STATISTICS IN MEDICINE, VOL. 16, 2157–2175 (1997)

EFFECTS OF DEPENDENT ERRORS IN THE ASSESSMENT OF DIAGNOSTIC TEST PERFORMANCE

VICKI L. TORRANCE-RYNARD* AND STEPHEN D. WALTER

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON L8N 3Z5, Canada

SUMMARY

Latent class models can be used to assess diagnostic test performance when there is no perfectly accurate gold standard test available for comparison. These models usually assume independent errors between the tests, conditional on the true disease state of the subject. Maximum likelihood estimates of the prevalence of the disease and the error rates of diagnostic tests are then obtained. This paper examines the impact of error dependencies between binary diagnostic tests on the parameter estimates obtained from the latent class models. The independence model often gives parameter estimates having relatively small bias, but in some situations (for example, when disease prevalence is low and the tests have low specificity, such as in population screening) the bias may be more serious. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2157–2175 (1997)
No. of Figures: 3 No. of Tables: 3 No. of References: 13

1. INTRODUCTION

Diagnostic tests and observations use clinical information to classify subjects by disease status. In the simplest situation, the classification is binary (disease/no disease). To assess the performance of a given diagnostic method, one requires probabilistic estimates of the accuracy of the method conditional on the true disease state of the subject. For a binary test, one may express test performance in terms of sensitivity (the probability of a correct, positive result for a true disease case) and specificity (the probability of a correct, negative result for a true non-case). The complements of these parameters are known as the false negative and false positive rates, respectively.

Clinicians using a diagnostic test may prefer to estimate the conditional probability of being a case, given that a positive test result has been obtained; this parameter is known as the positive predictive value and can be expressed as a simple function of the sensitivity, specificity and disease prevalence (or the prior probability of being a case).[1] Similarly, one may obtain the negative predictive value, the probability of being a non-case, conditional on having a negative test result. Generalizations of test error rates and predictive values to multi-level tests are possible.[1]

The preferred situation for the estimation of diagnostic test performance is when samples of disease cases and non-cases are available, whose status are known without error. The estimation

* Correspondence to: Vicki L. Torrance-Rynard, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON L8N 3Z5, Canada

CCC 0277–6715/97/192157–19$17.50

Received July 1996
Revised December 1996

of sensitivity and specificity is then straightforward, being based on binomial sampling theory. Such error-free classification requires a 'gold standard' measurement of disease status, but in many practical clinical situations, a gold standard method does not exist; the assessment of each subject must then remain probabilistic. Commonly, one assumes that the best available method is the gold standard. Comparison of the diagnostic test to a method that is not truly a gold standard leads to biased results.[2] Even if there is a perfect gold standard available, it may involve invasive, expensive and potentially hazardous techniques, for example, tissue biopsy of a suspected cancer. Ethical considerations often preclude the use of such measurements on all subjects, especially on individuals intended to serve as non-cases.

The use of latent class models has been proposed to deal with the common situation when there is no gold standard available.[2,3] In the typical application of the method, one makes several observations or tests on each subject, presumes all of the observations are potentially subject to error, and in particular, assumes none represents a gold standard. Examples include the use of three different tests in screening for colorectal cancer,[4] several alternative diagnostic criteria for the diagnosis of autism,[5] evaluation of several tissue biopsy samples from heart transplant patients to determine the occurrence of cardiac tissue rejection,[6] and the evaluation of chest X-rays for evidence of pleural thickening in hard-rock miners.[7]

Three observations per subject is the minimum amount of information needed for unconstrained estimation of the relevant test parameters, with each observation classifying subjects as diseased or not diseased. One presumes that each subject is indeed truly in one or the other of these states, but direct observation is not possible. The unobserved, true state is known as the latent variable. The likelihood can be written in terms of seven parameters (three sensitivity values, three specificity values and the disease prevalence in the sample). With the total sample size regarded as fixed, there are seven degrees of freedom available, and seven parameters to estimate, so the likelihood model is saturated.

One of the key assumptions in almost all the theoretical development and diagnostic applications to date is conditional independence of the errors. While an independence model may be appropriate in some situations, in other cases it is a strong assumption that may not be justified.

We consider the impact of the error dependencies between binary tests on parameter estimates obtained from latent class models based on the usual (but now false) assumption of independent errors. Using simulation results, we assess the bias in the latent class parameters as a function of the pattern and strength of the error dependence, and of the parameter values themselves. We also evaluate the accuracy of their asymptotic standard errors. Recommendations are made concerning the situations in which the independence model may produce badly biased results.

2. METHODS

2.1. Terminology and notation

The notation developed here closely follows that presented by Walter and Irwig.[2] The true prevalence of the disease is θ. The false negative rate for the ith test is β_i; this is the probability that the ith test classifies a subject who is truly positive for the disease as not having the disease. The probability that the ith test classifies a subject who truly does not have the disease as a disease case is the false positive rate, given by α_i. Thus, the sensitivity and specificity of the ith test are $1 - \beta_i$ and $1 - \alpha_i$, respectively.

If there are r binary tests that classify each subject, then there are 2^r possible combinations of classifications for each subject. We refer to these combinations as outcome categories. For $r = 3$ there are $2^3 = 8$ outcome categories. We represent the classification of the ith test by the random variable X_i where

$$X_i = \begin{cases} 0 & \text{if the classification is negative,} \\ 1 & \text{if the classification is positive.} \end{cases}$$

We represent the true disease status of the subject by Y where

$$Y = \begin{cases} 0 & \text{if the subject is negative,} \\ 1 & \text{if the subject is positive.} \end{cases}$$

The number of outcome categories determines the number of statistical degrees of freedom, d.f., available for estimation. For a fixed number of subjects there are $2^r - 1$ d.f.

2.2. Maximum likelihood parameter estimation

We show probabilities and likelihoods for $r = 3$; extensions to $r > 3$ are straightforward. We calculate the probability that a subject is in a particular outcome category conditional on the subject's true disease status. The unconditional probability that a subject is classified as negative by all three tests is the sum of the two conditional probabilities:

$$\begin{aligned} \Pr(X_1 = 0, X_2 = 0, X_3 = 0) &= \Pr(Y = 1)\Pr(X_1 = 0, X_2 = 0, X_3 = 0 \mid Y = 1) \\ &\quad + \Pr(Y = 0)\Pr(X_1 = 0, X_2 = 0, X_3 = 0 \mid Y = 0) \\ &= \theta\beta_1\beta_2\beta_3 + (1-\theta)(1-\alpha_1)(1-\alpha_2)(1-\alpha_3). \end{aligned}$$

There are eight such unconditional probabilities

$$\Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \theta \prod_{i=1}^{3} \beta_i^{1-x_i}(1-\beta_i)^{x_i} + (1-\theta) \prod_{i=1}^{3} \alpha_i^{x_i}(1-\alpha_i)^{1-x_i}. \quad (1)$$

In general, for r tests and n subjects where $n(\mathbf{x})$ are classified as the vector $\mathbf{x} = (x_1, x_2, \ldots, x_r)$ such that $\sum_{\mathbf{x}} n(\mathbf{x}) = n$, the likelihood function for the whole data set is proportional to the joint probability function

$$L \propto \prod_{\mathbf{x}} \left[\theta \prod_{i=1}^{r} \beta_i^{1-x_i}(1-\beta_i)^{x_i} + (1-\theta) \prod_{i=1}^{r} \alpha_i^{x_i}(1-\alpha_i)^{1-x_i} \right]^{n(\mathbf{x})}.$$

For estimation purposes, it is more convenient to maximize $\ln(L)$ rather than L. We can use direct numerical maximization, or we can adopt the EM algorithm.[8] We can obtain large sample variances and covariances from the inverse of the expected Fisher information matrix; details appear elsewhere.[9] The maximization provides the maximum likelihood estimates of the parameters.

2.3. Development of dependence parameters

We quantify the conditional dependence of classification errors between tests as terms to add to the independent conditional probabilities, following the approach used by Vacek.[10] Because the dependence is conditional, we define separate terms for truly positive subjects and truly negative subjects. The dependence between tests i and j is δ_{ij} for truly positive subjects and ε_{ij} for truly negative subjects. We consider only positive associations since negative associations are unlikely to occur in practice; hence, $\delta_{ij} \geqslant 0$ and $\varepsilon_{ij} \geqslant 0$, for all $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, r$, $i \neq j$.

