Bayesian estimation of the false negative fraction in screening tests

Leonhard Held[*]   Argwings Otieno Ranyimbo

March 26, 2004

1 Introduction

Over recent years the application of diagnostic tests for the detection and evaluation of various diseases has considerably increased (Begg (1983)). Two important summary measures of the accuracy of a screening or diagnostic test are sensitivity and specificity. Sensitivity is the probability of a test-identified case among the diseased while specificity is the probability of a test-negative result among the disease-free. It is also possible to give a summary of the test accuracy in terms of diagnostic or screening errors. The probabilities of these errors are by definition the false negative fraction (FNF) which is one minus the sensitivity and the false positive fraction (FPF) which is one minus the specificity (Lloyd and Frommer (2003)). Definitive assessment or verification of the true disease status is normally provided by a gold standard procedure. It is quite simple to estimate the above mentioned error rates when all the study subjects are verified. However, the gold standard procedure may be more invasive, costly or risky. In addition it is unethical to perform such procedures on individuals who have initial negative test results (Walter (1999)). As a consequence only a subsample of those who were initially tested receive the definitive assessment for their disease status. Usually verification is done only for individuals whose initial test results are positive. This leads to the problem of estimation of sensitivity and specificity or alternatively FNF and

[*]Address for correspondence:Institute of Statistics,Ludwig-Maximilians-University Munich, Ludwig str. 33, 80539 Munich, Germany.

FPF under partial verification. We consider the problem of estimating the FNF in *multiple screening test*, whereby the screening test comprises k repeated applications of a dichotomous kit test. Individuals who tested negative in all the k tests were not verified.

Our approach is motivated by data from Lloyd and Frommer (2003). In total 38000 patients were voluntarily screened for bowel cancer at St. Vincent's Hospital in Sydney, Australia. About 3000 patients tested positive at least once and their true disease status was verified. Only 196 had the disease. The primary data consisted of the count, between 1 and 6, of positive tests results for each of the 196 patients. Of the 196 patients 122 agreed to be screened further several weeks after the primary data was collected. Results from this second screening phase constitute the secondary data set. The problem is to estimate the FNF.

In Section 2 the beta-binomial model is derived while Section 3 describes the Bayesian approach as used in this paper. Section 4 gives a brief description of a Bayesian logistic regression approach to estimate the FNF as proposed in Lloyd and Frommer (2003). In Section 5 we present our results in the estimation of the FNF using both the beta-binomial model as well as the Bayesian logistic regression approach while in Section 6 we validate both the beta-binomial as well as the Bayesian logistic models and finally give the concluding remarks in Section 7.

2 Beta-binomial model

Let Y_{ij} ($i = 1, ..., n$ and $j = 1, ..., k$) be a random variable from a Bernoulli distribution with parameter p, $0 < p < 1$. Thus p is the probability that a randomly chosen individual tests positive. Defining $Y_i = \sum_{j=1}^{k} Y_{ij}$ it follows that Y_i is the count of the number of positive tests for the i-th individual and so has a binomial distribution. We assume that the population of the individuals is heterogenous so that p has a $beta(\alpha, \beta)$ distribution;

$$f(p|\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \quad \alpha > 0, \beta > 0, 0 < p < 1$$

where $B(\alpha, \beta)$ is the complete beta function. This assumption may be valid when the data exhibits overdispersion. For the i-th randomly chosen individual the distribution of Y_i for the case of k trials is the following mixture model

$$Pr(Y_i = y_i|\alpha, \beta, k) = \int_0^1 Pr(y_i|p, k) f(p|\alpha, \beta) dp$$

Hence the distribution of the count of the positive tests for the i-th randomly chosen individual is beta-binomial and is given as

$$Pr(Y_i = y_i|\alpha, \beta, k) = \pi_{y_i k}(\alpha, \beta) = \begin{pmatrix} k \\ y_i \end{pmatrix} \frac{B(\alpha+y_i, \beta+k-y_i)}{B(\alpha,\beta)},$$

$y_i \in \{0, ..., k\}$ and $i = 1, ..., n$.

The beta-binomial model has been applied widely in different studies (Williams (1975), Lee and Sabavala (1987)). In the bowel cancer data considered in this paper the individuals who tested negative in all the $k = 6$ initial tests were not verified. For fixed p it then follows that the distribution of the positive counts Y_i for the individuals who underwent verification is a zero-truncated binomial distribution. As a consequence the beta-binomial model is also truncated at zero value. Truncation essential means that the denominators for sensitivity and specificity remain unkown and so one cannot estimate these summary measures. Putting $y_i = 0$ in the beta-binomial distribution given above for the case of $k = 6$ diagnostic tests leads to the expression for the FNF as

$$\pi_{06}(\alpha, \beta) = \frac{B(\alpha,\beta+6)}{B(\alpha,\beta)}$$

If the responses of the individuals are considered independent then the likelihood for the primary data is

