A Capture–Recapture Approach for Screening Using Two Diagnostic Tests With Availability of Disease Status for the Test Positives Only

Dankmar BÖHNING and Valentin PATILEA

The article considers screening human populations with two screening tests. If any of the two tests is positive, then full evaluation of the disease status is undertaken; however, if both diagnostic tests are negative, then disease status remains unknown. This procedure leads to a data constellation in which, for each disease status, the 2×2 table associated with the two diagnostic tests used in screening has exactly one empty, unknown cell. To estimate the unobserved cell counts, previous approaches assume independence of the two diagnostic tests and use specific models, including the special mixture model of Walter or unconstrained capture–recapture estimates. Often, as is also demonstrated in this article by means of a simple test, the independence of the two screening tests is not supported by the data. Two new estimators are suggested that allow associations of the screening test, although the form of association must be assumed to be homogeneous over disease status. These estimators are modifications of the simple capture–recapture estimator and easy to construct. The estimators are investigated for several screening studies with fully evaluated disease status in which the superior behavior of the new estimators compared to the previous conventional ones can be shown. Finally, the performance of the new estimators is compared with maximum likelihood estimators, which are more difficult to obtain in these models. The results indicate the loss of efficiency as minor.

KEY WORDS: Capture–recapture; Capture–recapture estimator under screening test dependence; Diagnostic test accuracy; Testing independence.

1. INTRODUCTION

Screening a population for a specific disease is a fundamental aspect of disease surveillance and health care. Screening programs for cancer are a well-established component of health care in many countries. Examples of screening tests for cancer include mammography for breast cancer, serum prostate-specific antigen (PSA) levels for prostate cancer, and the pap smear for cervical cancer. Screening is considered highly important in other diseases as well. In cardiovascular diseases, screening typically focuses on risk factors. According to Galen and Gambino (1975), screening for heart disease focuses on blood pressure, electrocardiography, and X-ray. Other risk factors include smoking and high serum cholesterol levels. Chaisiri et al. (1997) used the urine stick and fasting blood sugar to screen a rural population in northeast Thailand for non–insulin-dependent diabetes mellitus. Due to the popularity of these screening devices and the arising data from numerous studies, determining the appropriate methodology for modeling screening studies with incomplete evaluation of disease status is a matter of concern.

1.1 Setting and Notation

In particular, we are interested in the following setting. A population of known size n is screened using two screening tests, T_1 and T_2, for having a specific disease D with d different states of disease. In the simplest case, there are two states: healthy and nonhealthy. To give a simple example, we might think of screening for prostate cancer using the digital rectal examination and PSA level. In these screening situations, the point is to find cancer cases in an early stage in the development of any disease or disease-related symptoms. This means that the screening is applied to large populations. In most cases, the results of both diagnostic tests are negative, and no further medical diagnostics are applied. If one of the tests leads to a positive finding, then further medical diagnostics are usually applied to evaluate the correct disease status. This diagnostic situation is usually referred to as *verification restricted to screen positives* (Pepe 2003). This setting leads to the following data constellation: the frequency of screened persons with outcome $j = 0, 1$ for test 1 and outcome $k = 0, 1$ for test 2 and disease status $i = 1, \ldots, d$ is denoted by $x_{jk}^{(i)}$. Note that $x_{00}^{(i)}$ is unknown for $i = 1, \ldots, d$. In addition, let $p_{jk}^{(i)}$ denote the probability that test 1 is having outcome $j = 0, 1$ and test 2 is having outcome $k = 0, 1$ conditional on being in disease state i. Furthermore, at various times we use the abbreviation $a_{i+} = a_{i0} + a_{i1}$ or $a_{+i} = a_{0i} + a_{1i}$ for 2×2 matrices a_{ij}. If there are only two ($d = 2$) disease states (healthy, 1; diseased, 2), then $p_{1+}^{(2)}$ is the sensitivity of test 1 and $p_{+1}^{(2)}$ is the sensitivity of test 2, whereas $p_{0+}^{(1)}$ is the specificity of test 1 and $p_{+0}^{(1)}$ is the specificity of test 2. In addition, let q_i denote the proportions of the disease states in the population, $i = 1, \ldots, d$. Again, if there are only two states, then q_2 is just the disease prevalence.

If all frequencies $x_{jk}^{(i)}$ are observed, then all conditional probabilities $p_{jk}^{(i)}$ can be estimated simply by their corresponding observed proportion $\hat{p}_{jk}^{(i)} = x_{jk}^{(i)}/n_i$, and the true proportions q_i can be estimated by their corresponding observed proportion $\hat{q}_i = n_i/n$, where $n_i = x_{++}^{(i)}$. However, $x_{00}^{(i)}$ is not known for all i, and developing estimates for it is the focus of this article. To do so, some modeling is required.

1.2 The Latent Class Model for Partially Available Disease Status

Latent class models for estimating diagnostic error have a long tradition (see Hui and Walter 1980; Hui and Zhou 1998; Albert and Dodd 2004). For the special situation of availability of the gold standard for the test positives only, Walter (1999)

Dankmar Böhning is Professor and Chair for Applied Statistics in the Life Sciences, School of Biological Sciences, University of Reading, Reading, RG6 6FN, U.K. (E-mail: *d.a.w.bohning@reading.ac.uk*). Valentin Patilea is Professor, CREST–ENSAI, Campus de Ker Lann, 35172 Bruz Cedex, France (E-mail: *valentin.patilea@ensai.fr*). The research of D. Böhning is supported by the German Research Foundation (DFG). The authors are grateful to the editor, the associate editor, and three referees for their helpful comments and suggestions.

Journal of the American Statistical Association
March 2008, Vol. 103, No. 481, Applications and Case Studies
DOI 10.1198/016214507000000383

developed a particular latent class model with a required latent class approach for the test negatives only. The Walter model (Walter 1999) uses independence of the screening tests conditional on disease status and assumes that $p_{11}^{(i)} = p_{1+}^{(i)} p_{+1}^{(i)}$ for all $i = 1, \ldots, d$. For a 2×2 table, this implies that all of the (conditional) joint probabilities are the product of the respective two (conditional) marginal probabilities, $p_{jk}^{(i)} = p_{j+}^{(i)} p_{+k}^{(i)}$. In general, this means that there are $2d$ parameters determining the conditional probabilities and $d - 1$ proportion parameters, in total $3d - 1$. If there are only two disease states, then we have only two sensitive parameters, two specificity parameters, and one prevalance parameter, leading to the Walter model in five parameters. In general, the likelihood is provided by

$$\prod_{i=1}^{d} \left(p_{1+}^{(i)} p_{+1}^{(i)} q_i\right)^{x_{11}^{(i)}} \prod_{i=1}^{d} \left(p_{0+}^{(i)} p_{+1}^{(i)} q_i\right)^{x_{01}^{(i)}} \times \prod_{i=1}^{d} \left(p_{1+}^{(i)} p_{+0}^{(i)} q_i\right)^{x_{10}^{(i)}} \left(\sum_{i=1}^{d} p_{0+}^{(i)} p_{+0}^{(i)} q_i\right)^{x_{00}^{(+)}}. \quad (1)$$

