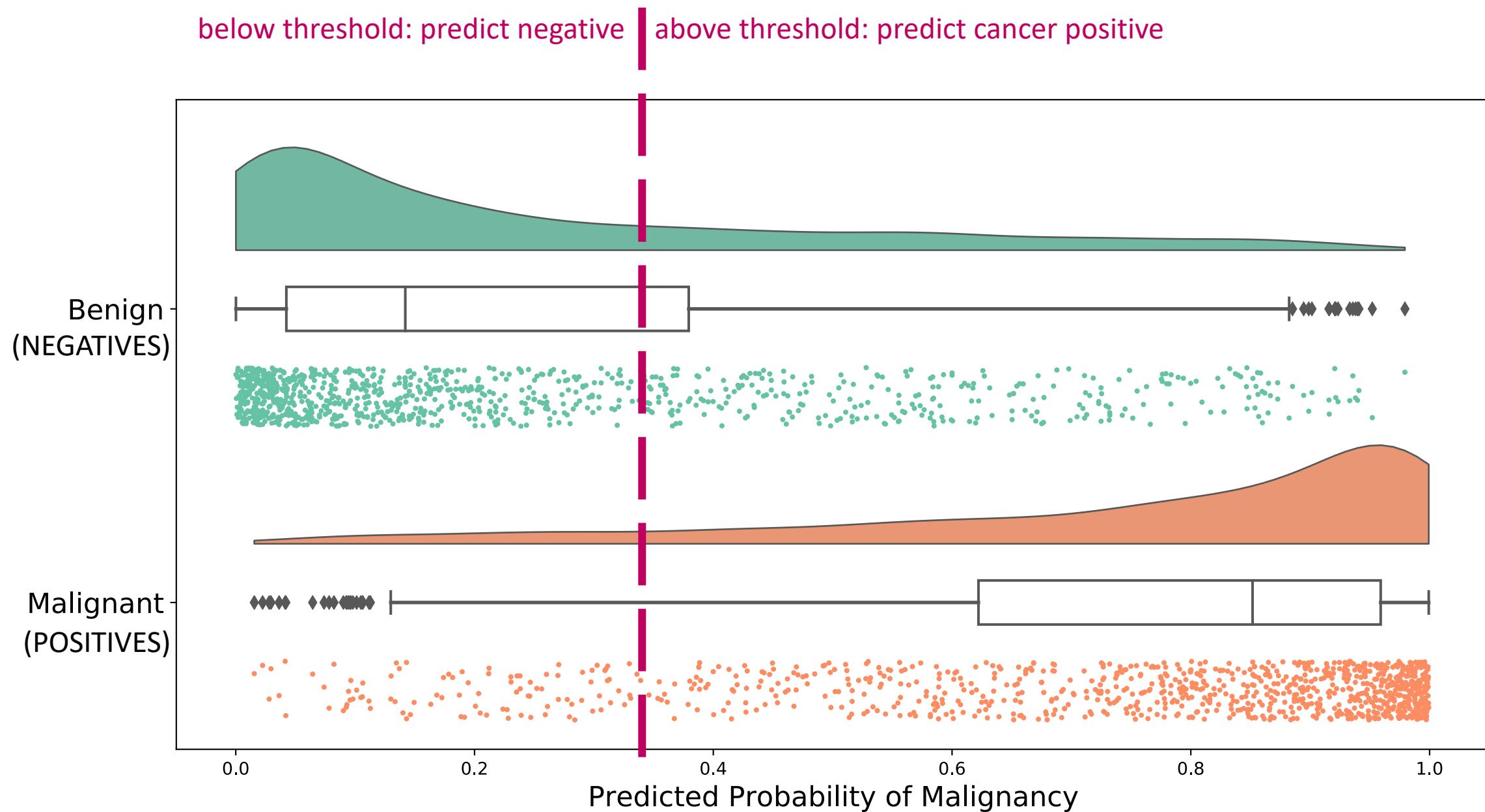


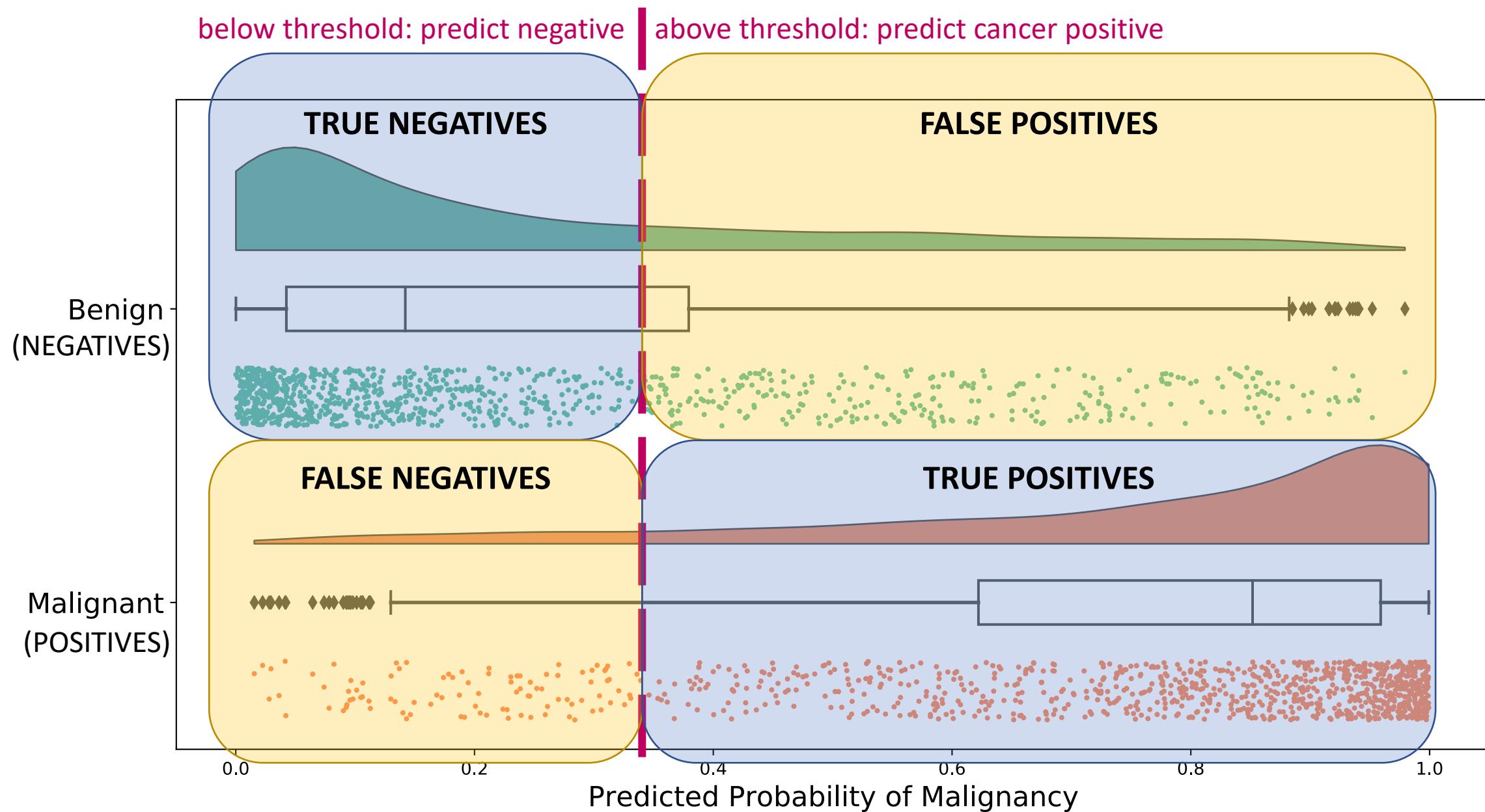
Back to cancer prediction. Suppose our features are highly informative. What might our model's predictions look like?



Back to cancer prediction. Suppose our features are highly informative. What might our model's predictions look like?

