Optimal diagnosis: direct estimation of expansion curves in cost-effectiveness space

Authors and Affiliations:

Joanne Lord, Tanaka Business School, Imperial College London

George Laking, Manchester Molecular Imaging Centre, Christie Hospital

Alastair Fischer, Dept of Community Health Sciences, St. George's Hospital Medical School; and NICE

Author for Correspondence & Contact Details:

Joanne Lord,

Tanaka Business School,

Imperial College London

South Kensington campus

London SW7 2AZ

0207 594 9214

j.lord@imperial.ac.uk

Status of the work:

Work in progress / likely to be substantive changes

ABSTRACT

Diagnostic tests are conventionally evaluated with reference to some ‘gold standard’. The trade-off between false positives and false negatives can then be represented in an ROC curve. We have previously shown how this can be translated into cost-effectiveness space, tracing out a non-linear expansion path. This approach can be used to determine an economically optimal cut-off for treatment, and to evaluate alternative tests. We now show how a CE expansion curve may be estimated directly from trial data, without recourse to a gold standard test.

The method uses patient-level observations of costs and effects for a treatment and some comparator, along with an index D that reflects diagnostic or prognostic information. With two-sample data, as from a standard RCT, D can be used to match individuals from the experimental and control groups. Alternatively, group matching may be used. The matched data can then be used to estimate the association between D and the *incremental* costs and effects of the treatment. Hence we can estimate the diagnostic threshold that maximises net benefits. We use a simulation approach to demonstrate this approach.

Our method provides an alternative to standard non-economic methods for determining treatment thresholds (such as the maximum accuracy method). It avoids the need for a gold standard by applying an empirical and economic concept of ‘diagnosis’. The method may also be seen as an alternative to conventional stratified sub-group analysis used in economic evaluation.

Key Issues for Discussion by HESG Audience:

The concept of empirical diagnosis, role of the ROC curve in economic evaluation of tests, statistical aspects of curve estimation, and comparisons with stratified sub-group analysis

1 Introduction

The cost-effectiveness of therapeutic technologies and the cost-effectiveness of diagnostic technologies are interdependent [1]. A treatment can only be cost-effective if the 'right' patients are selected to receive it. Conversely, a diagnostic test can only be cost-effective if it helps to channel patients towards the 'right' treatment. Despite this, most economic evaluations of therapeutic interventions start only after the point of diagnosis, assuming that any procedures required to divide the patient population into the relevant groups and subgroups have already occurred. For example, many studies of the cost-effectiveness of cholesterol-lowering drugs ('statins') that stratify populations according to cardiac risk factor do not include the costs of ascertaining individuals' risk factors. In contrast, economic evaluations of diagnostic interventions start further 'up stream', and often explicitly model the relationships between diagnostic and therapeutic choices. For example, Garber and Solomon [2] used a Markov model to estimate the impact of coronary angiography and non-invasive tests for patients with chest pain. The tests differ not only in terms of their costs and risks, but also in their accuracy. They lead to different numbers and types of patients being referred on for the various medical and surgical treatments, with consequent effects on treatment costs and outcomes. Evaluation of the performance of diagnostic tests is thus an important stage on the path towards the evaluation of test-treatment pairs.

Diagnostic tests are conventionally evaluated with reference to some 'gold standard'. The trade-off between false positives and false negatives can then be represented in an ROC curve. ROC curves can be used to compare the performance of different tests and also to select a diagnostic cut-off for a given test. To maximise expected net benefits, one must operate at the point on the ROC where the gain from a marginal decrease in false positives is just balanced by the loss from the associated marginal increase in false negatives [3]. This economic decision rule can be illustrated by drawing indifference lines in ROC space (along which expected net benefits are constant) and finding the point of tangency with the ROC curve [4;5]. In another paper submitted for publication [6], we have shown that this same problem can be approached by translating the ROC to cost-effectiveness space. This traces out a non-linear cost-effectiveness expansion curve, which shows the locus of expected costs and effects as the diagnostic cut-off is progressively relaxed.

In this paper we use a simulation model to demonstrate how such a cost-effectiveness expansion curve could be estimated directly from clinical trial data, without recourse to a gold standard test. In the next section we review the foundations of the ROC, explain how it can be used to compare tests and to select treatment cut-offs, and introduce the idea of ROC-based cost-effectiveness expansion curves. In section three we explain our method for direct estimation of cost-effectiveness expansion curves from clinical trial data, and describe

the simulation model that we use to illustrate our approach. The results of the model are presented in section four. Finally, we discuss the strengths and weaknesses of our method and draw some conclusions (section five).

2 The theory of test evaluation and optimisation

2.1 Definition of the ROC

The conventional approach to evaluating diagnostic tests is based on comparison with some 'gold standard', which is assumed to perfectly divide the given population into two groups, those with a disease (Group 1, say) and those without (Group 0). Suppose that the outcome of a test is given by a diagnostic index D, such that a higher value of D indicates that the patient is more likely to belong to Group 1. Above some threshold D_T we say that the test result is 'positive', and conclude that the patient is a member of Group 1. Otherwise the result is 'negative' and the patient is assumed to belong to Group 0. Unless the test happens to be a perfect 'gold standard' itself, there must be some overlap between the distributions of D for the two groups (as illustrated in Figure 1). Hence, it is not possible to sort Group 0 from Group 1 purely on the basis of D. Whatever the chosen value for the threshold D_T there must always be some mistakes; false positive (FP) and/or false negative (FN) results. The performance of a test for a given threshold is usually summarised in terms of its sensitivity (the proportion of Group 1 patients who test positive) and its specificity (the proportion of Group 0 patients who test negative).

There is generally an inverse relationship between sensitivity and specificity. As the threshold is increased, sensitivity falls towards zero and specificity rises towards one. Conversely, as the threshold declines, sensitivity approaches one and zero. This relationship is traced out in the Receiver Operating Characteristic (ROC) curve, which is a technique developed in the early 1950s in the field of signal detection theory [7] (see Figure 2). The ROC shows all combinations of sensitivity ('hits') and (1-specificity) ('false alarms') that can be attained by varying the diagnostic threshold for a given test. The better the test, the closer the ROC curve lies towards the top left corner of ROC space, where both sensitivity and specificity equal one and diagnosis is perfect. The worse the test, the closer the curve approaches the diagonal line, which could be achieved by random guessing. A variety of parametric, non-parametric and semi-parametric methods have been proposed for ROC estimation [8-11].