Note that $x_{00}^{(+)} = x_{00}^{(1)} + \cdots + x_{00}^{(d)}$ is known even though the individual frequencies $x_{00}^{(i)}$ are unknown. To find maximum likelihood estimates (MLEs), the likelihood (1) must be maximized in $p_{1+}^{(i)}$, $p_{+1}^{(i)}$, and q_i. The last term in (1) is a mixture likelihood and inhibits a closed-form solution for the maximum likelihood estimation. Walter (1999) used the Newton–Raphson procedure to maximize (1), which provides as a byproduct an estimate of the covariance matrix of the parameter estimates. Alternatively, we might use the EM algorithm (McLachlan and Krishnan 1997; McLachlan and Peel 2000), which delivers as an advantageous byproduct an estimate of $x_{00}^{(i)}$. Whatever algorithm is used, let $\hat{p}_{1+}^{(i)}$, $\hat{p}_{+1}^{(i)}$, and $\hat{q}_i$ denote the MLEs under the Walter model. Then,

$$\hat{n}_i^* = n \times \hat{q}_i \quad (2)$$

gives the needed size estimates for $i = 1, \ldots, d$. To provide valid inference, it is crucial that the model of screening test independence conditional on disease status is acceptable. Pepe (2003), among others, expressed concerns about the realistic nature of the conditional independence assumption. This assumption can be relaxed to develop realistic estimators of n_i. This is the major focus of this article, which is organized as follows. In the next section we present four study data sets that we used for illustration throughout the remainder of this article. In Section 1.4 we provide a simple approach to test the assumption of screening test independence and show that the hypothesis of screening test independence is not supported for the four data sets. In Section 2 we relax the assumption of screening test independence conditional on disease status and suggest a capture–recapture estimate of $x_{00}^{(i)}$ based on two homogeneity models. In Section 3 we introduce the two models for capturing screening test dependence more formally and discuss maximum likelihood estimation. In Section 4 we present a simulation study that compares maximum likelihood with capture–recapture estimation, and in Section 5 we conclude with a discussion that puts the findings in perspective.

1.3 Case Studies Considered

Before we proceed with modeling, we present four screening studies in which the disease status has been evaluated only for those persons who tested positive for at least one of the two tests. These case studies will serve as examples for illustration and evaluation throughout the remainder of the article. The first study has been discussed by numerous authors, including Walter (1999), Strax, Venet, Shapiro, and Gross (1967), Schatzkin, Connor, Taylor, and Bunnag (1987), and Cheng and Macaluso (1997). It consists of data from the Health Insurance Plan Study (HIP) for breast cancer screening in New York. A total of 20,211 women were screened for breast cancer using physical examination (test 1) and mammography (test 2). The observed frequencies are provided in Table 1 under study 1. In brief, 307 women who tested positive for breast cancer by either mammography or physical examination underwent biopsy, according to which they were classified into two ($d = 2$) disease states: no cancer and cancer. Smith, Bullock, and Catalona (1997) screened 18,527 white men (study 2) and 949 black men (study 3) for prostate cancer using digital rectal examination (test 1) and PSA (test 2). According to Pepe and Alonzo (2001), the PSA level was considered suspicious for cancer if it exceeded 4.0 ng/mL. Persons with positive screening test results on either DRE or PSA were referred for ultrasound-guided needle biopsy, which is considered a gold standard in this setting. According to biopsy, the men were classified into two ($d = 2$) disease states: no cancer ($i = 1$) and cancer ($i = 2$). De Sutter et al. (1998) conducted a multicenter study to compare cervicography with the standard pap smear cytology test for detecting cervical cancer. This study will serve here as study 4 (Table 1). According to Pepe and Alonzo (2001), subjects who were positive for either test were referred for colposcopy with directed biopsy, which is considered the gold standard in this example. A total of $n = 5{,}192$ women completed the protocol,

Table 1. Observed frequencies in four screening studies

Observed frequency	Study 1	Study 2	Study 3	Study 4
$x_{11}^{(1)}$	13	138	3	11
$x_{10}^{(1)}$	144	717	38	20
$x_{01}^{(1)}$	95	976	26	81
$x_{00}^{(1)}$				
$x_{11}^{(2)}$	10	179	10	6
$x_{10}^{(2)}$	24	264	28	29
$x_{01}^{(2)}$	21	137	8	48
$x_{00}^{(2)}$				
$x_{11}^{(3)}$				14
$x_{10}^{(3)}$				15
$x_{01}^{(3)}$				4
$x_{00}^{(3)}$				
n	20,211	18,527	949	5,192

NOTE: Study 1, Health Insurance Plan Screening Study for Breast Cancer in New York (Strax et al. 1967). Studies 2 and 3, Screening for prostate cancer (Smith et al. 1997) in white men (study 2) and black men (study 3). Study 4, Screening for cervical cancer (De Sutter et al. 1998).

of whom 228 underwent biopsy. Histological examination of the biopsy was used to classify the disease into three ($d = 3$) states: not present ($i = 1$), low grade (condyloma) ($i = 2$), and high grade (invasive cancer) ($i = 3$).

1.4 Testing Conditional Independence

Application of the Walter model for these kinds of screening studies is popular in epidemiology. However, valid application of the Walter model requires independence conditional on disease status.

A Test Based on Maximum Likelihood. To test conditional independence, we could follow the conventional chi-squared approach, as suggested by Walter (1999). After having achieved MLEs for $\hat{p}^{(i)}_{j+}$, $\hat{p}^{(i)}_{+k}$, and $\hat{q}_i$, we could compute expected values $e^{(i)}_{11} = p^{(i)}_{1+}p^{(i)}_{+1}\hat{q}_i n$, $e^{(i)}_{10} = p^{(i)}_{1+}p^{(i)}_{+0}\hat{q}_i n$, and $e^{(i)}_{01} = p^{(i)}_{0+}p^{(i)}_{+1}\hat{q}_i n$, as well as $e^{(+)}_{00} = \sum_{i=1}^{d} p^{(i)}_{0+}p^{(i)}_{+0}\hat{q}_i n$, to form

$$\chi^2 = \sum_{i=1}^{d}\left(\frac{(x^{(i)}_{11} - e^{(i)}_{11})^2}{e^{(i)}_{11}} + \frac{(x^{(i)}_{10} - e^{(i)}_{10})^2}{e^{(i)}_{10}} + \frac{(x^{(i)}_{01} - e^{(i)}_{01})^2}{e^{(i)}_{01}}\right) + \frac{(x^{(+)}_{00} - e^{(+)}_{00})^2}{e^{(+)}_{00}}. \quad (3)$$

There are $3d + 1$ observed frequencies in the table and $3d - 1$ parameters to be estimated, which leaves 1 degree of freedom for the null distribution of (3). Applying the test statistic (3) to the HIP study yields a chi-squared value of 177.565 (see also Walter 1999). This very large value indicates that the conditional independence model is inappropriate. Walter (1999) noted that "one interpretation of this pattern is that there are correlated errors between the two tests." We have applied the test statistic to the other data sets mentioned earlier. The results, given in Table 2, show empirical evidence for departure from the null hypothesis of conditional independence, with the exception of the screening study on prostate cancer in black men, which has a chi-squared value of 2.185, which is not significant.