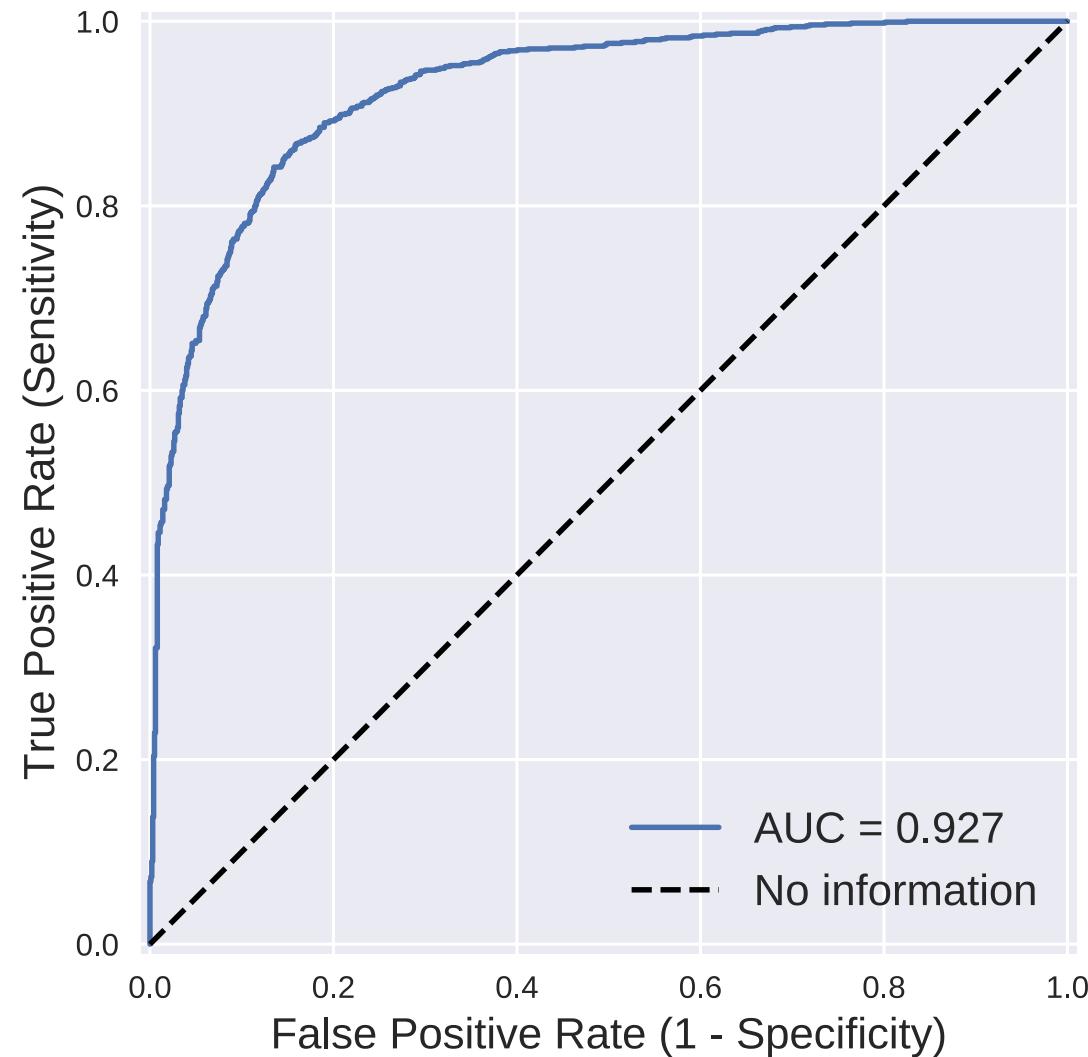






Receiver Operating Characteristic Curve

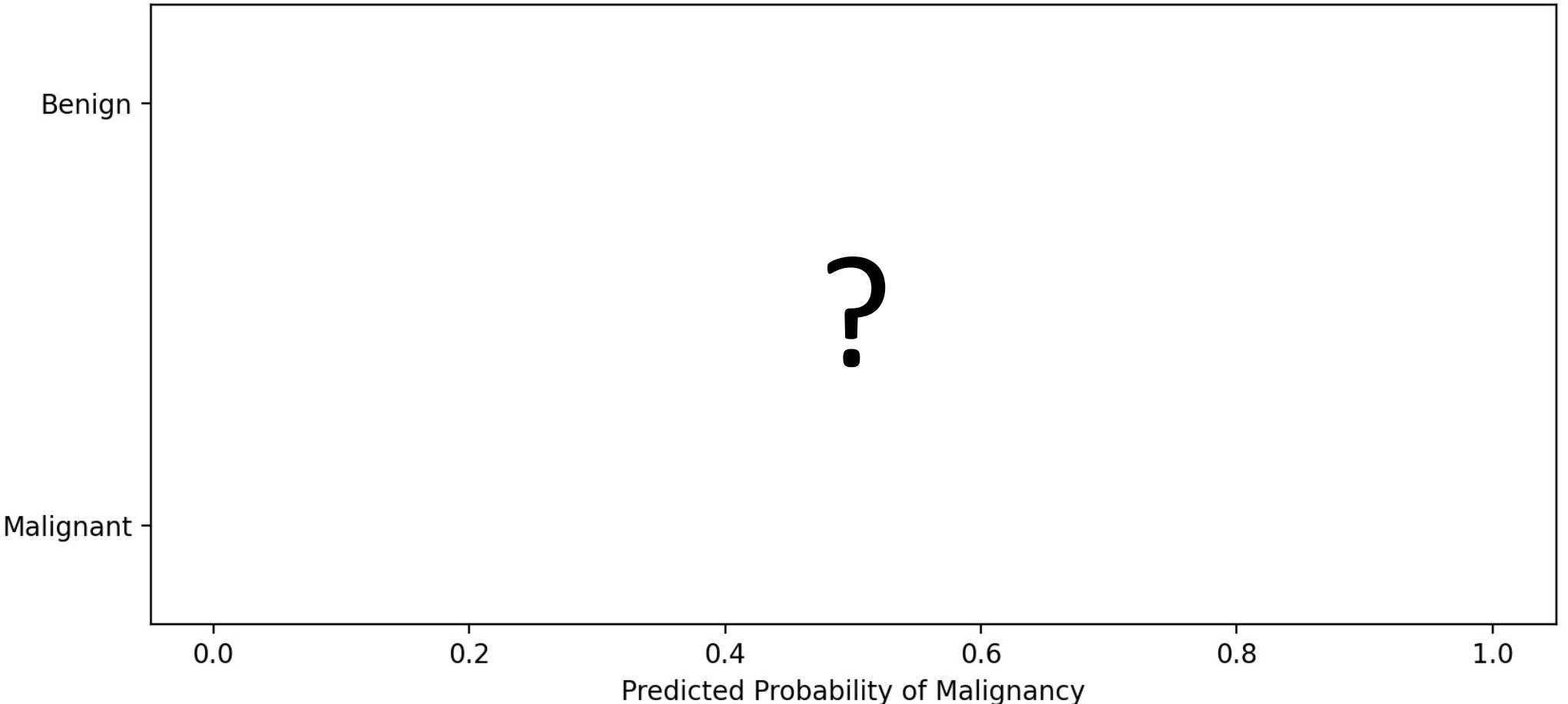
- Illustrates the tradeoff between the true positive rate (i.e., sensitivity) and the false positive rate (i.e., $1 - \text{specificity}$) as we vary the threshold.
- The area under this curve (AUC) provides a single summarizing this tradeoff.
- Note that to get the sensitivity versus specificity curve, we simply rotate the ROC curve clockwise by 90 degrees. The areas under the two curves are the same.



So, what's a *good* AUC value?
(i.e., *good* performance)?

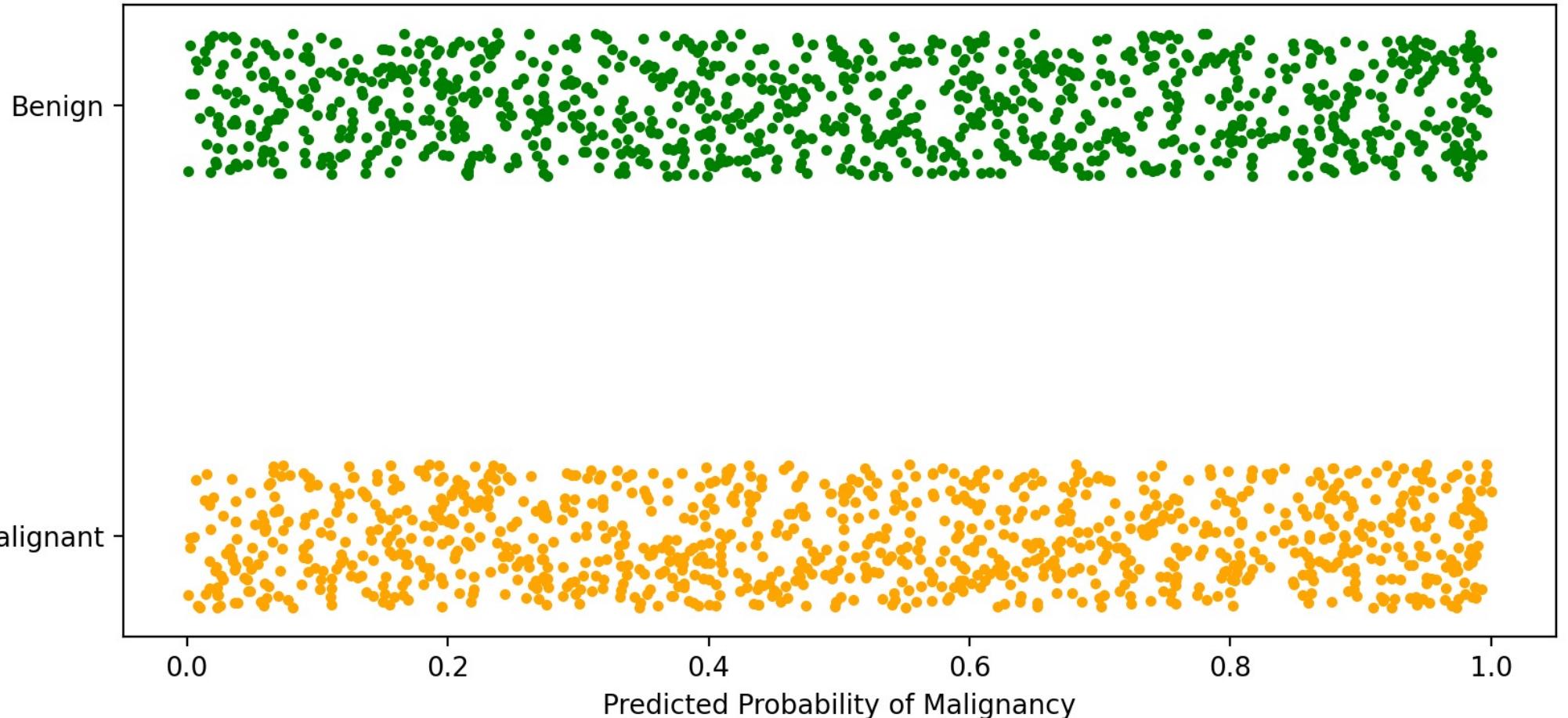
We'll start to answer this question by taking a look at *bad* performance.

