Fairness Worksheet

Solution Sheet

Zak Varty

The questions on this sheet are designed to let you test your own understanding of the course content on fairness. Some questions will test basic notions, while others will encourage you to think more deeply about some of the concepts introduced this week.

A pharmaceutical company has developed three new tests to screen patients for Coeliac disease (a condition where your immune system attacks your own tissues when you eat gluten), which are rapid and cost-effective. Receiving a diagnosis of Coeliac disease usually requires multiple visits to a general practitioner or specialist. This leads to long delay times between the first appointment and diagnosis. It is hoped that this delay can be reduced by the new tests, if they are effective.

Each of the three tests were administered to n people of whom m had been diagnosed and $n - m$ had no negative reactions to consuming gluten. For each individual $i = 1, \dots, n$ and each test $j = 1, 2, 3$, let:

- Y_i be an indicator that individual i has Coeliac disease;
- $D_{j,i}$ be an indicator that test j returns a positive diagnosis for individual i;
- A_i be a random variable denoting whether individual i belongs to subgroup "a" (children under 18), "b" (adults over 18 but under 65), or "c" (adults over 65).

Question 1: Confusion matrix terminology

Using the notation introduced previously for test 1, define and interpret in plain language:

1. The true positive rate (TPR) of the test,
2. The false positive rate (FPR) of the test,
3. The true negative rate (TNR) of the test,
4. The false negative rate (FNR) of the test.

Solution 1

Let $k = \sum_{i=1} D_1, i$ be the number of positive diagnoses made by test 1 and $\mathbb{I}\{\mathbb{A}\}$ be an indicator of event A.

1. The true positive rate of test 1 is the probability of a person with Coeliac disease receiving a positive test result from test 1.

$$\text{TPR} = \Pr(D = 1|Y = 1) \approx \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 1 \ \& \ Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 1\}}.$$

2. The false positive rate of test 1 is the probability of a person without Coeliac disease receiving a positive test result from test 1.

$$\text{FPR} = \Pr(D = 1|Y = 0) \approx \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 1 \ \& \ Y_i = 0\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 0\}}.$$

3. The true negative rate of test 1 is the probability of a person without Coeliac disease receiving a negative test result from test 1.

$$\text{TNR} = \Pr(D = 0|Y = 0) \approx \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 0 \ \& \ Y_i = 0\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 0\}}.$$

4. The false negative rate of test 1 is the probability of a person with Coeliac disease receiving a negative test result from test 1.

$$\text{FNR} = \Pr(D = 0|Y = 1) \approx \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 0 \ \& \ Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 1\}}.$$

Question 2: PPV and NPV

How do the positive predictive value and negative predictive value relate to the rates defined in Question 1?

Solution 2

The positive predictive value is the the proportion of people who have a positive test result that have truly got Coeliac disease:

$$\text{PPV} = \Pr(D = 1|Y = 1) \approx \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 1 \ \& \ Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 1\}}.$$

The negative predictive value is the the proportion of people who have a negative test result that truly have not got Coeliac disease:

$$\text{NPV} = \Pr(D = 0|Y = 0) \approx \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 0 \ \& \ Y_i = 0\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 0\}}.$$

The positive and negative predictive values are complementary measures to the TPR and TNR, where the order of conditioning has been reversed.

Question 3: Calculating with Confusion Matrices

The confusion matrices for the three tests are given below.

Test 1	D = 1	D = 0
Y = 1	81	24
Y = 0	24	382

Test 2	D = 1	D = 0
Y = 1	73	55
Y = 0	32	351

Test 3	D = 1	D = 0
Y = 1	5	3
Y = 0	100	403

a) Calculate the TPR, FPR, TNR and FNR for test 1, showing your working clearly.

