Biostatistics (2003), **4**, 3, *pp.* 341–351
Printed in Great Britain

Assessing the diagnostic accuracy of a sequence of tests

MARY LOU THOMPSON
Department of Biostatistics, Box 357232, University of Washington,
Seattle, WA 98195, USA
mlt@u.washington.edu

SUMMARY

We consider the assessment of the overall diagnostic accuracy of a sequence of tests (e.g. repeated screening tests). The complexity of diagnostic choices when two or more continuous tests are used in sequence is illustrated, and different approaches to reducing the dimensionality are presented and evaluated. For instance, in practice, when a single test is used repeatedly in routine screening, the same screening threshold is typically used at each screening visit. One possible alternative is to adjust the threshold at successive visits according to individual-specific characteristics. Such possibilities represent a particular slice of a receiver operating characteristic surface, corresponding to all possible combinations of test thresholds. We focus in the development and examples on the setting where an overall test is defined to be positive if *any* of the individual tests are positive ('believe the positive'). The ideas developed are illustrated by an example of application to screening for prostate cancer using prostate-specific antigen.

Keywords: Receiver operator characteristic curve; Sensitivity; Sequential diagnostic tests; Specificity.

1. INTRODUCTION

Methods for characterizing the accuracy of a single diagnostic test are well established. For tests whose outcome is measured on a continuous scale, for instance, the receiver operating characteristic (ROC) curve provides a summary which displays, for given diseased and non-diseased populations, the sensitivity and (1-) specificity associated with each possible test threshold. With repeated diagnostic tests as in, for example, routine screening, each individual test is associated with an ROC curve but the overall accuracy of, say, a sequence of T tests is not clearly defined and this issue has not received much attention in the literature. Our focus in considering the accuracy of a test or sequence of tests will be on the sensitivity and specificity associated with the test(s).

Our particular focus here will be the setting where tests are implemented *sequentially* and where there is interest in arriving at a diagnosis without necessarily carrying out the full array of tests. This may arise, for instance, when a battery of increasingly expensive/invasive tests is used towards diagnosis or in a routine screening programme where an individual may be screened 'positive' at any screening occasion without the full complement of screening tests (over future years) being available.

Two main approaches to defining an overall positive test have been considered in these settings. In the first example above it might be appropriate to define an individual as testing overall positive if the results of *all* their tests are positive. In the literature this is referred to as 'believe the negative' (BN). Here individuals who test negative on any particular test will not receive subsequent tests. Alternatively, as would be appropriate in a screening programme, an individual might be defined to be overall positive if they test positive on *any* test ('believe the positive', BP). Here, individuals who test positive on any particular test will not receive subsequent tests.

Consider, for instance, screening tests for prostate cancer using prostate-specific antigen (PSA) which are repeated over time. Etzioni *et al.* (1999) have considered estimating the ROC curve for PSA at a single time point (u years prior to clinical diagnosis). We are concerned here rather with *overall* sensitivity and specificity, by which we mean the true positive and true negative rates associated with an entire body of testing. Elmore *et al.* (1998), in the context of breast cancer screening, note the problem of accrual of false positives over time. In their retrospective study over a 10 year period, one-third of women screened had an abnormal test result that required additional evaluation, even though no breast cancer was present. The authors comment that 'Techniques are needed to decrease false positive results while maintaining high sensitivity'.

The literature on multiple testing has focused primarily on multiple binary tests (see, for example, Politser (1982), Marshall (1989), Lau (1991), Ten Have and Bixler (1997)). Kraemer (1992), in her book on evaluating medical tests, considers combining continuous tests, implemented in sequence or in parallel, and the consequences in terms of overall costs and benefits. Tolley *et al.* (1991) and Murtaugh (1995) consider repeated applications of the same continuous test where a positive test outcome is associated with the same threshold at each test application. As we will discuss, this is but one special case of the more general range of positivity definitions that are available in this setting. We are otherwise aware of very little in the literature that considers the the general question of the evaluation of accuracy in the context of sequential *continuous* tests.

There has been some focus in recent literature (Su and Liu, 1993; Pepe and Thompson, 2000) on identifying the optimal linear combination of a set of test outcomes. This differs from what we will consider here in two respects: firstly, these approaches require *all* tests to be applied to each individual, which need not be the case in the kind of setting that we consider here. If one is screening for cancer, for instance, one is not necessarily going to wait until one has accrued ten years of annual test results before deciding on an individual's screening status. Secondly, and as a consequence of the first point, the decision rules that we will consider will not necessarily be restricted to linear combinations of the individual test results.

