Evaluating diagnostic tests: the area under the ROC curve and the balance of errors

David J. Hand
Imperial College, London
{d.j.hand@imperial.ac.uk}

Abstract:

Because accurate diagnosis lies at the heart of medicine, it is important to be able to evaluate the effectiveness of diagnostic tests. A variety of accuracy measures are used. One particularly widely used measure is the *AUC*, the area under the Receiver Operating Characteristic (ROC) curve. This measure has a well-understood weakness when comparing ROC curves which cross. However, it also has the more fundamental weakness of failing to balance different kinds of misdiagnosis effectively. This is not merely an aspect of the inevitable arbitrariness in choosing a performance measure, but is a core property of the way the *AUC* is defined. This property is explored, and an alternative, the *H* measure, is described.

Keywords: diagnosis, ROC curve, area under the curve, *H* measure

1. Introduction

Diagnosis is the first step in medical care. Correct diagnosis can lead to cure, and mistaken diagnosis to incorrect treatment, sometimes with serious consequences. For this reason, there is now an extensive literature on the evaluation of diagnostic tests: that is, on methods and measures to determine how effective a given test is. This literature is spread throughout the medical specialities, and also within the more general statistical and biostatistical literature. There are now several books devoted entirely to evaluating diagnostic tests [1,2].

The need for evaluation of tests arises from the fact that diagnosis is all too often not perfect. This may because accurate diagnosis is not possible at an early stage of a disease, or for a wide variety of other reasons. Often speed of diagnosis is important, so that treatment can begin as soon as possible. If a definitive diagnosis requires a lengthy procedure, then a quick and less accurate procedure may be used initially. Likewise, in many situations a definitive diagnosis may rely on an invasive procedure, and a premium is therefore placed on approximate non-invasive methods.

For expositional convenience, in this paper we will assume that there is a single disease which we are interested in diagnosing, referring to patients with this disease as ‘cases’, and contrasting it with others not suffering from the disease, whom we will refer to as ‘non-cases’ or ‘normals’. We will label the disease class as class 1, and the non-disease class as class 0. Generalisations to more elaborate situations, involving more than a simple case/non-case diagnosis are possible, but this is the most important special case and consideration of these other situations would detract from the essence of the discussion.

Implicit in many diagnostic tests is a measurement which is compared with a ‘classification threshold’. If the value of the measurement exceeds the threshold, then

the patient is diagnosed as belonging to the disease class, and otherwise to the non-disease class. More elaborate statistical diagnostic procedures may derive the 'measurement' as a statistical combination of a collection of other measurements - biological measurements, responses to items on a questionnaire, physical characteristics of the patient, and so on - but the result is the same: a score on a univariate scale, which is compared with a threshold to yield a diagnosis. The score is a random variable, varying from patient to patient, which we will denote by *Z*, with particular values *z*. We will denote particular values of the threshold by *t*. Patients in the disease class will then yield a distribution of scores $f_1(z)$ and those in the non-disease class a distribution $f_0(z)$, with respective corresponding cumulative distribution functions $F_1(z)$ and $F_0(z)$. Without loss of generality, we will assume that diseased patients tend to have higher scores and non-diseased patients lower scores.

The fact that diagnostic tests are generally not perfect means that they will misclassify some cases as non-cases and some non-cases as cases. In particular, all diseased patients with scores $z \le t$ will be misclassified, as will all non-diseased patients with $z > t$. These two types of misclassification, and their complements, lead to a variety of ways of capturing aspects of the diagnostic performance of the test. Important amongst these are *sensitivity* and *specificity*:

- the *sensitivity*, $S_e(t)$, or *true positive rate*, is the probability that a true case will be correctly classified as a case: $S_e(t) = 1 - F_1(t)$.

- the *specificity*, $S_p(t)$, or *true negative rate*, is the probability that a true normal will be correctly classified as a normal: $S_p(t) = F_0(t)$.

To produce a single measure of performance of a diagnostic test (by means of which tests can be evaluated and compared), sensitivity and specificity need to be combined. This can be done in an unlimited number of different ways. Since each way represents a different aspect of performance, it would be incorrect to assert that any one of these ways was 'wrong' and others 'right' - they merely measure different things - but, in general, one should match one's measure of performance to what one is trying to achieve. These and related issues are discussed further in [3, 4, 5, 6].

A very common approach to combining sensitivity and specificity is to add them in a weighted sum, so striking an appropriate balance between them - see, for example, [7, 8]. In fact, this balance is usually described in terms of the complements of sensitivity and specificity, with the weights taking into account the relative sizes of the case and normal classes as well as the relative severities of the two types of misclassification. That is, the numbers misclassified from each class are weighted according to the balance one wishes to strike between the two types of misdiagnosis, and then added together. This yields an overall measure of the effectiveness of the diagnostic test:

$$Q(t; c_0, c_1) = c_0(1-\pi)\left(1 - S_p(t)\right) + c_1\pi\left(1 - S_e(t)\right). \qquad (1)$$

Here π is the prevalence of the disease (that is, the proportion in the population who suffer from the disease) and c_0 and c_1 are the weights, representing the importance of

misclassifying a class 0 and a class 1 patient respectively. For convenience of exposition, we will refer to the c_i as misclassification *costs* and the weighted sum *Q* as a loss. The smaller the value of, *Q* the better is the diagnostic test.