Solution 3(a)

Using notation as defined previously:

$$\text{TPR}_1 = \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 1 \ \& \ Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 1\}} = \frac{81}{105} \approx 77.1\%.$$

$$\text{FPR}_1 = \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 1 \ \& \ Y_i = 0\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 0\}} = \frac{24}{406} \approx 5.9\%.$$

$$\text{TNR}_1 = \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 0 \ \& \ Y_i = 0\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 0\}} = \frac{382}{406} \approx 94.1\%.$$

$$\text{FNR}_1 = \frac{\sum_{i=1}^{n} \mathbb{I}\{D_{1,i} = 0 \ \& \ Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = 1\}} = \frac{24}{105} \approx 22.9\%.$$

b) Calculate and state the TPR, FPR, TNR, and FNR for tests 2 and 3.

Solution 3(b)

By similar calculations we find that for test 2:

$$\text{TPR}_2 \approx 57.0\%, \ \text{FPR}_2 \approx 8.4\%, \ \text{TNR}_2 \approx 91.6\%, \ \text{FNR}_2 \approx 43.0\%.$$

And for test 3:

$$\text{TPR}_3 = 62.5\%, \ \text{FPR}_3 \approx 20.0\%, \ \text{TNR}_3 \approx 80.0\%, \ \text{FNR}_3 = 37.5\%.$$

c) Is it important that the same group of people took each of the three tests? Why or why not?

Solution 3(c)

The test rates calculated in 3(b) use sample proportions to estimate population probabilities that we are truly interested in. If a different group of people (sample from the population) were used for each test then differences in test rates may be attributable to either differences between the sampled groups or between the tests themselves. To control for this confounding effect, the same group should be used for each test.
Note: Using sampled groups with the same proportions of measured attributes would also be effective but would not control for differences in unmeasured attributes such as sex and age, which might impact test performance.

d) Calculate the sensitivity and specificity for each of these tests.

Solution 3(d)

Sensitivity and specificity are alternative names for the true positive rate (TPR) and true negative rate (TNR) that are often used when referring to binary classification problems in a medical setting. Therefore we have already calculated these in Question3(b).

Question 4: ROC curves

The receiver operating characteristic for a test plots the sensitivity of the test against 1 - specificity. This pair can be used to compare different tests for the same binary outcome. The ROC for test 1 is shown in Figure 1.

a) Which region of the ROC plot corresponds to a near-optimal classifier?

Solution 4(a)

Note that $1 - \text{specificity} = 1 - \text{TPR} = \text{FPR}$, and $\text{sensitivity} = \text{TPR}$. A near-optimal classifier will have a high TPR and a low FPR. This corresponds to a ROC value close to (1,0), i.e. in the upper left corner of the ROC plot.

b) What does it mean for a classifier to have an ROC on/above/below the line $y = x$?

Solution 4(b)

A classifier on the line $y = x$ has the same diagnostic ability as guessing randomly whether an individual has Coeliacs disease. A classifier above the line has better diagnostic ability than randomly guessing, while a classifier below the line performs worse than guessing randomly.
Note: A classifier that is worse than randomly guessing must be able to discriminate between the two groups, but is getting them the wrong way around. This means that it could be improved by simply reversing the diagnosis of each individual!

c) Add the ROC for tests 2 and 3 to Figure 1.

