Analytic methods for comparing two dichotomous screening or diagnostic tests applied to two populations of differing disease prevalence when individuals negative on both tests are unverified

Geoffrey Berry¹,∗,†, Catherine L. Smith¹,‡, Petra Macaskill¹ and Les Irwig¹

¹Department of Public Health and Community Medicine, University of Sydney, New South Wales 2006, Australia

SUMMARY

Two dichotomous screening tests may be compared by applying both tests to all members of a sampled population. For individuals with a positive result on either test the disease status may be verified by a reference standard, but for individuals negative on both tests the disease status may be unverified because the probability of disease is so low that further investigation is costly, unacceptable and perhaps unethical. If the tests have been applied to samples from two populations which have different disease prevalences then unbiased estimates of the true positive and false positive rates of each test, the prevalences in the two populations, and two parameters representing dependence between the two tests can be estimated using maximum likelihood methods. The methods are based on the assumption that the sensitivities and specificities of the two tests, and the dependencies between the tests, are independent of prevalence. A test of goodness of fit provides a test of this. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: sensitivity and specificity; diagnostic tests; statistical epidemiological methods

1. INTRODUCTION

The performance of a diagnostic test is usually expressed by the true positive rate (TPR, sensitivity), the probability of a positive test result given that the person has the disease, and the false positive rate (FPR, 1-specificity), the probability of a positive test result given that the person does not have the disease.

Two alternative dichotomous tests may be compared by applying both tests to each person. If true disease status may be ascertained using a reference standard, the TPR and FPR of each test may be estimated without bias if the reference standard is applied to a representa-
Table I. Sampling scheme where only those testing positive on either test are investigated with the reference standard. The unknown variables are represented by the square brackets.

<table>
<thead>
<tr>
<th>Diseased (reference standard positive)</th>
<th>Non-diseased (reference standard negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1+</td>
<td>Test 1−</td>
</tr>
<tr>
<td>Test 2+</td>
<td>$a$</td>
</tr>
<tr>
<td>Test 2−</td>
<td>$c$</td>
</tr>
<tr>
<td>Total</td>
<td>$a+c$</td>
</tr>
</tbody>
</table>

tive sample of those who have been assessed by both tests. In a screening situation and in many diagnostic situations the proportion of diseased people is small. Anyone with a positive result will be investigated further but asymptomatic people with negative results may not be, particularly if the reference standard is invasive. Hence information on those negative on both tests is missing, as shown in Table I [1]. Variables $a, b, c, A, B, C$ are observed and the total number sampled, $T = ([n] + [N])$, and hence $([d] + [D])$, are also known.

When two dichotomous tests are applied to a random sample from a population and only those with a positive result on either test are investigated with the reference standard, then the relative TPR and the relative FPR of the two tests may be estimated [1, 2]. When one test is more accurate than the other on both measures, and if cost and feasibility are neutral to the choice, then the more accurate test would be chosen. When one test has a higher TPR but also a higher FPR, then there is a trade-off between the tests measured by the FP:TP ratio [2]. This ratio depends on the prevalence of the outcome in the population to which the tests are to be applied. The choice between the tests then depends on the consequences of both false negatives and false positives. For the lowest prevalences the test with the lower FPR is likely to be preferred, but as the prevalence increases the test with the higher TPR would become more appropriate.

For situations where the true disease state is undetermined, Walter and Irwig considered the general case, with multiple tests and multiple populations, in terms of models that recognize that the true disease status is unknown; such models are referred to as latent class models [3]. Torrance-Rynard and Walter considered latent class models for a single population and evaluated the bias arising through assuming conditional independence between tests, that is, that the two tests are independent in both those with and without disease, when there is truly dependence [4]. They pointed out that the bias is more serious when disease prevalence is low, as happens with screening. Walter considered the case when disease status is known for those positive to one or both of two tests, but when the double negatives are unverified for disease status, for a single population assuming conditional independence between tests [5]. Qu et al. [6] considered the situation where there are multiple tests but no reference standard and where dependence between the tests is determined by an unobserved continuous variable which varies between subjects. They then applied a random effects model with latent classes, diseased and non-diseased.