Two versions of this measure, using default values of the (c_0, c_1) pair, are particularly widely used. Misclassification rate or error rate uses $(c_0, c_1) \propto (1,1)$ and is reported in [9], in a meta-analysis of classification studies, as being used in 'the vast majority' of comparative studies of classification rules (p19). The Kolmogorov-Smirnov statistic uses $(c_0, c_1) \propto (\pi, 1-\pi)$ and is especially popular in certain domains (it is, for example, the most popular measure for evaluating credit scorecards in the US).

An alternative to explicitly balancing sensitivity and specificity against each other is to fix one and let the other be the measure of performance. For example, one might choose to fix sensitivity at 0.80 and use the corresponding specificity as the measure of performance of the diagnostic test. While this is certainly sometimes done, sticking rigidly to this approach can be inappropriate, since the fixed value of sensitivity is an explicit operational choice, not something determined by considerations such as the relative importance of the two kinds of misdiagnosis. To see this, consider the following example.

Suppose we have two diagnostic tests, *T1* and *T2*, which give the following possible (S_e, S_p) pairs, the first pair of which, in each case, is the pair arising when sensitivity is fixed at 0.80:

T1: $(0.80, 0.90)$ and $(0.85, 0.10)$

T2: $(0.80, 0.90)$ and $(0.95, 0.89)$.

Using test *T1*, setting the sensitivity at 0.80 seems a reasonable choice, which one might well make, since the alternative of setting the sensitivity at the larger value of 0.85 incurs a dramatic reduction in specificity. Conversely, however, with *T2* it would seem harder to justify fixing sensitivity at 0.80, since a substantially larger value of 0.95 is associated with a tiny reduction in specificity. There is an intrinsic arbitrariness in setting sensitivity at a particular value.

The balance between sensitivity and specificity implicit in the definition of *Q* indicates what one believes are the relative misclassification severities. In contrast, if one chooses the sensitivity a priori, this choice is an indication of how one wishes to use the diagnostic test - it is not a question of belief about what the 'true' value of sensitivity might be, since this is a meaningless concept. The two approaches thus have rather different philosophical bases. That using the cost ratio is based on a ***belief*** about the nature of the problem, while that using sensitivity is based on a ***choice*** of an operating characteristic of the problem.

If the circumstances under which a diagnostic test is to be used are completely known, then the evaluation should take them into account. Often, however, the circumstances are not known and then a more general measure, based solely on the properties of the diagnostic tests and the diagnostic problem is needed. For example, one might be evaluating different diagnostic tests with a view to choosing one for future application, so that the circumstances of application are unknown, and a measure

which is invariant to different circumstances is preferable. That is the perspective we adopt in this paper: that the measure should be based solely on information about the diagnostic tests themselves, and about the diagnostic problem, as reflected by the relative severities of the two kinds of misdiagnosis, and not on what particular individuals might decide to adopt as values of operating characteristics.

Section 2 defines ROC curves, gives the background, and introduces our notation. Section 3 shows that a common measure of diagnostic performance based on the ROC curve, the *area under the curve*, or *AUC*, can be written as an average overall loss arising from misdiagnosis. However, it also shows that this interpretation of the *AUC* implies one must hold different beliefs about the relative severities of the consequences of misdiagnosis for different diagnostic tests, even though they are choosing between the same diagnostic classes. Section 4 presents an alternative measure which overcomes the problem. Section 5 gives an example, and Section 6 summarises the conclusions.

2. ROC curves and the area under the curve

The sensitivity and specificity of a diagnostic test are functions of the chosen threshold *t*, and, in general, changing *t* so as to increase the sensitivity will decrease the specificity, and vice versa. A very widely used way of displaying the values of the sensitivity and specificity as *t* is varied is by means of the *Receiver Operating Characteristic* (ROC) curve. This is a plot of sensitivity on the vertical axis and (1-specificity) on the horizontal axis (though variants of these axes are also sometimes used). ROC curves have the functional form

$$S_e = 1 - F_1\left[F_0^{-1}\left(S_p\right)\right] \qquad (2)$$

or, setting $y = S_e$ as the vertical axis and $x = 1 - S_p$, as the horizontal axis, $y = 1 - F_1\left[F_0^{-1}\left(1-x\right)\right]$. The ROC curve and its properties have been extensively studied and are well understood - see [10,11] for example.