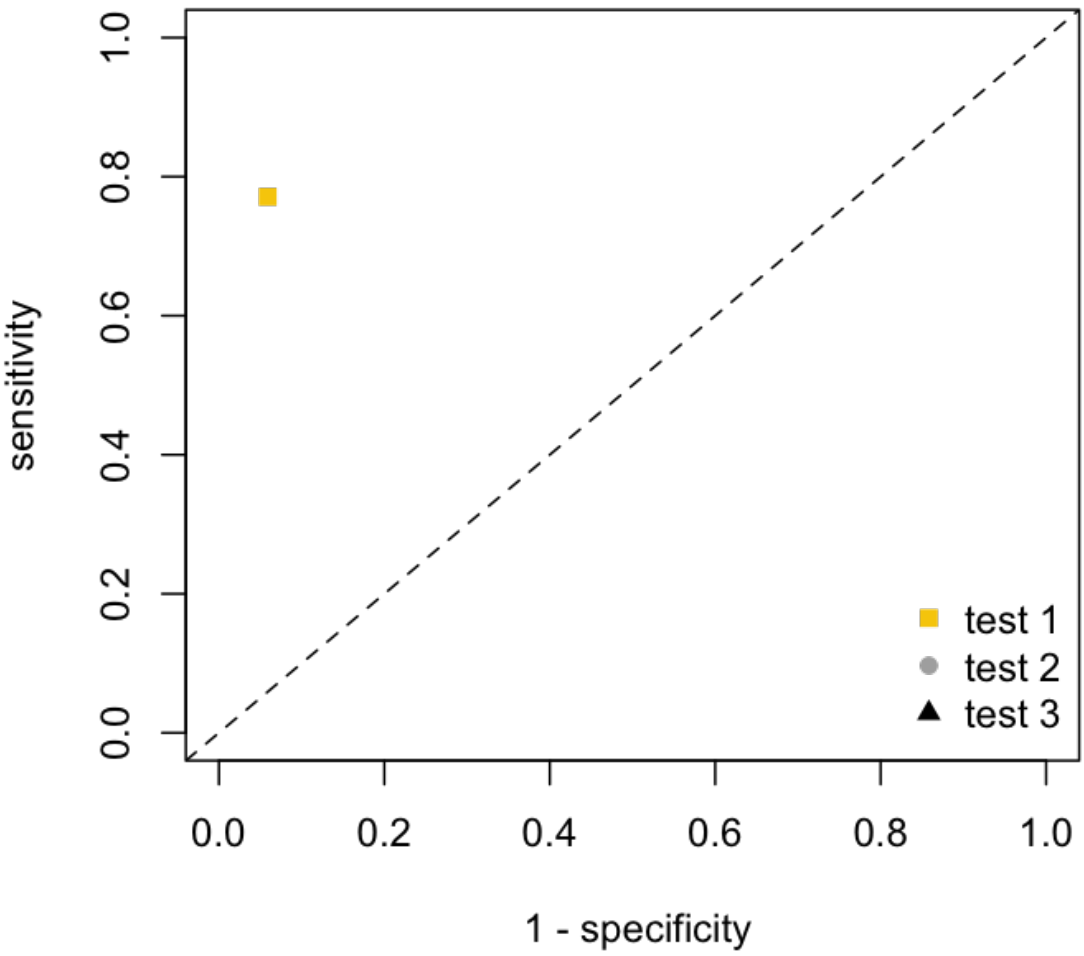

Figure 1: ROC plot for test 1

Solution 4(c)