In an earlier paper we developed methods for comparing two screening tests where there is a reference standard applied to those positive to one or both tests [2]. It was assumed that the prevalence in the population sampled was unknown; otherwise $n$ and $N$, and hence the TPR and FPR of each test, could be estimated. Secondly, it was assumed that the TPR
and FPR are independent of disease prevalence and it was then possible to estimate the relative TPR and the relative FPR of the two tests. Attention was focused on the situation where one test has both a higher TPR and a higher FPR than the other so that there is a trade-off between the two tests denoted by the FP:TP ratio, the ratio of extra false positives (FPs) to extra true positives (TPs) detected. The FP:TP ratio, and its confidence interval, can be estimated for the sampled population but the ratio is dependent on prevalence. For a target population with a different prevalence we developed estimates and confidence intervals.

2. SAMPLE DATA ON TWO DISTINCT POPULATION GROUPS

In this paper we consider the situation where the same two diagnostic tests have been compared in two population groups of different disease prevalence, using the sampling scheme given in Table I. This may arise if the tests are applied both to a sample from a low risk population and one from a higher risk population, or may occur when a sample from a single population is stratified, for example by age, to produce two subpopulations with differing risks. It is shown that it is possible to estimate the TPR and FPR of each test, the prevalence of disease in each population and two parameters representing the association between the two tests in the diseased and non-diseased groups (conditional dependence). The sensitivities, specificities and parameters representing dependence between the tests are assumed to have identical values in each group, that is, they are independent of disease prevalence (a common assumption which underlies the application of Bayes’ theorem [7]). The methodology provides a test for the validity of this assumption.

The assumption of constant sensitivity and specificity across subgroups has been questioned when the groups are defined according to clinical features of the patients [3, 8]. For example, test accuracy is higher in people with severe urinary symptoms than in those with mild symptoms [9]. Variability in test accuracy in such settings may be explained by differences in the spectrum of disease across groups defined by factors such as the severity of disease or whether the patient was seen in a clinical or primary setting. However, variability in test accuracy is less likely to occur when the groups are defined by a demographic criterion (such as age), particularly for a screening test where the people being tested are asymptomatic. When there are no strong \textit{a priori} grounds for expecting the test accuracy to vary across groups, our method provides a goodness-of-fit test to assess whether the TPR and FPR of the tests and the conditional dependence between them are constant across groups. If the fit is adequate, the model provides estimates of the (common) underlying test accuracy.

Hui and Walter [10] used maximum likelihood estimation to estimate the TPR and FPR of both tests and the two prevalences by assuming the tests are independent conditional on the reference standard, that is, independent within both the diseased group and the non-diseased group. However, in their case they had information on the test results only and no information on any definitive diagnoses. Walter \textit{et al.} [11] extended this approach to estimate the TPR of screening tests where the reference standard result was known in those who were positive on any test. However this approach still assumed that test errors were independent of each other. Vacek [12] tried to improve on this situation by allowing dependency between tests conditional on the reference standard. To do this a covariance term (dependent on the reference standard)
was added to the concordant probability cells given by Hui and Walter. As Vacek had no more information on the sampled groups than Hui and Walter, the covariance terms were parameterized as a proportion of their maximum covariance which can be written as a function of the TPRs or the FPRs depending on whether the covariance term is for the diseased or non-diseased group. In our case, the extra information (that is, the definitive diagnoses) on those who are positive on either test, allows us to estimate a dependency parameter conditional on the reference standard. In general, this parameter represents the inflation in the probability of obtaining positives with both tests, over and above that which occurs when the tests are independent.

3. MODELS AND MAXIMUM LIKELIHOOD ESTIMATION

3.1. Notation

Symbolically, the data are represented using the notation of Table I with a subscript \( g \) to indicate the population group number \((g = 1, 2)\) for each frequency. The prevalence of disease in group \( g \) is denoted by \( p_g \), \( T_g \) is the total sample size \([T_g = n_g + N_g \text{ where } n_g = p_g T_g, \ N_g = (1 - p_g) T_g]\) and \( M_g = d_g + D_g \) (the missing data corresponding to those negative on both tests). In addition the following notation is used, where subscript \( i \) represents the tests \((i = 1, 2)\):

\[
\begin{align*}
  s_i &= \text{true positive rate (TPR) of test } i \text{ (the sensitivity)} \\
  f_i &= \text{1-specificity, the false positive rate (FPR)}
\end{align*}
\]