Although monotonically increasing, ROC curves can have convex regions. That is, regions in which

$$S_e\left(\lambda t_1 + (1-\lambda)t_2\right) \le \lambda S_e\left(t_1\right) + (1-\lambda) S_e\left(t_2\right),$$

where t_1 and t_2 represent two values of the threshold. When such convex regions occur, it is possible to define a 'randomised' diagnostic test which yields a ROC curve corresponding to a dominating curve, and which has sensitivity which is always equal to or greater than that of the original test, for all values of specificity. We will not go into details here, since it will distract us from the core of the argument. Interested readers can refer to [12] (though note that, in accordance with machine learning conventions, that paper uses the terms 'convex' and 'concave' in the sense opposite to that used above - see the paper for details, and explanations). For simplicity of exposition, in what follows we will assume that the ROC curve is strictly monotonic increasing, and that the first derivative is strictly monotonic decreasing. We will also assume that the ROC curve is continuous and everywhere differentiable. These simplifying assumptions do not detract from the generality of the conclusions, and the program described at the end of Section 4 can handle general cases.

The area under the ROC curve, the *AUC*, can be expressed in various ways, including

$$AUC = \int_0^1 S_p dS_e = \int_0^1 F_0\left(F_1^{-1}(1-S_e)\right) dS_e \,, \qquad (3)$$

and it is clear that this is the mean specificity value, assuming a uniform distribution for the sensitivity.

The *AUC* has some attractive properties. It lies between 0 and 1, taking the value 1 for a perfect test and the value 0.5 for one which gives random diagnoses. It is non-subjective in the sense that all researchers, working with the same data, would obtain the same *AUC* measure. Note, though, that the choice of a uniform distribution over sensitivity implicit in (3) is an arbitrary choice: one could choose a different distribution. Indeed, one might feel that using the uniform distribution here is not just arbitrary, but is inappropriate, since it means one believes that the probability that very small values of sensitivity might be chosen is the same as the probability that very large values might be adopted. The choice of this distribution is an important point, to which we return below. The *AUC* has other attractive intuitive interpretations as well as being the mean specificity assuming uniform sensitivity. For example, it is equivalent to the Mann-Whitney-Wilcoxon test statistic, that is the probability that a randomly chosen member of class 0 (the healthy class) will produce a score lower than a randomly chosen member of class 1 (the disease class).

On the other hand, by definition, all summary statistics aggregate data in some way, and so sacrifice details. In particular, a well-known deficiency of the *AUC* arises when it is used to compare ROC curves which cross. If ROC curves cross, then one curve has larger specificity for some choices of sensitivity, and the other has larger specificity for other choices of sensitivity. Aggregating over all choices, as the *AUC* does, is all very well, but it could clearly lead to conclusions which misrepresent the performance of the diagnostic test as it is actually used (when some particular threshold value, and hence sensitivity and specificity *must* be chosen).

Partly in an attempt to overcome this problem, and partly in recognition of the fact that it is likely that not all values of sensitivity or specificity will be regarded as relevant, various researchers have proposed the use of the *partial AUC*, *PAUC*, in which the integration in (3) is not over the entire range of sensitivity (or specificity), but over some interval $[a,b]$ within $[0,1]$, regarded as of particular relevance (see, for example, [13]). Of course, this requires the user to specify *a* and *b*, which means that the non-subjective merit of the *AUC* is lost. In any case, it is entirely possible that the interval $[a,b]$ will include a point where the ROC curves cross. Furthermore, the *PAUC* has the unfortunate implication that values of sensitivity just outside the interval are discounted, whereas those just inside are included. An alternative solution would be to choose a smooth non-uniform distribution with support $[0,1]$.

3. *AUC* as average loss

In Section 1, we described the practice of striking a balance between sensitivity and specificity by weighting them in terms of the relative severities of the two kinds of misclassification. We supposed that misdiagnosing a sick person as healthy incurred a cost c_1, and that misdiagnosing a healthy person as having the disease incurred a

cost c_0. Implicit in the balance between sensitivity and specificity in the definition of Q is that assigning either a healthy or a sick person to the correct class incurs no cost. One can generalise the diagonal cost matrix implied by c_0 and c_1 to include costs associated with making correct classifications, but it is easy to show that this can be simplified to the diagonal case. It is also possible that different costs will be incurred when different healthy people are misdiagnosed, and that different costs will be incurred when different ill people are misdiagnosed. One can thus seek to generalise the following exposition by modelling cost in terms of the characteristics of the patients (see, for example, [6]), but here we stick to the simple case.

Since, for a given pair (c_0, c_1) the choice of threshold which minimises the balanced misclassification loss depends only on their ratio, and not on their absolute values, we can simplify things by requiring $c_0 + c_1 = 1$ and defining $c = c_0$ and $c_1 = 1 - c$.

In the Appendix we show that

$$AUC = \frac{1}{2\pi(1-\pi)} \int_0^1 Q\left(P^{-1}(c)\right) w(c)\, dc$$

where

$$w(c) = \left\{(1-\pi)\, f_0\left(P^{-1}(c)\right) + \pi f_1\left(P^{-1}(c)\right)\right\} \left| \frac{dP^{-1}(c)}{dc} \right|, \qquad (4)$$

$$Q(t) = \left\{(1-\pi) P(t)\left(1 - F_0(t)\right) + \pi\left(1 - P(t)\right) F_1(t)\right\},$$

and the relationship between c and t is given by P, defined as

$$c = \frac{\pi f_1(t)}{(1-\pi) f_0(t) + \pi f_1(t)} \equiv P(t). \qquad (5)$$

As the Appendix shows, $P^{-1}(c)$ gives the value of t which minimises Q for given c. (The assumption that the ROC curve is concave implies that P can be inverted.)