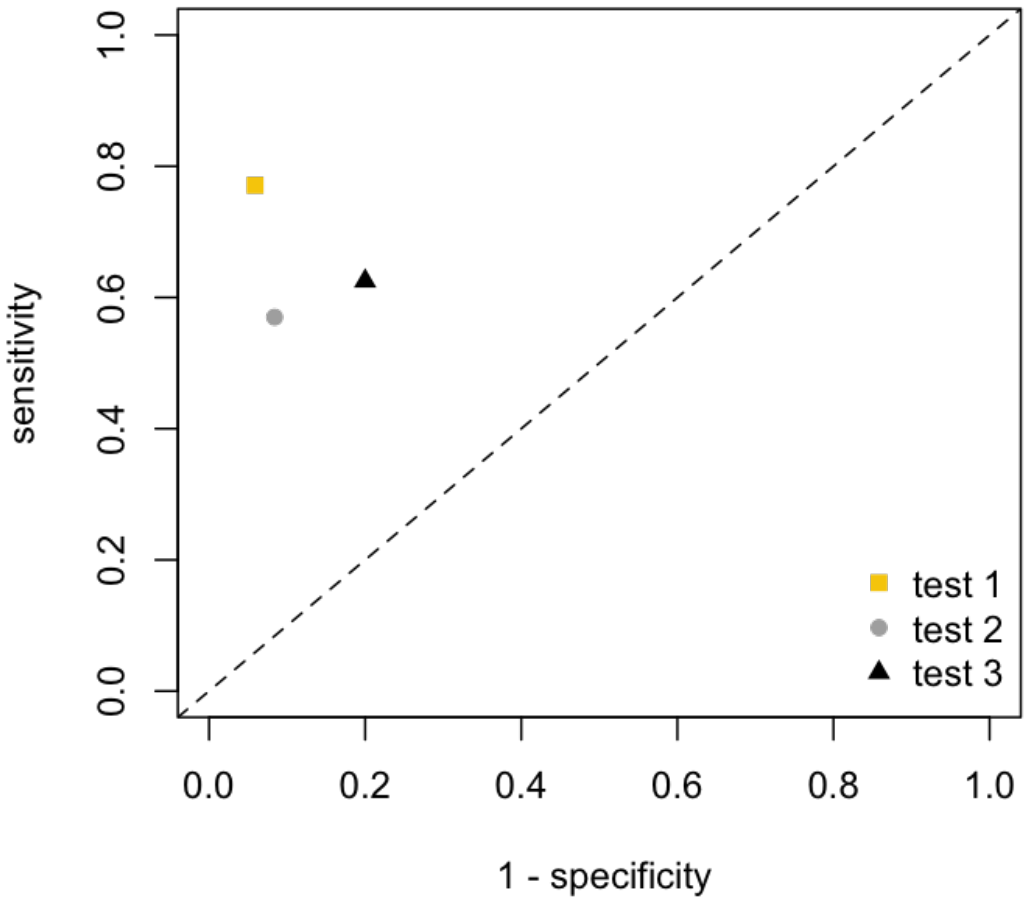

Figure 2: ROC plot for tests 1, 2 and 3

d) Which, if any, of the tests would you recommend that the company develops further?

Solution 4(d)

All three tests have some predictive power (and are therefore better than guessing at random). Based on these results, test 1 gives the best predictive performance and would be the most promising for further development.

e) By altering the concentration of an enzyme in the tests, the pharmaceutical company can change the threshold at which each test gives a positive diagnosis. The ROC curve for a test interpolates ROC values for each test at a range of enzyme concentrations. The ROC curves for tests 1,2, and 3 are shown in Figure 3.

When enzyme levels are optimised for predictive performance, which test has the best results? Does this change your earlier recommendation?

Solution 4(e)

By appropriately selecting the enzyme concentration used in each test, test number 2 can be made to have greater diagnostic ability than test 1.
Based on prediction considerations alone, this suggests that we should take test 2 forward for further development. However, this improved prediction might be out-weighed by other considerations such as the additional cost or supply problems if more enzyme is required per test.

f) The company decides to further develop test 2, having considered the cost, logistics and predictive performance of all three tests. A follow-up study is used to establish the effects of age on test outcomes. Three age groups are considered: group “a” represents people under 18, group “b” represents people 18-65 years old, and group “c” those over 65.

The ROC curves for test 2 are shown for people in age-groups “a”, “b” and “c” in Figure 4.

Interpret the ROC curves shown. You should use plain language suitable for an executive summary to the directors of the pharmaceutical company.

Solution 4(f)

At any enzyme concentration, test 2 provides the best diagnostic results to group a, the people under 18.
We might therefore pick the enzyme concentration to optimise the diagnostic ability of the test for either children or working age adults. Alternatively, enzyme concentration can be selected to satisfy a given trade-off between the diagnostic ability in these two groups.