3.2. Model

The model treats the dependency factor as multiplicative. The cell probabilities are given in terms of eight parameters, the sensitivities \((s_1, s_2)\), false positive rates \((f_1, f_2)\), prevalences \((p_1, p_2)\), and \( \lambda \) and \( \delta \), the dependency factors between the tests in the diseased (reference standard positive) and non-diseased (reference standard negative) groups, respectively. The TPRs, FPRs and dependency factors were assumed independent of prevalence. The cell probabilities are defined as follows:

\[
\begin{align*}
  a_g &= p_g s_1 s_2 \lambda \\
  b_g &= p_g s_2 (1 - s_1 \lambda) \\
  c_g &= p_g s_1 (1 - s_2 \lambda) \\
  d_g &= p_g (1 - s_1 - s_2 + s_1 s_2 \lambda) \\
  A_g &= (1 - p_g) f_1 f_2 \delta \\
  B_g &= (1 - p_g) f_2 (1 - f_1 \delta) \\
  C_g &= (1 - p_g) f_1 (1 - f_2 \delta) \\
  D_g &= (1 - p_g) (1 - f_1 - f_2 + f_1 f_2 \delta)
\end{align*}
\]

The model is specified in terms of a symmetric parameterization. It can easily be shown that the correlation between the tests for reference standard positive is given by

\[
\text{corr(Test 1, Test 2|Ref +ve)} = (\lambda - 1) \sqrt{\frac{s_1 s_2}{(1 - s_1)(1 - s_2)}}
\]
with a similar expression in \( f_1, f_2 \) and \( \delta \) for reference standard negative. A positive correlation indicates that the proportion of concordant test results, that is both tests positive or both tests negative, is higher than for the independence model, whilst a negative correlation indicates an excess of discordant test results.

This model is equivalent to the model proposed by Vacek [12] except that Vacek expressed the dependencies in additive form with cell probabilities for \( a_g, b_g, c_g, \) and \( d_g \) of \( p_g(s_1s_2+e_b), p_g(s_2(1-s_1)-e_b), p_g(s_1(1-s_2)-e_b) \) and \( p_g((1-s_1)(1-s_2)+e_b) \), respectively, for the test positives, where \( e_b \) is a dependency parameter, and equivalent expressions for the test negatives with a dependency parameter \( e_a \). The independence model is given by \( e_b = e_a = 0 \) and Vacek’s model is related to the model used in this paper by \( e_b = s_1s_2(\lambda - 1) \) and \( e_a = f_1f_2(\delta - 1) \). It is also equivalent to an asymmetric parameterization proposed by Flanders et al. [13] in which the dependencies are the ratios of the sensitivities of test 2 according as test 1 is positive or negative. In fact as there are three parameters representing three independent effects in each \( 2 \times 2 \) table then all models including dependencies between the tests must be equivalent.

3.3. Fitting the model

3.3.1. Log-likelihood. The cells of Table I form a multinomial distribution. The probabilities of these cells are expressed in terms of the eight parameters in the model, and the log-likelihood is expressed in terms of the observed frequencies and expected cell probabilities summed over the observed cells as

\[
\ln(L) = \sum_{g} \{ a_g \ln(p_g s_1 s_2 \lambda) + b_g \ln(p_g s_2(1-s_1 \lambda)) + c_g \ln(p_g s_1(1-s_2 \lambda)) \\
+ A_g \ln((1-p_g)f_1 f_2 \delta) + B_g \ln((1-p_g)f_2(1-f_1 \delta)) + C_g \ln((1-p_g)f_1(1-f_2 \delta)) \\
+ M_g \ln[1-p_g(s_1+s_2-s_1 s_2 \lambda) - (1-p_g)(f_1 + f_2 - f_1 f_2 \delta)] \}
\]

3.3.2. Fitting by maximum likelihood. The parameter values which maximize the log-likelihood and satisfy all the necessary parameter constraints (for example, \( 0 < p_1, p_2 < 1 \)) are the maximum likelihood estimates. These estimates were obtained using the EM algorithm as follows:

1. Set \( d_1 + d_2 \) at some initial value.
2. Conditional on \( d_1 + d_2 \), set starting values for \( d_1, d_2, D_1 \) and \( D_2 \)
   \[
d_i = \frac{(a_i + b_i + c_i) d_i}{a_i + b_i + c_i} \quad D_i = T_i - d_i
   \]
   where \( a_i = a_1 + a_2 \) etc.
3. Set the eight parameters at their conditional maximum likelihood estimates of

\[
p_1 = \frac{n_1}{T_1} \quad p_2 = \frac{n_2}{T_2} \\
s_1 = \frac{a_1 + c_1}{n} \quad s_2 = \frac{a_2 + b_2}{n}
\]
$$f_1 = \frac{A+C}{N}, \quad f_2 = \frac{A+B}{N}$$

$$\lambda = \frac{a \cdot n \cdot}{(a+b)\cdot(a+c)}, \quad \delta = \frac{A \cdot N}{(A+B)\cdot(A+C)}$$

4. Calculate the fitted values of $d_1$, $d_2$, $D_1$ and $D_2$, using the parameter estimates from step 3 and the known total sample sizes $T_1$ and $T_2$.