Figure 1 – Distribution of the diagnostic index D for two sub-groups

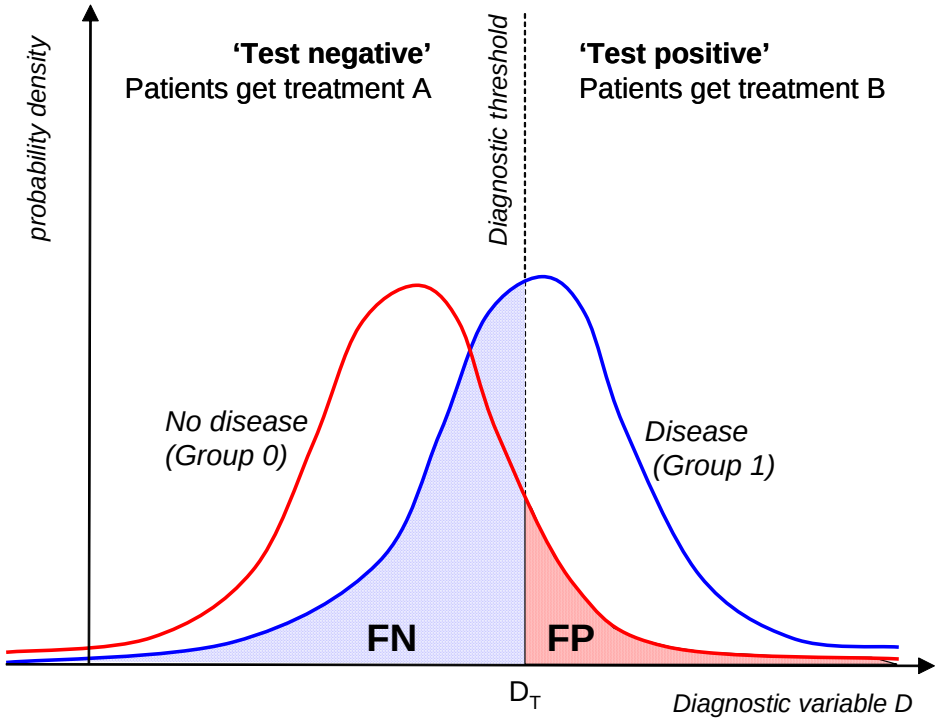
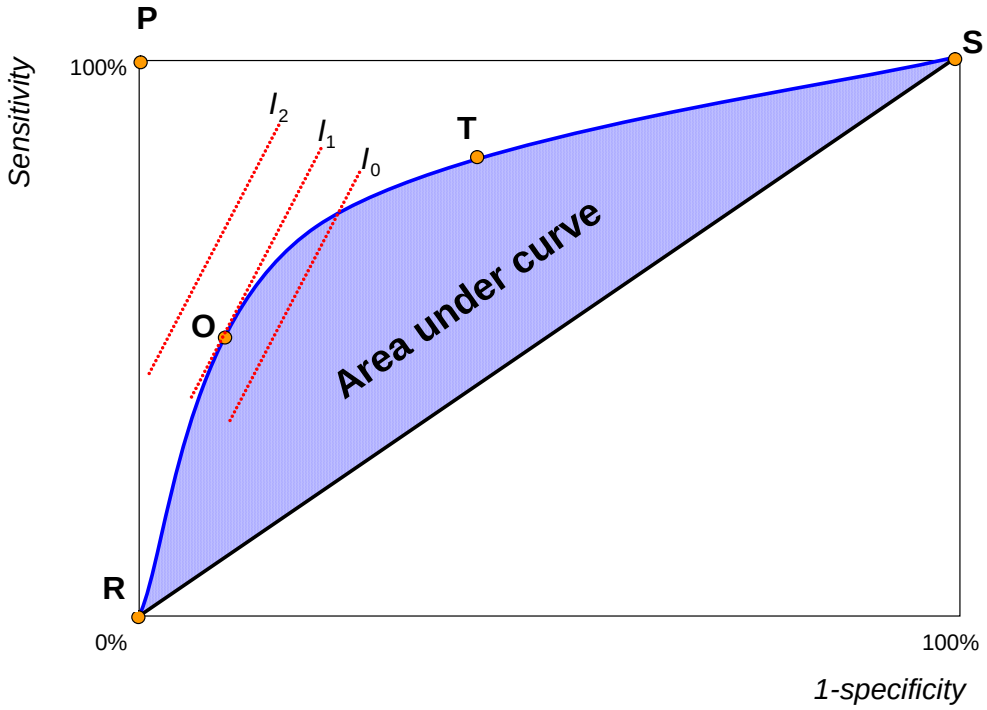


Figure 2 – The ROC curve for a diagnostic test



2.2 Decision rules for ROCs

ROC curves can be used to compare tests. If the curve for Test 1 lies entirely above and to the left of the curve for Test 2, then Test 1 is unequivocally better than Test 2. It is less clear

which test to use when the ROC curves cross. A rule of thumb that is often used in this situation is the Area Under the Curve (AUC) rule: that is, choose the test that contains the greatest area between its ROC curve and the diagonal line. This area can be interpreted as the probability that a randomly chosen person from Group 1 has a higher value of D than a randomly chosen person from Group 0 [12]. Comparison of AUC can be insensitive when two curves cross [13]. To avoid this problem, and also to avoid values of sensitivity and/or specificity that can be ruled out *a priori*, evaluation of partial AUC for selected regions of specificity has been used as an alternative decision criterion. However, neither total nor partial AUC bear any clear relation with the economic consequences of a test.

Even when a test has been chosen, the task of selecting a cut-off value for the test remains. A commonly used method is to identify the threshold that corresponds to the point of maximum accuracy on the ROC, where 'accuracy' is defined as the proportion of decisions that are correct. But like the AUC rule, the maximum-accuracy rule neglects the economic consequences of different types of mistake. Despite the lack of economic rationale for many such commonly used rules, they continue to be used in economic evaluations of diagnostic technologies [14-16]

As far back as 1978, Metz argued that the test cut-off should be selected so as to maximise economic net benefits [3]. He showed that this optimal cut-off corresponds to the point on the ROC with slope:

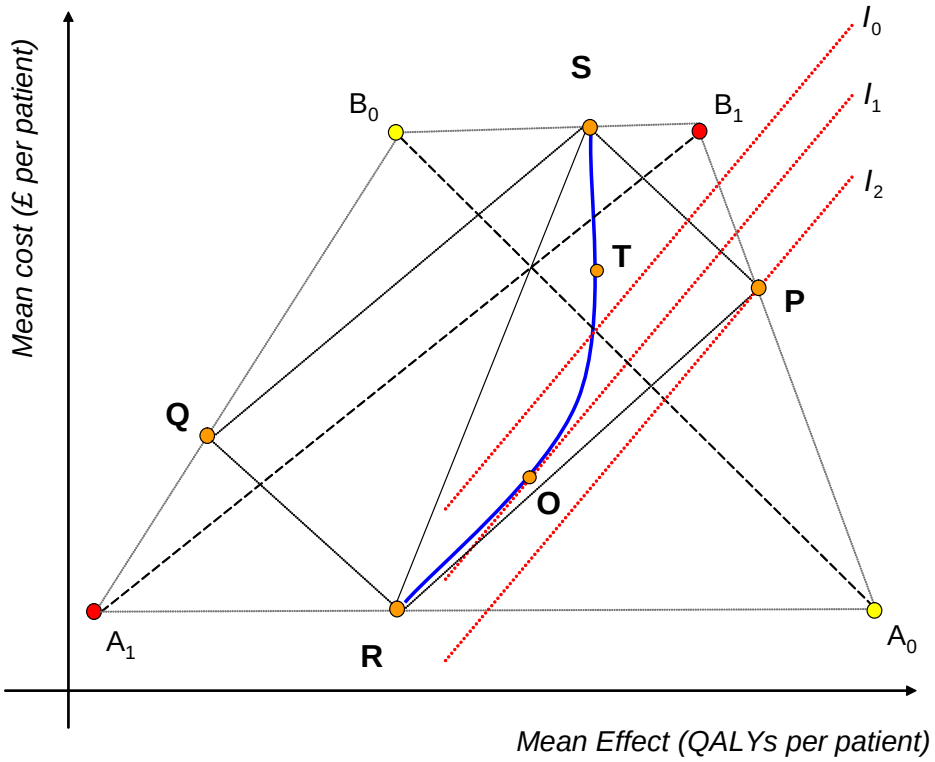
$$\frac{1-p}{p}\times\frac{L_{FP}}{L_{FN}} \qquad (1)$$