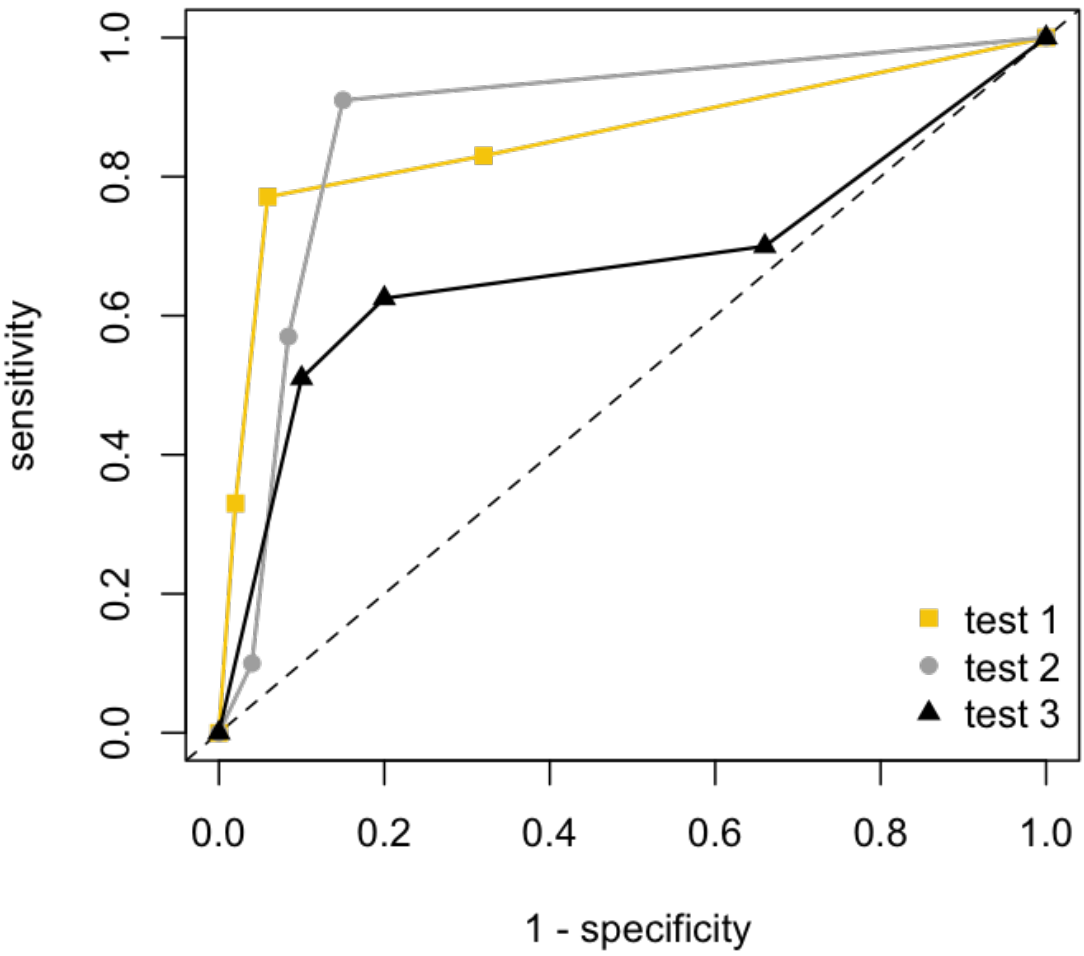

Figure 3: ROC curves for tests 1, 2 and 3

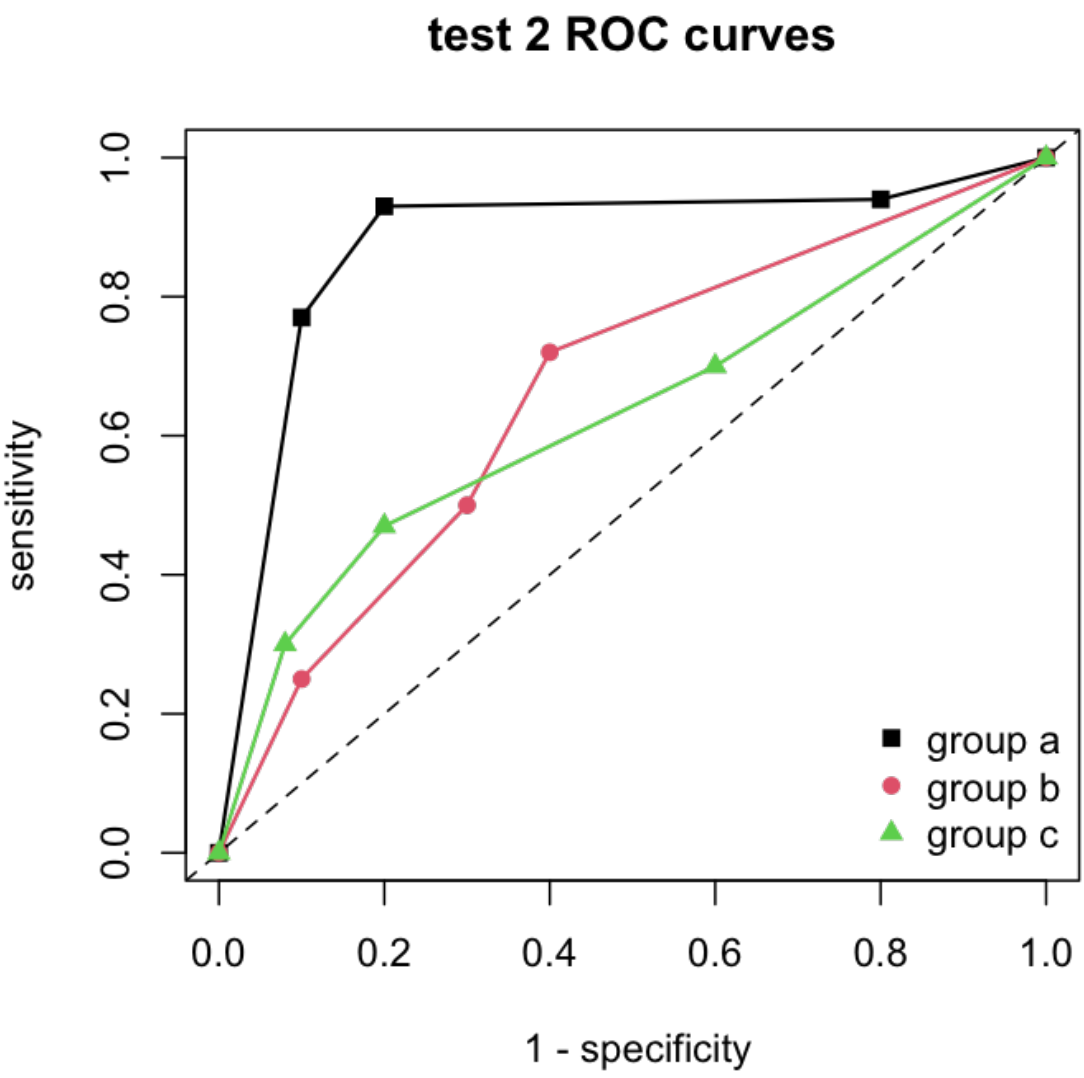

Figure 4: Sub-group specific ROC curves for test 2

g) The fairness condition of error-parity ensures that the false positive rates and the false negative rates of a test are equal for all sub-groups of a protected characteristic, such as age.

Shade the area of the ROC plot in which tests satisfying the error-parity fairness condition will be located. Explain why this is the case.