Suppose our features contain *no information* about the label.
What might our model's predictions look like?



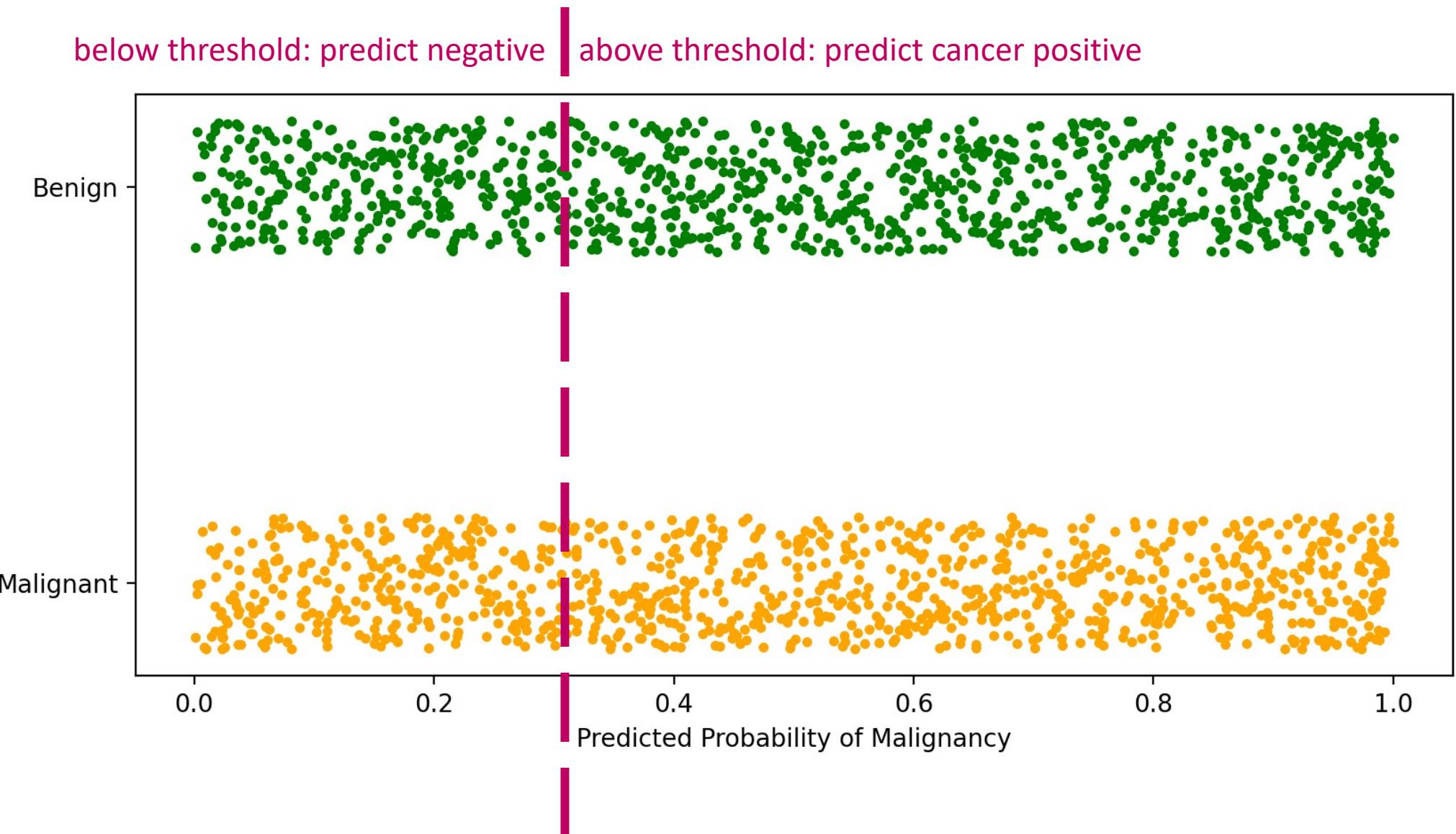
Suppose our features contain no information about the label.
What might our model's predictions look like?

- Similar distributions between positive and negative cases.
- The predicted value tells you nothing about which one it's more likely to be.



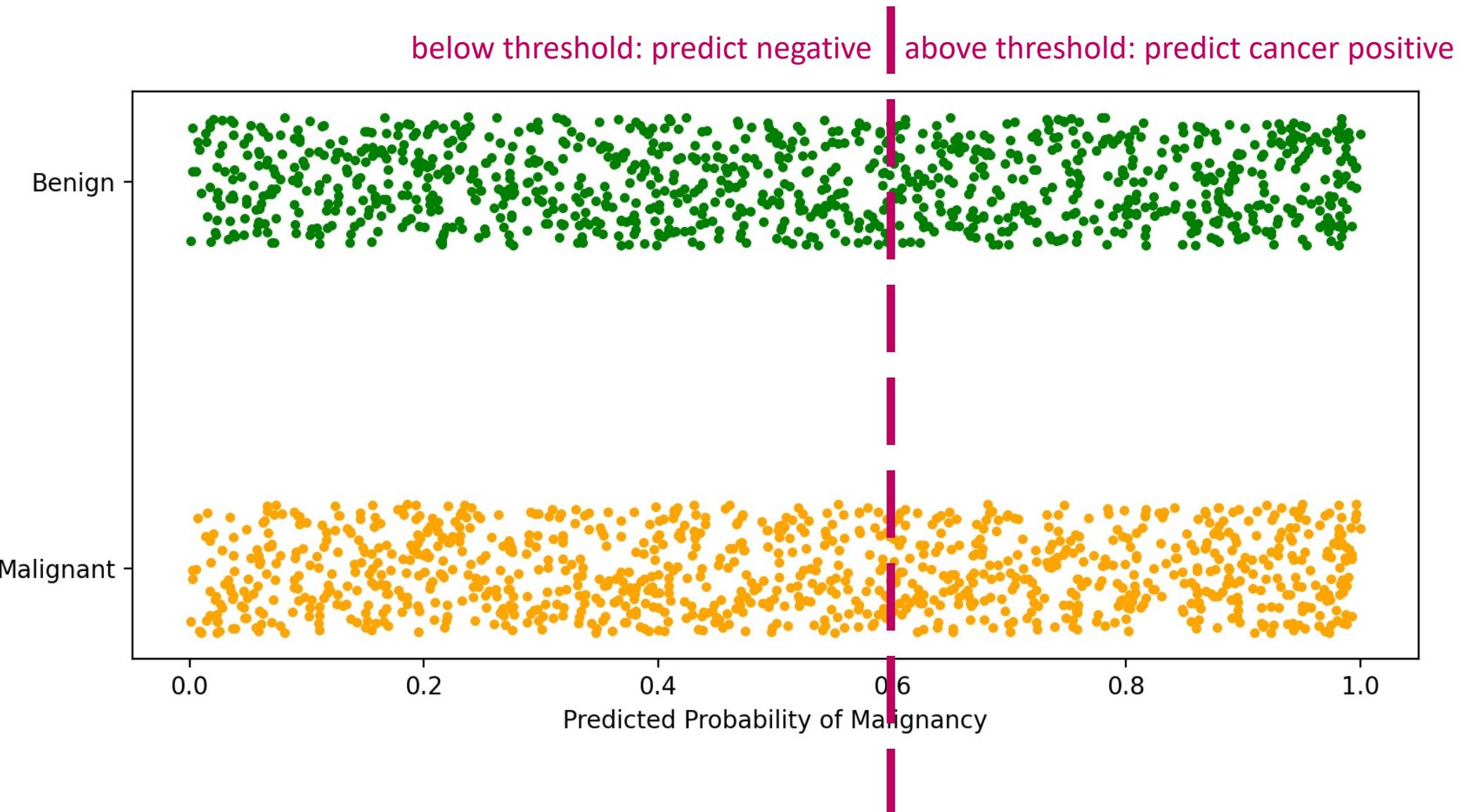
We'll try placing a threshold just like before

What is the:
(a) Sensitivity?
(b) Specificity?
(c) Positive
predictive
value?