5. Calculate the log-likelihood. Return to step 3 and continue repeating steps 3 to 5, unless the process has converged, defined as a change since the previous cycle of less than 0.000001 in ln($L$).

6. (a) If $\lambda$ is fixed, as in the independence model where $\lambda = 1$, then the maximum likelihood solution has been reached.

(b) If $\lambda$ is not fixed, the solution given by the above process remains conditional on $d_1 + d_2$, that is, at step 4 the fitted values sum to the starting total in step 1. The maximum likelihood solution is obtained by repeating the process from step 1 and searching, over a grid of values of $d_1 + d_2$, for the value giving the maximum likelihood.

3.4. Goodness of fit

As the total sample size in each population group is fixed, there are twelve independent counts (six from each population), and hence twelve degrees of freedom available to estimate the eight parameters. This leaves four degrees of freedom for a test of goodness of fit. A test of goodness of fit follows from the maximum likelihood solution of any of the three models. The log-likelihood of a model fitting the observed data exactly is

$$\ln(L_o) = \Sigma_g \{a_g \ln[a_g/T_g] + b_g \ln[b_g/T_g] + c_g \ln[c_g/T_g] + A_g \ln[A_g/T_g] + B_g \ln[B_g/T_g]$$

$$+ C_g \ln[C_g/T_g] + M_g \ln[M_g/T_g]\}$$

where $T_g = n_g + N_g$.

Then a test of goodness of fit is

$$G = 2[\ln(L_o) - \ln(L_{ML})]$$

distributed approximately as a $\chi^2$ with 4 degrees of freedom. Alternatively an approximate $\chi^2$ test can be constructed from the observed and expected frequencies $a_g$, $b_g$, $c_g$, $A_g$, $B_g$, $C_g$ and $M_g$ using the usual Pearson formula.

4. EXAMPLE

The data in Table II are based on the data of Castiglione et al. [14] comparing rehydrated guaiac testing (Hemoccult) on three consecutive bowel movements and immunochemical testing using reversed passive haemagglutination (RPHA—Hemeselect) on the first bowel movement only to detect colorectal cancers or adenomas $\geq 10$ mm in 8353 males and females aged 40–70.
years recruited between March 1992 and September 1995. Further details on study methods are provided by Castiglione et al. [15]. The data were made available to us by the authors to extract the scenario of Table II. In fact the values of $d_1$ and $d_2$ are known to be at least 9 and 8, respectively, from follow-up studies of test negatives, but this information was not used to fit the models.

Two groups with different prevalences were formed by subdividing the total sample into two age groups, 40–59 years and 60–70 years. The division into two groups at age 60 was governed by the need to balance the positive test results between the two groups whilst subdividing at a ‘round’ age. In this scenario, test1+ corresponds to at least one positive Hemoccult test and test2+ is a positive(+) Hemeselect test.

The model was fitted to these data with the model in four versions: (a) with independence between the two tests, $\lambda = \delta = 1$; (b) with independence in the non-diseased group $\delta = 1$; (c) with independence in the diseased group $\lambda = 1$; (d) without restriction. The fitted parameters are given in Table III and the observed and fitted frequencies for versions (c) and (d) are in Table IV.