Our aim here is to take a more general approach to the evaluation of accuracy of a sequence of tests, thereby highlighting some of the complexities which should be considered (and exploited) in determining how multiple tests should be implemented and evaluated. We will focus primarily on the BP rule, which is appropriate for repeated screening tests, but the general ideas that we introduce apply equally to BN rules or other positivity definitions. In Section 2 we present expressions for sensitivity and specificity corresponding to two tests and illustrate the ideas with a hypothetical example. Section 3 suggests alternatives to the conventional use of constant thresholds for defining positive tests. Section 4 contains an example of application to screening for prostate cancer using PSA and Section 5 concludes with a discussion.

2. TWO TESTS

We will use the following notation throughout: D is a binary random variable denoting disease status. We assume that some definitive gold standard is used to ascertain D. Let Se_X and Sp_X denote respectively the sensitivity and specificity of a diagnostic test, X. Although our primary interest is the sequential administration of continuous tests, it is informative to examine first somewhat simpler settings which are special cases of the more general one. We are concerned throughout with theoretical relationships, rather than questions of inference.

2.1 *Two binary tests*

Let X_1 and X_2 denote two binary diagnostic tests. One setting might be that where X_2 is the same test as X_1, administered at a subsequent time point. Here, as we have discussed, a positive outcome of the combined screening $X_1 + X_2$ might be defined to occur if *either* X_1 *or* X_2 were positive ('believe the positive', BP). We will denote this particular combination of individual tests by $X_1 \vee X_2$, where

$$Se_{X_1 \vee X_2} = 1 - P(X_2 = \text{ negative } | X_1 = \text{ negative}, D)(1 - Se_{X_1})$$

$$Sp_{X_1 \vee X_2} = \quad P(X_2 = \text{ negative } | X_1 = \text{ negative}, \bar{D}) Sp_{X_1}.$$

It is easily seen that the sensitivity of the combined test $X_1 \vee X_2$ is greater than or equal to that of either test administered on its own and, similarly, the specificity is less than or equal to application of an individual test. Similarly, if the combined test is $X_1 \wedge X_2$ (BN), it is again easily seen that the sensitivity of the combined tests is less than or equal to and the specificity greater than or equal to that of either test on its own.

In either setting, the accuracy of the combined test is well defined and can be summarized by its sensitivity and specificity. The comparison of application of a single test with both tests involves a comparison of different points in ROC space: those associated with the tests X_1, X_2 on their own, $X_1 \vee X_2$ (BP) and $X_1 \wedge X_2$ (BN). Any preference of one point over the other will involve a consideration of the costs associated with false positives and false negatives and with application of the tests.

2.2 *One test binary, one continuous*

Assume now that X_1 is a binary test, but that it is followed by a continuous test, X_2, where large values of X_2 are more indicative of disease, i.e. a positive test outcome is defined to occur when $X_2 > c$, for some threshold c. We will again consider the combined test, $X_1 \vee X_2$ (BP). Results for the setting where a positive outcome is associated with *both* tests being positive again follow similarly.

The diagnostic accuracy for the binary test X_1 may be summarized by the ordered pair (Sp_{X_1}, Se_{X_1}). For the combined test there will be a different accuracy for each possible threshold, c, of X_2. Note that, when $c = \infty$, $Se_{X_1 \vee X_2}(c) = Se_{X_1}$ and $Sp_{X_1 \vee X_2}(c) = Sp_{X_1}$ so that the diagnostic performance of both X_1 alone and X_1 and X_2 combined can be described by a single ROC curve with lower left-hand bound $(1 - Sp_{X_1}, Se_{X_1})$. It is clear, then, that application of X_1 alone corresponds to min(Sensitivity) and max(Specificity) of the possible implementations of the combined tests. The combination of two binary tests discussed above is clearly a special case of this.

Letting $F_{2.1D}$ and $F_{2.1\bar{D}}$ denote the conditional distribution functions for X_2 when X_1 is negative in the diseased and non-diseased populations respectively, then

$$Se_{X_1 \vee X_2}(c) = 1 - F_{2.1D}(c)(1 - Se_{X_1})$$

where $F_{2.1D}(c)$ then represents the fraction of false negatives which are maintained, relative to application of X_1 alone, and

$$Sp_{X_1 \vee X_2}(c) = F_{2.1\bar{D}}(c) Sp_{X_1}$$

where $F_{2.1\bar{D}}(c)$ represents the fraction of specificity that is maintained, relative to application of X_1 alone.