We modify the independent unconditional probabilities (1) to incorporate dependence between pairs of tests as follows:

$$\Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \theta \left[\prod_{i=1}^{3} \beta_i^{1-x_i}(1-\beta_i)^{x_i} + \sum_{i,j,i<j}^{3} (-1)^{|x_i - x_j|}\delta_{ij} \right]$$

$$+ (1-\theta)\left[\prod_{i=1}^{3} \alpha_i^{x_i}(1-\alpha_i)^{1-x_i} + \sum_{i,j,i<j}^{3} (-1)^{|x_i - x_j|}\varepsilon_{ij} \right]. \quad (2)$$

Note a constraint on the probabilities for a true positive case when X_1 is held constant (while the other tests may vary):

$$\Pr(X_1 = 0 \mid Y = 1) = \beta_1 \quad \text{and} \quad \Pr(X_1 = 1 \mid Y = 1) = 1 - \beta_1.$$

First, the probability that test 1 classifies a truly positive subject as negative is equated to β_1. The second equation is a consequence of the first. There are similar constraints for negative subjects and for the other tests. The development of the dependence terms δ_{ij} and ε_{ij} respect these constraints. When dependence is present we have from (2):

$$\begin{aligned} E(X_1 \mid Y = 1) &= (1-\beta_1)\beta_2\beta_3 - \delta_{12} - \delta_{13} + \delta_{23} + (1-\beta_1)\beta_2(1-\beta_3) - \delta_{12} + \delta_{13} - \delta_{23} \\ &\quad + (1-\beta_1)(1-\beta_2)\beta_3 + \delta_{12} - \delta_{13} - \delta_{23} \\ &\quad + (1-\beta_1)(1-\beta_2)(1-\beta_3) + \delta_{12} + \delta_{13} + \delta_{23} \\ &= 1 - \beta_1. \end{aligned}$$

Similarly, $E(X_2 \mid Y = 1) = 1 - \beta_2$, and so on. Thus, the addition of the dependence terms leaves the expected values of the observations unchanged and we do not violate the constraints.

We now consider the interpretation of the dependence terms in terms of corresponding covariances. For instance, it can be shown that $E(X_1X_2 \mid Y = 1) = (1-\beta_1)(1-\beta_2) + 2\delta_{12}$. So, $\operatorname{cov}(X_1, X_2 \mid Y = 1) = 2\delta_{12}$. Thus, $\delta_{12} = 1/2 \operatorname{cov}(X_1, X_2 \mid Y = 1)$. Hence, δ_{12} is one half of the conditional covariance between the classifications of tests 1 and 2 for true cases. Similarly, ε_{12} is one half of the conditional covariance between the classifications of tests 1 and 2 for the true non-cases. We can formulate the other dependence parameters that involve test 3 similarly. When $r = 4$ there are four constraints. One can show that $\delta_{12} = 1/4 \operatorname{cov}(X_1, X_2 \mid Y = 1)$ and similarly for the other dependence parameters.

This paper concentrates on situations with $r = 3$ and $r = 4$, and restricts attention to pairwise dependence terms that, as shown above, have an intuitive interpretation in terms of covariances. Higher order dependencies are possible in principle, but they are more difficult to rationalize, and less likely to occur in practice.

Others have recently begun to investigate conditional dependence between diagnostic tests using other approaches to representing the dependence. Qu *et al.* have developed random effects models to model the conditional dependence between tests in latent class analysis.[11] Brenner has assessed the correlation of diagnostic test errors and has shown that the correlation can bias reliability indices using numerical examples.[12]

2.4. Bounds on dependence parameter values

The dependence parameters are bounded because they are defined in terms of probabilities. We consider positive associations, so we need only consider the maxima for the dependence parameter values. These are determined by θ, α and β for four cases of interest:

(i) dependence between tests 1 and 2 when $r=3$;
(ii) equal dependence between all pairs of tests when $r=3$;
(iii) dependence between tests 1 and 2 when $r=4$;
(iv) equal dependence between all pairs of tests when $r=4$.

We consider α, $\beta < 50$ per cent, which covers most practical situations.

For case (i), we consider the dependence between tests 1 and 2 when classifying true cases. For a true case the probability of an outcome is

$$\Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3 \mid Y = 1) = \prod_{i=1}^{3} [\beta_i^{(1-x_i)}(1-\beta_i)^{x_i}] + (-1)^{|x_1 - x_2|}\delta_{12}$$

and for each (x_1, x_2, x_3), $0 \leqslant \Pr(x_1, x_2, x_3 \mid Y = 1) \leqslant 1$. So one can calculate restrictions for δ_{12} for each outcome. One can show (see the Appendix) that δ_{12} is constrained by

$$\delta_{12} \leqslant \beta_1(1-\beta_2)\beta_3 \quad \text{and} \quad \delta_{12} \leqslant (1-\beta_1)\beta_2\beta_3.$$

Similarly, one can determine

$$\varepsilon_{12} \leqslant \alpha_1(1-\alpha_2)\alpha_3 \quad \text{and} \quad \varepsilon_{12} \leqslant (1-\alpha_1)\alpha_2\alpha_3.$$

For case (ii) there is equal dependence, δ or ε, between all three pairs of tests. We can follow a similar procedure to determine the maximum value for the dependence parameters. For a true case

$$\Pr(\mathbf{x} = (x_1, x_2, x_3) \mid Y = 1) = \prod_{i=1}^{3} \beta_i^{(1-x_i)}(1-\beta_i)^{x_i} + [(-1)^{|x_1-x_2|} + (-1)^{|x_1-x_3|} + (-1)^{|x_2-x_3|}]\,\delta.$$

One can show that the maximum value of δ is determined by requiring

$$\delta \leqslant \beta_1\beta_2(1-\beta_3), \delta \leqslant \beta_1(1-\beta_2)\beta_3 \quad \text{and} \quad \delta \leqslant (1-\beta_1)\beta_2\beta_3.$$

Similarly, we determine the maximum value of ε by requiring

$$\varepsilon \leqslant \alpha_1\alpha_2(1-\alpha_3), \varepsilon \leqslant \alpha_1(1-\alpha_2)\alpha_3 \quad \text{and} \quad \varepsilon \leqslant (1-\alpha_1)\alpha_2\alpha_3.$$

We determine the maximum values for $r=4$ similarly. Consider dependence between tests 1 and 2 in case (iii). We use the constraints

$$\delta_{12} \leqslant \beta_1(1-\beta_2)\beta_3\beta_4 \quad \text{and} \quad \delta_{12} \leqslant (1-\beta_1)\beta_2\beta_3\beta_4$$

to determine the maximum value of δ_{12} when $r=4$ with dependence only between tests 1 and 2. The constraints

$$\varepsilon_{12} \leqslant \alpha_1(1-\alpha_2)\alpha_3\alpha_4 \text{ and } \varepsilon_{12} \leqslant (1-\alpha_1)\alpha_2\alpha_3\alpha_4$$

determine the maximum value for ε_{12} in this situation.

In case (iv) there is equal dependence, δ or ε, between all paris of tests. Using a similar (although lengthier) method, one can show that the maximum for δ is determined by restricting

$$\delta \leqslant \frac{1}{2} \sum_{i=1}^{4} \beta_i^{x_i}(1-\beta_i)^{1-x_i}$$

for all configurations where $\sum_{i=1}^{4} x_i = 2$, that is, for individuals with two positive and two negative observations. Similarly, one can show that

$$\varepsilon \leqslant \frac{1}{2} \sum_{i=1}^{4} \alpha_i^{x_i}(1-\alpha_i)^{1-x_i}$$

for all configuration where $\sum_{i=1}^{4} x_i = 2$.

2.5. Outcome summaries

For each set of parameter values (α, β, θ, δ and ε) we replicated the simulation to generate a sample distribution of estimates for α, β and θ and calculated the estimated bias, by subtracting the input parameter value from the sample mean, and standard deviation of each parameter. We summarized results in tables and graphs to show trends in bias for various combinations of α, β and θ.

We compared the true standard deviation of the simulated estimates with the estimated standard error from large sample theory obtained from the Latent program (this is a software program developed by S. D. Walter) using expected frequencies as input. We use the ratio of standard error to standard deviation as an outcome summary for this comparison.