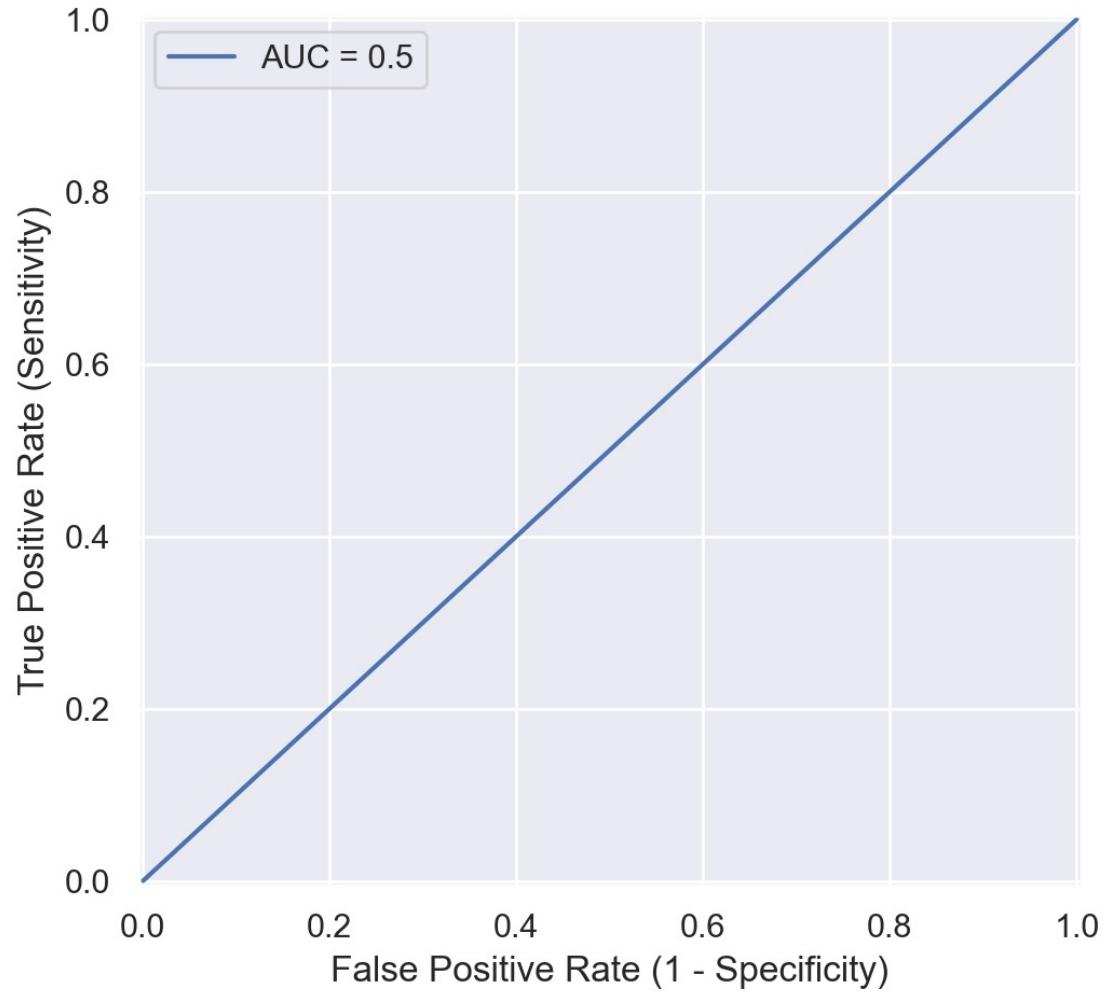
We'll try placing a threshold just like before

What is the:
(a) Sensitivity?
(b) Specificity?
(c) Positive
predictive
value?



Our no information predictive model:

- Place threshold at .3
 - Sensitivity = .3
 - False positive rate = .3
 - Specificity = 1-.3
- Place threshold at p
 - Sensitivity = p
 - False positive rate = p
 - Specificity = $1-p$



Let's think about it a different way.

- Suppose we have no predictors – again, no information – so we decide we'll just flip a coin instead of building a model.
 - If the coin comes up *heads*, we'll predict *positive*.
 - If the coin comes up *tails*, we'll predict *negative*.



Fair Coin: $P(\text{heads}) = .5$

- Sensitivity = ?
- False positive rate = ?
- Specificity = ?

Let's think about it a different way.

- Suppose we have no predictors – again, no information – so we decide we'll just flip a coin instead of building a model.
 - If the coin comes up *heads*, we'll predict *positive*.
 - If the coin comes up *tails*, we'll predict *negative*.



Fair Coin: $P(\text{heads}) = .5$

- Sensitivity = .5
- False positive rate = .5
- Specificity = .5

Let's think about it a different way.

- Suppose we have no predictors – again, no information – so we decide we'll just flip a coin instead of building a model.
 - If the coin comes up *heads*, we'll predict *positive*.
 - If the coin comes up *tails*, we'll predict *negative*.



Fair Coin: $P(\text{heads}) = .5$

- Sensitivity = .5
- False positive rate = .5
- Specificity = .5



Biased Coin: $P(\text{heads}) = p$

- Sensitivity = ?
- False positive rate = ?
- Specificity = ?

Let's think about it a different way.

- Suppose we have no predictors – again, no information – so we decide we'll just flip a coin instead of building a model.
 - If the coin comes up *heads*, we'll predict *positive*.
 - If the coin comes up *tails*, we'll predict *negative*.



Fair Coin: $P(\text{heads}) = .5$

- Sensitivity = .5
- False positive rate = .5
- Specificity = .5

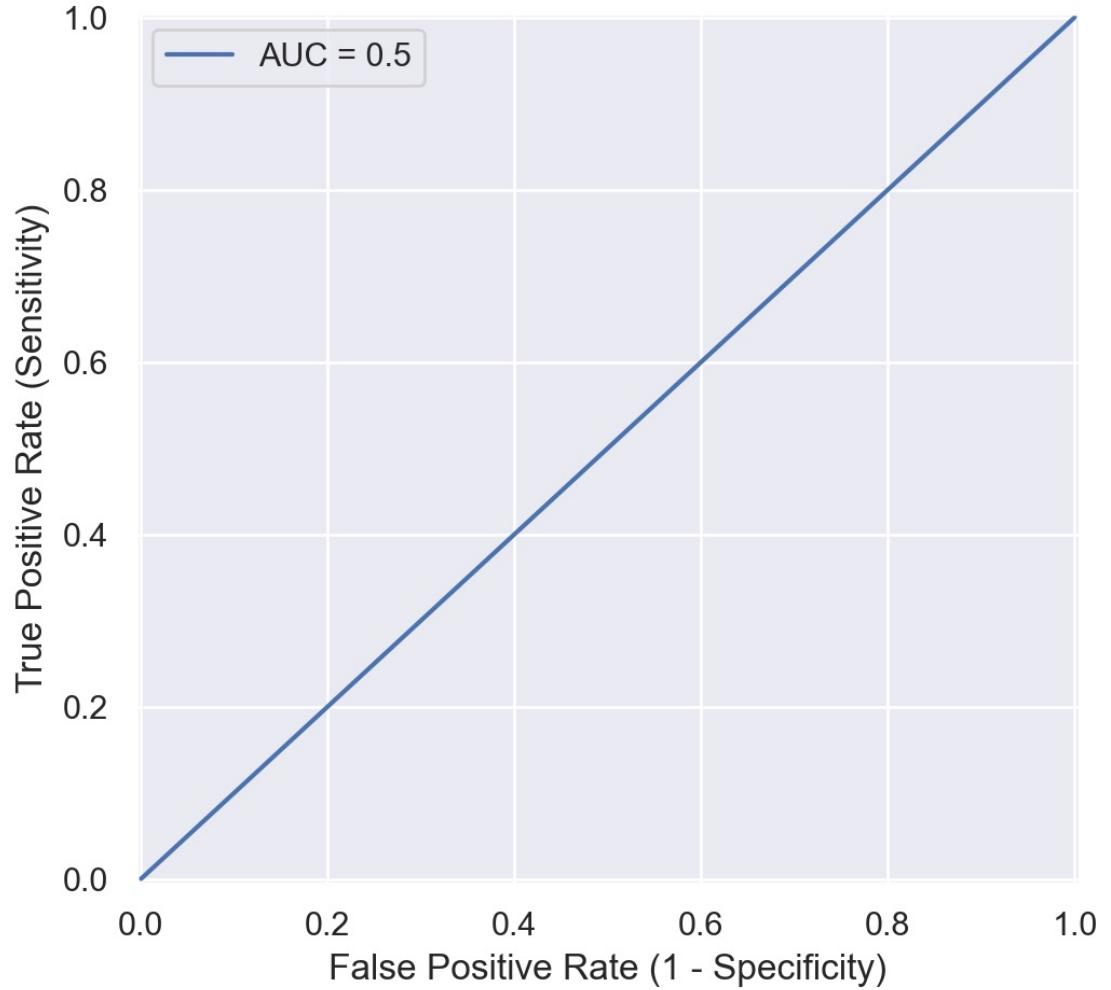


Biased Coin: $P(\text{heads}) = p$

- Sensitivity = p
- False positive rate = p
- Specificity = $1-p$

Again, we arrive at the following *no information* curve

- We may choose any p between 0 and 1 to get:
 - Sensitivity = p
 - False positive rate = p
 - Specificity = $1-p$
- What's the area under this curve (AUC)? --> 0.5



So, what's a *good* AUC value?
(i.e., *good* performance)?

It depends.

- Are predictions better than random?
- Are predictions than the previous best performing model?
- Are predictions better than expert performance?
- Does performance exceed our (informed) expectations?
- **Is the model clinically useful?**

OK, we've quantified performance across all thresholds. But how do we use the model?