The ROC curve for the combined test may be expressed as

$$\mathrm{ROC}_{(X_1 \vee X_2)}(s) = 1 - F_{2.1D}\left(F^{-1}_{2.1\bar{D}}\left(\frac{1-s}{Sp_{X_1}}\right)\right)(1 - Se_{X_1}) \qquad (1)$$

for $s \geqslant 1 - Sp_{X_1}$ and where s is the overall false positive fraction for the combined test.

As noted above, the accuracy characteristics of application of X_1 alone are represented by the lower left-hand endpoint of this curve ($s = 1 - Sp_{X_1}$). An example illustrating this setting is provided in Section 2.4. It is clear, then, that a comparison of the performance of X_1 with $X_1 \vee X_2$ involves a comparison of different points on the same ROC curve and will hence depend on the associated costs of the tests and of the consequences of mis-diagnoses. Similar remarks apply for the setting where the combined test is defined to be $X_1 \wedge X_2$ and that where X_1 is assumed to be continuous and X_2 binary.

2.3 *Two continuous tests*

All of the above are special cases of the application of two continuous diagnostic tests. These may, for instance, be sequential applications of the same test in a routine screening programme or a series of different tests implemented in sequence towards diagnosis. We will again focus here on the setting where a positive outcome is defined to occur when *either one* of the tests is positive. Similar lines of argument follow for the BN setting. The diagnostic accuracy of any single continuous test, X_t, may be described by the corresponding ROC curve. When one considers a similar summary for, say, $X_1 \vee X_2$, the situation is more complex.

With continuous tests, test thresholds are frequently defined in terms of percentiles of the distribution in the non-diseased and we will use the notation c_{tp} to denote the pth percentile of X_t in the non-diseased. We note again that we focus here on illuminating theoretical relationships and hence we are referring to the percentiles of the underlying theoretical distributions rather than to empirical percentiles.

We will consider initially the definition of a positive test which is such that, for any given c_{1p_1} and c_{2p_2}, an individual may be defined to test positive if $X_1 \geqslant c_{1p_1}$ or $X_2 \geqslant c_{2p_2}$. Note that here at $p_2 = 1.0$ ($p_1 = 1.0$), this just reduces to an application of $X_1(X_2)$. Essentially, each particular choice of threshold for X_1, say, results in an ROC curve of the form in equation (1) above. For fixed p_1, the combined ROC curve may be expressed in terms of the distribution functions F_{1D} and $F_{1\bar{D}}$ of X_1 in the diseased and non-diseased populations and the conditional (given $X_1 < c_{1p_1}$) distribution functions for X_2, $F_{2.1D}$ and $F_{2.1\bar{D}}$:

$$\mathrm{ROC}_{X_1 \vee X_2}(s|p_1) = 1 - F_{2.1D}\left(F_{2.1\bar{D}}^{-1}\left(\frac{1-s}{p_1}\right)\right) F_{1D}(c_{1p_1}) \tag{2}$$

for $s \geqslant 1 - p_1$ and where s is the false positive fraction for the combined test. Note that specificities above p_1 are not achievable for the combined test. Again, the situations described in Sections 2.1 and 2.2 above are special cases of this result.

It is clear, then, that the notion of an ROC curve for a single continuous test is replaced by that of an ROC surface, when there are two (or more) tests. Typically, in current practice, when the same screening test is implemented sequentially, the same threshold, or the same percentile (in the non-diseased), is used to define positivity at each application of the test (i.e. $p_1 = p_2$). In this setting, the summary of accuracy may be represented by a single ROC curve, but the above discussion illustrates that the use of equal thresholds represents only one of an infinite number of possible (p_1, p_2) choices and not necessarily one that is optimal in terms of accuracy characteristics.

One might consider, instead, optimizing over p_1 to define a single ROC curve for $X_1 \vee X_2$ of the form

$$\mathrm{MaxROC}(s) = \max_{p_1} \mathrm{ROC}(s|p_1).$$

Each point on the MaxROC curve would be associated with thresholds (p_1, p_2) which are such that they maximize the overall sensitivity at that overall specificity (and vice versa). The sensitivity for any given specificity will be at least as high as that for either of the individual tests, applied on its own. It should also be noted that this is equivalent to identifying the thresholds (p_1, p_2) that are associated with the ROC curve for the sequence of tests which has maximal area under the curve.

2.4 *Hypothetical example*

To illustrate these ideas, consider repeated application at $t = 1, 2$ of a diagnostic test which is $N(0, 1)$ distributed in the non-diseased population, with correlation $\rho = 0.4$ between successive measurements on the same individual and $N(0.25, 2)$, $N(1, 1)$ in the diseased population, with $\rho = 0.8$. This corresponds to the setting where it is anticipated that progression of disease will lead to an increase in the level of the test and where successive measurements are more highly correlated in diseased individuals.