2.6. Selection of parameter values

The input parameters required for each simulation are: θ; $\beta_1, \beta_2, \ldots, \beta_r$; $\alpha_1\alpha_2, \ldots, \alpha_r$; δ_{ij} and ε_{ij}, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, r$ and $i \neq j$; the number of subjects in the sample, n; and the number of replications, R. Explanations of how we chose or calculated these values follow:

Prevalence: We set prevalence, θ, at 5, 15 and 40 per cent, considered to be low, medium and high. Values over 50 per cent were excluded, but one can infer their expected results by symmetry. Suppose $\theta = 60$ per cent for a given condition and $\alpha_1 = 10$ per cent and $\beta_1 = 25$ per cent. We can infer results for this scenario by considering the complementary event where $\theta = 40$ per cent and $\alpha_1 = 25$ per cent and $\beta_1 = 10$ per cent.

Error rates: We set the error rates (α and β) at 10 per cent and 25 per cent. Their combinations represent situations where all tests are equally accurate.

Dependence parameters: We express the dependence parameters as proportions of their upper bounds. We set the dependence values at no dependence, one half of maximum dependence and maximum dependence.

Sample size: We set the sample size, n, at 2000 for simulations with $r = 3$ and at 4000 for simulations with $r = 4$. These numbers were sufficiently large to avoid small outcome category frequencies (<5).

Number of replications: We determined the number of replications, R, needed for each combination of the above parameter values so that at the 5 per cent significance level there is at least 80 per cent power to detect a meaningful difference between the mean of the parameter estimates and the true parameter value, if such difference truly exists. We defined a meaningful difference as a bias of 20 per cent or more. Let ϕ be a general parameter of interest where ϕ_0 is its true value and ϕ_1 is a meaningfully different result for the mean of the estimates. Then

$$R = \left[\frac{\sigma(z_{\alpha/2} + z_{\beta})}{\phi_0 - \phi_1} \right]^2$$

where $z_{\alpha/2} = 1.96$ and $z_{\beta} = 0.84$.[13] We estimated σ from its large sample expected value. For any particular set of parameter values we determined R by the single parameter value that required the largest sample size providing that this size was at least 30, to avoid unstable results due to small sample sizes.

2.7. Simulation runs

We carried out simulations for all selected combinations of α, β and θ with each simulation replicated R times to obtain the means and standard deviations of the parameter estimates. We show typical results indicating general effects of various types of error dependence for $r = 3$ or $r = 4$. We concentrate on the situation where there exists a pair of dependent tests while other tests are independent.

We ran the simulations on a 486DX computer with 4 Mbytes of memory and a 220 Mbyte hard drive using programs written in Borland Pascal, version 7.0. The values for θ, α, β, δ, ε and n were input to the simulation program, which calculated the probability of a subject being in each outcome category, using equation (2). We used these probabilities to classify subjects into the outcome categories. We then input the resulting total frequencies of subjects in each outcome category into a latent class model to obtain estimates of the parameters. For the purpose of this study the Latent program was translated to Borland Pascal from FORTRAN by adapting the logic and rewriting the code. The latent class model estimates were obtained under an assumption of independence of errors between tests.

3. RESULTS

3.1. Bias

Dependence between two tests means that they classify subjects equivalently more often than when they are independent and the frequency of concordant errors will increase. If one mistakenly assumes independence, then the estimated probability that two concordant misclasifications are correct will increase. Thus, we expect the error rates for the dependent tests to be underestimated in the latent class model assuming independence. We anticipate similar effects with more than two dependent tests.

3.1.1. Three test case ($r = 3$)

Table I shows, for $\theta = 15$ per cent only, the results from a simulation in which $r = 3$ and where there was dependence between tests 1 and 2 when classifying both positive and negative subjects (δ and ε terms). Figures 1(*a*) to (*d*) show graphs of the bias in the error rate estimates by level of dependence, for all values of θ.

We will focus our description of the results mainly on the error rates. The true error rates for the two dependent tests are, by definition, identical. The only differences in their results were due to random variation, so, arbitrarily, only the error rates for test 1 are shown in the tables and

Table I. Means and standard deviations of parameter estimates in the case of three tests ($r = 3$), with dependence between tests 1 and 2 when classifying cases and non-cases, true prevalence (θ) = 15 per cent

True error rates	Dep.*	Mean of estimates (%) (standard deviation (%))				
		θ	β_1	β_3	α_1	α_3
$\beta = 10\%$	0	15·09 (1·05)	10·72 (2·64)	8·93 (2·49)	10·14 (0·92)	10·09 (0·86)
$\alpha = 10\%$	0·5	14·82 (1·03)	6·01 (2·90)	12·73 (2·82)	9·40 (0·93)	10·72 (0·91)
	1	15·00 (0·82)	2·32 (2·20)	16·06 (2·32)	8·86 (0·77)	10·95 (0·65)
$\beta = 10\%$	0	15·09 (2·23)	9·91 (4·94)	9·98 (4·65)	24·98 (1·37)	24·94 (1·65)
$\alpha = 25\%$	0·5	18·55 (1·10)	0·62 (1·41)	26·67 (2·82)	19·93 (1·17)	25·99 (1·09)
	1	24·07 (0·82)	0·00 (0·00)	34·45 (2·53)	13·74 (0·80)	24·81 (0·88)
$\beta = 25\%$	0	15·55 (1·77)	25·38 (4·05)	26·18 (6·05)	9·73 (0·96)	9·72 (1·24)
$\alpha = 10\%$	0·5	13·90 (1·00)	14·73 (3·51)	29·71 (3·98)	9·18 (1·13)	11·47 (0·86)
	1	13·02 (0·95)	5·81 (3·68)	32·41 (3·67)	8·75 (0·85)	12·62 (0·83)
$\beta = 25\%$	0	14·95 (4.56)	25·17 (7·62)	21·79 (9·12)	24·88 (1·82)	24·85 (2·09)
$\alpha = 25\%$	0·5	15·57 (1·30)	1·44 (2·34)	39·32 (3·09)	20·68 (1·39)	27·28 (0·96)
	1	21·76 (1·10)	0·00 (0·00)	44·38 (2·71)	14·04 (1·30)	26·49 (1·12)

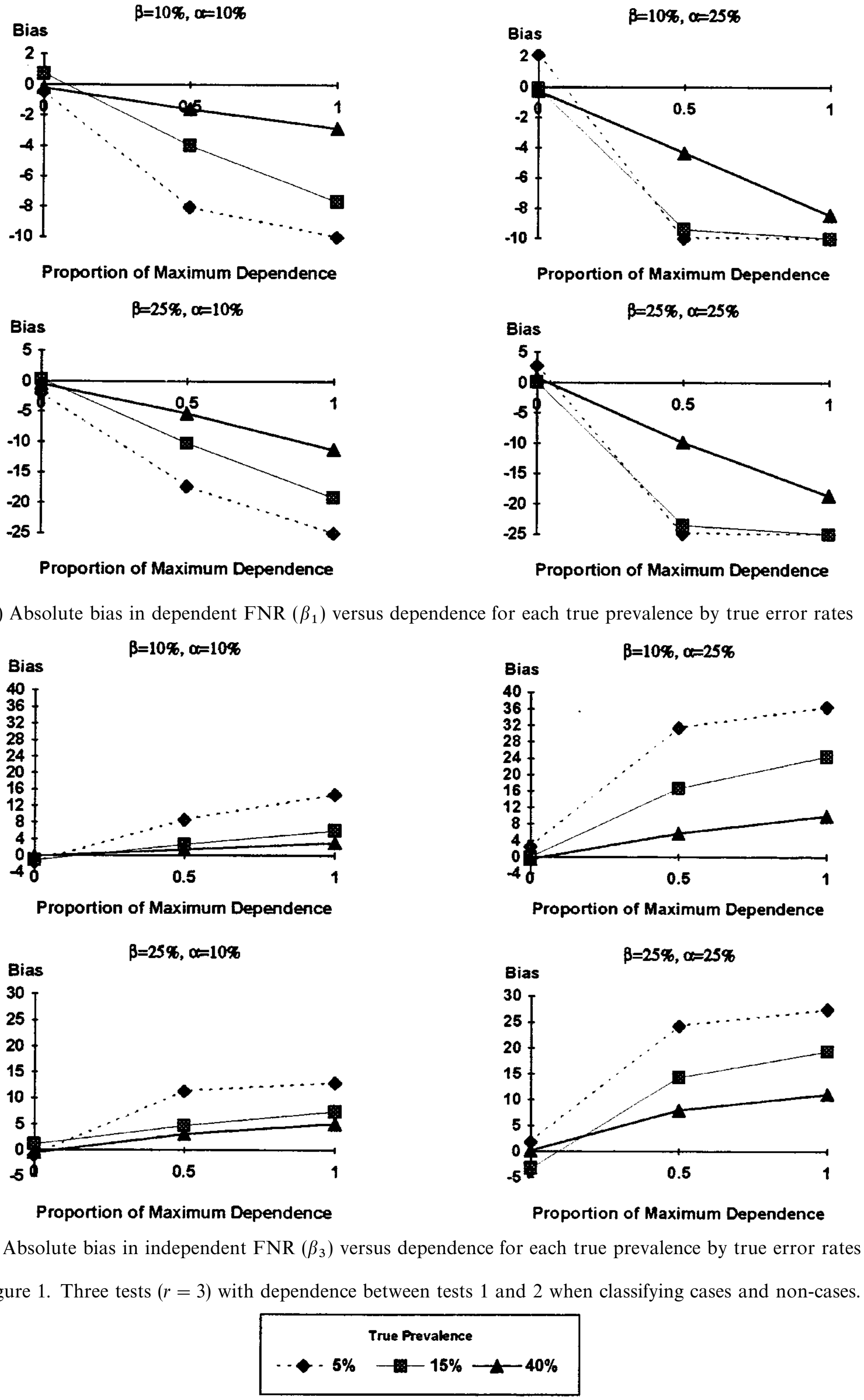
*Proportion of maximum dependence between tests 1 and 2 for both cases and non-cases
β represents the true false negative rates, $\beta_1 = \beta_2 = \beta_3$
α represents the true false positive rates, $\alpha_1 = \alpha_2 = \alpha_3$