We see from this that, if we choose a distribution given by $w(c)$ for the cost c, and then choose the threshold (and hence sensitivity and specificity) to minimise the overall loss $Q\left(P^{-1}(c)\right)$ for each value of c, we obtain (a measure proportional to) the *AUC*. This is simply a mathematical fact. It provides another way of interpreting the *AUC*.

However, this way of looking at the *AUC* has implications. In particular, $w(c)$ in (4) depends on the empirical score distributions, via the mixture distribution and via the function P. That is, the distribution $w(c)$ will be different for different diagnostic tests. Now, as we saw in Section 1, c gives us a way of striking a suitable balance between sensitivity and specificity. One might therefore choose the distribution to reflect one's beliefs about how likely different c values were. But such a distribution cannot depend on the empirical score distributions. Doing so would be analogous to saying 'if you use test 1 to make a diagnosis then your misclassifications of the disease as normal are twice as serious as the reverse, while if you use test 2 they are

three times as serious.' This is clearly inappropriate: misdiagnoses incur the same relative cost no matter by what route they are arrived at.

Of course, for any given value of c it is not *mandatory* to choose the threshold (and hence sensitivity and specificity) to minimise the overall cost. That is, one does not have to choose the threshold to strike a balance between the two kinds of misclassification: one need not relate t to c via $t = P^{-1}(c)$, and sensitivity to c via $S_e = 1 - F_1\left(P^{-1}(c)\right)$. For example, in a screening application in which one had limited resources to treat identified cases, one might want to treat only a given number of patients, and hence (assuming known prevalence) fix the sensitivity to be used. In this example, the specificity is irrelevant, and no balance is made between the two kinds of misclassification. It follows that in this example the value of c would be irrelevant.

More generally, if one has a distribution of values for sensitivity (and hence of the threshold) which one feels one might choose, then this can be used, regardless of the distribution one feels would be appropriate for c. As we have already noted, the *AUC* does precisely this: it takes a uniform distribution over sensitivity. However, this choice hinges on considerations beyond the empirical score distributions and the balance between the severities of the two different kinds of misclassification. It requires an explicit choice of an operating characteristic such as sensitivity, or distributions of such characteristics. That is, it depends on how the diagnostic test is to be *used*.

If one has sufficient detail of a particular application to be able to choose the operating characteristics on external grounds (for example, one knows the desired sensitivity or its distribution) then such measures are appropriate. But even then note that, as the numerical example in Section 1 showed, the choice of operating characteristic might vary between diagnostic tests. In general, when the performance of a test is to be evaluated out of context of a particular application (for example, when developing a test which might be used in future applications), then we believe it is preferable to adopt a performance criterion which is independent of predetermined choices of operating characteristics (or their distributions). Rather, the criterion is better based solely on properties of the empirical score distributions and the relative severities of the two kinds of misdiagnosis. That is, we believe it is better to work with a particular $w(c)$ distribution and base the performance criterion on minimising the expected overall loss. Such a criterion is described in the next section.

4. An invariant alternative to the *AUC*

In the preceding section we argued that, in many situations, an appropriate measure of diagnostic test performance should be based on one's beliefs about the balance to be struck between the two different kinds of misclassification. We defined this balance in terms of c, and suggested that beliefs about likely values of c could be articulated in terms of a distribution over values of c, $w(c)$. But we then saw that the *AUC*, when expressed as an integral of the expected minimum overall loss, averaged over the distribution of c, led to different $w(c)$ distributions for different diagnostic tests. We

pointed out that this may not be desirable, since one might argue that the belief distribution over *c*, which is measure of the relative severity of the two kinds of misclassification, should not depend on the diagnostic test used.

This leads us to suggest an alternative performance measure, in which the $w(c)$ derived in Section 3 is replaced by a common belief distribution, $v(c)$, for all diagnostic tests. This leads to the measure

$$L = \int_0^1 Q\left(P^{-1}(c)\right) v(c)\, dc\,. \qquad (6)$$

For a given diagnostic problem, any particular researcher will choose the same $v(c)$, regardless of which diagnostic test is being evaluated. Different researchers, of course, may well choose different distributions for $v(c)$. L has the interpretation that it is the expected minimum loss if the value of *c* is unknown, but is to be chosen at random from the distribution $v(c)$.

We propose that two forms of $v(c)$ are used. Sometimes researchers may have beliefs about an appropriate shape for the distribution, and then a distribution which reflects these beliefs should be used. Indeed typically, they *will* have some beliefs, which can be used to select a distribution from some family (in a way very similar to choosing a prior in Bayesian analysis). For example, medical diagnostic problems often have a known asymmetry, in that misclassifying membership of one class is known to be more serious than the reverse kind of misclassification. Misdiagnosing a potentially fatal but easily treatable disease is more serious than misdiagnosing a harmless condition. If class 1 misdiagnoses are more serious than class 0 misdiagnoses, then $c < 1/2$, and $v(c)$ can be chosen to reflect this.