A Test Based on the Lincoln–Petersen Estimator. The conventional chi-squared approach requires knowledge of the MLEs under the null hypothesis of conditional independence. We now look into a setting in which algorithm construction by means of the EM algorithm or Newton–Raphson is not required. Evidently, only the entry $x^{(i)}_{00}$ is unknown, and we might consider the Lincoln–Petersen estimate for it. Using independence and replacing parameters by their sample estimates, we have $x^{(i)}_{11}/n_i = (x^{(i)}_{1+}/n_i)(x^{(i)}_{+1}/n_i)$, which is easily solved for n_i to provide

$$\hat{n}_i = \frac{x^{(i)}_{1+}x^{(i)}_{+1}}{x^{(i)}_{11}}. \quad (4)$$

This estimator is connected with the work of Petersen (1896) and Lincoln (1930), who derived it independently for different settings (see also Bishop, Fienberg, and Holland 1975, pp. 232–233, where large-sample variances are also given). However, for reasons of stability, we use the more robust estimator

$$\hat{n}_i = \frac{(x^{(i)}_{1+} + 1)(x^{(i)}_{+1} + 1)}{x^{(i)}_{11} + 1} - 1, \quad (5)$$

as suggested by Chapman (1951), and used and recommended by others, including Borchers, Buckland, and Zucchini (2002). This estimator was also used by Goldberg and Wittes (1978) to estimate the false negatives in screening situations. It is approximately unbiased in general and exactly unbiased if $x^{(i)}_{+1} + x^{(i)}_{1+} \geq n_i$ (Seber 1970; Wittes 1972). Summing the estimator (5) over all disease states gives $\hat{n} = \sum_i \hat{n}_i$. Because the size of the screened population is known to be n, we can compare $\hat{n}$ with n to form

$$Z = \frac{\hat{n} - n}{\sqrt{\widehat{\mathrm{Var}}(\hat{n})}}, \quad (6)$$

with

$$\widehat{\mathrm{Var}}(\hat{n}) = \sum_{i=1}^{d} \frac{(x^{(i)}_{1+} + 1)(x^{(i)}_{+1} + 1)x^{(i)}_{10}x^{(i)}_{01}}{(x^{(i)}_{11} + 1)^2(x^{(i)}_{11} + 2)}. \quad (7)$$

The variance estimator (7) is a stratified version of the variance estimator provided by Seber (1970) for the unstratified case. We apply the test to the four screening studies of Section 2. Evidently, in all studies $\hat{n}$ is below n, leading to significant test results. An exception is again the prostate cancer study for black men, in which Z is smallest in absolute value although still significant, whereas the corresponding chi-squared value is not (see Table 2). This might indicate that the statistical test based on Z is more liberal than that based on the conventional chi-squared test. Also, the Z value provides insight into the type of dependence, in the sense that estimates lower than the size of

Table 2. Values of the test statistics in the four screening studies

Screening study	Chi-squared	Z	$\hat{n}$	n	$\hat{\theta}$ (95% confidence interval)	$\hat{\alpha}$
HIP study	177.565	−66.537	1,330	20,211	15.169 (10.716, 26.120)	19.06
Prostate cancer						
White men	148.442	−21.858	7,646	18,527	2.423 (2.149, 2.777)	3.07
Black men	2.185	−4.541	380	949	2.497 (1.517, 7.031)	2.92
Cervical cancer	83.062	−47.176	566	5,192	9.175 (6.849, 13.895)	13.84

the population indicate a positive dependence of the two screening tests. To sketch the argument for this phenomenon, we follow Hook and Regal (1995) and consider the expected value of (4), which can be roughly approximated by

$$n_i \frac{p_{1+}^{(i)} p_{+1}^{(i)}}{p_{11}^{(i)}} = n_i \frac{p_{1+}^{(i)}}{p_{11}^{(i)}/p_{+1}^{(i)}} = n_i \frac{p_{1+}^{(i)}}{p_{1|1}^{(i)}} = \frac{n_i}{\theta}, \tag{8}$$

where $p_{1|1}^{(i)}$ is the probability for the test 1 being positive, conditional on test 2 being positive, and $\theta = p_{1|1}^{(i)}/p_{1+}^{(i)}$ is the ratio of conditional and unconditional probabilities of test 1 being positive, which we call the *dependence parameter*. If the tests are not associated, then $p_{1|1}^{(i)} = p_{1+}^{(i)}$, or $\theta = 1$, and the estimator is approximately unbiased. However, if there is a positive association, then $p_{1|1}^{(i)} > p_{1+}^{(i)}$, or $\theta > 1$ and the expected value will be below n_i. The amount of underestimation is determined by the value of θ; the higher the value of θ, the larger the underestimation. Similarly, for negative associations of the screening procedures, we will observe values of $\theta < 1$. Looking again at Table 2, we see that positive associations have occurred in all four cases. The last column in the Table also provides an estimate of the dependence factor θ, which is significantly above 1 in all four cases. Note the similarity of the dependence parameter for the two screening studies for prostate cancer in white men and black men. The likely interpretation for this is that the two screening procedures worked similarly in both populations.

2. ESTIMATORS FOR THE FREQUENCIES OF TEST NEGATIVES UNDER DEPENDENCE

The data structure contains $3d + 1$ observed frequencies. The capture–recapture model under conditional independence requires estimation of the parameters $p_{1+}^{(i)}, p_{+1}^{(i)}, n_i$ for $i = 1, \ldots, d$, in total $3d$ parameters. This means that we can include one additional, estimatable parameter into the model.

2.1 The Model With a Homogeneous Dependence Parameter

Let us first consider the model of *homogeneous dependence*, $E(\hat{n}_i) \approx n_i/\theta$ for $i = 1, \ldots, d$. In more detail, $\theta = p_{1|1}^{(i)}/p_{1+}^{(i)}$, the ratio of conditional (conditional on test 2 being positive) and unconditional probabilities of test 1 being positive, is allowed to differ from 1 but is assumed homogeneous over all disease states. This model allows the association of the screening procedures, but this association is assumed similar in all disease states. Models implying heterogeneity in the dependence parameter will not be estimatable from the data structure available here. This model is more flexible than the conditional independence model but it is also limited in allowing dependencies. To accomplish estimation for θ in this model, we simply consider the moment estimator, $\theta\hat{n}_i$, which makes the sum of the estimated sizes of the disease states equal to the size of the screened population, $\sum_{i=1}^d E(\theta\hat{n}_i) = \theta \sum_{i=1}^d E(\hat{n}_i) = \sum_i n_i = n$. Replacing the theoretical expected value $E(\hat{n}_i)$ by its "observed value" $\hat{n}_i$ gives $\theta \sum_{i=1}^d \hat{n}_i = n$, or

$$\hat{\theta} = n \Big/ \sum_{i=1}^{d} \hat{n}_i. \tag{9}$$

From (9), an adjusted disease state size estimator can be constructed,

$$\hat{v}_i = \hat{n}_i \hat{\theta} = \frac{\hat{n}_i}{\sum_{i=1}^d \hat{n}_i} n \tag{10}$$

for $i = 1, \ldots, d$, which evidently meets $\sum_i \hat{v}_i = n$. Different versions of (9) and (10) will occur according to the form of estimator $\hat{n}_i$. We use (5) to estimate n_i. A direct argument using Seber's variance estimator (7) provides the estimated variance for $1/\hat{\theta}$,

$$\widehat{\mathrm{Var}(1/\hat{\theta})} = \frac{1}{n^2} \sum_{i=1}^{d} \frac{(x_{1+}^{(i)} + 1)(x_{+1}^{(i)} + 1) x_{10}^{(i)} x_{01}^{(i)}}{(x_{11}^{(i)} + 1)^2 (x_{11}^{(i)} + 2)}. \tag{11}$$

Finding an analytical expression for the variance of $\hat{v}_i$ is more elaborate. We use bootstrap approach to provide estimates for it.