figures. No meaningful bias occurred in the false negative rate for the independent test (β_3) when $\alpha = 10$ per cent. In addition, at high prevalences and low levels of dependence the dependent tests' false negative rates (β_1 and β_2) were not bias. However, these rates for the dependent tests are seriously underestimated and the independent test's false negative rate (β_3) is overestimated when θ is low and when $\alpha = 25$ per cent. Figures 1(*a*) and (*b*) show the trends in bias for the dependent and independent tests, respectively.

The effects on the false positive rates are weaker. The false positive rates for the dependent tests (α_1 and α_2) are sometimes underestimated. At $\theta = 5$ per cent and $\theta = 15$ per cent this underestimation is only strong when $\alpha = 25$ per cent (Figure 1(*c*)). The independent test's false positive rate (α_3) is mostly not affected by the dependence except when $\theta = 40$ per cent and $\beta = 25$ per cent which results in slight overestimation (Figure 1(*d*)).

The results of simulations that involved only one kind of dependence between tests 1 and 2 (that is, when classifying true non-cases or when classifying true cases) are not shown here in detail, but we describe their main features below. These features help in the interpretation of the trends found for the earlier simulation where both types of dependency were present.

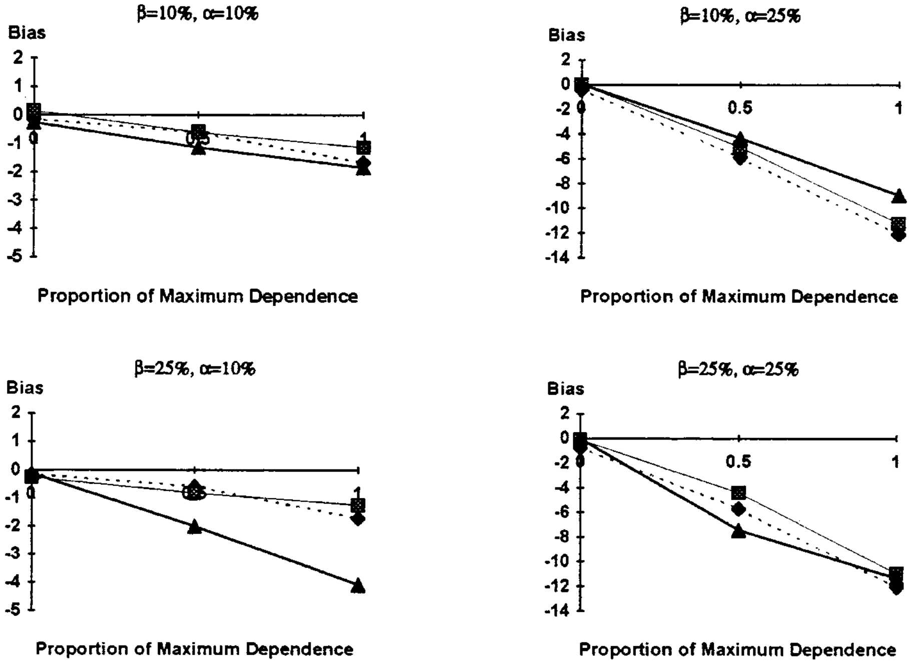
Dependence when classifying truly negative subjects (ε terms only) causes overestimation of θ (when there is a non-zero false positive rate) since simultaneous misclassifications, by tests 1 and 2, of negative subjects as positive are more likely. This leads to an overabundance of concordant positive classifications, implying positive bias in the estimate of θ and negative bias in the false positive rate estimates for the dependent tests (α_1 and α_2). When α is high, the biases are larger since more false positive errors occur. The biases are also large if θ is low because there are more negative subjects available to be classified equivalently by the dependent tests. The false negative rates for the dependent tests (β_1 and β_2) are underestimated while the false negative rate for the independent test (β_3) is slightly overestimated, especially for low θ and high α.



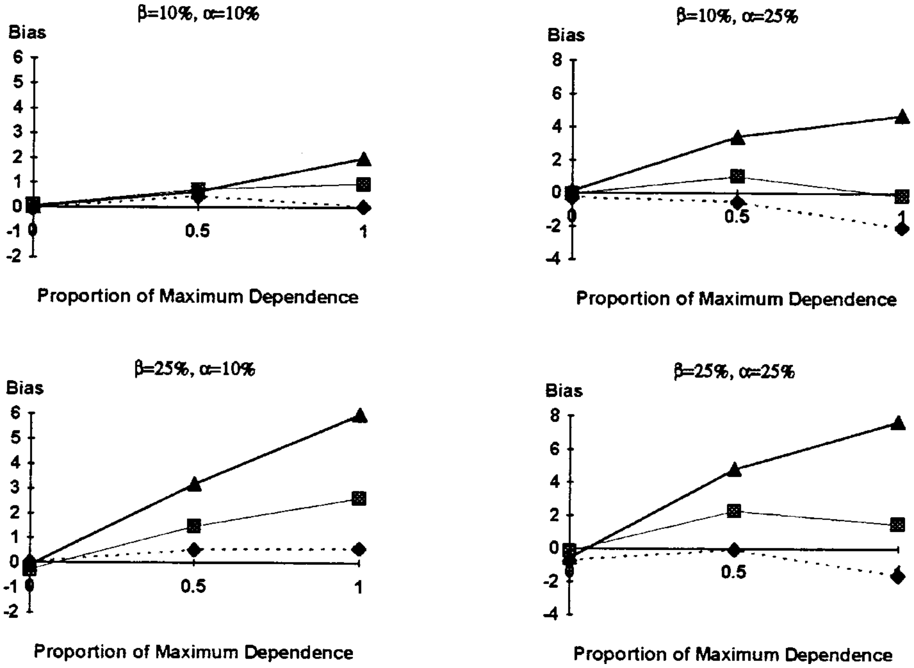
(a) Absolute bias in dependent FNR (β_1) versus dependence for each true prevalence by true error rates

(b) Absolute bias in independent FNR (β_3) versus dependence for each true prevalence by true error rates

Figure 1. Three tests ($r = 3$) with dependence between tests 1 and 2 when classifying cases and non-cases.