$$L_y(\alpha, \beta) = \frac{\prod_{j=1}^{k} \pi_{jk}^{n_j}}{(1-\pi_{0k})^n}$$

where $n = \sum_{j=1}^{k} n_j$ and $n_j (j = 1, ..., k)$ is the count of individuals with j positive tests while $y = (y_1, ..., y_n)$ is the vectors of positive observations for all the n individuals. If y_i^* denotes the count of the positive tests for the i-th randomly chosen individual who had a further k^* tests in the secondary phase after registering y_i positive tests in the first k trials we have that

$$Pr(Y_i^* = y_i^* | y_i, \alpha, \beta, k^*) = \pi_{y_i^* k^* | y_i}(\alpha, \beta) = \begin{pmatrix} k^* \\ y_i^* \end{pmatrix} \frac{B(\alpha+y_i+y_i^*, \beta+k-y_i+k^*-y_i^*)}{B(\alpha+y_i, \beta+k-y_i)},$$

$y_i^* = 0, ..., k^*$ and $y_i = 1, ..., k$.

This is another beta-binomial distribution. The likelihood for this secondary data is

$$L_{y^*}(\alpha, \beta) = \prod_{i=1}^n \pi_{y_i^* k^* | y_i}(\alpha, \beta) 1_S(i) = \prod_{l=1}^k \prod_{j=0}^{k^*} \{\pi_{jk^*|l}(\alpha, \beta)\}^{n_{lj}}$$

where S is the index set of all the individuals who participated in the secondary phase and n_{lj} is the number of individuals with j positive tests in the secondary phase given that they had l positive tests in the primary tests.

3 Reparameterization and Bayesian approach

Instead of using the parameters α and β we use here the reparameterization $\mu = \alpha/(\alpha + \beta)$ and $\rho = 1/(\alpha + \beta + 1)$. The parameter μ is the expectation of the $beta(\alpha, \beta)$ distribution. Due to multiple diagnostic tests being done on the same individual there is bound to be dependency in the response. The parameter ρ gives a measures of the correlation of the responses of an individual. We assume that μ and ρ have each *a priori* the $beta(0.5, 0.5)$ distribution, which turn out to be the reference prior in this setting (van der Linde (2003)). The joint posterior distribution of μ and ρ is then a complex two-dimensional distribution. Attempts to obtain conditional posterior distributions does not give distributions with well known form. This then means that Gibbs sampling methodology is not applicable in this case. However, other Markov Chain Monte Carlo (MCMC) methods such as Metropolis-Hastings algorithm may be used. Basically MCMC simulates a Markov chain whose stationary or limiting distribution is the posterior distribution of interest.

4 Bayesian logistic regression

An alternative approach to estimate the FNF is based on logistic regression (Lloyd and Frommer (2003)). This approach is based on modeling and extrapolating patterns in the FNF conditional on individual histories. In case of heterogeneity or if there is positive correlation, then individuals who test positive more often in the past are more likely to test positive in the future, and hence are likely to return false negatives. Lloyd and Frommer (2003) shows that the probability of m consecutive negative test results conditional on x prior positive results out of k (denoted $\gamma_{m|xk}$) is a decreasing function of x and an increasing function of k. Letting γ_k denote the FNF it can be shown that

$$\gamma_k = \frac{\gamma_{1|11}}{1-\gamma_{1|01}+\gamma_{1|11}} \prod_{j=1}^{k-1} \gamma_{1|0j}$$

Starting with the basis functions $x/k, x/k^2, x/k^3, 1/k, 1/k^2$ and using the minimum AIC Lloyd and Frommer (2003) selects the following model

$$\log\left(\frac{\gamma_{1|xk}}{1-\gamma_{1|xk}}\right) = \beta_0 + \beta_1 \frac{x}{k} + \beta_2 \frac{x}{k^3}$$

We fit the same model but in a Bayesian setup using the Metropolis-Hastings algorithm and incorporating the Gamerman (1997) iterative weighted least squares algorithm.We assume that β_0, β_1 and β_2 have each a uniform prior distribution. For each posterior draw of $(\beta_0, \beta_1, \beta_2)$ we obtain a posterior sample of $\gamma_{1|xk}$. In particular for the values $(x = 0, k = 1)$ and $(x = 1, k = 1)$ posterior draws of γ_k for any value of k can be obtained. We use an automatic and efficient algorithm proposed by Gamerman (1997) for MCMC simulation.

5 Results

The MCMC algorithm was tuned so that an acceptance rate of between 35% and 40% was obtained (Gelman, Roberts, Gilks (1995)). In addition, thinning was done so as to eliminate serial correlation in the draws of μ and ρ. Table 1 displays the estimates of the FNF for different models. In model 1 only the primary data is used. In both models 2 and 3 both primary and secondary data are used.