2.2 The Model With Homogeneous Odds Ratio

The association of the two tests conditional on disease status also can be measured using the odds ratio. The odds ratio parameter α_i is defined as $\alpha_i = p_{11}^{(i)} p_{00}^{(i)}/(p_{10}^{(i)} p_{01}^{(i)})$ in the ith disease state; we let $\hat{\alpha}_i = x_{11}^{(i)} x_{00}^{(i)}/(x_{10}^{(i)} x_{01}^{(i)})$ denote its estimate. Under homogeneity, all disease status–specific odds ratios should coincide: $\alpha_i = \alpha$ for all i. Clearly, using $\hat{\alpha}_i = \hat{\alpha}$ and solving for $x_{00}^{(i)}$, we find that

$$\hat{x}_{00}^{(i)} = \hat{\alpha} \frac{x_{10}^{(i)} x_{01}^{(i)}}{x_{11}^{(i)}}, \tag{12}$$

where $\hat{\alpha}$ in (12) is found as

$$\hat{\alpha} = \frac{x_{00}^{(+)}}{\sum_{i=1}^d x_{10}^{(i)} x_{01}^{(i)}/x_{11}^{(i)}}, \tag{13}$$

using the fact that $\sum_{i=1}^d \hat{x}_{00}^{(i)} = x_{00}^{(+)}$; for example, the total of screened negative persons is known. If we have independence conditional on disease status (e.g., $\hat{\alpha} = 1$ and $\hat{\theta} = 1$), then

$$\begin{aligned} &\hat{x}_{00}^{(i)} + x_{11}^{(i)} + x_{10}^{(i)} + x_{01}^{(i)} \\ &\quad = \frac{x_{10}^{(i)} x_{01}^{(i)} + x_{11}^{(i)} x_{11}^{(i)} + x_{10}^{(i)} x_{11}^{(i)} + x_{01}^{(i)} x_{11}^{(i)}}{x_{11}^{(i)}} \\ &\quad = \frac{x_{1+}^{(i)} x_{+1}^{(i)}}{x_{11}^{(i)}}, \end{aligned}$$

so that (12) and (10) agree, if we use $\hat{n}_i = x_{1+}^{(i)} x_{+1}^{(i)}/x_{11}^{(i)}$ in (10); otherwise, they will be different. To avoid undefined cases and to achieve a stabilized estimator, we again use pseudovalues in the construction process. Let $\hat{\alpha}_i = (x_{11}^{(i)} + 1)(x_{00}^{(i)} + 1)/((x_{10}^{(i)} + 1)(x_{01}^{(i)} + 1))$ denote the estimate of the odds ratio in the ith disease state. Solving for $x_{00}^{(i)}$ under odds ratio homogeneity leads to

$$\hat{x}_{00}^{(i)} = \hat{\alpha} \frac{(x_{10}^{(i)} + 1)(x_{01}^{(i)} + 1)}{x_{11}^{(i)} + 1} - 1, \tag{14}$$

with

$$\hat{\alpha} = \frac{x_{00}^{(+)} + d}{\sum_{i=1}^{d} (x_{10}^{(i)} + 1)(x_{01}^{(i)} + 1)/(x_{11}^{(i)} + 1)}.$$

From (14), we can construct an estimator of n_i as

$$\hat{\upsilon}_i = x_{11}^{(i)} + x_{10}^{(i)} + x_{01}^{(i)} + \hat{\alpha} \frac{(x_{10}^{(i)} + 1)(x_{01}^{(i)} + 1)}{x_{11}^{(i)} + 1} - 1. \quad (15)$$

We use this form of the estimator in all data analysis.

2.3 An Illustration of the Estimators in the Four Screening Studies

Table 3 gives all estimated sizes of the disease classes for the four studies mentioned in Section 1.3 according to the four methods (2), (5), (10), and (15). First, all estimators meet the constraint that their sum equals the size of the screened population except for the unconstrained Lincoln–Petersen estimator (5), which underestimates the size of the screened population in all four studies. For the two estimators under independence according to the Walter model (2) and the unconstrained Lincoln–Petersen estimator (5), the results are similar for the diseased population, although considerably different for the healthy population. The two capture–recapture estimators under dependences (10) and (15) differ in size, although their dimensions are similar. It can be expected that the latter two are more flexible in coping with potential dependence structures, although this question can be answered only be looking at data structures with completely evaluated disease status. We consider this issue in the next section.

2.4 Estimators in a Screening Study With Completely Known Disease Status

Here we illustrate the performance of the estimators in studies with completely evaluated disease status. According to Galen and Gambino (1975), screening for cardiovascular diseases focuses on risk factors, such as systolic blood pressure, smoking, serum cholesterol, and body mass index. Data for these risk factors were found from a subset of the Framingham Heart Study (Shurtleff 1974). Four study sets were formed from these four risk factors by defining a combination of risk factor 1 (test 1) and risk factor 2 (test 2) as follows: systolic blood pressure and smoking (set 1), serum cholesterol and smoking (set 2), serum cholesterol and body mass index (set 3), and systolic blood pressure and body mass index (set 4). Their frequencies are provided in Table 4. Note that for the first two sets, conditional on disease status, the two risk factors under consideration are negatively associated, although not significantly as when measured by the Mantel–Haenszel odds ratio with summary taken over disease status. In the last two sets, conditional on the disease status, the risk factors are positively and significantly associated. For the latter two sets, we can expect stronger differences among the estimators under consideration. For this evaluation, we assume that $x_{00}^{(i)}$ is not known and use the estimators $\hat{n}_i$, $\hat{\upsilon}_i$, and $\hat{\upsilon}_i$ for predicting the known n_i. We can evaluate these estimators by comparing predicted and known n_i. Table 5 provides the estimates for all four sets with 95% confidence intervals. Not surprisingly, the confidence intervals for the unconstrained capture–recapture estimator do not include the true size of the screened population for the significantly associated screening factors in sets 3 and 4. In all other cases, the confidence intervals cover the sizes of the screened population. The estimator $\hat{\upsilon}_i$, based on (10), appears to be closer to the observed values in all four sets.

Table 3. Estimated sizes of disease classes in four screening studies

Screening study	$\hat{n}_i^{*}$[a]	$\hat{n}_i$[b]	$\hat{\upsilon}_i$[c]	$\hat{\upsilon}_i$[d]
HIP				
1	20,105	1,229	18,679	19,204
2	106	101	1,532	1,007
Prostate cancer				
White men				
1	17,740	6,685	16,635	17,324
2	787	781	1,892	1,203
Black men				
1	880	314	783	835
2	69	66	166	114
Cervical cancer				
1	264	247	2,266	2,097
2	4,891	282	2,586	2,989
3	37	37	340	106

[a] n_i estimated using (2) with the EM algorithm in the Walter model.
[b] n_i estimated using (5).
[c] n_i estimated using (10).
[d] n_i estimated using (15).