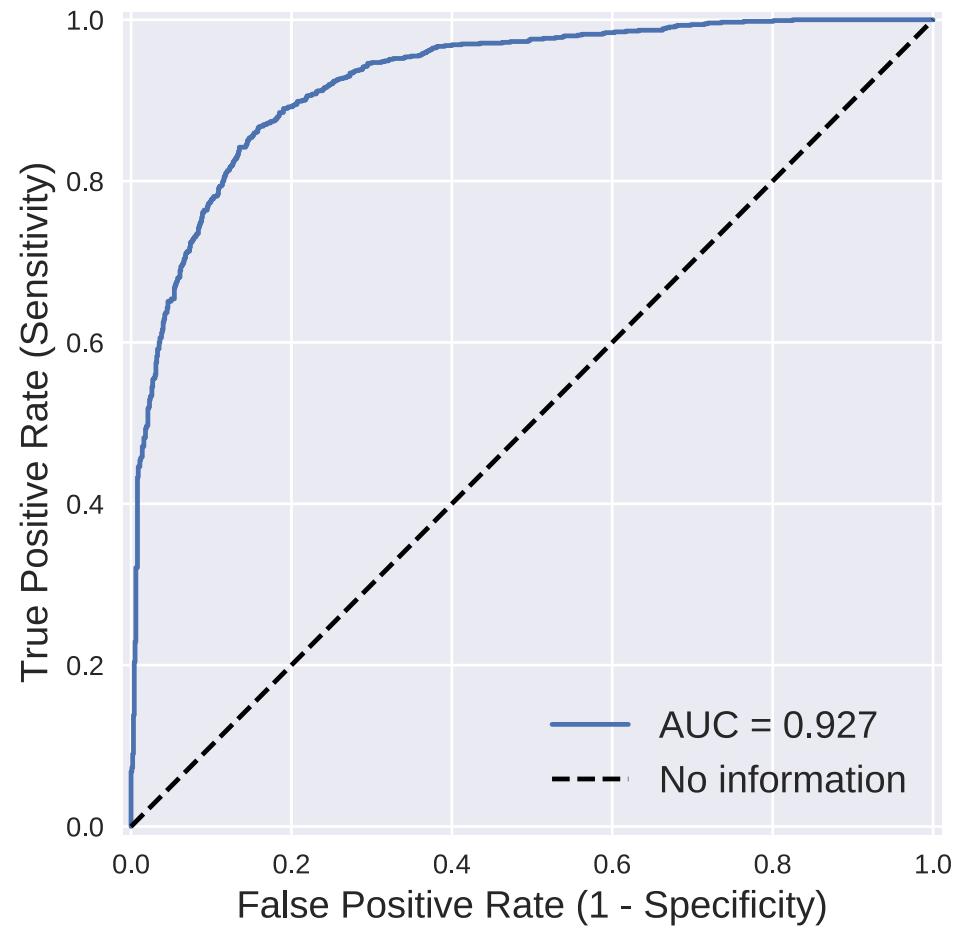
Sometimes the predicted probability really is what we care about.

- *Example*: probability of heart attack
- If so, we need to make sure our model is *calibrated*

More often, we need to pick a threshold so we can decide whether to:

- Alert a provider
- Get a biopsy
- Refer the patient
- etc

What threshold should we pick? What's the right tradeoff?



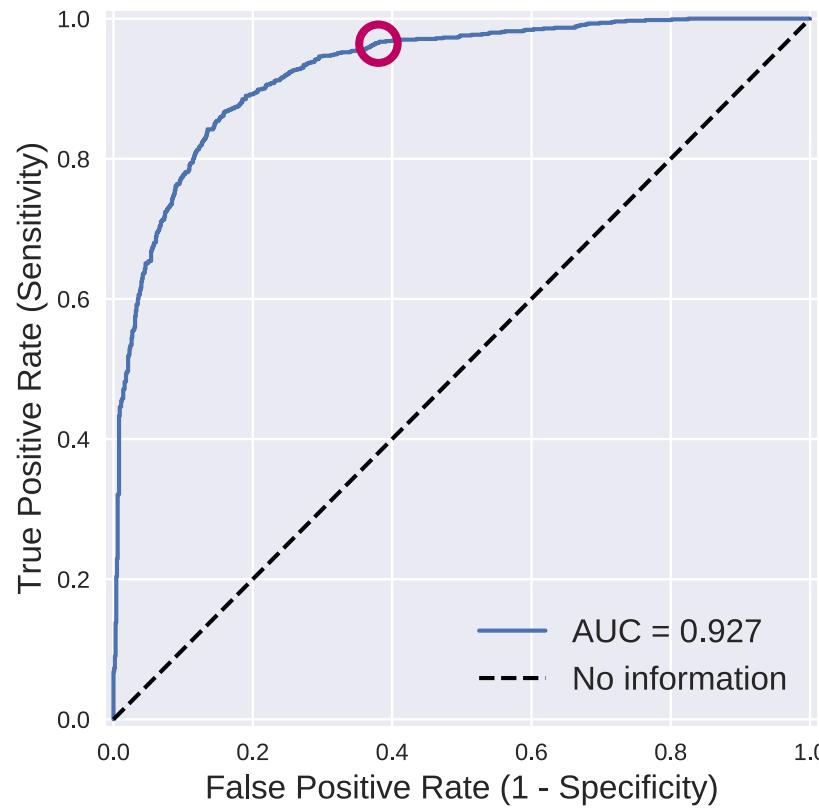
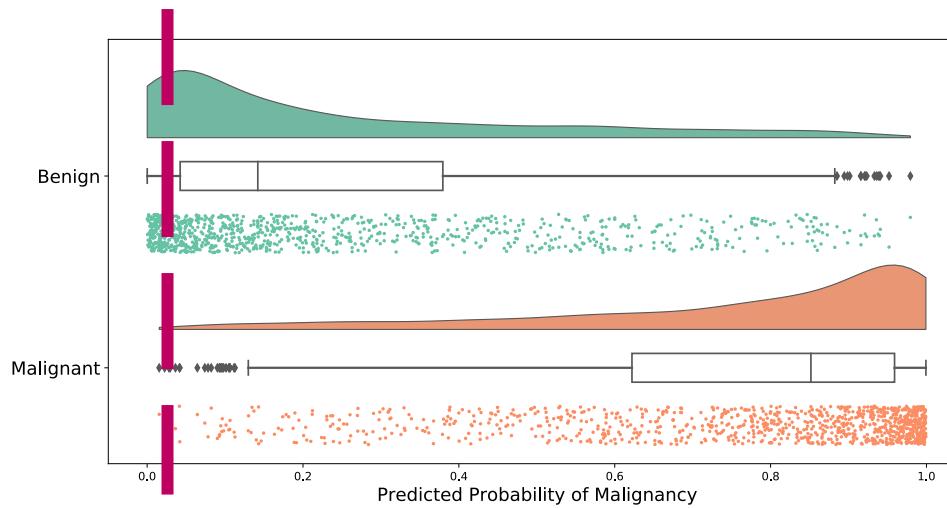
Healthcare Scenarios

Which performance measure is most important?

1. A computer vision model that detects carcinoma

Operating Point:

high sensitivity

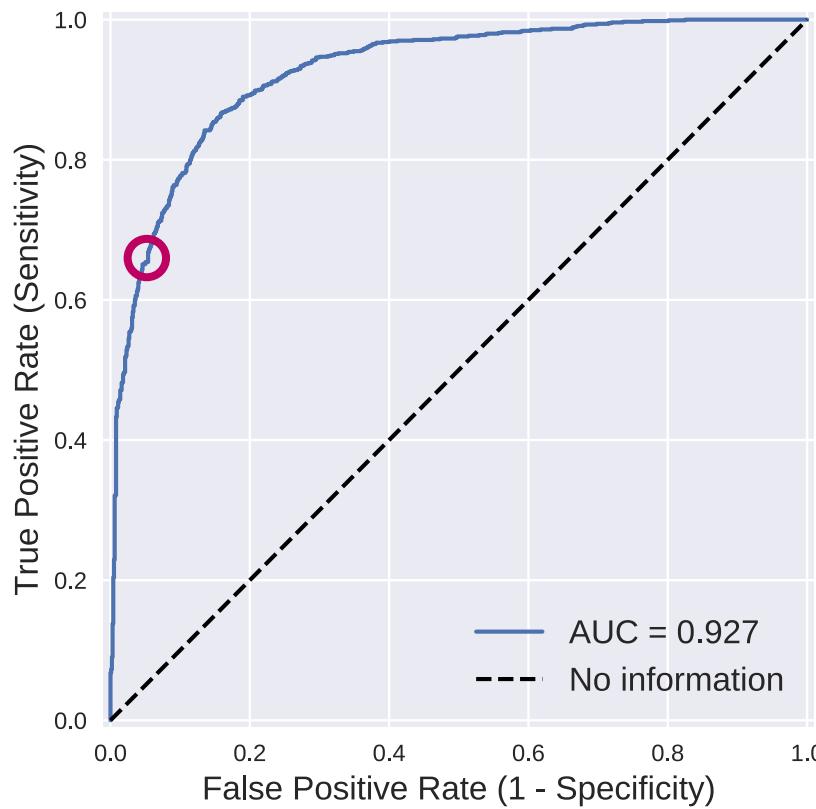
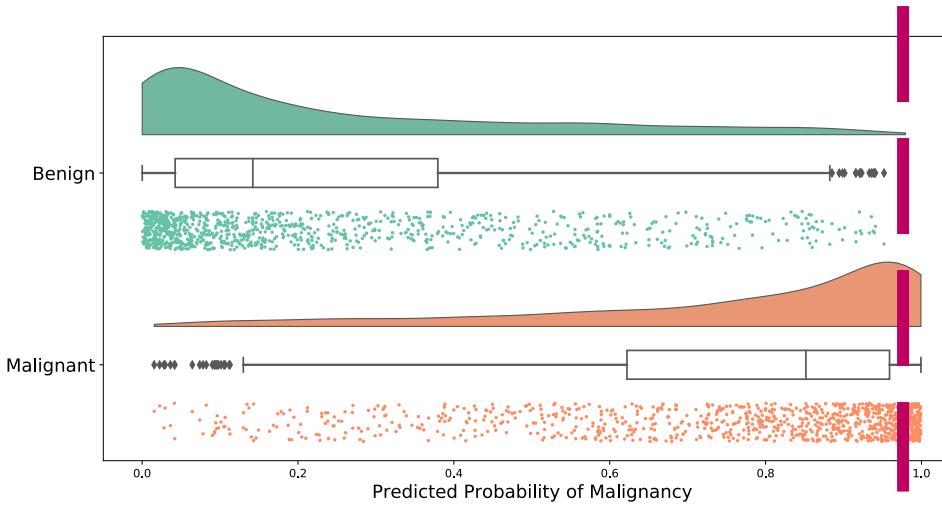