Comparing versions (a) and (b), and (c) and (d), it is clear that in this example where the prevalence of disease is low in both groups, the association between the two tests in the diseased ($\lambda$) cannot be estimated reliably since correlations from $-0.38$ to zero give very similar fits to the data; the former is the minimum possible with these data since it corresponds to $d_1 = d_2 = 0$. Consequently no weight can be attached to the negative correlation between tests in the diseased at the boundary maximum likelihood solution, and the estimates of the sensitivities of the two tests are not reliably estimated without knowledge of the dependency between the tests in the diseased. Examination of Table IV shows that, although the fitted frequencies for the unobserved cells ($d_1$ and $d_2$) differ between versions (c) and (d), the values of the observed totals, $d_g + D_g$, are very similar. The prevalence of disease is low and consequently the frequencies in the non-diseased groups predominate in the maximum-likelihood fitting. Since $d_1$ and $d_2$ are known to be at least 9 and 8, respectively, version (c) is a better fit that model (d), but this information would not usually be
Table III. Fitted parameter values for data of Table II: (a) independence model, $\lambda = \delta = 1$; (b) independence in non-diseased group $\delta = 1$; (c) independence in diseased group $\lambda = 1$; (d) without restriction. The figures in parentheses below $\lambda$ and $\delta$ are the corresponding correlations between the diagnostic tests in the diseased and non-diseased, respectively.

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0.6226</td>
<td>0.726</td>
<td>0.6226</td>
<td>0.726</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0.6226</td>
<td>0.726</td>
<td>0.6226</td>
<td>0.726</td>
</tr>
<tr>
<td>$f_1$</td>
<td>0.0559</td>
<td>0.0558</td>
<td>0.0555</td>
<td>0.0554</td>
</tr>
<tr>
<td>$f_2$</td>
<td>0.0245</td>
<td>0.0244</td>
<td>0.0243</td>
<td>0.0243</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0.0059</td>
<td>0.00506</td>
<td>0.0059</td>
<td>0.00506</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.01954</td>
<td>0.01676</td>
<td>0.01954</td>
<td>0.01676</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1.0</td>
<td>0.8576</td>
<td>1.0</td>
<td>0.8576</td>
</tr>
<tr>
<td>$\delta$</td>
<td>(0)</td>
<td>(−0.3774)</td>
<td>(0)</td>
<td>(−0.3774)</td>
</tr>
<tr>
<td>$G$ (log-likelihood test)</td>
<td>172.31 (6 d.f.)</td>
<td>172.14 (5 d.f.)</td>
<td>48.46 (5 d.f.)</td>
<td>48.13 (4 d.f.)</td>
</tr>
</tbody>
</table>

Table IV. Fitted frequencies for data of Table II: (c) independence in diseased group $\lambda = 1$; (d) without restriction.

<table>
<thead>
<tr>
<th>Cell</th>
<th>Observed</th>
<th>(c) $\lambda = 1$</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>11</td>
<td>13.11</td>
<td>13.11</td>
</tr>
<tr>
<td>$b_1$</td>
<td>9</td>
<td>7.95</td>
<td>7.95</td>
</tr>
<tr>
<td>$c_1$</td>
<td>9</td>
<td>7.95</td>
<td>7.95</td>
</tr>
<tr>
<td>$d_1$</td>
<td>—</td>
<td>4.81</td>
<td>0</td>
</tr>
<tr>
<td>$A_1$</td>
<td>28</td>
<td>41.32</td>
<td>41.29</td>
</tr>
<tr>
<td>$B_1$</td>
<td>82</td>
<td>97.09</td>
<td>97.03</td>
</tr>
<tr>
<td>$C_1$</td>
<td>228</td>
<td>274.75</td>
<td>274.58</td>
</tr>
<tr>
<td>$D_1$</td>
<td>—</td>
<td>5280.03</td>
<td>5285.1</td>
</tr>
<tr>
<td>$d_1 + D_1$</td>
<td>5360</td>
<td>5284.84</td>
<td>5285.1</td>
</tr>
<tr>
<td>$a_2$</td>
<td>22</td>
<td>19.89</td>
<td>19.89</td>
</tr>
<tr>
<td>$b_2$</td>
<td>11</td>
<td>12.05</td>
<td>12.05</td>
</tr>
<tr>
<td>$c_2$</td>
<td>11</td>
<td>12.05</td>
<td>12.05</td>
</tr>
<tr>
<td>$d_2$</td>
<td>—</td>
<td>7.31</td>
<td>0</td>
</tr>
<tr>
<td>$A_2$</td>
<td>32</td>
<td>18.68</td>
<td>18.71</td>
</tr>
<tr>
<td>$B_2$</td>
<td>59</td>
<td>43.91</td>
<td>43.97</td>
</tr>
<tr>
<td>$C_2$</td>
<td>171</td>
<td>124.25</td>
<td>124.42</td>
</tr>
<tr>
<td>$D_2$</td>
<td>—</td>
<td>2387.85</td>
<td>2394.9</td>
</tr>
<tr>
<td>$d_2 + D_2$</td>
<td>2320</td>
<td>2395.16</td>
<td>2394.9</td>
</tr>
</tbody>
</table>