Table 5. Estimated sizes (with 95% confidence intervals) of disease classes in the four studies with completely known disease status

Data set	n_i[a]	$\hat{n}_i$[b]	$\hat{\upsilon}_i$[c]	$\hat{\upsilon}_i$[d]
Set 1				
1	226	234 (203, 264)	226 (204, 249)	225 (204, 247)
2	191	197 (174, 220)	191 (168, 213)	192 (171, 214)
Set 2				
1	226	238 (202, 274)	227 (203, 252)	226 (203, 249)
2	188	196 (170, 221)	187 (162, 211)	188 (166, 212)
Set 3				
1	226	166 (131, 201)	225 (195, 255)	242 (209, 277)
2	187	138 (113, 163)	188 (158, 218)	171 (137, 205)
Set 4				
1	226	199 (162, 225)	223 (198, 249)	231 (205, 259)
2	190	172 (143, 190)	193 (167, 218)	185 (158, 212)

[a] n_i is known for this evaluation.
[b] n_i estimated using (5).
[c] n_i estimated using (10).
[d] n_i estimated using (15).

3. MAXIMUM LIKELIHOOD FOR MODELS WITH DEPENDENCE

Here we provide a more formal investigation of the two models of screening test dependency using likelihood methods. As will be seen in the following, maximum likelihood estimation is cumbersome in both models, and MLEs are used here only as a benchmark method for efficiency considerations.

We recall that the marginal probabilities are $p_{j+}^{(i)} = P(T_1 = j|D = i)$ and $p_{+k}^{(i)} = P(T_2 = k|D = i)$, whereas $p_{jk}^{(i)} = P(T_1 = j, T_2 = k|D = i)$ denotes the cell probabilities ($j, k = 0$ or 1).

Table 4. Frequency data constellation for four data sets defined by the risk factor (RF) combination in screening for coronary heart disease (CHD) with Mantel–Haenszel summary odds ratios (summing over disease status) and test for homogeneity of odds ratios between disease status

CHD	RF 1	RF 2	Set 1 $x_{jk}^{(i)}$	Set 2 $x_{jk}^{(i)}$	Set 3 $x_{jk}^{(i)}$	Set 4 $x_{jk}^{(i)}$
1	1	1	65	50	51	54
1	1	0	24	20	19	35
1	0	1	106	121	70	67
1	0	0	31	35	86	70
2	1	1	95	69	69	78
2	1	0	28	20	20	45
2	0	1	57	83	38	31
2	0	0	11	16	60	36
OR_{MH}			.73	.74	4.18	1.78
(95% CI)			(.44, 1.22)	(.42, 1.16)	(2.69, 6.77)	(1.17, 2.72)
χ^2_{hom}			.14	.03	1.23	.29
p value			.71	.87	.27	.59

NOTE: OR, odds ratio.

Unfortunately, there is no unique way to model dependency. In Section 2 we distinguished between two different dependency measures, one based on the ratio of conditional and unconditional probability for one test being positive and the other based on the odds ratio. Here we provide the associated likelihoods for both models.

3.1 The θ Model

The dependency parameter is defined as

$$\theta = \frac{p_{1|1}^{(i)}}{p_{1+}^{(i)}} = \frac{P(T_1 = 1|T_2 = 1, D = i)}{P(T_1 = 1|D = i)},$$

assumed to be identical for all i. We must express $p_{11}^{(i)}$, $p_{10}^{(i)}$, $p_{01}^{(i)}$, and $p_{00}^{(i)}$ as functions of $p_{1+}^{(i)}$, $p_{+1}^{(i)}$, and θ. By easy algebra, we obtain

$$\begin{aligned} p_{11}^{(i)} &= P(T_1 = 1|T_2 = 1, D = i)P(T_2 = 1|D = i) \\ &= \theta P(T_1 = 1|D = i)P(T_2 = 1|D = i) \\ &= \theta p_{1+}^{(i)} p_{+1}^{(i)} \end{aligned}$$

and

$$p_{10}^{(i)} = p_{1+}^{(i)}\big[1 - \theta p_{+1}^{(i)}\big], \qquad p_{01}^{(i)} = p_{+1}^{(i)}\big[1 - \theta p_{1+}^{(i)}\big],$$

and

$$p_{00}^{(i)} = 1 - p_{1+}^{(i)} - p_{+1}^{(i)} + \theta p_{1+}^{(i)} p_{+1}^{(i)}.$$

Let us also define $a_i = \theta p_{1+}^{(i)} = P(T_1 = 1|T_2 = 1, D = i)$, $b_i = \theta p_{+1}^{(i)} = P(T_2 = 1|T_1 = 1, D = i)$, and $\eta = 1/\theta$, and rewrite the cell probabilities as

$$p_{11}^{(i)} = \eta a_i b_i, \qquad p_{10}^{(i)} = \eta a_i (1 - b_i), \qquad p_{01}^{(i)} = \eta (1 - a_i) b_i,$$

and

$$p_{00}^{(i)} = 1 - p_{11}^{(i)} - p_{10}^{(i)} - p_{01}^{(i)} = \eta(1 - a_i)(1 - b_i) + 1 - \eta.$$

Note that this gives a reparameterization of the four cell probabilities $p_{jk}^{(i)}$ for each i in three independent parameters η, a_i, and b_i. The log-likelihood function with incomplete observations (i.e., $x_{00}^{(i)}$ is not known but only $x_{00}^{(+)} = x_{00}^{(1)} + \cdots + x_{00}^{(d)}$ is known) is

$$\begin{aligned} &\sum_{i=1}^{d} \big\{ x_{11}^{(i)} \ln(\eta a_i b_i q_i) \\ &\quad + x_{10}^{(i)} \ln[\eta a_i (1 - b_i) q_i] + x_{01}^{(i)} \ln[\eta (1 - a_i) b_i q_i] \big\} \\ &\quad + x_{00}^{(+)} \ln\left[\sum_{i=1}^{d} \big\{ q_i [\eta (1 - a_i)(1 - b_i) + 1 - \eta] \big\} \right]. \end{aligned} \tag{16}$$

For maximum likelihood estimation, (16) must be maximized in $\eta, a_1, \ldots, a_d, b_1, \ldots, b_d, q_1, \ldots, q_d$, which turns out to be tedious. Formulation of the EM algorithm proves unfruitful in this problem, because the complete likelihood has no closed-form maximizers. Thus we argue that we should work directly with the observed log-likelihood and use the readily available routines for maximization.

3.2 The α Model

In the model with the homogeneous odds ratio, the parameter α is defined as

$$\alpha = \frac{p_{11}^{(i)} p_{00}^{(i)}}{p_{10}^{(i)} p_{01}^{(i)}}$$

and is considered to have the same value for all i. Using the parameterization of the other model, we can write

$$\begin{aligned} \frac{p_{11}^{(i)} p_{00}^{(i)}}{p_{10}^{(i)} p_{01}^{(i)}} &= \frac{\eta a_i b_i [\eta (1 - a_i)(1 - b_i) + 1 - \eta]}{\eta a_i (1 - b_i) \eta b_i (1 - a_i)} \\ &= 1 + \frac{1 - \eta}{\eta (1 - a_i)(1 - b_i)}. \end{aligned}$$

This relationship between α and η shows that in general, the two types of homogeneity hypothesis (i.e., η constant or α constant) are different in nature. More precisely, both, only one, or no homogeneity conditions may be satisfied.