Consider, for example, the situation where the positivity threshold for the first test is set at the 80th percentile of the non-diseased (i.e. $p_1 = 0.8$). Then the overall specificity of the sequence of two tests will be *at most* 80%. If the second test also has positivity threshold set at the 80th percentile ($p_2 = 0.8$), it can be shown that in this setting the overall specificity is 68% and the overall sensitivity is 60%. If, instead, the second test is implemented at a positivity threshold corresponding to the 90th percentile ($p_2 = 0.9$), then the overall specificity of the two tests will be 74% and the overall sensitivity 48%. These two points on the general ROC *surface* are also points on the ROC *curve* for the sequence of tests corresponding to fixed $p_1 = 0.8$ which is shown in Figure 1(a) (‘ConROC’). As discussed in Section 2.2, the lower endpoint of this curve corresponds to the implementation of X_1 alone (at $p_1 = 0.8$).

Alternatively, say one has a target of an overall false positive fraction of 10%. It can be shown that this might, for instance, be achieved in this setting by choice of thresholds $p_1 = 0.91$ and $p_2 = 0.99$ with associated overall sensitivity 30%, or $p_1 = p_2 = 0.943$ with associated overall sensitivity 35% and so on. The latter case is a point on the ROC curve corresponding to all possible choices of $p_1 = p_2$ (‘EqROC’) and is also shown in Figure 1(a). The maximum sensitivity that is possible with this specificity is 39%, corresponding to a choice of thresholds $p_1 = 0.996$ and $p_2 = 0.903$. This is a point on the ‘MaxROC’ curve, also shown in Figure 1(a). Because the expression for the ROC surface in this setting cannot be expressed in closed form (see (2) above), these maxima were identified by a grid search.

Note that, if the tests were independent, 90% overall specificity for two tests using equal thresholds would be achieved by fixing this threshold at the 94.9th percentile of the non-diseased at each test. With $\rho > 0$, the threshold corresponding to a given overall specificity will decrease with increasing ρ.

This example illustrates some of the additional choices that arise in sequential application of tests. The same *overall* false positive rate of 10% can be achieved by accruing varying false positives at each time of testing. Each choice of (p_1, p_2) will be associated with a different overall sensitivity. It is seen that, in this example, choice of equal thresholds ($p_1 = p_2$) is sub-optimal for most specificities.

3. Alternative approaches to defining positive tests

3.1 *Adaptive thresholds*

The above discussion assumes that, while positivity thresholds may differ from one test application, X_t, to the next, at any particular application (i.e. fixed t), the same threshold is used to define a positive test on *all individuals*. In practice, even with once-off implementation of a test, one might consider varying the threshold according to an individual’s personal characteristics (such as age, ethnicity, height and weight). With sequential testing, one can also consider the idea of *adaptive* thresholds, where the threshold implemented at time t depends on the prior measurements on the individual at times $1, 2, \ldots, t - 1$. The rationale for this would be that there is likely to be greater between-individual than within-individual variability and hence improved screening characteristics might be expected when one takes the prior *path* of the individual into account. These ideas are developed in the context of adaptive reference ranges by Thompson and Fatti (1997) and Fatti *et al.* (1998).

In practice, as discussed above, the choice of threshold often corresponds to some percentile of the distribution in the non-diseased. Analogously, in the multivariate normal setting considered by Thompson

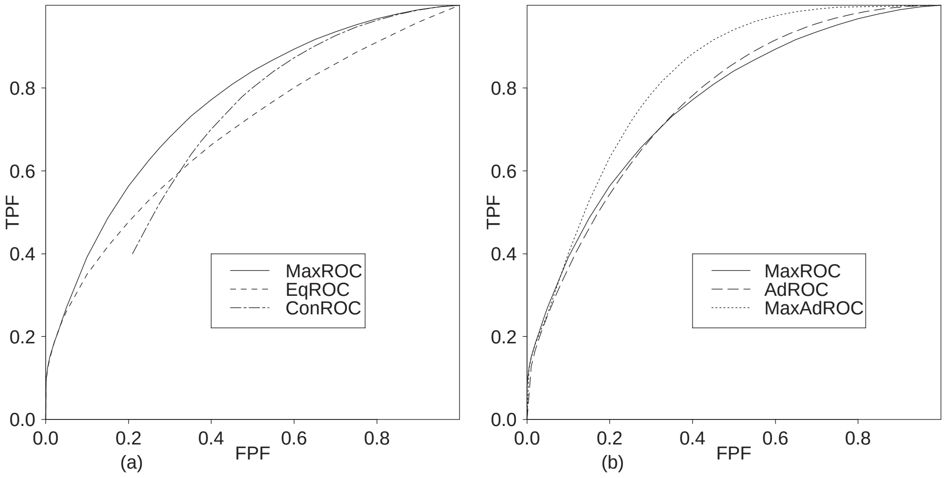

Fig. 1. ROC curves for two sequential tests.