(c) Absolute bias in dependent FPR (α_1) versus dependence for each true prevalence by true error rates



(d) Absolute bias in independent FPR (α_3) versus dependence for each true prevalence by true error rates

Figure 1. (Continued)

Table II. Means and standard deviations of parameter estimates in the case of three tests ($r = 3$), with dependence between tests 1 and 2 set such that the covariance between test 1 and 2 is equal to the four test ($r = 4$) case when classifying cases and non-cases, true prevalence (θ) = 15 per cent

True error rates	Dep.* ($2 \times$ Dep for $r = 4$)	Mean of estimates (%) (standard deviation (%))				
		θ	β_1	β_3	α_1	α_3
$\beta = 10\%$	0	15·09 (0·98)	9·92 (2·41)	10·38 (2·14)	10·08 (0·70)	9·74 (0·64)
$\alpha = 10\%$	0·5	14·95 (1·00)	9·88 (2·40)	11·31 (2·36)	9·44 (0·79)	10·26 (0·76)
	1	15·00 (1·07)	8·14 (2·20)	11·54 (2·62)	9·70 (0·86)	10·28 (0·94)
$\beta = 10\%$	0	15·00 (1·98)	9·15 (4·66)	9·35 (5·38)	24·91 (1·56)	25·09 (1·31)
$\alpha = 25\%$	0·5	15·67 (1·56)	2·91 (3·51)	19·37 (3·87)	23·25 (1·26)	26·09 (1·31)
	1	18·29 (1·05)	0·41 (0·90)	25·59 (2·64)	20·24 (1·12)	25·78 (1·25)
$\beta = 25\%$	0	14·73 (1·25)	24·20 (5·05)	23·50 (3·73)	10·03 (0·89)	9·94 (0·89)
$\alpha = 10\%$	0·5	14·58 (1·27)	21·73 (4·67)	28·17 (3·90)	9·59 (0·73)	10·39 (1·00)
	1	13·91 (1·33)	18·27 (4·78)	27·09 (3·78)	9·42 (0·94)	11·12 (1·06)
$\beta = 25\%$	0	15·22 (3.51)	25·23 (5·97)	24·78 (5·50)	24·66 (1·56)	25·06 (1·66)
$\alpha = 25\%$	0·5	14·46 (3·21)	10·89 (7·76)	33·40 (5·11)	23·25 (1·93)	27·03 (1·45)
	1	15·14 (0·90)	0·98 (2·08)	38·43 (3·03)	20·66 (0·95)	27·61 (0·81)

*Proportion of maximum dependence between tests 1 and 2 for both cases and non-cases in the corresponding four test case ($r = 4$), that is, $\delta_{12}(r = 3) = 2\delta_{12}(r = 4)$

β represents the true false negative rates, $\beta_1 = \beta_2 = \beta_3$

α represents the true false positive rates, $\alpha_1 = \alpha_2 = \alpha_3$

Dependence when classifying positive subjects (δ terms) causes the dependent tests to classify truly positive subjects equivalently more often than under independence. Thus, the effects are 'symmetric' to those just described. The bias is made stronger by high β. High θ also leads to large bias because of the presence of more true cases to misclassify.

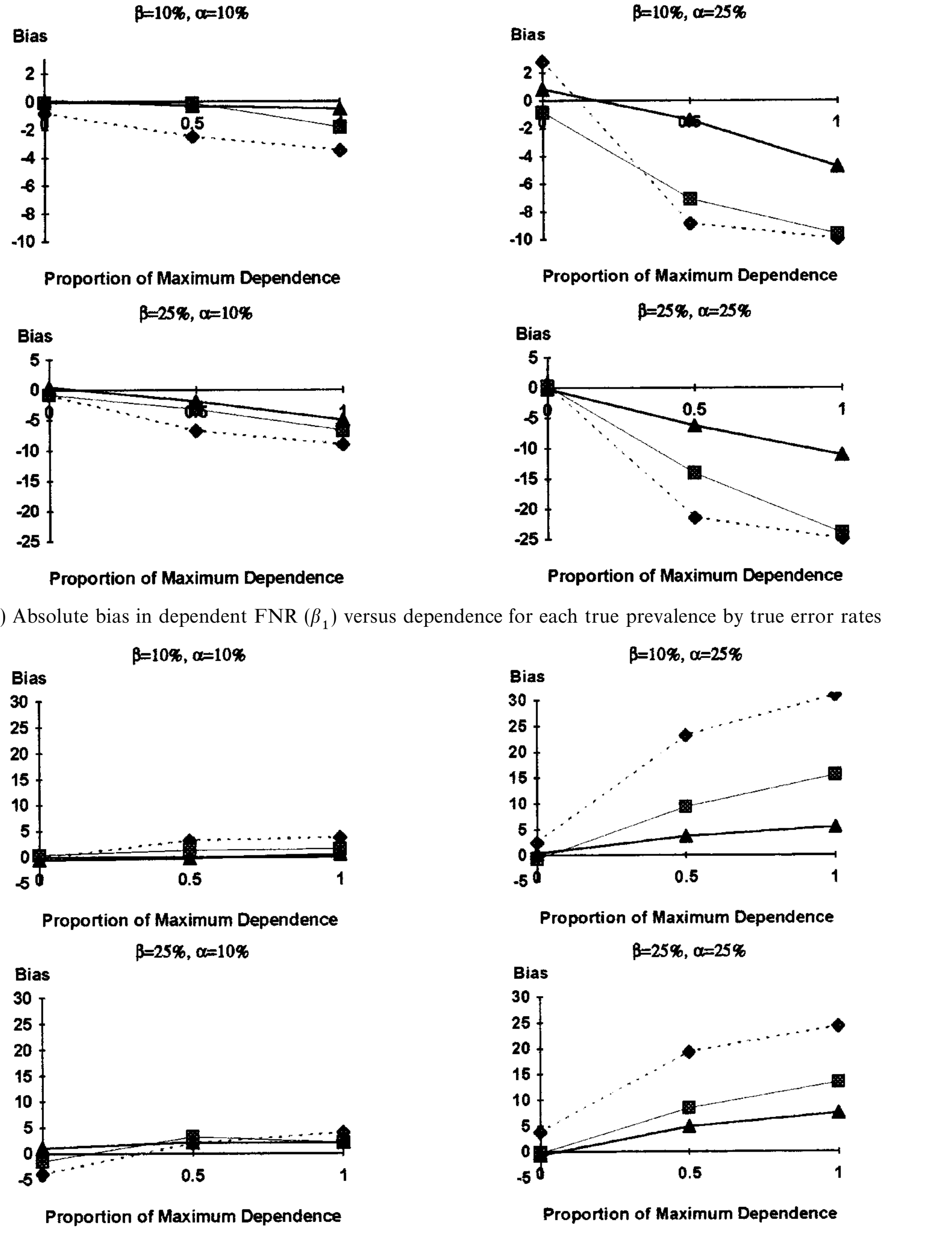
The type of dependence that leads to the larger bias, when both types are present, is determined by θ. If $\theta < 50$ per cent then the strongest bias comes from the ε terms as seen in Table I and Figures 1(*a*) to (*d*). If $\theta > 50$ per cent then greater bias results from the δ terms.

3.1.2. Comparison of three tests ($r = 3$) to four tests ($r = 4$)

When r changes, δ and ε represent different proportions of the conditional covariance between the two dependent tests. For $r = 3$, $\delta_{12}(3 \text{ obs}) = 1/2\text{cov}(X_1, X_2 \mid Y = 1)$, but for $r = 4$, $\delta_{12}(4 \text{ obs}) = 1/4\,\text{cov}(X_1, X_2 \mid Y = 1)$. To assess validly the effect of adding a further independent test to a situation where there is dependence between tests 1 and 2, the covariance between tests 1 and 2 should remain constant. To achieve this objective, we constrained δ for further simulations with $r = 3$ such that

$$\delta_{12}(r = 3) = 2\delta_{12}(r = 4) \tag{3}$$

with a similar constraint on ε_{12}. We ran the simulations for $r = 4$ using the three dependence levels as before: no dependence; half of maximum dependence, and maximum dependence. We then set dependence values for corresponding simulations with $r = 3$ using equation (3). Table II and Figure 2 show the results for $r = 3$ and Table III and Figure 3 show the corresponding results for $r = 4$.



