1. Performance metrics are better on training data than test data because the model has been explicitly designed to model the training data. In extreme cases the results are so much better on the training data than the test data that we say we have overfitted the model.

2. Training and test data should be selected from the same data source. You should select the data to be used for training at random.

3.

	Predicted healthy	Predicted ill
Actual healthy	85	5
Actual ill	3	7

3a.

$$True\ Positive\ Rate = Sensitivity = \frac{True\ Positive}{Positive} = \frac{85}{90}$$

3b.

$$True\ Negative\ Rate = Specificity = \frac{True\ Negative}{Negative} = \frac{7}{10}$$

3c.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Population} = \frac{92}{100}$$

3d.

$$Positive\ Predictive\ Value = Precision = \frac{True\ Positive}{Predicted\ Positive} = \frac{85}{88}$$

3e. According to Wikipedia, recall = sensitivity, so see 3a.

$$\text{True positive rate (TPR, Sensitivity, Recall)} = \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$$

4a. A threshold of zero means all inputs result in a prediction of 1. This means we capture all the actual cases but also falsely categorise all the 0 cases as true as well. This puts us at 1,1.

4b. A threshold of zero means we assert that all inputs result in a prediction of 0. This means we do not correctly detect any true 1s, but also do not incorrectly label any 0s as 1s. This puts us at 0,0.

4c. The threshold of 0.5 could be in many places on the graph, depending on the classifier used. The straight line from 0,0 to 1,1 is the line of no discrimination, so one would hope the 0.5 threshold was above this line, otherwise the classifier might be outperformed by a random guesser!

5a. Given accuracy of 0.7, we have:

$$True\ Positives + True\ Negatives = 700$$

$$False\ Positives + False\ Negatives = 300$$

We can substitute these into equations for the true positive rate and the false positive rate:

$$True\ Positive\ Rate = Sensitivity = \frac{True\ Positive}{Positive} = 0.8 = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate = \frac{False\ Positive}{Positive} = 0.4 = \frac{FP}{TN + FP}$$

We now have 4 equations and 4 unknowns. Solving them, we get:

$$TP = 400, TN = 300, FP = 200, FN = 100$$

	Predicted positive	Predicted negative
Actual positive	400	100
Actual negative	200	300

5b. We can't say anything about the probability threshold. The model assigns each observation a probability, and we move the threshold to move along the ROC curve. Without the ROC curve we cannot determine what threshold was set that resulted in these numbers.