In addition, however, it is useful also to have a universal standard form for $v(c)$, which can be used as well as any particular chosen form. This will result in a measure which has the attractive property, like the *AUC*, that all researchers would obtain the same result from the same data. In [12] it is suggested that a beta distribution be used: $v(c) = c^{\alpha-1}(1-c)^{\beta-1} \Big/ \int_0^1 c^{\alpha-1}(1-c)^{\beta-1}\, dc$, with $\alpha = \beta = 2$. (In fact, by varying the parameters, the beta distribution gives a nice flexible family which may also be used for many situations when something is known or believed about likely values of *c*.) The choice of the beta distribution is arbitrary, but it is impossible to avoid some arbitrariness in the choice: different measures reflect different aspects of performance, and there is no absolute best choice. The choice of $\alpha = \beta$ is a deliberate attempt to avoid injecting beliefs about which of the two misdiagnoses is more serious - since these may differ between researchers. The choice of $\alpha = 2$ is arbitrary, but again, since there is no absolute way of choosing a 'best' value, some arbitrariness must remain. However, these arbitrarinesses - the shape of the distribution and the choice of the value of the common parameter - are fundamentally different from the implications of using the *AUC*, with its implication of integrating over a relative cost distribution which varies between diagnostic tests.

Finally, it is convenient to standardise L, dividing it by its maximum possible value and subtracting it from 1. In [12] this standardised measure is called the H measure. That paper discusses estimation of H and issues associated with non-concave ROC curves. An R program for calculating the H measure (along with the AUC, the area under the concave hull, and the Kolmogorov-Smirnov measure) is given in http://stats.ma.ic.ac.uk/d/djhand/public_html//. (Note that the program treats class 0 as the 'case' class, with class 0 tending to take smaller values. To use it for data in which class 1 represents cases, and where cases tend to take higher scores, it is necessary to invert the score ordering (e.g. score → max(score)-score) and to relabel the classes (0 → 1, 1→ 0).)

5. Example

As we have noted above, different criteria measure different aspects of performance, so that there is no absolutely 'right' one. Since the AUC and the H measure are defined in different ways, it follows that they will give different results, and perhaps different rank orders, to different diagnostic tests. In a sense, then, an example is unnecessary: it will merely show that different measures lead to different results. Nonetheless, for completeness, and to illustrate that the AUC and the H measure can indeed rank diagnostic instruments differently, hence leading to adoption of different methods, and thus to consequences in terms of misdiagnosis, we provide one such simple illustration.

The data analysed here describe 846 women between the ages of 48 and 81, 65 of whom were suffering from osteoporosis and 781 of whom were not, according to lateral thoracolumbar spine radiography. The aim was to construct a non-invasive screening questionnaire, for use as a preliminary diagnostic instrument. Nine variables were available for this analysis, these being age, height, and the answers to the questions listed below

1) age
2) height
3) after the age of 45 have you ever broken (excluding severe trauma) a bone in your back?
4) have you lost any height over the last 20 years?
5) were you ever given hormone replacement therapy at any time after your periods had stopped?
6) how many children have you had?
7) has your thyroid gland ever been overactive?
8) how many pints of beer would you drink in an average week?
9) how many cups of tea and/or coffee do you drink per day?

Logistic regression models were fitted to all 511 possible non-null subsets of variables from this list of nine variables, simply using additive models and ignoring the possibility of predictive interactions. The AUC and the H measure were computed for each of these models. A scatterplot of the results is shown in Figure 1.

It is clear from this plot that, although there is correlation between the two measures, it is not dramatically strong - it is in fact 0.58. It is also clear that models with a very large range of *H* values have *AUC* values near to the model with largest *AUC*.

The model which leads to greatest *AUC* is not the same as the model which leads to greatest *H* value. In fact, in terms of the variables given above, the best *AUC* and *H* models are

$$AUC \sim 1 + 2 + 4 + 5 + 6 + 7 + 8 + 9 \quad (7a)$$

$$H \sim 2 + 3 + 5 + 6 + 9, \quad (7b)$$

respectively, and their ROC curves are shown in Figure 2. For these two models respectively, the *AUC* and *H* values are

$AUC = 0.694$ $H = 0.062$

$AUC = 0.645$ $H = 0.084$.

This example is interesting because a first glance at the ROC curves for the models in (7), shown in Figure 2, suggest that the curve for the first model dominates the curve for the second model. If one ROC curve does dominate another, then all performance measures give the same order of merit to the models. However, since all ROC curves necessarily coincide at the points $(0,0)$ and $(1,1)$, there is a good chance that they will cross near these points. This is the case in Figure 2, where it can be seen that the two curves do cross over near the bottom left of the plot. The crossing of the curves near this end point is crucial, and is sufficient for the *AUC* and *H* measure to give a different relative order for these two diagnostic tests. This merits some further explanation.