available. Versions (c) and (d) fit the data much better than (a) and (b), providing evidence of dependence between the tests in the non-diseased groups. For the non-diseased the dependence parameter is greater than 1 corresponding to correlated false positives (correlation 0.17). Although there is evidence of dependence between tests in the non-diseased, estimates of the sensitivities and specificities of the two tests are scarcely influenced by this dependence (Table III). Versions (c) and (d) do not fit well based on a test statistic of 48 approximately distributed as a $\chi^2$ with 5 and 4 degrees of freedom, respectively. Comparing the fitted frequencies with the observed (Table IV) in group 1, the fitted frequencies for $A$, $B$ and $C$ are
all greater than the observed, and the opposite occurs in group 2. Clearly the specificities and
dependence between tests in the non-diseased are not the same in groups 1 and 2, that is
they are not independent of prevalence. If version (c) is fitted to each age group separately
then there is a perfect fit since there are six parameters fitted to six independent frequencies. For the 40–59
year age group the parameter estimates were \( s_1 = 0.5501, s_2 = 0.5501, f_1 = 0.0450, f_2 = 0.0193, p_1 = 0.00635 \) and \( \delta = 5.658 \) (correlation = 0.14), and for the 60–70
year age group, \( s_1 = 0.6667, s_2 = 0.6667, f_1 = 0.0788, f_2 = 0.0353, p_2 = 0.01885 \) and \( \delta = 4.463 \)
(correlation = 0.19). The false positive rates are almost twice as high in the older age group
than in the younger age group.

5. DISCUSSION

In this paper we have discussed the situation when the two tests have been applied to two
distinct populations with those negative on both tests not verified by the reference stan-
dard. In this case it has been shown that the TPR and FPR of both tests, the population
prevalences and the conditional dependence between tests, may be estimated using maximum
likelihood.

We have not developed methods for estimating confidence intervals for the parameters.
For some of the parameters, asymptotic standard errors may be produced from the inverse
of the matrix of second derivatives of the log-likelihood. This would be improved by a logit
transformation of those parameters restricted to lie in the range 0 to 1, but would not be
valid when a dependency parameter is estimated on the boundary of its permitted range,
as occurred in the example we used. Lower limits of the standard errors could be obtained
from the expressions in step 3 of Section 3.3.2 ignoring the uncertainty in the numbers of
diseased and non-diseased. This may be satisfactory for the specificities when most of the
group are non-diseased since then the denominators will have little variability compared with
the numerators. Perhaps more important are the non-sampling errors arising from uncertainty
on the model and this is particularly so for the sensitivities when only a minority are in the
diseased group.

The application of the methods depend on their distinctiveness in terms of the prevalence of
disease. If the populations have similar prevalences then the estimates will not be very accurate
and in the limiting situation when the two prevalences are equal then the estimation procedure
breaks down. Thus the methods are likely to be most useful when the two populations are
chosen to have a relatively low and a relatively high prevalence. One situation which may
lead to this is when the populations are different age groups within the same population.

In our example, the results showed that the specificities of the two tests are likely to
vary with disease prevalence. Hence, the model did not provide a simplifying description,
indicating that it is difficult to generalize the performance of these tests to a wider set of
circumstances. The goodness-of-fit test in our methodology is thus useful in determining the
extent to which it is possible to generalize test performance characteristics to populations with
different underlying disease rates.

The method could be extended to a situation with more than two groups and examination
of the estimates of the sensitivities and specificities in each group may then provide insight
into what factors determine these test characteristics. If sensitivity or specificity, or both,
are influenced by a continuous variable, such as age, then a way forward may be to model
these parameters as a function of age, perhaps a linear function of the logit of sensitivity or specificity on age.

ACKNOWLEDGEMENTS

We wish to thank the Public Health Research Development Committee (PHRDC) of the National Health and Medical Research Committee (NHMRC) for financial support. We are grateful to Dr G. Castiglione and colleagues who provided the data used in the example, and to Dr Stephen Walter for discussions and suggestion of the use of the EM algorithm. Finally we are grateful to the referees who suggested how the methods could be extended.

REFERENCES