where p is the proportion of the relevant population who have the disease, and L_{FP} and L_{FN} are the expected opportunity losses for false positive and false negative cases respectively. The opportunity losses are defined as $L_{FP} = N_{TN} - N_{FP}$ and $L_{FN} = N_{TP} - N_{FN}$, where N_{TN}, N_{FP}, N_{TP} and N_{FN} are the expected net benefits for true negative, false positive, true positive and false negative cases respectively. Metz's optimal point can be illustrated as the point of tangency between the ROC curve and an indifference line with slope given by the above expression, along which the net benefit is constant (I_0, I_1 and I_2 in Figure 2) [4;5]. Higher indifference lines, which represent a greater level of net benefit and so are preferred, lie further towards to top left corner of ROC space ($I_0 \, B \, I_1 \, B \, I_2$). Halpern and colleagues used this method to compare alternative diagnostic tests on the basis of the maximum net benefit line that they could reach at their respective optimal operating points. For test comparison, the loss functions in Metz's expression must include the expected costs and health side effects of the tests, as well as those of the treatments.

2.3 Translating ROCs to cost-effectiveness space

In a recently submitted paper [6], we suggested that this same method could be translated to cost-effectiveness space. This has the advantage of being familiar ground for health economists, and now also for many decision-makers (for example, for the NICE appraisals committees). This method is illustrated in **Figure 3**.

Figure 3– Diagnosis in cost-effectiveness space



Suppose that there are two alternative (mutually exclusive) treatments available for the population of interest. The mean costs and effects for treatment A and B are shown at point R and point S respectively. Hence the standard incremental cost-effectiveness ratio (ICER) for B compared with A is the slope of the line RS. Now suppose that there are two sub-groups (Group 0 and Group 1) who differ in terms of how likely they are to benefit from A and B: on average the Group 1 patients attain a better outcome with treatment B (at point B_1) than they do with treatment A (point A_1), whereas the Group 0 patients benefit more from treatment A (point A_0) than from treatment B (point B_0). The slopes of the lines A_0B_0 and A_1B_1 are the ICERs for Group 0 and Group 1 respectively. Point P shows the mean costs and effects of treating all of the Group 1 patients with B and all the Group 0 patients with A. Conversely, point Q shows the mean costs and effects if all the Group 1 patients are given A, and all the Group 0 patients are given B. The area RPSQ thus represents the set of mean costs and effects that can be attained by treating the two subgroups with different combinations of A and B.

The standard method of choosing which treatments to offer to which patients would then depend on the threshold cost-effectiveness ratio ∂. This can be illustrated by means of net benefit isoquants (I_0, I_1, I_2), which show all combinations of mean costs and mean effects that yield a fixed level of net benefit for a given ∂[17]. These net benefit isoquants have slope equal to ∂. In cost-effectiveness space, isoquants further to the southeast reflect a greater level of net benefit and so are preferred (I_0 B I_1 B I_2). Suppose that ∂ lies between the ICER for Group 1 and the ICER for Group 0 (as illustrated in **Figure 3**). Then clearly the optimal point is P, where the Group 1 patients are given treatment B and the Group 0 patients treatment A.

However, point P is only attainable if the doctors can correctly classify the two groups of patients *prior* to the treatment decision. Consider the situation where there is a diagnostic test that can provide information about whether a particular patient is most likely to belong to Group 0 or to Group 1, and that is to be used to allocate patients to treatment A or B. Patients who pass the test ($D \varnothing D_T$) will be classified as Group 1 patients and allocated to treatment B. Patients who fail the test ($D \{ D_T$) are classified as Group 0 patients and allocated to treatment A.

The ROC curve for this test has an analogue within the RPSQ region in cost-effectiveness space, which is the curved line ROTS in **Figure 3**. Point R equates to a very high treatment threshold, such that all patients fail the test and are given treatment A, hence Se=0 and Sp=1 (the bottom left corner of the ROC). Conversely, at point S there is a very low treatment threshold, all patients pass and are given B, so Se=1 and Sp=0 (the top right corner of the ROC). Each intermediate point on the ROTS curve equates to a particular treatment threshold, and hence to a particular point on the ROC. The curvature of the ROTS curve is related to the curvature of the ROC. If the test were no better than random, we would be constrained to the straight line RS (the diagonal in the ROC). The better the test, the closer the ROTS curve approaches point P. If the test were perfect we correctly categorise all patients and reach point P where Se=1 and Sp=1 (the top left corner of the ROC).

The optimal point on a ROTS curve (point O) is the point of tangency with a net benefit isoquant, which is simply the point where the slope is equal to the cost-effectiveness threshold ∂. Here the gain in net benefit due to a marginal increase in the proportion of Group 1 patients correctly given treatment B is just balanced by the loss of net benefit due to the associated increase in the proportion of Group 0 patients incorrectly given treatment B.

ROTS curves can also be used to compare tests. The more cost-effective test is the one that reaches further towards the southwest, attaining the highest net benefit isoquant at its

optimal point. However, when comparing tests, it is important that the expected costs and any health side effects of the tests are taken into account.

3 Methods

3.1 Aims and objectives

The aim of this paper is to describe and test a method for directly estimating a cost-effectiveness expansion curve based on information from a diagnostic test (the ROTS curve) using patient-level randomised controlled trial data. The method is non-parametric and does not rely on the existence of a gold standard test. We use simulated data to illustrate our ROTS estimation method and to show how it can be used to estimate an optimal operating point for a given test and to compare alternative tests. We also use the simulation to test the performance of the direct ROTS estimation method in a variety of situations in comparison with a conventional ROC rule-of-thumb (the maximum accuracy rule) and Metz's method.