In general, for the α model, the (incomplete, observed) likelihood is

$$\left[\prod_{i=1}^{d}\{p_{11}^{(i)}q_i\}^{x_{11}^{(i)}}\right]\times\left[\prod_{i=1}^{d}\{p_{10}^{(i)}q_i\}^{x_{10}^{(i)}}\right]\times\left[\prod_{i=1}^{d}\{p_{01}^{(i)}q_i\}^{x_{01}^{(i)}}\right]\times\left[\sum_{i=1}^{d}p_{00}^{(i)}q_i\right]^{x_{00}^{(+)}}. \quad (17)$$

What remains to be done is to express $p_{kl}^{(i)}$ as functions of a set of independent parameters and the likelihood maximized with respect to these parameters. For comparison purposes (the simulations in the next section are performed in this way), we take as fixed parameters $p_{1+}^{(i)}$, $p_{+1}^{(i)}$, α, and q_i, that is, $3d+1$ parameters with one constraint $q_1+\cdots+q_d=1$. The Appendix provides some details of this process.

4. SIMULATION–BASED COMPARISON OF ESTIMATORS

Because the likelihood approach turns out to be tedious, the simpler and easier to perform and understand capture–recapture estimators might be of interest. We compare these estimators with the maximum likelihood approach by means of a simulation study.

4.1 Design of the Simulation Study

The design of the simulation study chooses values for the sample size n, the disease strata weights q_i, the marginal conditional probabilities $p_{1+}^{(i)}$ that test T_1 is positive given disease status i, the marginal conditional probabilities $p_{+1}^{(i)}$ that test T_2 is positive given disease status i, and α or θ. All other model parameters are functions of these.

For each design considered, we generated 2,000 samples, and for each sample, we computed the estimators under study using a SAS/IML code. For the optimization problems, we used the NLPNMS function of SAS/IML (version 8.2) based on the Nelder–Mead simplex optimization method. We considered two disease strata, and fixed the parameters as

$$\left(p_{1+}^{(1)}, p_{1+}^{(2)}\right)=(.1,.3),\quad \left(p_{+1}^{(1)}, p_{+1}^{(2)}\right)=(.05,.25),\quad \text{and}\quad (q_1,q_2)=(.85,.15). \quad (18)$$

In the θ model (resp. α model), we used three different values for θ (resp. α): 1.1, 2, and 3. The sample sizes considered were $n=1{,}000$, 5,000, 10,000, and 25,000. We used the true values of the parameters as starting values in the optimization procedures.

4.2 Results

The simulation study was designed with $d=2$ disease stata. Consequently, we consider, for given n, $n_1=q_1\times n$ and $n_2=q_2\times n$ to be the true values and define the bias as $(E(\hat{n}_i)-n_i)/n$, where $\hat{n}_i$ is $n\times\hat{q}_i$ for the MLEs, and otherwise is the capture–recapture estimate as defined in (10) and (15). The mean $E(\hat{n}_i)$ is estimated by the average of the 2,000 values of $\hat{n}_i$ computed. Note that the bias is measured in a normed way, independent of the size of n, to enable comparisons over different values of n. Because of the symmetry, we can restrict our discussion to the results for one of the n_i's, say n_1. Two situations should be clearly kept apart. In the first case, data are generated under the θ model and MLEs are computed under this model, whereas capture–recapture estimates (10) and (15) are easily available for both models. In addition, MLEs for the Walter model have been included for comparison. As can be seen in Figure 1(a), the bias of the MLE and the capture–recapture estimate (10) are close and get closer when the sample sizes increases. For sample size >5,000, differences in bias appear negligible. Not surprisingly, the bias for the capture–recapture (15) is large and remains large (because it assumes the wrong model in this situation). Similarly, the bias of the Walter model is severe in all situations. Look at the standard errors of estimates (again normed by the sample size), it appears from Figure 1(b) that the capture–recapture estimate (10) has a standard error close to the standard error of the MLE even for the smaller sample sizes considered. The standard errors in the Walter model are small, in fact smaller than in all other models, which this seems reasonable if we recall that the Walter model can be viewed as a constrained θ or α model.

In the second case, data are generated under the α model and MLEs are computed under this model, whereas capture–recapture estimates (10) and (15) are again computed for both models. As it turns out (omitting details here for the sake of brevity; available on request), the bias of MLE and the capture–recapture estimate (15) are close and grow closer with increasing sample sizes. Again for large sample sizes, differences in bias appear negligible. Not surprisingly, the bias for the capture–recapture (10) is large and stays large (because it assumes the wrong model in this situation). Looking at the standard errors of estimates (again normed by sample size), it turns out (omitting details to save space; available on request) that the capture–recapture estimate (10) has a standard error close to the standard error of the MLE for even the smaller sample sizes considered. In summary, only a small loss of efficiency results from using the capture–recapture estimates (10) and (15).

Finally, we considered a configuration in which the cell probabilities $p_{jk}^{(i)}$ do not satisfy any of the two models under study. To compute the cell probabilities in this case, we used the parameterization described in Section 3.1 with the same values as in (18) but with different values θ for each i, denoted by θ_1 and θ_2. Two cases are reported: $\theta_1=2$, $\theta_2=1.5$ and $\theta_1=1.5$, $\theta_2=2$. In addition to the MLE for θ and capture–recapture estimators (10) and (15), we computed the MLE in the Walter model. The results are shown in Figure 2. In the first case ($\theta_1=2, \theta_2=1.5$), the bias of the MLE in the Walter model is close to the bias obtained in the θ model (with MLE and capture–recapture estimates) in absolute value but with the opposite sign. The very small bias obtained with the capture–recapture estimator in the α model can be explained by the configuration that is close to homogeneity of the odds ratios, indeed, $p_{00}^{(i)}p_{11}^{(i)}/\{p_{10}^{(i)}p_{01}^{(i)}\}$ is equal to 2.389 for $i=1$ and to 2.455 for $i=2$. The standard error of the MLE in the Walter model is smaller than that for the other estimates. In the second case ($\theta_1=1.5$, $\theta_2=2$), the θ model produces the smallest bias, whereas the Walter model yields the largest. When n increases, the estimate in the α model has a standard error close to that of the MLE in the Walter model, which is again the smallest among the four estimates.