Solution 4(g)

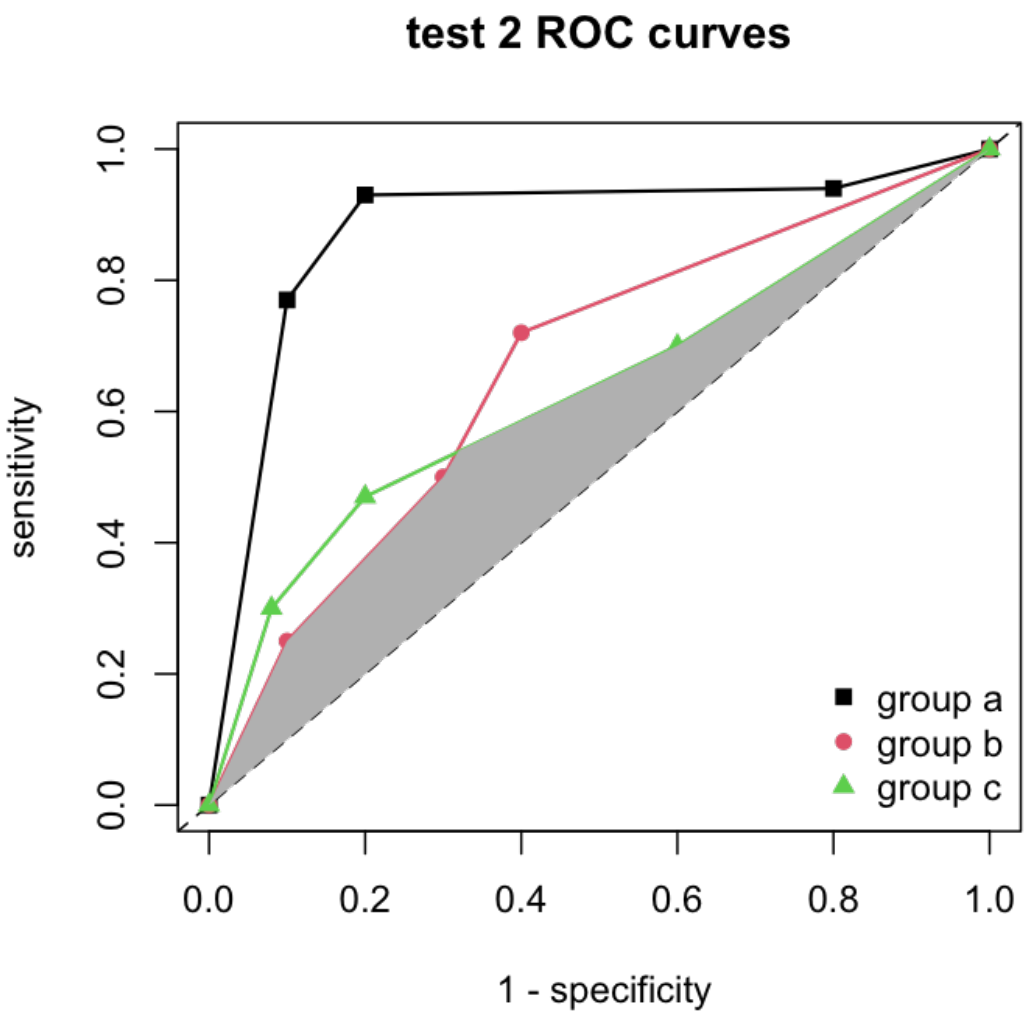

Figure 5: Sub-group specific ROC curves for test 2 with equal error region shaded

A test should not have a FPR > TPR, or else it is worse than random guessing and could be improved by switching all diagnoses.
If a test satisfies error parity then the false positive rate (and therefore the true positive rate) in each protected sub-group must be equal. This means the the combined false positive rate must be at least as large as in the worst-classified subgroup, or equivalently the true positive rate must be at least as low as in the worst-classified sub-group.
Therefore tests satisfying error parity fall between the line $y = x$ and point-wise minimum of the sub-group specific ROC curves.

h) What does your answer to 4(g) imply about the relative predictive performance of classifiers with and without an error-parity condition?

Solution 4(h)

A sub-group specific test is at least as good at diagnosing individuals within that subgroup as a combined test with error parity. By requiring equal error rates in all sub-groups we may be degrading diagnostic performance in one or more subgroups.