and Fatti (1997), for instance, the adaptive threshold at t might be the corresponding percentile of the distribution of X_t in the non-diseased, *conditional* on past measurements at times $1, 2, \ldots, t-1$. When $t = 2$, this corresponds to choice of threshold for X_2 of $\mu_{2\overline{D}} + \rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_{1\overline{D}}) + z_p\sigma_2\sqrt{1-\rho^2}$, where $\mu_{t\overline{D}}$ and σ_t^2 are the mean and variance of X_t, ρ is the correlation between X_1 and X_2 and z_p is the pth percentile of the standard normal distribution.

For two tests and the BP rule, for any given (p_1, p_2), a positive outcome might be defined to occur when X_1 exceeds its p_1th percentile (in the non-diseased) or X_2 exceeds the p_2th percentile of the *conditional* distribution of $X_2|X_1$ (in the non-diseased). Hence, as in Section 2.3, we have here another ROC surface which, in the bivariate normal setting with equal variance–covariance structure in the diseased and non-diseased populations, corresponds to

$$\mathrm{ROC}_{X_1 \vee X_2}(s|p_1) = 1 - \Phi\left(\frac{\mu_{1\overline{D}} - \mu_{1D}}{\sigma_1} + z_{p_1}\right)\Phi\left(\frac{\mu_{2\overline{D}} - \mu_{2D} - \rho\frac{\sigma_2}{\sigma_1}(\mu_{1\overline{D}} - \mu_{1D})}{\sigma_2\sqrt{1-\rho^2}} + z_{\frac{1-s}{p_1}}\right)$$

for $s \geqslant 1 - p_1$, where s is the false positive fraction of the combined test and μ_{1D}, μ_{2D} denote the test means in the diseased population at times $t = 1, 2$.

It should be noted that, for any given sequence of positivity thresholds, $p_1, p_2, \ldots, p_T$, for the sequence of conditional distributions, the associated overall specificity is given by $\Pi_{t=1}^{T} p_t$. In contrast, the overall specificity associated with a similar sequence of unconditional thresholds cannot be uniquely specified (see examples in Section 2.4) and will depend on the variance–covariance structure of the T tests.

As in Section 2.3, one can also consider in this setting the special cases corresponding to choice of *equal* percentile thresholds ($p_1 = p_2 = \sqrt{1-s}$, for $T = 2$ and overall specificity $1-s$) and the MaxROC curve, where, for each s, p_1 is chosen so as to maximize $\mathrm{ROC}(s|p_1)$ (again, for $T = 2$). Each of these represents a particular two-dimensional slice of the multivariate ROC surface, which we will denote by the 'AdROC' and 'MaxAdROC' curves respectively. Hence for $T = 2$ and overall specificity $1 - s$, the AdROC represents the overall sensitivity that is attained when each test has positivity threshold equal to the $\sqrt{1-s}$th conditional percentile and the MaxAdROC curve represents the maximum possible overall sensitivity when conditional percentiles are used as thresholds, but the percentiles may differ between the tests.

Figure 1(b) shows the theoretical AdROC and MaxAdROC curves resulting from implementing adaptive thresholds in the hypothetical example presented in Section 2.4. Also shown is the constant (non-adaptive) threshold MaxROC curve. Consider again, for instance, a target overall specificity of 90%. This can be achieved using the $\sqrt{90} = 94.9$th conditional percentile as the positivity threshold for each test. The associated overall sensitivity is 37%. For X_1 this will simply involve defining a positive outcome when $X_1 > 1.64$; for X_2, the threshold will be individual-specific and will depend on the measurement at $t = 1$. The maximum possible overall sensitivity in this setting is 40%, achieved using $p_1 = 0.986$, $p_2 = 0.913$. The maximisation was carried out using the command 'optimize' in Splus (Statistical Sciences, 1995).