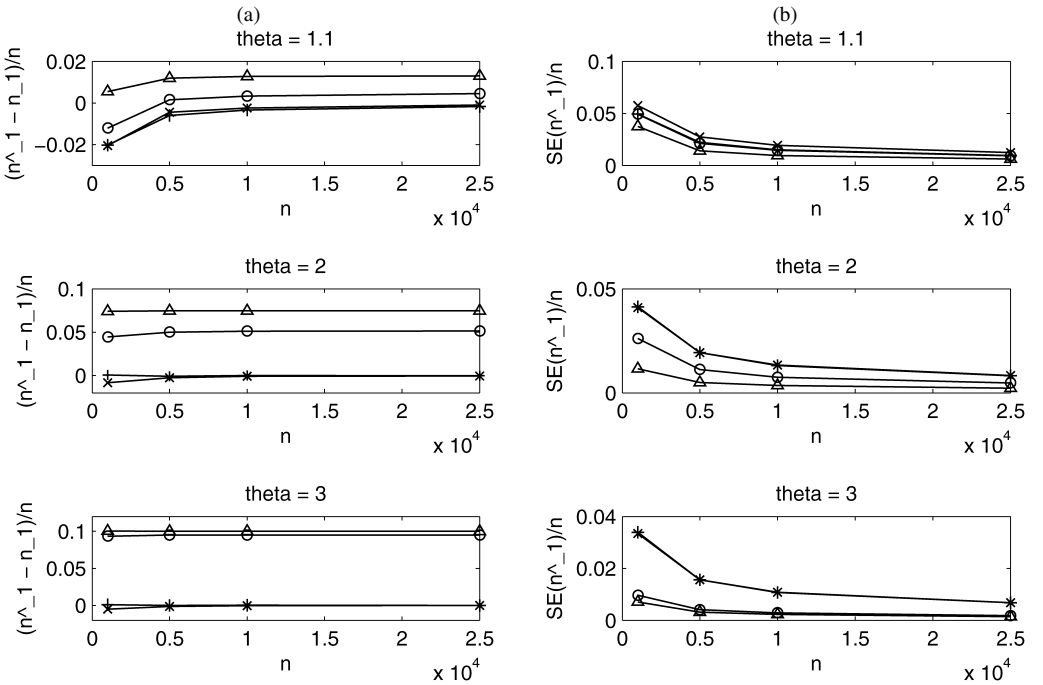

Figure 1. Bias (a) and standard error (b) of the estimates for n_1 based on the MLE for θ (+), on the CRE for θ (×), on the CRE for α (∘), and MLE in the Walter model (△). The true model is the θ model.

5. DISCUSSION

The methods of estimation described in the previous sections are applicable to many practical screening situations in which evaluation of the disease status is restricted to those for which at least one of the screening tests is positive. Screening is most effective when applied to large populations where evaluation of the disease status is limited to the test positives. Thus large data bases exist that allow assessment of diagnostic measures as well as disease prevalence estimates with good precision.

Previous approaches have used the assumption of screening test independence to a large extent. However, as we have seen here, this assumption appears to be questionable in many data sets. Procedures using this assumption of independence conditional on disease status can be grouped into two categories. The first group is a variant of the latent-class model that uses the axiom of local independence of the diagnostic tests conditional on disease status. In full generality, these models require no evaluation of the disease status. Estimation in this specific

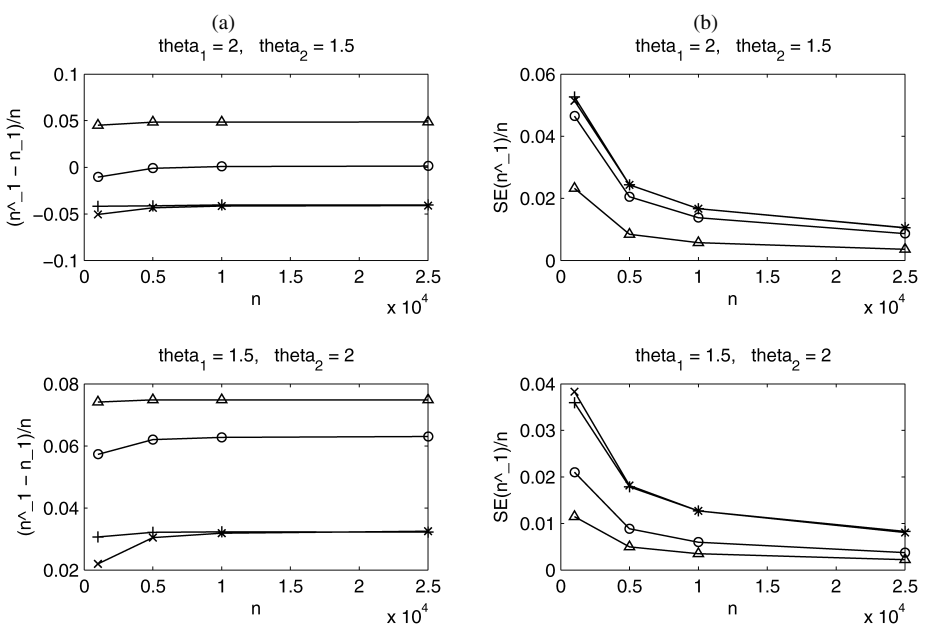

Figure 2. Bias (a) and standard error (b) of the estimates for n_1: MLE for θ (+), CRE for θ (×), CRE for α (∘), and MLE in the Walter model (△). The true model is neither the θ model nor the α model.

mixture model is frequently done using the EM algorithm or the Newton–Raphson algorithm. (For an overview of these models, see Greiner 2003.) The model of Walter (1999) uses the additional partial knowledge from the evaluated screened persons to achieve a specific latent class model. The other category comprises of capture–recapture procedures, such as the proposal by Goldberg and Wittes (1978). However, Goldberg and Wittes (1978) targeted only procedures for estimating the size of the diseased population (false negatives), and thus their procedures do not use knowledge on the test positives with negative disease status evaluation (false positives). In cases of independence, the results of their procedure (in fact, it corresponding to $\hat{n}_i$ here) and the modified estimators suggested here will be similar. In cases of dependence, the results will differ.

Indeed, the screening tests in HIP study were associated, as shown in Section 3. Consequently, the estimates provided by Goldberg and Wittes (1978) likely were to be too low.

The main focus of the article was to develop estimators under dependency. Section 1 mentioned simple statistical tests to illustrate that the assumption of screening test independency is violated in many case studies. It might well be possible to construct that more powerful statistical tests. On the other hand, in many application studies the independence of screening tests will not be of substantial interest itself, but rather interest will lie in constructing estimators of the sizes of the disease classes.

In Section 2 two simple estimators for the diseased and screened-negatives were suggested, one based on the θ model and the other based on the α model. The ability to discriminate between these two models would be of interest. This is impossible, however, because both provide a full model; therefore, some uncertainty remains when using any of the models. But which is better, using an independence model which in most cases fails to be valid, or using a model that allows homogeneous forms of dependencies, even though failure of homogeneity might not be detectable? We would argue that the latter approach should be preferred, because it is more likely to hold than the independence model.

Turning to the results of the simulation study (a full documentation of all results is available on request), it appears that no great loss of efficiency occurs when using the capture–recapture estimates instead of the MLEs. This is good news, because MLEs are more difficult to obtain. The Walter model has surprisingly small standard error in all situations, although the bias is, as expected, severe. Here there seems to be a slight tendency for the α model to perform better than the θ model (mostly smaller in standard error, and often smaller in bias is as well), but in general this depends on the true data-generating process, of course.

But a major question remains: How can the homogeneity assumption within the θ or α model be justified? One possibility is to look at empirical cases where the disease status is completely known, such as in the situation of Section 2.4 and Table 4. None of the four constellations considered in Table 4 shows any evidence of violation of the assumption of homogeneity as indicated by the test of homogeneity (the last two lines in Table 4).

Another strategy appears to be to incorporate a design element into the screening study that allows testing of the homogeneity model. This element could be a small subset of the screened population with gold standard evaluation of all units in the subsample. A further strategy (particularly if the determination of the gold standard for healthy individuals might be considered unethical) might involve sampling from a population of confirmed cases and determine the two tests for this sample. This would seem particularly attractive if the two tests were noninvasive. Both strategies would allow estimation of a heterogeneous θ or α model.