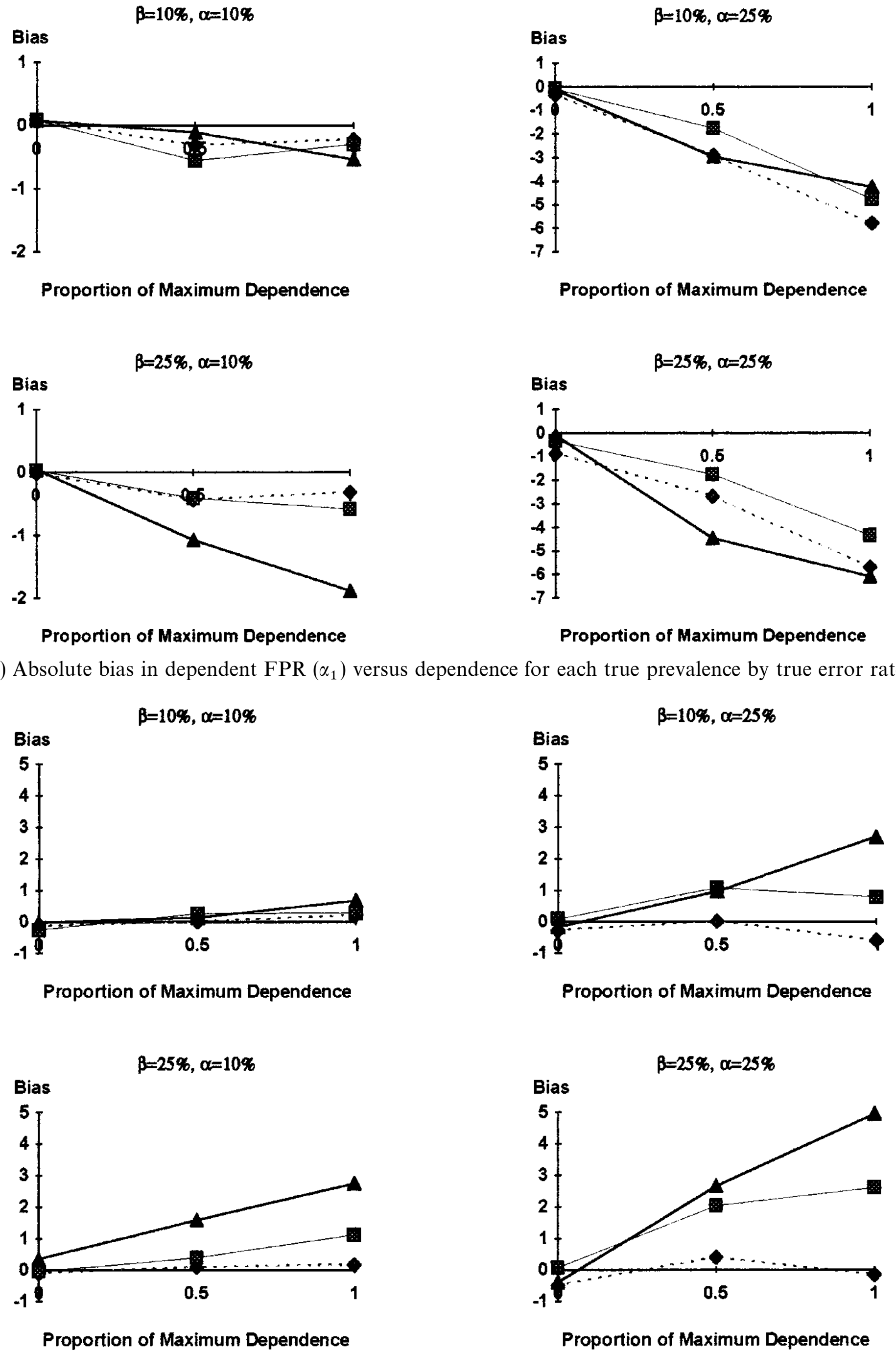
(a) Absolute bias in dependent FNR (β_1) versus dependence for each true prevalence by true error rates

(b) Absolute bias in independent FNR (β_3) versus dependence for each true prevalence by true error rates

Figure 2. Three tests ($r = 3$) with dependence between tests 1 and 2 set such that the covariance between test 1 and 2 is equal to the four test ($r = 4$) case when classifying cases and non-cases

True Prevalence

- - ◆ - 5% —■— 15% —▲—40%



(c) Absolute bias in dependent FPR (α_1) versus dependence for each true prevalence by true error rates

(d) Absolute bias in independent FPR (α_3) versus dependence for each true prevalence by true error rates

Figure 2. (Continued)

Table III. Means and standard deviations of parameter estimates in the case of four tests ($r = 4$), with dependence between tests 1 and 2 when classifying cases and non-cases, true prevalence (θ) = 15 per cent

True error rates	Dep.*	Mean of estimates (%) (standard deviation (%))				
		θ	β_1	β_3	α_1	α_3
$\beta = 10\%$	0	14·99 (0·60)	10·04 (0·88)	9·66 (1·59)	9·81 (0·47)	9·97 (0·50)
$\alpha = 10\%$	0·5	14·90 (0·67)	9·25 (1·25)	10·09 (1·54)	10·03 (0·61)	10·07 (0·45)
	1	15·08 (0·51)	9·36 (1·74)	10·27 (1·46)	9·82 (0·55)	10·05 (0·48)
$\beta = 10\%$	0	15·07 (0·87)	9·44 (2·08)	10·25 (2·25)	24·82 (0·76)	25·05 (0·80)
$\alpha = 25\%$	0·5	16·06 (0·82)	5·25 (1·55)	15·25 (1·97)	23·26 (0·75)	25·17 (0·83)
	1	17·15 (0·79)	0·84 (1·03)	20·73 (2·34)	21·25 (0·69)	25·69 (0·84)
$\beta = 25\%$	0	14·80 (0·75)	23·90 (2·29)	23·88 (2·52)	9·94 (0·66)	10·18 (0·51)
$\alpha = 10\%$	0·5	15·11 (0·76)	22·95 (2·28)	25·43 (2·44)	9·66 (0·53)	10·25 (0·66)
	1	15·07 (0·76)	20·87 (2·44)	26·14 (2·93)	9·37 (0·61)	10·41 (0·56)
$\beta = 25\%$	0	15·01 (1.84)	25·23 (4·33)	24·53 (4·19)	25·00 (0·76)	25·01 (0·93)
$\alpha = 25\%$	0·5	15·37 (1·73)	13·46 (3·61)	31·69 (3·44)	22·64 (1·15)	25·84 (0·83)
	1	16·16 (0·95)	3·59 (2·45)	38·25 (2·53)	20·26 (0·85)	26·99 (0·71)

*Proportion of maximum dependence between tests 1 and 2 for both cases and non-cases
β represents the true false negative rates, $\beta_1 = \beta_2 = \beta_3 = \beta_4$
α represents the true false positive rates, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$

The patterns of bias are similar to those described earlier in Table I and Figure 1. In general, the bias is slightly greater when $r = 3$ than when $r = 4$. For example, at maximum dependence when $\beta = 10$ per cent and $\alpha = 10$ per cent, the mean of the estimates of β_1 was 8·14 per cent when $r = 3$ (Table II) and less biased at 9·36 per cent when $r = 4$ (Table III). Exceptions occur for the false negative rate estimates for the dependent tests (β_1 and β_2) when θ is low at 5 per cent and α is high at 25 per cent, and for the false positive rate estimates for the dependent tests (α_1 and α_2) when θ is high at 40 per cent and both α and β are high at 25 per cent.