We see, from (1), that the loss for a particular S_e, S_p pair and value *c* is given by the inner product of $\left(c(1-\pi),-(1-c)\pi\right)$ and $\left(\left(1-S_p\right),-\left(1-S_e\right)\right)$. If we take the top left corner $(0,1)$ of the ROC square as the origin, then this loss is proportional to the length of the projection of the line from $(0,1)$ to $\left(\left(1-S_p\right),-\left(1-S_e\right)\right)$ onto the line through $(0,1)$ in direction $\left(c(1-\pi),-(1-c)\pi\right)$. This means that we can determine the loss (1) associated with any point on the ROC curve by the length of the projection of the point onto this latter line. In particular, for any given value of *c* (and π) we can find the score threshold - the point on the ROC curve - which minimises the loss. The slope of the line we project onto is $-(1-c)\pi/c(1-\pi)$, and, for our example, $\pi = 65/846 = 0.077$, so that the slope is equal to $D = -0.083(1-c)/c$.

The *H* measure and the *AUC* use different distributions over *c* when calculating an overall measure of performance (and, as we have seen, the *AUC* distribution differs between diagnostic tests). The standard *H* measure uses the beta(2,2) distribution, which is symmetric, with a mode at $c = 0.5$. A value of $c = 0.5$ corresponds to a *D* value of -0.083, a very shallow negative slope. The points on the ROC curves which minimise the length of projections onto such a line are indicated by the solid circles in Figure 2. That is, threshold values in the neighbourhoods of those indicated by the solid circles contribute most substantially to the loss (1) when *H* is used.

The *AUC* on the other hand, averages over a distribution of *c* given by (4). This is equivalent to averaging over a distribution of threshold values given by the mixture distribution of the scores for the two classes. For the two models given in (7), the threshold values corresponding to the modes of the mixture distributions are shown by the solid diamonds in Figure 2. These threshold values minimise the loss (1) for values of *c* given by 0.018 and 0.034, corresponding to *D* values of -4.479 and -2.332, respectively. We see that the *AUC* and *H* place greatest weight on very different ranges of costs *c* (and, indeed, that the two diagnostic rules place emphasis on different cost ranges). We also see from this example that the apparent dominance of one ROC curve can be misleading.

Figure 3 shows a scatterplot of the log scores, *z*, for the subjects, produced by these two models. If one chooses a classification threshold value for the horizontal (best *AUC* model) axis and also a classification threshold value for the vertical (best *H* model), then the scatterplot is divided into four quadrants. Points in the upper left and lower right quadrants constitute what is known as the *swap set*. This is the set of points corresponding to patients who would be assigned to different diagnoses by the two classifiers. The wide dispersion of the points (with a correlation of only 0.78) shows that the swap sets would be quite significant, unless the thresholds were placed very near to the extremes. That is, the two models would make different incorrect diagnoses.

In this example, the *AUC* implicitly uses a cost ratio weight distribution which is heavily skewed towards weighting case misclassifications more heavily than non-case misclassifications. This is not unreasonable - one would often regard misclassifying cases as non-cases as more serious than the reverse - and one might deliberately choose a distribution which would weight things in this way. However, what may not be so reasonable is that the *AUC* uses *different* weight distributions for the two models. This paper takes the view that, to make valid comparisons between diagnostic instruments, one needs to evaluate them in the same way. We therefore recommend (a) using a common weight distribution which reflects one's beliefs about the relative severity of the misclassifications, if one can choose a good one, and also (b) always including a standard distribution (the beta(2,2) distribution underlying the standard *H* measure), so that researchers have a common currency for discussion.

Incidentally, it has been pointed out in [14] that estimates of the area under the ROC curve based on the data used to derive the score function will tend to be optimistically biased. This is similar to the more well-known optimism of estimates of misclassification rate resulting when the same data set is used to construct a diagnostic rule and estimate its likely future misclassification rate. The same effect will apply for the *H* measure.

6. Conclusion

Diagnostic tests, of the kind considered in this paper, can make two kinds of misclassification: they can misdiagnose cases as normals, or the reverse. To enable straightforward comparisons to be made between the diagnostic effectiveness of different tests (and to estimate parameters when constructing diagnostic rules by combining symptom indicators) some way is needed to reduce things to a single

numerical measure of performance. This can be done in various ways. One way is to combine the two kinds of misclassification, in terms of their relative severity. Another is to fix one of them and let the other be the performance measure. We have argued that the second method requires arbitrary choices about the value at which to fix the first type of misclassification, and that this can depend on the diagnostic test (the numerical example in Section 1) and also on external peculiarities of the application (the screening example). We therefore prefer the first method - striking a balance between the severities of the two kinds of misdiagnosis.

Often (perhaps almost always), however, deciding exactly what this balance should be is difficult. We therefore suggested that, instead of picking a particular balance, one should take an expectation of the minimum weighted misclassification loss over a distribution of values of the weights defining the balance. This is analogous to the *AUC*, which is an average of specificity over a uniform distribution of sensitivity, except that we take the expectation over the severity balance between the two kinds of misdiagnosis, instead of over sensitivity. We then showed that the *AUC* itself could be expressed as an expectation over the minimum balanced misclassification loss. However, it turns out that the expectation is with respect to distributions of the relative severity of the two kinds of misclassification which differ between different diagnostic tests. This is simply a consequence of the way the *AUC* is defined. This seems inappropriate: one's beliefs about the relative severity of the consequences of the two kinds of misclassification cannot depend on the diagnostic test one happens to have chosen. It would mean that one could alleviate suffering simply by choosing a different diagnostic instrument. Although the *AUC* has been criticised on various methodological grounds (see, for example, [15, 16]) this interpretation suggests that it also has a core theoretical weakness, at least when viewed from some perspectives.