We would like to conclude by discussing one with a final issue. As pointed out by a referee, the idea of incorporating some form of dependency of the two screening tests had been mentioned previously. Van der Merve and Maritz (2002) modeled the odds ratio as 1 (independence of the two tests) for the disease population, whereas they allow an extra parameter γ for the odds ratio of the two tests within the healthy population. This idea could be extended and combined with previous idea of estimating the odds ratio first for a confirmed case population (with all four cells defined by the combination of the two tests available) and then estimating a ratio parameter γ of the two odds ratio by means of the screening sample.

APPENDIX: SOME DETAILS FOR THE α MODEL

We need to write the cell probabilities $p_{kl}^{(i)}$ involved in likelihood (17) as functions of a set of independent parameters. For comparison purposes, let us take as fixed parameters $p_{1+}^{(i)}$, $p_{1+}^{(i)}$, α, and q_i, that is, $3d+1$ parameters with one constraint, $q_1+\cdots+q_d=1$. We may write the following obvious relationships: $p_{10}^{(i)} - p_{01}^{(i)} = p_{1+}^{(i)} - p_{+1}^{(i)}$ and $p_{00}^{(i)} - p_{11}^{(i)} = 1 - p_{1+}^{(i)} - p_{+1}^{(i)}$. Moreover, by definition, $p_{00}^{(i)} p_{11}^{(i)} = \alpha p_{10}^{(i)} p_{01}^{(i)}$. Clearly, all of the probabilities $p_{kl}^{(i)}$ can be expressed as functions of the fixed parameters as soon as this can be done for $p_{00}^{(i)}$. By simple algebra, for each i, the probability $p_{00}^{(i)}$ is a solution of the equation

$$(\alpha-1)\big[p_{00}^{(i)}\big]^2 + (\alpha B_i + A_i - 2\alpha C_i)p_{00}^{(i)} + \alpha(C_i^2 - B_i C_i) = 0, \tag{A.1}$$

with $A_i = 1 - p_{1+}^{(i)} - p_{1+}^{(i)}$, $B_i = p_{1+}^{(i)} - p_{1+}^{(i)}$, and $C_i = 1 - p_{+1}^{(i)}$. It easy to see that (A.1) has only one solution in the interval (0, 1) which can be found by elementary algebra. Next, we write $p_{11}^{(i)}$, $p_{10}^{(i)}$, and $p_{01}^{(i)}$ as functions of $p_{1+}^{(i)}$, $p_{1+}^{(i)}$, and α. Finally, the (incomplete, observed) likelihood (17) can be written as a function of the parameters $p_{1+}^{(i)}$, $p_{1+}^{(i)}$, α, and q_i.

[Received September 2003. Revised August 2006.]

REFERENCES

Albert, P. S., and Dodd, L. E. (2004), "A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error Without a Gold Standard," *Biometrics*, 60, 427–435.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Borchers, D. L., Buckland, S. T., and Zucchini, W. (2002), *Estimating Animal Abundance: Closed Populations*, Heidelberg: Springer.

Chaisiri, K., Tungtrongchitr, R., Kulleap, S., Sutthiwong, P., Mahaweerawat, U., Sanchaisuriya, P., Merkle, A., Pongpaew, P., Phonrat, B., Kuhathong, C., Intarakhao, C., Khongdee, W., Saowakontha, S., and Schelp, F.-P. (1997), "Prevalence of Abnormal Glucose Tolerance in Khon Kaen Province and Validity of Urine Stick and Fasting Blood Sugar as Screening Tools," *Journal of the Medical Association of Thailand*, 80, 363–370.

Chapman, D. G. (1951), "Some Properties of the Hypergeometric Distribution With Applications to Zoological Censuses," *University of California Publications in Statistics*, 1, 131–160.

Cheng, H., and Macaluso, M. (1997), "Comparison of the Accuracy of Two Tests With a Confirmatory Procedure Limited to Positive Results," *Epidemiology*, 8, 104–106.

De Sutter, P. H., Coibion, M., Vosse, M., Hertens, D., Huet, F., Wesling, F., Wayembergh, M., Bourdon, C., and Autier, P. H. (1998), "A Multicenter Study Comparing Cervicography and Cytology in the Detection of Cervical Intraepithelial Neoplasia," *British Journal of Obstetrics and Gynecology*, 105, 613–620.

Galen, R. S., and Gambino, S. R. (1975), *Beyond Normality*: *The Predictive Value and Efficiency of Medical Diagnosis*, New York: Wiley.

Goldberg, J. D., and Wittes, J. T. (1978), "The Estimation of False Negatives in Medical Screening," *Biometrics*, 34, 77–86.

Greiner, M. (2003), *Serodiagnostische Tests. Evaluierung und Interpretation in der Veterinärmedizin and anderen Fachgebieten*, Heidelberg: Springer.

Hook, E. B., and Regal, R. (1995), "Capture–Recapture Methods in Epidemiology: Methods and Limitations," *Epidemiologic Reviews*, 17, 243–264.

Hui, S. L., and Walter, S. D. (1980), "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36, 167–171.

Hui, S. L., and Zhou, X. H. (1998), "Evaluation of Diagnostic Tests Without a Gold Standard," *Statistical Methods in Medical Research*, 7, 354–370.

Lincoln, F. C. (1930), "Calculating Waterfowl Abundance on the Basis of Banding Returns," *United States Department of Agriculture Circular*, 118, 1–4.

McLachlan, G., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford, U.K.: Oxford University Press.

Pepe, M. S., and Alonzo, T. A. (2001), "Comparing Disease Screening Tests When True Disease Status Is Ascertained Only for Screen Positives," *Biostatistics*, 2, 249–260.

Petersen, C. G. J. (1896), "The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea," *Report of the Danish Biological Station (1895)*, 6, 5–84.

Schatzkin, A., Connor, R. J., Taylor, P. R., and Bunnag, B. (1987), "Comparing New and Old Screening Tests When a Reference Procedure Cannot Be Performed on All Screenees," *American Journal of Epidemiology*, 125, 672–678.

Seber, G. A. F. (1970), "The Effects of Trap Response on Tag-Recapture Estimates," *Biometrika*, 26, 13–22.

Shurtleff, D. (1974), "Some Characteristics Related to the Incidence of Cardiovascular Disease and Death: 18-Year Follow-up," in *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*, eds. W. B. Kannel and T. Gordon, Washington, DC: U.S. Department of Health, Education, and Welfare, Public Health Service, National Institutes of Health, pp. 74–599.

Smith, D., Bullock, A., and Catalona, W. (1997), "Racial Differences in Operating Characteristics of Prostate Cancer Screening Tests," *Journal of Urology*, 158, 1861–1866.

Strax, P., Venet, L., Shapiro, S., and Gross, S. (1967), "Mammography and Clinical Examination in Mass Screening for Cancer of the Breast," *Cancer*, 20, 2184–2188.

Van der Merwe, L., and Maritz, J. S. (2002), "Estimating the Conditional False-Positive Rate for Semi-Latent Data," *Epidemiology*, 13, 424–430.

Walter, S. D. (1999), "Estimation of Test Sensitivity and Specificity When Disease Confirmation Is Limited to Positive Results," *Epidemiology*, 10, 67–72.

Wittes, J. T. (1972), "On the Bias and Estimated Variance of Chapman's Two-Sample Capture–Recapture Population Estimate," *Biometrics*, 28, 592–597.