Here, use of the adaptive thresholds with equal *conditional* percentiles at each time of testing represents accuracy approximately equal to that for the optimal constant thresholds (varying across times of testing, but the same for all individuals) and the maximum adaptive ROC curve represents an improvement over both of these for most specificities. Generally, the advantage of incorporating the prior path will increase with increasing ρ.

3.2 *Combining tests*

The issue of optimally combining tests at *a single point in time* is an area of active research in different contexts (Pepe and Thompson, 2000; Richards *et al.*, 1996; Skates *et al.*, 1995; Su and Liu, 1993; Kraemer, 1992). For instance, in screening for ovarian cancer using CA125 antigen (X_t), Skates *et al.* (1995) use a calculation of the risk of ovarian cancer at time t (Y_t) based on X_j, $j = 1, 2, \ldots, t$. Use of an adaptive threshold that depends on past test measurements (as in Section 3.1) can also be viewed as application of a new test at time t, which is a linear combination of the results of tests at times $1, 2, \ldots, t$. Our focus here, however, is rather on the *overall* accuracy of a *sequence* of tests, where decisions are made *after each time of testing*, rather than once-off after all test results are available.

Referring back to the hypothetical example introduced in Section 2.4, if both test results were simultaneously available on all individuals, then the optimal linear combination (as in the linear combination that maximizes the area under the ROC curve) of the test outcomes is well defined (Su and Liu, 1993). This would require, however, deferring decisions on all individuals until after the second time of testing. For illustration, application of a single test based on the optimal linear combination of X_1 and X_2 ($-0.25X_1 + 0.75X_2$) has sensitivity 34% at specificity 90%. At this specificity, the optimal linear combination has lower sensitivity than the sequential testing settings considered, although it has better accuracy for lower specificities. It is important to stress, however, that this comparison is not central to our development here where our focus is the *sequential* application of tests.

A final point to make here is that any of the methods considered in the literature to develop a combination of tests that could be implemented at a single time point, say $Y_t = g_t(X_1, X_2, \ldots, X_t)$, may be regarded as yielding a *sequence* of tests $Y_1, Y_2, \ldots, Y_t$ to which the same general principles developed above will apply. Whatever test sequence is implemented, it remains the case that evaluation of their overall accuracy must involve a consideration of a variety of ROC surfaces and, for instance, that optimal implementation does not necessarily involve choice of the same threshold at each time of testing.

4. Example of application

PSA has been considered as a screening tool towards early identification of prostate cancer. Etzioni *et al.* (1999) use retrospective data from the randomized Beta-Carotene and Retinol Efficiency Trial (CARET) (Thornquist *et al.* 1993) to model the distribution of log(total PSA) in non-diseased individuals and in diseased individuals as a function of their time (in years) prior to prostate cancer diagnosis (u).

Participants in the CARET study included men aged 50–65, who were at high risk of lung cancer. The intervention (beta-carotene and retinol) had no noticeable effect on the incidence of prostate cancer.

We use an approximation to the model proposed by Etzioni *et al.* (1999) namely we assume that test outcome (log(total PSA)) is distributed as $N(0.35, 0.6)$ in non-diseased individuals and $N(2.23 - 0.18u, 1.2)$ in diseased individuals at u years prior to diagnosis, with an exchangeable correlation structure in diseased and non-diseased, with $\rho = 0.85$. This is an approximation to the model developed by Etzioni *et al.* (1999) in that, for simplicity, we assume a constant variance in the diseased and a common correlation in diseased and non-diseased. The assumed values are, however, close to the modelled values for the range of times to diagnosis considered here.

We consider the following hypothetical setting. A five year screening programme using PSA and a BP positivity decision rule is being considered. For the evaluation of such a programme, diseased individuals are defined to be all those who will be clinically diagnosed with prostate cancer within 10 years of the start of the screening programme. We assume that the baseline diseased distribution is a mixture (with uniform weights) of individuals who are $u = 1, 2, \ldots, 10$ years from clinical diagnosis. Once an individual is diagnosed with prostate cancer, he is removed from follow-up and will remain as a false negative in the evaluation of PSA screening accuracy if he was not screened positive prior to clinical detection of the disease. Of interest here is the accuracy that is achievable by such a screening progamme under different choices of interval between successive screens. This practical setting differs from the hypothetical example considered above in that individuals do not remain available for testing after clinical diagnosis.