3.2 Outline of the model

We start by assuming that each member of a given population is to be allocated to one and only one of a pair of treatment options: treatment A or treatment B. The health effects of A and B for patients in this population are measured by two random variables E_A and E_B respectively (in QALYs, say). Similarly, C_A and C_B are random variables that measure the cost of A and B (in £, say). We assume that there are two sub-groups within the population, indicated by a binary random variable G=0,1, as measured by a 'gold standard' test. The proportion of group 1 patients in the population is p. Finally, the diagnostic test index is given by D. D does not have to be a continuous or cardinal measure, but must be measurable on at least an ordinal scale. Each member of the population is thus described by a vector of seven random variables: $(G, D, E_A, C_A, E_B, C_B)$.

3.3 Methods for non-parametric ROTS estimation

3.3.1 Estimation from cross-over data

Estimation is straightforward if we have access to patient level observations for the full set of variables $(G, D, E_A, C_A, E_B, C_B)$. Suppose that we have observations $(g_i, d_i, e_{Ai}, c_{Ai}, e_{Bi}, c_{Bi})$ for a sample of n patients from a cross-over trial (i=1,2,…,n). We assume initially that the diagnostic information contained in D is free and riskless.

To estimate the ROTS curve, we start by ranking the patients in order of decreasing D, so that if $d_i>d_j$ then i<j. In the case of a tie, $d_i=d_j$ the patients are ordered randomly.

The mean effectiveness of treatments selected by the test at threshold d_k, k=0,1,…,n, is:

$$\overline{E}_k = \begin{cases} \sum_{i=1}^{n} e_{Ai} \Big/ n, & \text{if } k = 0 \text{ (Point R)} \\ \left(\sum_{i=1}^{k} e_{Bi} + \sum_{i=k+1}^{n} e_{Ai} \right) \Big/ n, & \text{if } 0 < k < n \\ \sum_{i=1}^{n} e_{Bi} \Big/ n, & \text{if } k = n \text{ (Point S)} \end{cases}$$

Similarly, the mean cost with threshold d_k is:

$$\overline{C}_k = \begin{cases} \sum_{i=1}^{n} c_{Ai} \Big/ n, & \text{if } k = 0 \text{ (Point R)} \\ \left(\sum_{i=1}^{k} c_{Bi} + \sum_{i=k+1}^{n} c_{Ai} \right) \Big/ n, & \text{if } 0 < k < n \\ \sum_{i=1}^{n} c_{Bi} \Big/ n, & \text{if } k = n \text{ (Point S)} \end{cases}$$

The locus of points $\{\overline{E}_k,\overline{C}_k\}$ for k=0,1,…,n is our sample estimate of the ROTS for test 1.

The estimated net benefit with a test threshold of d_k, is $\overline{N}_k = \varsigma(\overline{E}_k - \overline{E}_0) - (\overline{C}_k - \overline{C}_0)$, which is measured with respect to point R. The maximum of $\overline{N}_k$ with respect to k is denoted by $\overline{N}^*$, and the corresponding value of the optimal test threshold by d^*.

Note that the above estimation procedure does not depend on the existence of a gold standard test (the estimates of $\{\overline{E}_k,\overline{C}_k\}$ are not dependent on the 'true group' variable G).

The above formulae assume that the diagnostic information can be obtained free of charge and at no risk to the patient. There are cases where this may be true, where D is patient age for example. In other cases, it may be approximately true. For example, some blood tests are very cheap and low-risk. However, there might well be significant costs and/or risks associated with the test. If so, then the above methods must be modified. Suppose, for instance, that the individual-level estimates of costs and effects already include the financial and health impacts of the diagnostic test. The ROTS estimates outlined above then automatically incorporate test-related costs and health effects. However, the two end-points (R and S) over-estimate the mean costs and under-estimate the effects that would be likely to occur in the absence of diagnostic information. To correct for this, we would have to shift these two points downwards by the expected cost of the test, and to the right according to its expected health risk.

3.3.2 Estimation from RCT data

We now consider the situation where we have a set of observations of (G,D,E_A,C_A) for one sample of 'control group' patients, and a set of observations of (G,D,E_B,C_B) for another sample of 'intervention' group patients, as in a randomised controlled trial. Estimation of the ROTS curves is more complicated than in the one-sample case discussed above. The method that we suggest treats the diagnostic index, D, as a matching variable. This variable is used to associate observations of the costs and effects of treatment A in one sample with the costs and effects of treatment B in the other sample.

Suppose that we had observations (g_i,d_i,e_{Ai},c_{Ai}) for a sample of n control patients (i=1,2,…,n) and observations (g_j,d_j,e_{Bj},c_{Bj}) for a sample of m intervention patients (j=1,2,…,m). In this case we start by ranking the patients *within each group* in order of increasing D. As before, patients are indexed in random order in the case of a tie. If n=m, we could individually match the patients in our two samples and proceed as before. However, the occurrence of exactly equal sample sizes is rare. If the samples are of different sizes, we could just ignore the extra data in the larger sample, picking out only the most closely matched pairs of control and intervention group patients. However, this is inefficient.

An alternative to individual matching is some form of 'group matching'. One simple option is to sub-divide each group into a given number of sections according to the value of the diagnostic variable of interest. For example, we could divide the control group by percentiles of decreasing D, then take the sample mean of the variables within each percentile group. Hence, we would summarise the control group data in the form of 100 percentile-group means for the diagnostic index $\bar{d}_p^C$, the health outcome $\bar{e}_{Ap}^C$ and the cost $\bar{c}_{Ap}^C$, p=1,2,…,100 such that if $\bar{d}_p^C \geq \bar{d}_s^C$ then p<s. Similarly for intervention group the data is summarised as 100 percentile-group means $\bar{d}_p^I$, $\bar{e}_{Bp}^I$ and $\bar{c}_{Bp}^I$.

The mean effect for the test at threshold d_k, k=0,1,…,100, can then be estimated by :

$$\bar{E}_k = \begin{cases} \sum_{p=1}^{100} n_p \bar{e}_{Ap}^C / n, & \text{if } k = 0 \text{ (Point R)} \\ \sum_{p=1}^{k} m_p \bar{e}_{Bp}^I + \sum_{p=k+1}^{100} n_p \bar{e}_{Ap}^C \Big/ (m+n), & \text{if } 0 < k < 100 \\ \sum_{p=1}^{m} m_p \bar{e}_{Bp}^I / m, & \text{if } k = 100 \text{ (Point S)} \end{cases}$$

where n_p and m_p are the number of people in the p[th] percentile of the control group and the intervention group respectively, $m = \sum_{p=1}^{100} m_p$ and $n = \sum_{p=1}^{100} n_p$. The value of threshold d_k can be estimated as $(n_k \bar{d}_k^C + m_k \bar{d}_k^I)/(n+m)$.