To overcome this problem with the *AUC*, we proposed an alternative measure, which fixes the distribution of the relative severity of the consequences of the two kinds of misdiagnosis. In fact, we suggested that two versions were used: (i) using expert or specialist knowledge about the implications of the two kinds of misdiagnoses when this is available; (ii) a universal standard, which could be used by everyone to give consistent and readily interpretable non-subjective results.

Although the word 'diagnosis' has been used throughout the discussion above, exactly the same issues arise in prognosis and other situations in which the aim is to assign a patient to one of two or more possible classes.

APPENDIX

We show that the *AUC* is a multiple (depending only on the disease prevalence and a constant) of the average minimum loss, for a particular distribution over the values of c.

By definition,

$$AUC = \int_{-\infty}^{\infty} S_p(t)\, dS_e(t).$$

This can be rewritten as

$$AUC = 1 - \int_{-\infty}^{\infty} F_0(t)\, dF_1(t)$$

$$= 1 - \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t} f_0(u)\, du \right) f_1(t)\, dt$$

$$= \int_{-\infty}^{\infty} \left(\int_{t}^{\infty} f_0(u)\, du \right) f_1(t)\, dt$$

which can be expressed as

$$= \frac{1}{2} \left\{ \int_{-\infty}^{\infty} \left(\int_{t}^{\infty} f_0(u)\, du \right) f_1(t)\, dt + \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t} f_1(u)\, du \right) f_0(t)\, dt \right\}$$

$$= \frac{1}{2} \left\{ \int_{-\infty}^{\infty} \left(1 - F_0(t)\right) f_1(t)\, dt + \int_{-\infty}^{\infty} F_1(t)\, f_0(t)\, dt \right\}$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} \left\{ \left(1 - F_0(t)\right) f_1(t) + F_1(t)\, f_0(t) \right\} dt$$

Now, simply multiplying and dividing by the mixture $W(t) = (1-\pi)\, f_0(t) + \pi f_1(t)$ within the integral, this is equal to

$$= \frac{1}{2\pi(1-\pi)} \int_{-\infty}^{\infty} \left\{ \begin{array}{l} \dfrac{\pi(1-\pi)}{(1-\pi)\, f_0(t) + \pi_1 f_1(t)}\, f_1(t)\left(1 - F_0(t)\right) \\ + \dfrac{\pi(1-\pi)}{(1-\pi)\, f_0(t) + \pi f_1(t)}\, f_0(t) F_1(t) \end{array} \right\} \left((1-\pi)\, f_0(t) + \pi f_1(t) \right) dt \,.$$

If we define

$$\frac{\pi f_1(t)}{(1-\pi)\, f_0(t) + \pi f_1(t)} \triangleq P(t)$$

and substitute it into the above, we can express the *AUC* as

$$\frac{1}{2\pi(1-\pi)} \int_{-\infty}^{\infty} \left\{ (1-\pi) P(t) \left(1 - F_0(t)\right) + \pi \left(1 - P(t)\right) F_1(t) \right\} \left((1-\pi)\, f_0(t) + \pi f_1(t) \right) dt \,.$$

Or, putting $Q(t) = \left\{ (1-\pi) P(t) \left(1 - F_0(t)\right) + \pi \left(1 - P(t)\right) F_1(t) \right\}$,

$$AUC = \frac{1}{2(1-\pi)\pi} \int_{-\infty}^{\infty} Q(t) W(t)\, dt \,.$$

Now, in fact a standard result shows that, for given *c*, the minimum loss is given by choosing $t = P^{-1}(c)$.

Finally, by making the change of variable from t to c using P, we obtain

$$AUC = \frac{1}{2\pi(1-\pi)} \int_{0}^{1} Q\left(P^{-1}(c)\right) w(c)\, dc$$

where $w(c) = \left\{(1-\pi)\, f_0\left(P^{-1}(c)\right) + \pi\, f_1\left(P^{-1}(c)\right)\right\}\left|\frac{dP^{-1}(c)}{dc}\right|$

That is, the *AUC* is proportional to the average minimum loss, if the distribution of values of c is given by $w(c)$.

Acknowledgements

This work was partially supported by a Wolfson Research Merit Award from the Royal Society.

References

1. Zhou X-H, Obuchowski NA, and McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.

2. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.

3. Hand DJ. Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society* 2002; **56**:1109-1117.

4. Hand DJ. Classifier technology and the illusion of progress (with discussion). *Statistical Science* 2006; **21**:1-34.

5. Hand DJ, Whitrow C, Adams NM, Juszczak P, and Weston D. Performance criteria for plastic card fraud detection tools. *Journal of the Operational Research Society* 2008; **59**:956-962.