Healthcare Scenarios

1. A computer vision model that detects carcinoma
2. An algorithm that detects atrial fibrillation in Apple Watch users

Operating Point:

high specificity

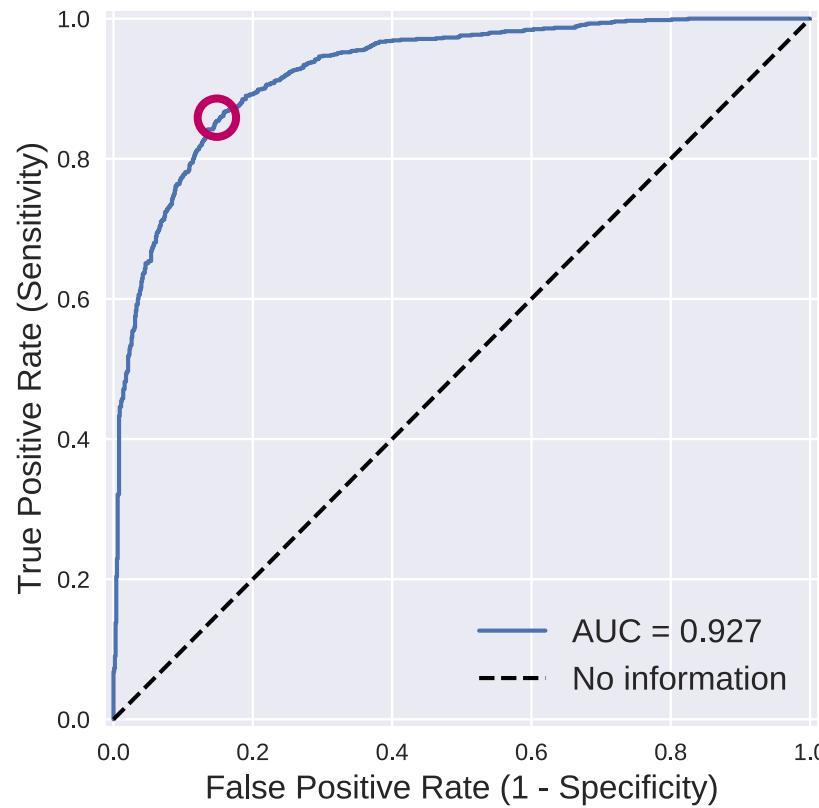
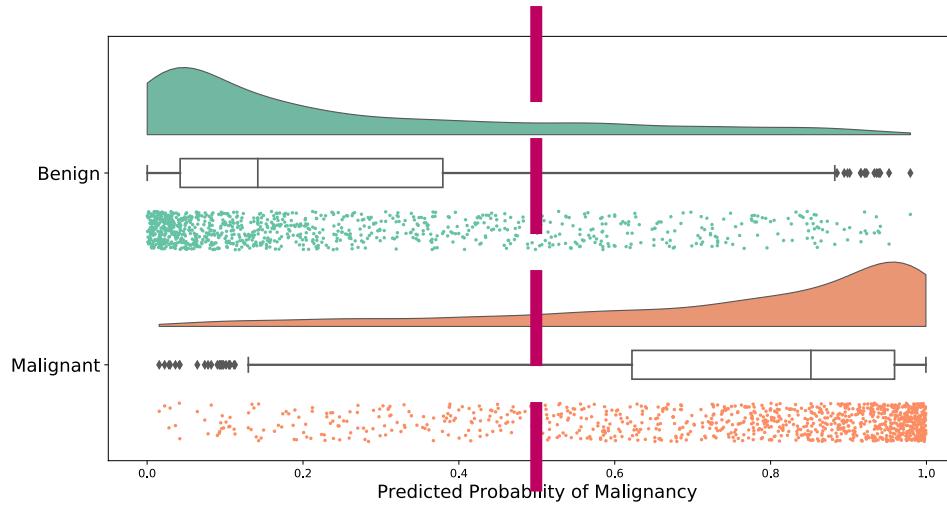


Healthcare Scenarios

1. A computer vision model that detects carcinoma
2. An algorithm that detects atrial fibrillation in Apple Watch users
3. An EHR-based model that monitors autism risk

Operating Point:

balanced



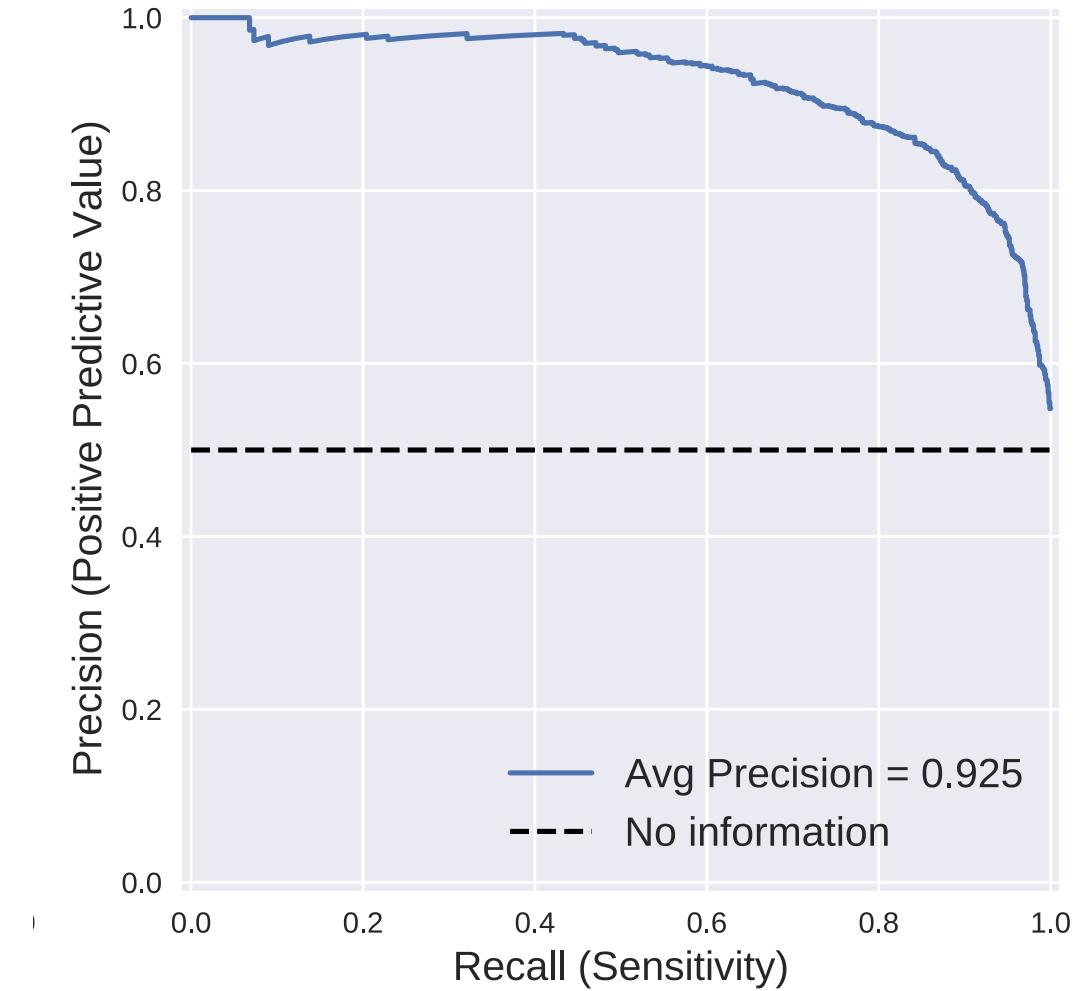
Healthcare Scenarios

1. A computer vision model that detects carcinoma
 2. An algorithm that detects atrial fibrillation in Apple Watch users
 3. An EHR-based model that monitors autism risk
-
- Sometimes specificity and sensitivity are difficult to interpret, particularly for rare conditions or events.
 - The most clinically relevant measure is often the positive predictive value (or negative predictive value).