The mean cost for the test at threshold d_k is estimated by:

$$\bar{C}_k = \begin{cases} \sum_{p=1}^{100} (n_p \bar{c}_{Ap}^C)/n, & \text{if } k = 0 \text{ (Point R)} \\ \left(\sum_{p=1}^{k} (m_p \bar{c}_{Bp}^I) + \sum_{p=k+1}^{100} (n_p \bar{c}_{Ap}^C) \right) / (m+n), & \text{if } 0 < k < 100 \\ \sum_{p=1}^{m} (m_p \bar{c}_{Bp}^I)/m, & \text{if } k = 100 \text{ (Point S)} \end{cases}$$

Once these one hundred points on the ROTS curve have been estimated, the associated estimates of net benefits can be obtained $\bar{N}_k$, and hence the optimal threshold d^* that yields the maximum net benefit $\bar{N}^*$ can be estimated. Just as with the one-sample procedure, this method does not depend on the existence of a gold standard test for comparison. The estimates of the end-points R and S also require adjustment for test-related costs and health risks.

We used this percentile estimation procedure in the analysis of the simulated data reported below. Evidently this is only appropriate for relatively large sample sizes (certainly greater than 100 in each group).

3.4 Simulation procedure

We started by using Monte Carlo simulation to draw a three random samples from a specified joint distribution function for the variables $(G, D, E_A, C_A, E_B, C_B)$ (see next section for population parameters).

- Sample I was a very large (n=20,000) sample of crossover data (including observations of all seven variables for each patient).
- Sample II was a smaller (n=1,000) crossover sample, randomly drawn from Sample I.
- Sample III was simulated RCT sample, with equal sample sizes in the two study groups (n=m=1,000). The control group was the same as Sample II and the intervention group was also drawn randomly from Sample I.

A standard incremental analysis was conducted using these three samples. This provided sample estimates of the expected net benefit with no diagnostic information (point S) and with the gold standard test (point P).

Estimates of the ROTS curves for were then obtained from each data sample (using the method outlined in section 3.3.1 for Sample I and II, and the method in section 3.3.2 for Sample II). An 'optimum' threshold d^* was then obtained from each curve.

Similarly, standard non-parametric methods were used to obtain ROC estimates, and hence estimates of the thresholds based on the maximum-accuracy (d^{ACC}) and Metz criteria (d^{METZ}), for each of the three samples.

Two different diagnostic tests with different levels of accuracy were examined for each method. These tests are called D1 and D2.

The sensitivity and specificity associated with the various threshold estimates were obtained from the ROC curve for Sample I – given the very large sample size, this provides a good estimate of the 'true' values for these output parameters. Similarly, we estimated the net benefit associated with each threshold using the Sample I ROTS curves. This procedure allows us to compare the relative efficiency of the estimation criteria, given different levels of diagnostic information and different quantities and types of trial data.

3.4.1 Baseline scenario

We assumed that the variables D_1, D_2, E_A and E_B were normally distributed, and that C_A and C_B were log-normally distributed (reflecting the positive skew of cost data that is frequently observed [18]). The parameters used in our baseline simulation are given in Table 1. Initially we assumed zero correlations between the variables $/D_1, D_2, E_A, C_A, E_B, C_B 0$. We also started by assuming that all diagnostic tests are free and riskless. We based our estimates of net benefit on a cost-effectiveness threshold ə of £20,000 per QALY.

Table 1 – Parameter values for baseline scenario

Parameter	Group 1	Group 2
Prevalence	0.5	0.5
D_1, mean (sd)	0.2 (0.4)	0.8 (0.4)
D_2, mean (sd)	0.4 (0.4)	0.6 (0.4)
E_A, mean (sd)	0.6 (0.2)	0.2 (0.2)
C_A, mean (sd)	200 (100)	200 (100)
E_B, mean (sd)	0.4 (0.2)	0.8 (0.2)
C_B, mean (sd)	2,000 (1,000)	2,000 (1,000)

An incremental analysis of the baseline case is shown in Table 2. In the absence of any diagnostic information the recommended action would be to provide B for all patients (the net

benefit for point S compared with point R is £2,200). With prior perfect diagnostic information, the recommendation would be to provide A for Group 0 and B for Group 1 (yielding a net benefit of £5,100 at point P). Hence, the expected value of perfect diagnostic information is £2,900; this is the maximum that it would be worth paying per patient for a gold standard test (allowing for the monetary value of any health risks as well as direct costs for the test).

Table 2 – Incremental analysis for baseline scenario

	Treatment for Group 0	*Treatment for Group*	*Mean Effect*	*Mean Cost*	*ICER**	*NB**
Group 0 only	A	-	0.6	£200	-	-
	B	-	0.4	£2,000	< 0	-£5,800
Group 1 only	-	A	0.2	£200	-	-
	-	B	0.8	£2,000	£3,000	£10,200
All patients						
Point R	A	A	0.4	£200	-	-
Point P	A	B	0.7	£1,100	£3,000	£5,100
Point S	B	B	0.6	£2,000	£9,000	£2,200
Point Q	B	A	0.3	£1,100	< 0	-£2,900

* ICER and NB calculated for treatment B with respect to treatment A for the sub-group analyses, and for point P, S and Q each with respect to point R for the analysis of all patients

3.4.2 Sensitivity analysis

We repeated our analysis for a number of different scenarios (see Table 3) to test the impact of some key parameters on the relative efficiency of the various test optimisation procedures.

Table 3 – Scenarios tested in the sensitivity analysis

Scenario	Change to baseline scenario
1	Higher cost-effectiveness threshold (~~a~~= £40,000 per QALY)
2	Lower proportion of Group 1 in the population (p = 0.25)
3	Greater effects for B for Group 1 (mean of 1.2 for G=1)
4	Greater variation in a diagnostic index (sd for D_1 = 0.8).
5	Positive correlation between diagnostic index and net benefit of B (correlation coefficient for D_1 and E_B = 0.8)

4 Results

The estimated ROC and ROTS curves for the baseline analysis are shown in Figure 4 and Figure 5 respectively. With a sample of 1,000 patients in a cross-over trial (Sample II) or 1,000 in each arm of an RCT (Sample II) the approximations of both curves to the large sample estimates (Sample I) appear to be quite good. Clearly, the curve estimates are subject to sample variation – different samples will give rise to curves of slightly different shape. The results presented here are merely illustrative.

Figure 4 - ROC curves estimated from Sample I, II and III (baseline analysis)