6. Eguchi S and Copas J. A class of logistic-type discriminant functions. *Biometrika* 2002; **89**:1-22.

7. Baker SG, Cook NR, Vickers A, and Kramer BS. Using relative utility curves to evaluate risk prediction. *J. R. Statist. Soc. A* 2009; **172**:729-748.

8. Gail MH and Pfeiffer RM. On criteria for evaluating models for absolute risk. *Biostatistics* 2005; **6**: 399-412.

9. Jamain A. *Meta-analysis of Classification Methods*. PhD thesis. Department of Mathematics, Imperial College, London, 2004.

10. Krzanowski WJ and Hand DJ. *ROC Curves for Continuous Data*. Chapman and Hall: London, 2009.

11. Gönen M. *Analyzing Receiver Operating Characteristic Curves with SAS.* SAS Institute: Cary, NC, 2007.

12. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* 2009; **77**:103-123.

13. McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making* 1989; **9**:190-195.

14. Copas JB and Corbett P. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* 2002; **89**:315-331.

15. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115**:928-935.

16. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, and Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2008; **27**:157-172.

Figure 1: Scatterplot of *AUC* by *H* measures for models based on all 511 non-null subsets of variables from the osteoporosis data.

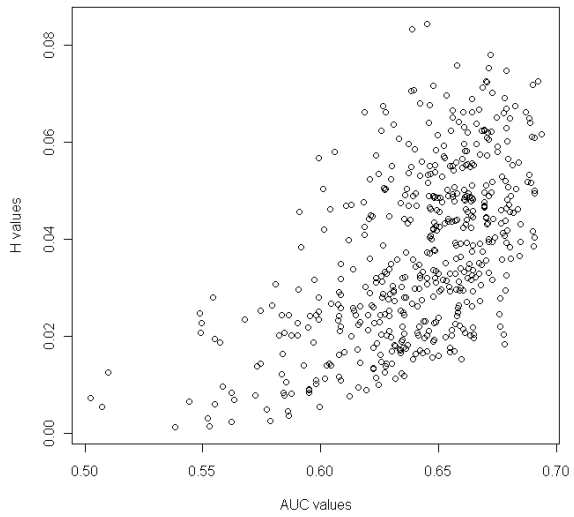

Figure 2: The ROC curves from the two diagnostic models

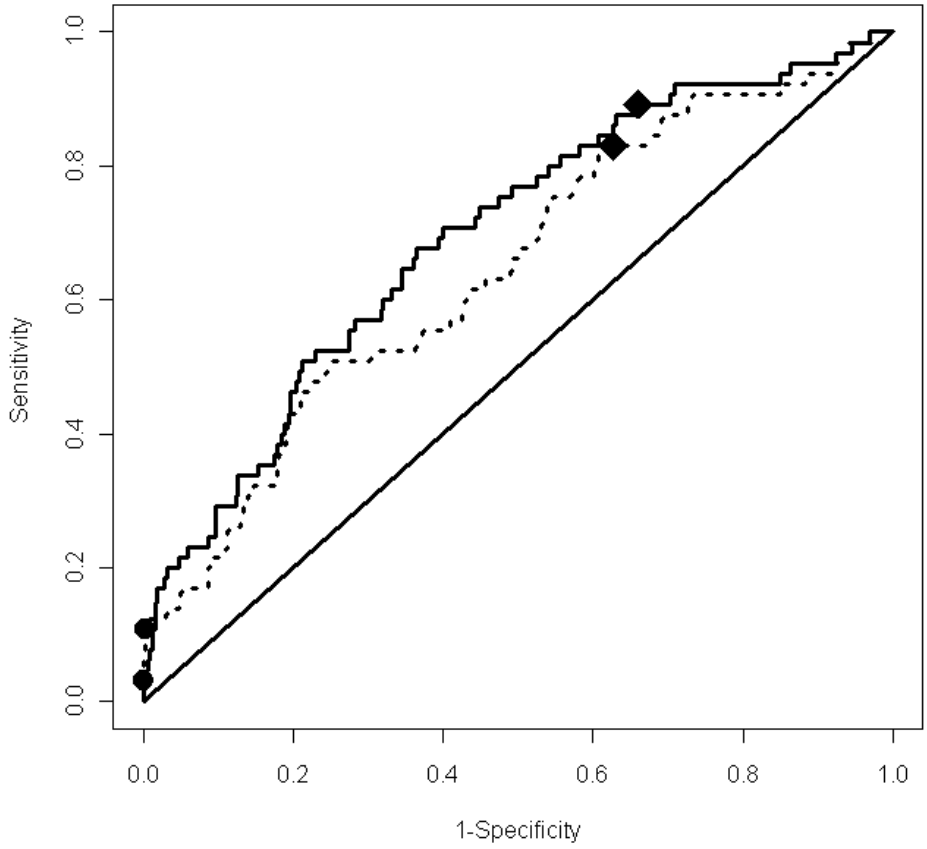

Figure 3: Plot of *H* vs *AUC* scores for the models in (7).