Figure 2 shows the equal threshold (adaptive and non-adaptive) theoretical ROC curves corresponding to five annual screens and three biannual screens (at years 1, 3, 5). We see that three screens perform almost as well as five and that the adaptive thresholds offer an improvement in screening accuracy over the use of constant thresholds. The ‘MaxAdROC’ curve for three biannual visits (not shown) offers only marginal improvement over the AdROC curve, e.g. at specificity $=$ 90%, the maximum sensitivity is 66.7% compared to the sensitivity of 66.1% using the same adaptive percentile at each screen. Similarly, the ‘MaxEqROC’ curve (also not shown) offers only marginal improvement over the EqROC.

While the trend in disease path might suggest potential advantage in reducing percentile thresholds at later testing points (i.e. postponing accrual of false positives), the incremental loss of true positives to clinical diagnosis prior to their being screened positive based on their PSA levels mitigates against this in this setting. Thus, choice of equal thresholds across screening occasions is close to optimal here, for each of the ROC surfaces considered. The absence of substantial improvement in accuracy by adjusting the percentiles (cross-sectional or adaptive) at successive screens has the advantage of simplifying the implementation of a screening mechanism. One does note, however, the gains in accuracy that can potentially be achieved by considering alternatives to the use of thresholds that ignore subject-specific information such as the prior path.

Consider, for example, a prostate cancer case from the CARET study, who was screened at 6, 4 and 2 years prior to diagnosis, with PSA levels recorded as 0.48, 0.75, 1.27. At 90% overall specificity, the man’s PSA values remain well below the corresponding constant (non-adaptive) threshold of 4.76. His adaptive thresholds for the same overall specificity would be (5.77, 1.19, 1.24) and hence he would screen positive at the third biannual screen, which is two years prior to clinical diagnosis.

One sees further that three biannual screens offer accuracy almost equal to that of five annual screens, particularly at lower false positive rates, which is typically where screening for a low-prevalence disease would be implemented. Of concern, however, might be the possibility that the use of biannual screens would result in delayed disease diagnosis relative to an annual screening programme. (This would, perhaps, be of greater concern with fast growing and highly fatal tumours such as ovarian cancer.) For 90% specificity, however, the mean time of positive screening prior to clinical diagnosis is about 4 years in both cases. With only biannual screens, the positivity thresholds can be more lenient (implemented at lower percentiles) and still be associated with the same overall false positive rate.

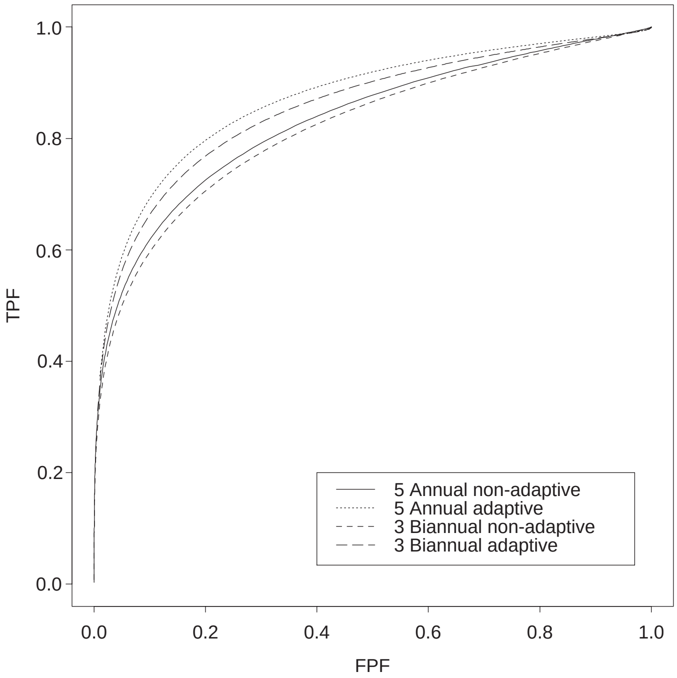

Fig. 2. ROC curves for five annual and three biannual PSA screens.

5. Discussion

We have considered evaluation of the *overall* accuracy of a *sequence* of tests and the complexity of the trade-offs of different test compositions in terms of the timing of accrual of false and true positives. Acknowledgement of the issues considered here can facilitate the design of screening programmes, for instance with regard to definitions of positivity and spacing of tests. The application to PSA considered above illustrates how knowledge of the disease path and distributional characteristics of the screening test may be used to this end.

We have focused here mainly on overall positivity defined according to the BP rule and we have only considered two of the many ways in which positivity of a sequence of tests may be defined, namely those where

- the same threshold is used for each individual at a particular time of testing, but the threshold may differ across times of testing, or
- the same *conditional* percentile is used at each time of testing, with the threshold corresponding to that percentile differing across individuals and where the percentile used may differ across times of testing.