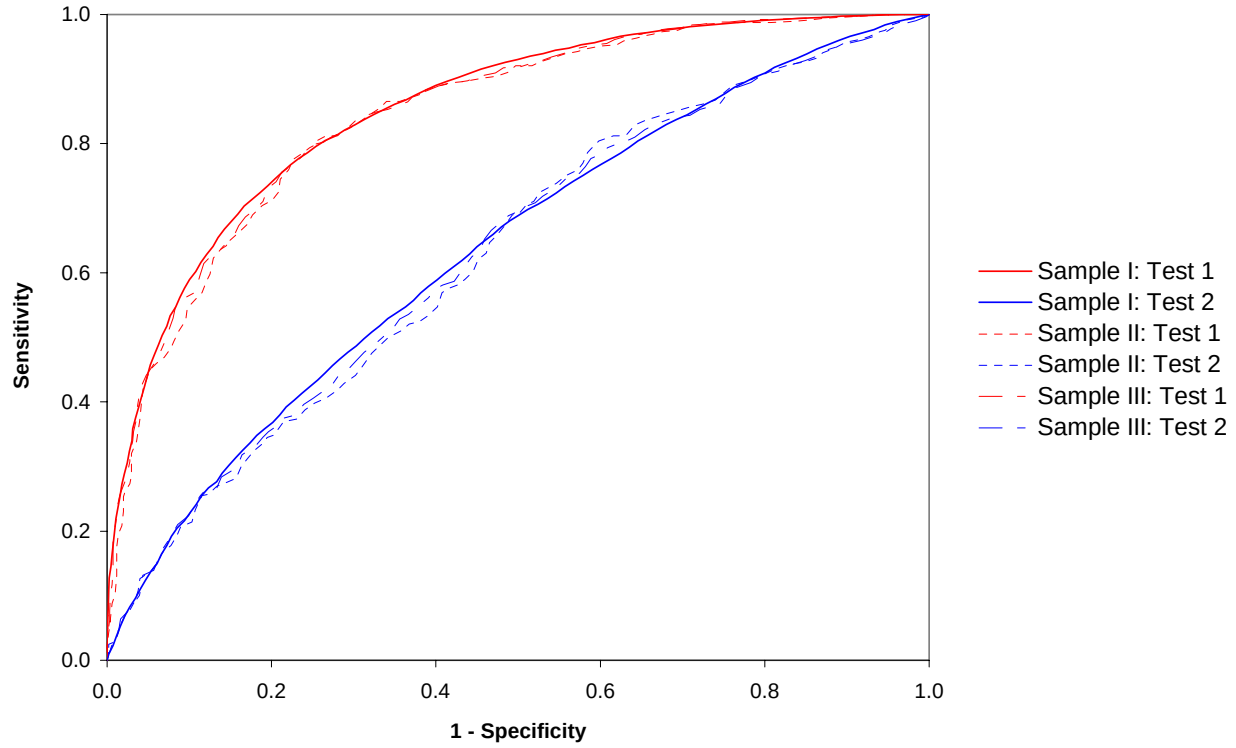
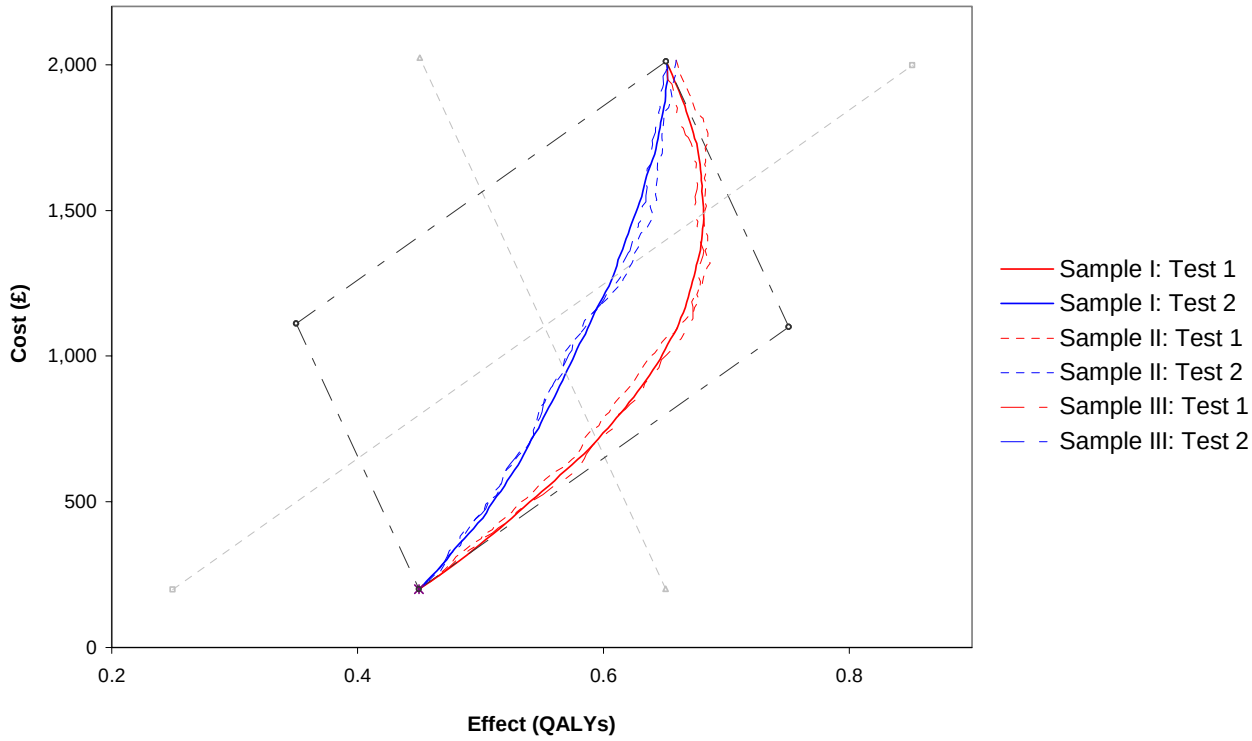


Figure 5 – ROTS curves estimated from Sample I, II and III (baseline)



Test thresholds associated with the above curves for three different threshold selection criteria are shown in Table 4. The test sensitivities and specificities reported in this table are

obtained from the Sample I estimates of the ROCs (the solid lines in Figure 4). Hence, they represent our best estimates of the 'true' population sensitivities and specificities associated with each threshold. These may deviate from sample estimates of the associated sensitivities and specificities obtained from the Sample II or Sample III ROCs (the dotted lines in Figure 4). Similarly, the mean costs and effects, and hence the net benefits, reported in Table 4 are obtained from the 'true' ROTS estimates (the solid lines in Figure 5) rather than from the sample ROTS estimates (the dotted lines in Figure 5). This allows us to compare the likely 'real-world' efficiency of the various threshold estimation procedures.

Sample I indicates that if we wish to maximise the accuracy of the test, we should set the Test 1 threshold so as to treat 51% of patients with A (hence 49% with B). This is estimated to yield a sensitivity of 77% and specificity of 78%, and a net benefit of £3,303 (compared with treating all patients with A – point R). This represents an expected gain of £1,103 over decision making in the absence of diagnostic information. However, if we select the threshold for Test 1 based on the maximum net benefit point on the ROC or the maximum net benefit point on the ROTS, we would treat 36% with A (64% with B), yielding a greater expected net benefit £3,412. Hence there is an expected gain of £109 per patient obtained by setting the test threshold by a net benefit rule rather than a maximum-accuracy rule. Similarly, for Test 2, the estimated net benefit associated with the maximum-accuracy rule (£2,047) is lower than that estimated by the maximum net benefit rules (£2,338 using the Metz ROC based method and £2,347 using the ROTS method). These figures compare with a net benefit of £2,200 in the absence of any diagnostic information (point S) and £5,100 with perfect diagnostic information (point P). Hence, Test 2 with the threshold determined by maximum-accuracy is worse than making the decision on the basis of no diagnostic information at all.