The Precision-Recall Curve

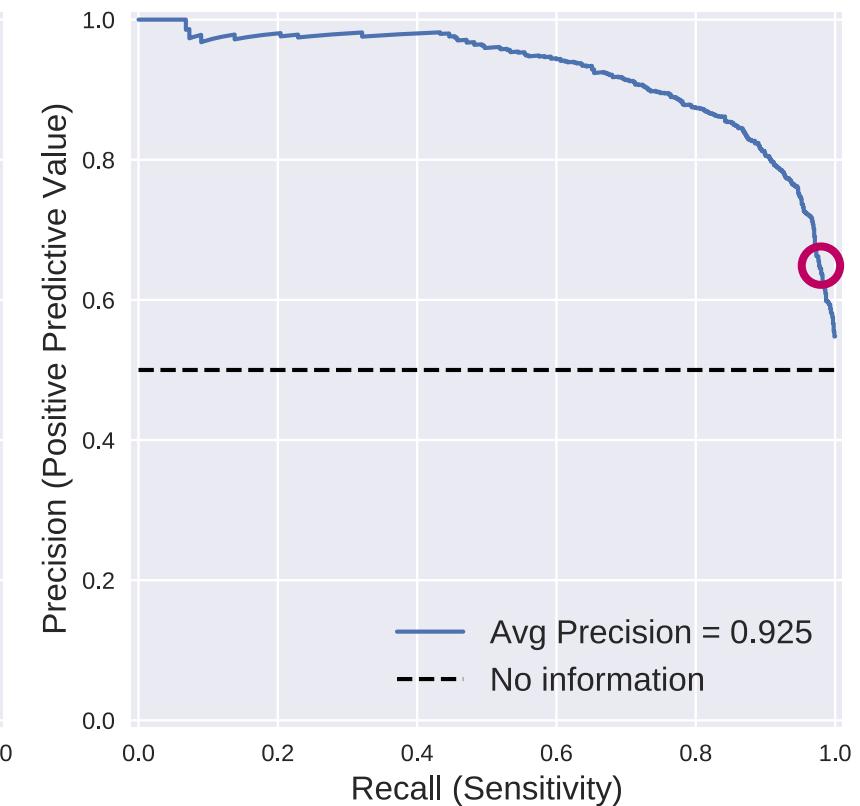
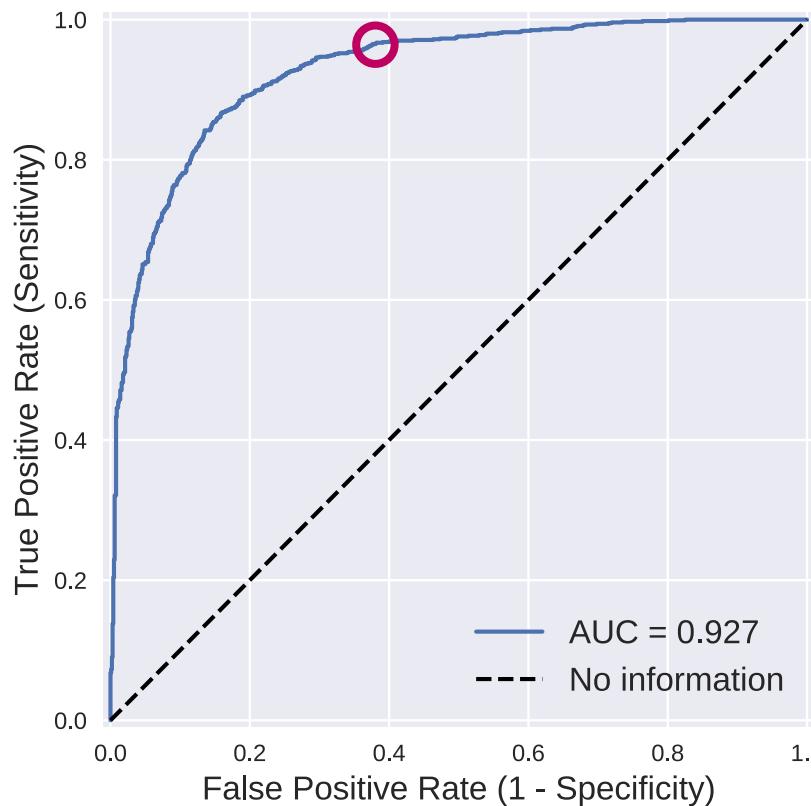
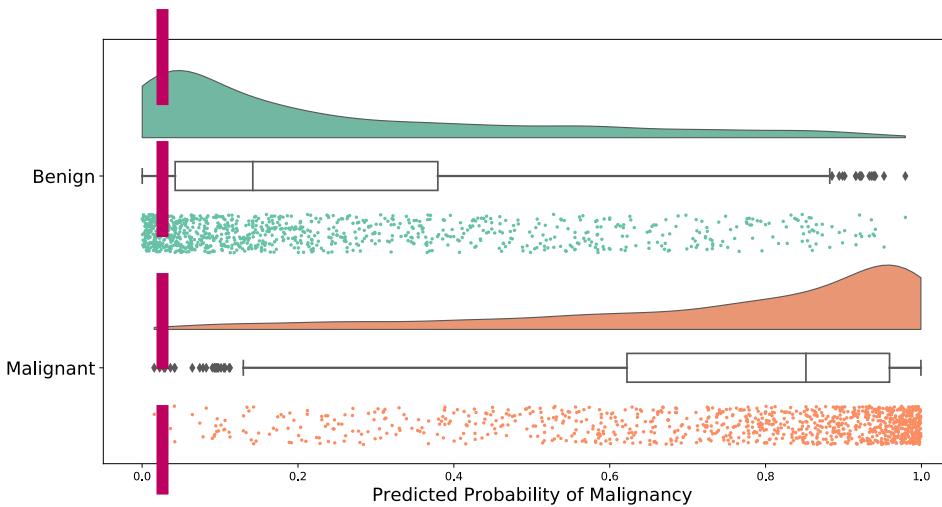
(i.e., PPV-Sensitivity Curve)

- Often has greater direct clinical relevance than the ROC curve
- The *no information* classifier always achieves PPV equal to the *base rate, or prevalence* (why?)
- PPV as well as the area under this curve (average precision) must be interpreted relative to prevalence



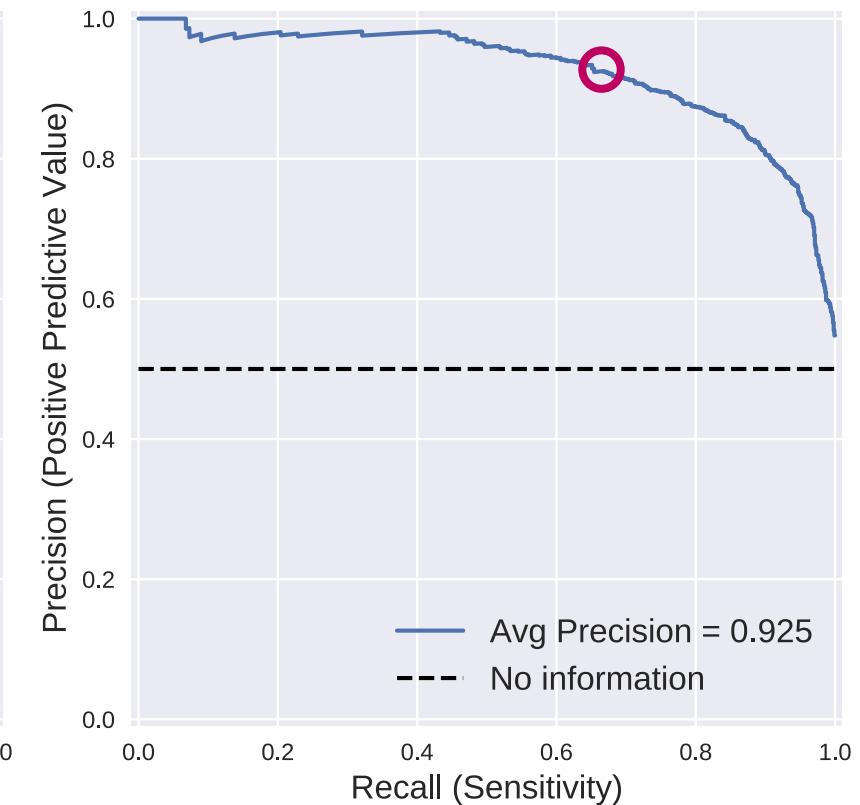
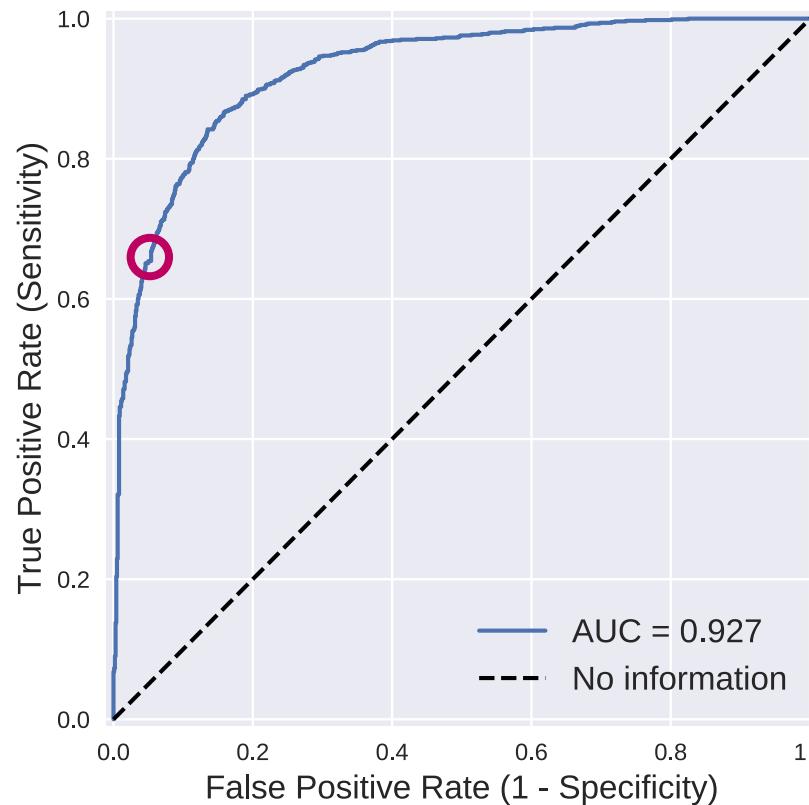
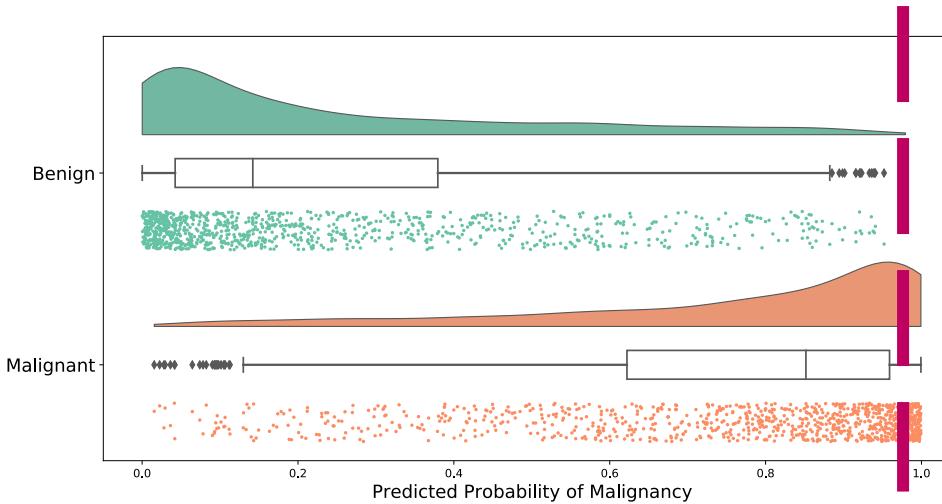
Operating Point:

high sensitivity



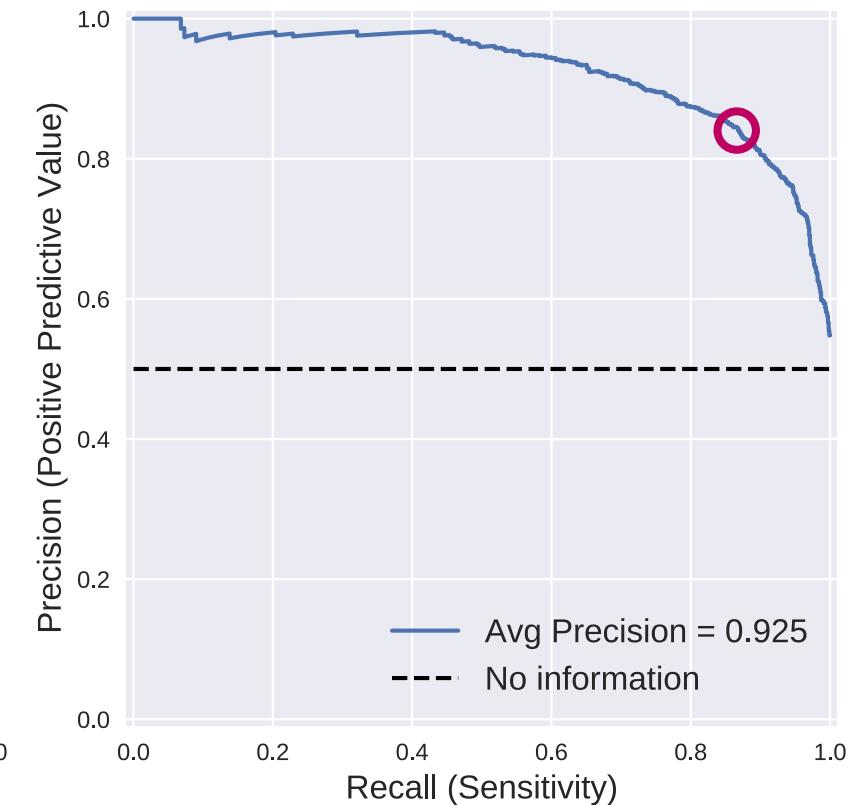
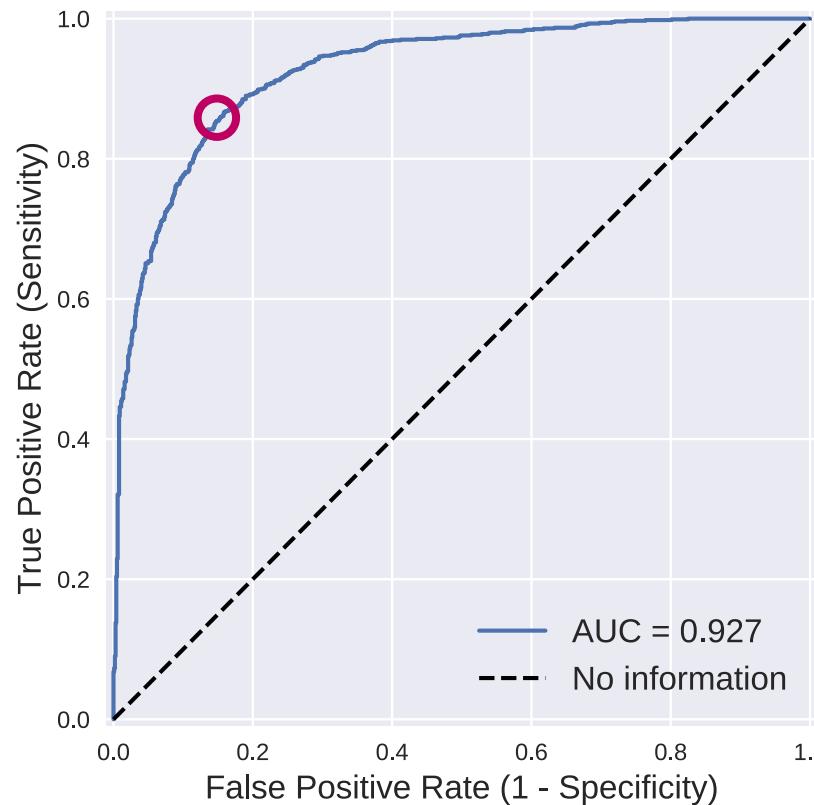
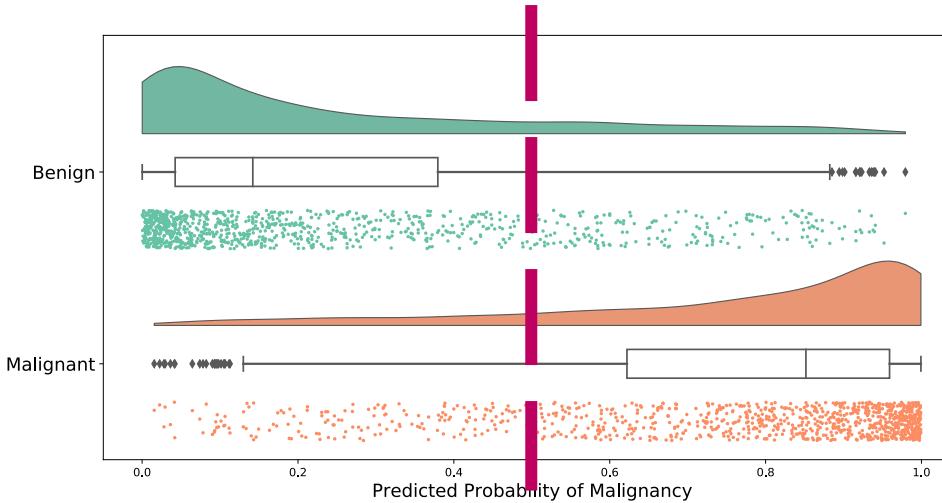
Operating Point:

high specificity



Operating Point:

balanced



Summary

- It is critical to understand performance measures in order to critically evaluate models and put them to clinical/healthcare use.
- To contextualize performance, we often compare models to a *no information* model whose predictions are random.
- However, *good* performance depends on existing alternative approaches, both tech- and non-tech-based, and the clinical scenario.
- Which measure is most important also depends on the clinical scenario.

Confusion Matrix for Multi-Class Problems

If time allows.

Multi-class problems: “Confusion Matrix”

		Predicted Label				
		1	2	3	4	5
True Label	1	True Positive	False Positive	False Positive	False Positive	False Positive
	2	False Negative	True Positive	False Positive	False Positive	False Positive
	3	False Negative	False Negative	True Positive	False Positive	False Positive
	4	False Negative	False Negative	False Negative	True Positive	False Positive
	5	False Negative	False Negative	False Negative	False Negative	True Positive

Multi-class problems: Binary for Label 1

		Predicted Label				
		0	1	2	3	4
True Label	0	True	False	False	False	False
	1	False	True	False	False	False
	2	False	False	True	False	False
	3	False	False	False	True	False
	4	False	False	False	False	True
	5	False	False	False	False	False

Multi-class problems: Binary for Label 2

		Predicted Label				
		0	1	2	3	4
True Label	0	Yellow	Blue	Blue	Blue	Blue
	1	Blue	Yellow	Blue	Blue	Blue
2	Blue	Blue	Yellow	Blue	Blue	Blue
3	Blue	Blue	Blue	Yellow	Blue	Blue
4	Blue	Blue	Blue	Blue	Yellow	Blue

There are many more, of course, but classification metrics go a long way.

- Regression
 - Mean squared error (MSE)
 - Mean absolute error (MAE)
 - R^2
- Survival Analysis (i.e. failure time)
 - Concordance index
 - MSE, MAE
 - Brier Score
 - AUC_t