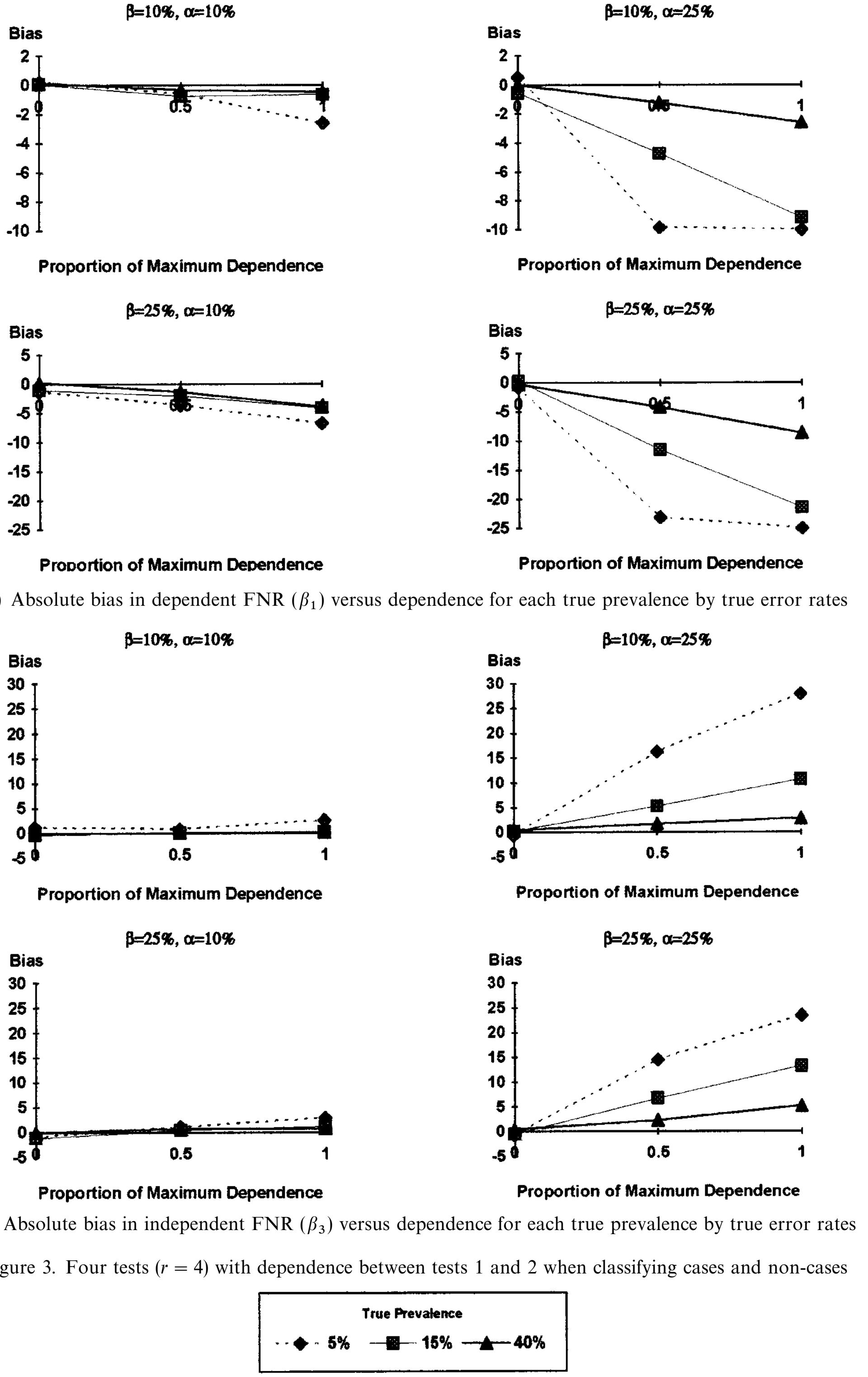
These exceptions aside, the addition of a fourth independent test helps to 'dilute' the effect of the dependence between tests 1 and 2. Concordant misclassifications by the dependent tests are less likely to be considered correct when a fourth independent observation is available, and bias is thus typically reduced.

3.1.3. Dependence between all pairs of tests

A positive dependence between all pairs of tests when classifying the cases and non-cases will result in similar underestimation of the error rates of all tests. There is more agreement between tests then under independence, leading to bias. High error rates (α when θ is low and β when θ is high) lead to larger bias.

3.2. Variability of estimates

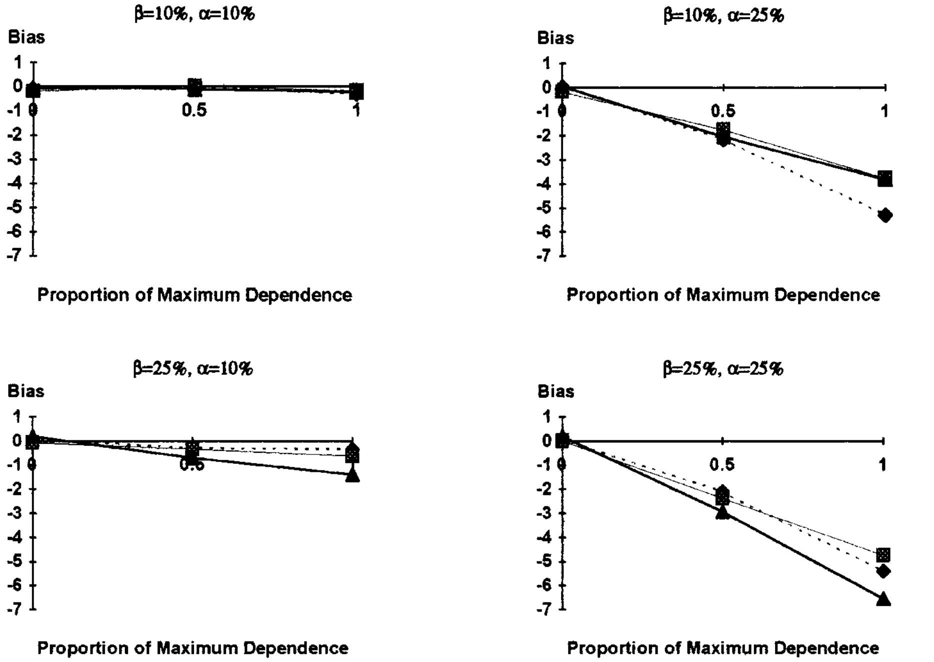
We compared the standard error estimates from the Latent program to the empirical standard deviations of the simulation estimates for $r = 3$ with dependence between two tests with respect to cases and non-cases. The estimates were very close, with most ratios of standard error to standard deviation being close to one. When the bias forced the error rate parameter estimates to zero



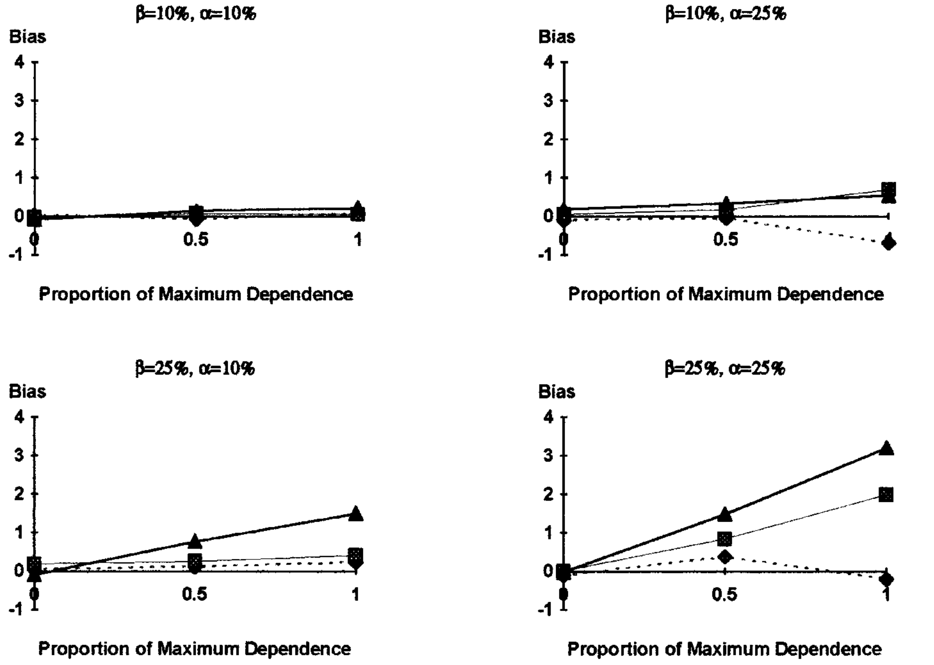
(a) Absolute bias in dependent FNR (β_1) versus dependence for each true prevalence by true error rates

(b) Absolute bias in independent FNR (β_3) versus dependence for each true prevalence by true error rates

Figure 3. Four tests ($r = 4$) with dependence between tests 1 and 2 when classifying cases and non-cases



(c) Absolute bias in dependent FPR (α_1) versus dependence for each true prevalence by true error rates



(d) Absolute bias in independent FPR (α_3) versus dependence for each true prevalence by true error rates

Figure 3. (Continued).

(when θ was small and error rates were high) the ratios became large, but in these cases of severe bias the standard error estimates are of less interest anyway. When $\theta = 40$ per cent, meaning that the bias is small, the ratios stayed close to 1, ranging from 0·68 to 1·38.