The results of the sensitivity analyses are summarised in Table 5. The net benefit maximising thresholds must clearly be sensitive to the cost-effectiveness threshold. An increase in the CE threshold (Scenario 1) is thus associated with a fall in the treatment thresholds, as it becomes cost-effective to treat a greater proportion of patients with B. A similar change is observed with a higher mean effect for treatment B (Scenario 3). Conversely, a reduction in the proportion of Group 1 people in the population (Scenario 2) raises the treatment thresholds, as does increased variation in the diagnostic index for test 1 (scenario 4).

Table 4 – Results of baseline analysis
For three data samples (I, II and III), four levels of diagnostic information (Gold standard, Test 1, Test 2, and none) and three test threshold criteria (ROTS, Metz and MaxAcc)

	Test 1			Test 2		
	MaxAcc	NB_{ROC}	NB_{ROTS}	*MaxAcc*	NB_{ROC}	NB_{ROTS}
I - Large cross-over sample *(20,000)*						
Threshold (percentile)	51%	36%	36%	43%	14%	17%
Threshold (value)	0.510	0.308	0.308	0.428	0.054	0.105
Sensitivity [1]	77%	89%	89%	67%	92%	90%
1-Specificity [1]	22%	40%	40%	48%	81%	78%
Mean effects [1]	0.610	0.629	0.629	0.554	0.595	0.593
Mean costs [1]	£1,095	£1,364	£1,364	£1,239	£1,766	£1,711
Net benefit [1,2]	£3,303	£3,412	£3,412	£2,047	£2,338	£2,347
Value of diagnosis [3]	£1,103	£1,212	£1,212	-£153	£138	£147
II - Smaller cross-over sample *(n=1,000)*						
Threshold (percentile)	48%	40%	40%	31%	26%	27%
Threshold (value)	0.469	0.362	0.362	0.300	0.237	0.248
Sensitivity [1]	80%	86%	86%	78%	82%	81%
1-Specificity [1]	25%	35%	35%	61%	67%	66%
Mean effects [1]	0.615	0.625	0.625	0.573	0.581	0.580
Mean costs [1]	£1,149	£1,295	£1,295	£1,456	£1,546	£1,528
Net benefit [1]	£3,349	£3,400	£3,400	£2,204	£2,272	£2,261
Value of diagnosis [3]	£1,149	£1,200	£1,200	£4	£72	£61
III - Smaller RCT sample *(n=m=1,000)*						
Threshold (percentile)	48%	41%	41%	43%	16%	30%
Threshold (value)	0.473	0.376	0.376	0.428	0.092	0.288
Sensitivity [1]	80%	85%	85%	67%	90%	79%
1-Specificity [1]	25%	34%	34%	48%	79%	63%
Mean effects [1]	0.615	0.624	0.624	0.554	0.594	0.575
Mean costs [1]	£1,149	£1,276	£1,276	£1,239	£1,729	£1,474
Net benefit [1,2]	£3,349	£3,404	£3,404	£2,047	£2,342	£2,216
Value of diagnosis [3]	£1,149	£1,204	£1,204	-£153	£142	£16

1 – Estimated from 'true' ROC and ROTS curves (data Sample I).
2 – Net benefits calculated with respect to treatment A for all patients (point R on the ROTS) using a cost-effectiveness threshold ∂ of £20,000 per QALY
3 – Value of diagnostic information calculated by subtracting the expected net benefit in the absence of diagnosis (£2,200).

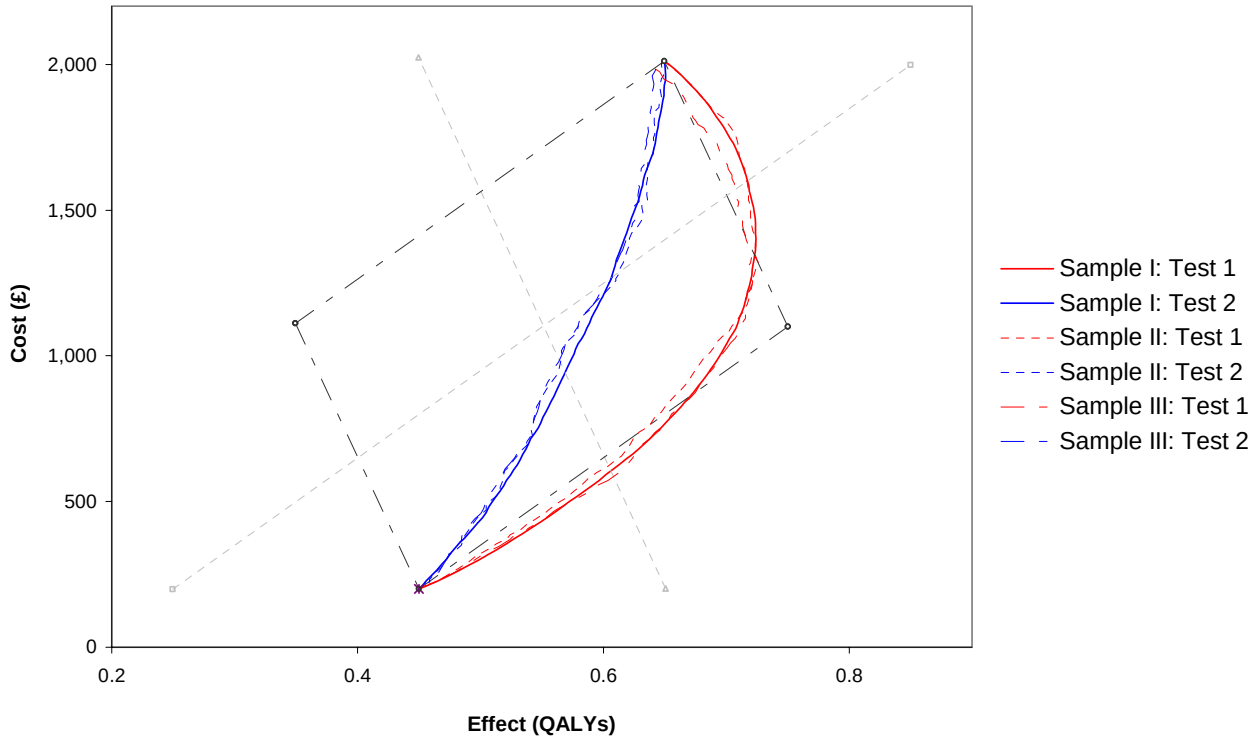
Finally, we tested the impact of introducing a positive correlation between the diagnostic index for Test 1 and the effect of treatment B (Scenario 5). This makes no difference to the ROC, or to the thresholds derived from the ROC, however it does impact on the ROTS, increasing its curvature (see Figure 6). With a strict diagnostic threshold (close to R), the people who pass the test are not only more likely to be true positives, but both true positives and false positives get a higher than average gain from treatment B. Conversely, with a lax threshold (close to S), both true and false positives tend to get a lower than average health gain. Hence the opportunity costs of false positives and false negatives vary with the treatment threshold. This means that the Metz method might not be optimal when there are correlations between the diagnostic index and the costs or effects of one or both of the

treatments. In our simulated example (scenario 5), the net-benefit threshold based on the ROC is at the 36th percentile, whereas that based on the ROTS is at the 41st percentile.