In each case, accuracy may be represented by a ROC surface which can be simplified to the customary two-dimensional representation of sensitivity versus 1-specificity either by choosing percentiles that are the same across repeated testings, or by implementing the tests at percentiles that correspond to the maximum sensitivity at any given specificity. The gains to be made by this latter choice will depend on test correlation structure and the anticipated nature of the disease path. Examination of the *maximum* achievable overall sensitivity for any given overall specificity also allows assessment of whether the test sequence is viable in terms of diagnostic accuracy *at all*. It may be that, even under optimal choices of threshold, the overall accuracy of the test sequence is not sufficiently high to make it feasible for practical implementation.

The development considered here has not addressed the choice of *which* threshold (under whatever definition of positivity) should be used in implementation of a test sequence. Such choices will depend on issues such as disease prevalence as well as the costs of false positives and false negatives and the costs of the tests themselves. In addition, with routine screening, one may wish to incorporate cost of delayed diagnosis which will depend on the rate of advancement of the disease and disease consequences. Formal incorporation of such factors would seem an interesting topic for further research.

ACKNOWLEDGEMENTS

The author acknowledges helpful comments on the manuscript by the Associate Editor and a reviewer, and by Tianxi Cai and colleagues of the Early Detection Research Network: Ziding Feng, Margaret Pepe and Yutaka Yasui. Thanks also to Ruth Etzioni for access to the CARET data. The information in this paper has been funded in part by the NIH grant CA-86368-01.

REFERENCES

ELMORE, J. G., BARTON, M. B., MOCERI, V. M., POLK, S., ARENA, P. J. AND FLETCHER, S. W. (1998). Ten-year risk of false positive screening mammograms and clinical breast examinations. *New England Journal of Medicine* **338**, 1089–1096.

ETZIONI, R., PEPE, M., LONGTON, G., HU, C. AND GOODMAN, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making* **19**, 242–251.

FATTI, L. P., SENAOANE, E. M. AND THOMPSON, M. L. (1998). Bayesian updating in reference centile charts. *Journal of the Royal Statistical Society (A)* **161**, 103–115.

KRAEMER, H. C. (1992). *Evaluating Medical Tests*. California: Sage Publications.

LAU, T. S. (1991). On dependent repeated screening tests. *Biometrics* **47**, 77–86.

MARSHALL, R. J. (1989). The predictive value of simple rules for combining two diagnostic tests. *Biometrics* **45**, 1213–1222.

MURTAUGH, P. A. (1995). ROC curves with multiple marker measurements. *Biometrics* **51**, 1514–1522.

PEPE, M. S. AND THOMPSON, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

POLITSER, P. (1982). Reliability, decision rules, and the value of repeated tests. *Medical Decision Making* **2**, 47–69.

RICHARDS, R. J., HAMMITT, J. K. AND TSEVAT, J. (1996). Finding the optimal multiple-test strategy using a method analagous to logistic regression. *Medical Decision Making* **16**, 367–375.

SKATES, S. J., XU, F. J., YU, Y. H., SJOVALL, K., EINHORN, N., CHANG, Y., BAST, R. C. JR AND KNAPP, R. C. (1995). Toward an optimal algorithm for ovarian cancer screening with longitudinal tumor markers. *Cancer* **76**, 2004–2010.

STATISTICAL SCIENCES, INC. (1995). *S-PLUS Guide to Statistical and Mathematical Analysis*. Seattle, WA.

SU, J. Q. AND LIU, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.

TEN HAVE, T. R. AND BIXLER, E. O. (1997). Modelling population heterogeneity in sensitivity and specificity of a multi-stage screen for obstructive sleep apnoea. *Statistics in Medicine* **16**, 1995–2008.

THOMPSON, M. L. AND FATTI, L. P. (1997). Construction of multivariate centile charts for longitudinal data. *Statistics in Medicine* **16**, 333–345.

THORNQUIST, M. D., OMENN, G. S., GOODMAN, G. E., GRIZZLE, J. E., ROSENSTOCK, L., BARNHART, S., ANDERSON, G. L., HAMMAR, S., BALMES, J. AND CHERNIACK, M. (1993). Statistical design and monitoring of the Carotene and Retinol Efficacy Trial (CARET). *Controlled Clinical Trials* **14**, 308–324.

TOLLEY, E. A., SOMES, G. W. AND WILLEY, E. S. (1991). Determining efficacy of monitoring devices: evaluating new technologies. *Statistics in Medicine* **10**, 351–360.

[*Received October 15, 2001; revised April 2, 2002; accepted for publication June 10, 2002*]