4. DISCUSSION

We feel the latent class method is useful in many situations, even in the presence of conditional dependence between tests. The method often provides unbiased estimates, but in some situations, substantial bias occurs and one must interpret the estimates from the independence model with caution.

Specifically, when the false positive rates are high and the prevalence is low, conditional dependence between two diagnostic tests can lead to serious bias in the estimates of all the parameters except the independent test's false positive rate. This situation occurs commonly in population screening where one desires tests to have high sensitivity and lower specificity. This kind of a test has been referred to as SnNout, meaning that it has a sufficiently high *Sen*sitivity so that a *N*egative result rules *out* the disease.[1] Therefore, one should approach the use of latent class methods to evaluate screening tests for a rare disease cautiously.

The presence of any dependence between the tests with respect to false positive classifications, meaning any commonalties when classifying non-cases that lead to concordant false positive classifications, cause seriously underestimated error rates for these tests, especially for the false negative rates. Equivalently, the specificity and especially the sensitivity of the two dependent tests are overestimated, leading to a false sense of confidence regarding their performance. Most care should be taken at very low prevalences where even a slight dependence results in very biased results. As the prevalence increases, the latent class estimates become relatively unbiased even in the presence of dependence, but even at prevalences near 50 per cent, if the false positive rate is high then dependence still causes some bias in most of the error rates. By symmetry, the latent class model estimates, under an assumption of independence, are also seriously biased when the prevalence is very high and the false negative rates are high.

The addition of one or more independent tests reduces the bias. One needs to balance this advantage against the extra cost and logistical difficulty of making the additional observations, and a possibly larger sample size to support the estimation of a larger number of error rate parameters. Although we have not considered cases with more than four tests, one may presume that inclusion of further independent diagnostic information will reduce bias even more; note, however, that the number of parameters involved increases rapidly with the number of tests.

The validity of the independence assumption depends on the circumstances. For example, test/retest observations by the same individual will often involve correlated errors. The errors of different observers may be less dependent, but even here there may be problems. Suppose, for example, that the observations are made by physicians who have undergone similar medical training. There may be certain types of disease cases that have a higher probability of being 'missed' by each of the observers, thus creating a dependence between the errors for true disease cases. A similar problem arises if the disease, rather than being binary, has a spectrum of severity; then patients close to the diagnostic threshold have a relatively high probability of misclassification, while patients far from the threshold are more easily classified correctly. Overall the spectrum of cases will have positive dependence between the errors associated with a discrete classification. On the other hand, clinical tests that are fundamentally different, such as a blood test and an X-ray, are more likely to satisfy the independence assumption, as their classifications are based on different biological characteristics of the disease manifestation.

As described earlier, when $r = 3$ the model is saturated. If $r > 3$, however, there are excess degrees of freedom which can be exploited to evaluate the goodness-of-fit (GOF) of the model. In particular, use of a GOF test is one mechanism to examine the evidence for a departure from the independence model. We did estimate the power of the GOF test in our simulations when $r = 4$. As expected, higher power was found when the degree of dependence, and thus the bias, was larger. Lower power was found if the bias in the estimates was small, which occurred mainly when the false positive rates were low. These results are preliminary, and further work is needed to establish the power curves and associated required sample sizes for a wider range of situations than those covered in our simulations.

APPENDIX

When tests 1 and 2 agree, δ_{12} adds to the probability that must remain less than or equal to 1. Therefore

$$\delta_{12} \leqslant 1 - \prod_{i=1}^{3} [\beta_i^{(1-x_i)}(1-\beta_i)^{x_i}].$$

Since $\beta_i \leqslant 0{\cdot}5$, $\beta_i \leqslant 1 - \beta_i$ the most strict restriction when tests 1 and 2 agree is

$$\delta_{12} \leqslant 1 - (1-\beta_1)(1-\beta_2)(1-\beta_3). \qquad (1)$$

When tests 1 and 2 disagree, δ_{12} subtracts from the probability that must remain greater than or equal to 0. Therefore

$$\delta_{12} \leqslant \prod_{i=1}^{3} [\beta_i^{(1-x_i)}(1-\beta_i)^{x_i}].$$

Thus, the most strict restrictions when tests 1 and 2 disagree are

$$\delta_{12} \leqslant \beta_1(1-\beta_2)\beta_3 \qquad (2)$$

and

$$\delta_{12} \leqslant (1-\beta_1)\beta_2\beta_3. \qquad (3)$$

Consider (1) − (2):

$$\begin{aligned}(1)-(2) &= 1-(1-\beta_1)(1-\beta_2)(1-\beta_3)-\beta_1(1-\beta_2)\beta_3\\ &= 1-(1-\beta_1-\beta_2-\beta_3+\beta_1\beta_2+\beta_1\beta_3+\beta_2\beta_3-\beta_1\beta_2\beta_3)-\beta_1\beta_3+\beta_1\beta_2\beta_3\\ &= \beta_1+\beta_2+\beta_3-\beta_1\beta_2-2\beta_1\beta_3-\beta_2\beta_3+2\beta_1\beta_2\beta_3\\ &= \beta_1(1-\beta_3)+\beta_2(1-\beta_1)+\beta_3(1-\beta_2)-\beta_1\beta_3+2\beta_1\beta_2\beta_3 \geqslant 0\end{aligned}$$

since $\beta_1\beta_3 \leqslant \beta_1(1-\beta_3)$ and all other terms are positive. Thus (2) is a stricter restriction than (1) and similarly (3) is also stricter than (1). Therefore, the maximum value for δ_{12} is determined by restricting

$$\delta_{12} \leqslant \beta_1(1-\beta_2)\beta_3 \quad \text{and} \quad \delta_{12} \leqslant (1-\beta_1)\beta_2\beta_3.$$

ACKNOWLEDGEMENTS

This work was part of the requirements for completion of a Masters of Science degree in Statistics and was supported in part by a scholarship from the Natural Sciences and Engineering Research Council and by a National Health Scientist award from Health and Welfare, Canada.

REFERENCES

1. Sackett, D. L., Haynes, R. B., Guyatt, G. H. and Tugwell, P. *Clinical Epidemiology: A Basic Science*, 2nd edn, Little, Brown and Company, Boston, 1991.
2. Walter, S. D. and Irwig, L. M. 'Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review', *Journal of Clinical Epidemiology*, **41** (9), 923–937 (1988).
3. Hui, S. L. and Walter, S. D. 'Estimation error rates of diagnostic tests', *Biometrics*, **36**, 167–171 (1980).
4. Walter, S. D., Frommer, D. J. and Cook, R. J. 'The estimation of sensitivity and specificity in colorectal cancer screening methods', *Cancer Detection and Prevention*, **15**(6), 465–470 (1991).
5. Szatmari, P., Volkmar, F. and Walter, S. D. 'Evaluation of diagnostic criteria for autism using latent class models', *Journal of the American Academy of Child and Adolescent Psychiatry*, **34**, 216–222 (1995).
6. Spreigelhalter, D. J. and Stovin, P. G. I. 'An analysis of repeated biopsies following cardiac transplantation', *Statistics in Medicine*, **2**, 33–40 (1983).
7. Irwig, L. M., duToit, R. S. J., Sluis-Cremer, G. K., Solomon, A., Glyn Thomas, R., Hamel, P. P. H., Webster, I. and Hastie, T. 'Risk of asbestosis in crocidolite and mosite mine in South Africa', *Annals of the New York Academy of Science*, **330**, 35–52 (1979).
8. Dawid, A. P. and Skene, A. M. 'Maximum likelihood estimation of observer error rates using the EM algorithm', *Applied Statistics*, **28**(1), 20–28 (1979).
9. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1979.
10. Vacek, P. M. 'The effect of conditional dependence on the evaluation of diagnostic tests', *Biometrics*, **41**, 959–968 (1985).
11. Qu, Y., Tan, M. and Kutner M. H. 'Random effects models in latent class analysis for evaluating accuracy of diagnostic tests', *Biometrics*, **52**, 797–810 (1996).
12. Brenner, H. 'How independent are multiple 'independent' diagnostic classificaitons?', *Statistics in Medicine*, **15**, 1377–1386 (1996).
13. Devore, J. L. *Probability and Statistics for Engineering and Sciences*, 2nd edn, Brooks/Cole Publishing Company, Monterey, California, 1987.