Table 5 – Summary of sensitivity analysis results (Sample A)

	Test 1			Test 2		
	MaxAcc	NB_{ROC}	NB_{ROTS}	*MaxAcc*	NB_{ROC}	NB_{ROTS}
Threshold (percentile)						
0 Baseline analysis	51%	36%	36%	43%	14%	17%
1 Higher CE threshold	51%	32%	31%	43%	7%	4%
2 Lower prevalence	83%	76%	76%	100%	86%	85%
3 Higher mean E_B	51%	30%	30%	43%	3%	4%
4 Greater variation in D_1	49%	22%	22%	43%	14%	17%
5 Positive D_1/E_B correlation	51%	36%	41%	43%	14%	17%
Net benefit						
0 Baseline analysis	£3,303	£3,412	£3,412	£2,047	£2,338	£2,347
1 Higher CE threshold	£7,502	£7,997	£8,004	£5,133	£6,335	£6,343
2 Lower prevalence	£956	£1,044	£1,044	£23	£144	£150
3 Higher mean E_B	£6,371	£7,062	£7,062	£4,729	£6,231	£6,233
4 Greater variation in D_1	£2,333	£2,583	£2,583	£2,047	£2,338	£2,347
5 Positive D_1/E_B correlation	£4,257	£4,308	£4,336	£2,029	£2,303	£2,310

Figure 6 – ROTS curves estimated from Sample I, II and III (Scenario 5)



5 Conclusion

In this paper we have described how an analogue of the ROC curve in cost-effectiveness space (the ROTS) can be estimated directly from one- or two-sample data at the individual patient level. In the two-sample case, as in a randomised controlled trial for example, the diagnostic index D is used to 'match' observations or groups of observations from the control group with those from the intervention group. By finding the point on this curve where net benefits are at a maximum we can estimate an 'optimal' cut-off for a given diagnostic index (and for a given cost-effectiveness threshold λ). Alternative tests can also be evaluated by comparing the estimated net benefits at the 'optimal' points on their respective ROTS curves.

We illustrated our method using simulated data, and compared the results with those of two other threshold selection criteria which are based on the ROC curve - the maximum-accuracy rule and Metz's net benefit criterion [3;5]. Our analysis illustrates the inefficiency of the maximum-accuracy rule. Across all of our samples and scenarios, this consistently gave the lowest estimated net benefit of the three threshold-selection criteria examined (Table 4 and Table 5). The results of the ROC-based and ROTS-based net benefit criteria, however, were very similar. This is not surprising given the duality of their underlying principles. Nevertheless, there are some differences that may lead us to prefer a ROC-based or a ROTS-based evaluation in different situations.

One might generally expect the ROC estimate to be more robust than the ROTS, since the latter incorporates additional variation over treatment costs and effects. In contrast, Metz's criterion requires estimates of only mean treatment costs and effects for the two subgroups. Hence, the threshold and its associated net benefits may be less vulnerable to sampling variation if estimated via the ROC rather than the ROTS. Further simulation studies are required to investigate the magnitude and importance of any such effects.

However, the ROTS does have some potential advantages. Firstly, placing the problem in cost-effectiveness space has the advantage of familiarity for health economists and, increasingly, for many health care decision-makers. Secondly, when the diagnostic index is correlated with the net benefits of treatment within diagnostic groups, the opportunity costs associated with false positive and false negative results vary with the diagnostic threshold. If so, then the Metz criterion will not necessarily yield a 'true' net benefit optimum. In the above analysis we showed how correlation between a diagnostic index and a treatment outcome led to divergence between the ROC-based and ROTS –based net benefit thresholds (Figure 6 and Scenario 5 in Table 5). Thirdly, estimation of the ROTS does not rely on the existence of a 'gold standard'. The absence of a perfect standard presents serious difficulties for test evaluation and optimisation [19]. And this is certainly not an uncommon phenomenon.

[Words 5,937]

References

1. Phelps CE. Good technologies gone bad: how and why the cost-effectiveness of a medical intervention changes for different populations. *Medical Decision Making* 1997;**17**:107-17.
2. Garber AM,.Solomon NA. Cost-effectiveness of alternative test strategies for the diagnosis of coronary artery disease. *Ann Intern Med* 1999;**130**:719-28.
3. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978;**VIII**:283-98.
4. Phelps CE,.Mushlin AI. Focusing medical technology assessment using medical decision theory. *Medical Decision Making* 1988;**8**:279-89.
5. Bosch JL, Halpern EF, Gazelle GS. Comparison of preference-based utilities of the Short-Form 36 health survey and Health Utilities Index before and after treatment of patients with intermittent claudication. *Medical Decision Making* 2002;**99**:403-9.
6. Laking, G., Lord, J., and Fischer, A. The economics of diagnosis. 2004.
   Ref Type: Unpublished Work
7. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;**240**:1285-93.
8. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Medical Decision Making* 1997;**17**:94-102.
9. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997;**16**:2143-56.
10. Qin J,.Zhang B. Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* 2003;**90**:585-96.
11. Zou KH, Warfield SK, Fielding JR, Tempany CM, William MW3, Kaus MR *et al*. Statistical validation based on parametric receiver operating characteristic analysis of continuous classification data. *Academic Radiology* 2003;**10**:1359-68.
12. Hanley JA,.McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:36.
13. Zhang DD, Zhou XH, Freeman DHJ, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine* 2002;**21**:701-15.
14. Angus DC, Marrie TJ, Obrosky DS, Clermont G, Dremsizov TT, Coley C *et al*. Severe community-acquired pneumonia: use of intensive care services and evaluation of American and British thoracic society diagnostic criteria. *American Journal of Respiratory & Critical Care Medicine* 2002;**166(5)**:717-23.
15. Kosuda S, Yoshimura I, Aizawa T, Koizumi K, Akakura K, Kuyama J *et al*. Can initial prostate specific antigen determinations eliminate the need for bone scans in patients with newly diagnosed prostate carcinoma? A multicenter retrospective study in Japan. *Cancer* 2002;**94(4)**:964-72.
16. Shirasaya K, Miyakawa M, Yoshida K, Takahashi E, Shimada N, Kondo T. Economic evaluation of alternative indicators for screening for diabetes mellitus. *Preventive Medicine* 1999;**29**:79-86.
17. Ament A,.Baltussen R. The interpretation of results of economic evaluation: explicating the value of health. *Health Econ* 1997;**6**:625-35.
18. Briggs AH,.Gray AM. Handling uncertainty when performing economic evaluation of health care interventions. *Health Technol Assess* 1999;**3**.
19. Bossuyt PMM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou P, Irwig LM *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;**326**:41-4.