In model 2 we assume one dropout probability while in model 3 we assume several dropout probabilities. Having one dropout probability is equivalent to assuming that all the individuals are equally likely not to take part in the secondary phase of the screening tests. Different dropout probabilities reflect the fact that the likelihood of an individual having secondary tests is dependent on the primary result. The median posterior estimate of the FNF from the primary data (model 1) was 26.4% with a 95% credible interval of $(0.123, 0.650)$. The median posterior estimates of μ and ρ in case of model 2 were 0.455 and 0.500 respectively. The corresponding 95% credible intervals were $(0.208, 0.569)$ and $(0.374, 0.660)$ respectively.

Estimation of the number of diseased individuals who were diagnosed as disease free depends of the draw of the FNF and whether we are considering the primary data only or both primary and secondary data. For example in the case of model 1 the data consists of 196 diseased patients so that solving the equation $FNF = m/(196 + m)$ for each draw of the FNF gives the posterior distribution of the number of missed cases. For the case of model 1 the median estimate of missed cases was approximately 71 with a 95% credible interval as (33,174)(see Table 1). Inclusion of the secondary data in the analysis (models 2 and 3) lead to lower estimates for the FNF with narrower credible intervals. There does not appear to be informative dropout as estimates of the FNF are similar in the last two models.

The Bayesian logistic model with covariates x/k and x/k^3 given in Section 3 was fitted to the data in Table 2 of Lloyd and Frommer (2003). The symbol x represents the number of positive tests out of a total of k tests in the primary phase. Table 2 gives both the Maximum Likelihood Estimates (MLE) and the corresponding MCMC ones. Both approaches give roughly similar estimates. With (x=0,k=1) posterior samples of $\gamma_{1|01}$ were obtained. This was repeated for (x=1,k=1) to obtain the posterior samples of $\gamma_{1|11}$. Posterior draws of γ_k for different values of k were obtained from the expression for γ_k in Section 3. For k=6 the mean estimate of the FNF based on MLE using the primary data is 23.6% and is similar to 23.8% obtained from Bayesian logistic regression.

6 Model Validation

We validate the beta-binomial model fitted to the primary data by predicting the expected marginal frequencies for the secondary data. Using the μ and ρ parameterization the predictions are obtained from the expression

$$Pr(Y^* = y^* \mid \alpha, \beta) = N_s \binom{k^*}{y^*} \frac{B(\alpha + y^*, \beta + k^* - y^*) - B(\alpha + y^*, \beta + k + k^* - y^*)}{B(\alpha, \beta) - B(\alpha, \beta + k)}$$

for $y^* \in \{0, ..., k^*\}$

where $\alpha = \mu(1/\rho - 1)$ and $\beta = (1-\mu)(1/\rho - 1)$. μ and ρ are replaced by the values from the posterior draws after the beta-binomial model is fitted to the primary data with $k = k^* = 6$. $N_s = 122$ is the number of individuals who had secondary tests. We also validate the Bayesian logistic model fitted above. To do this we need to evaluate the conditional probability of an individual having l positive tests given that the individual had x positive test out of k previous tests. For example the probability that an individual has one positive test out of six tests is the sum of the probabilities of each of the following 6-tuple binary vectors (1,0,0,0,0,0), (0,1,0,0,0,0), (0,0,1,0,0,0), (0,0,0,1,0,0,), (0,0,0,0,1,0,), (0,0,0,0,0,1), where 1 is for positive test and 0 otherwise. To evaluate the probability of the first vector say, we have

$$\begin{aligned} Pr(1,0,0,0,0,0) &= (1 - \gamma_{1|00})\gamma_{1|11}\gamma_{1|12}\gamma_{1|13}\gamma_{1|14}\gamma_{1|15} \\ &= (1 - \gamma_1)\gamma_{1|11}\gamma_{1|12}\gamma_{1|13}\gamma_{1|14}\gamma_{1|15}. \end{aligned}$$

All the $\gamma_{l|xk}$ can be obtained from the logistic model. However it is computationally tedious because the conditional probabilities have to be evaluated for all the different combinations of 0s and 1s. For example for two positive tests we have 10 different combinations while for 3 positive tests we have 20 different combinations. Table 3 gives the results based on the two models. Generally the two models do not predict the secondary observations quite well. The "p-value" is the probability of the predicted value being less than the observed count. Table 4 are the cell predictions based on the beta-binomial model. Again the model do not appear to predict the cell frequencies quite well.

7 Conclusion

We conclude that by using the Bayesian approaches on the observed diagnostic histories and patterns it is possible to estimate the FNF. Estimation from the primary data with $beta(0.5, 0.5)$ reference priors for μ and ρ leads to a median estimate of the FNF as 26.4% with a 95% credible interval of $(0.123, 0.650)$. The corresponding estimate based on the Bayesian logistic regression was 23.8%. There do not appear to be informative selection with regard to those who had the secondary tests because assuming that the individuals had varying probability of taking part in the secondary phase do not result to estimate of the FNF that is different from that obtained when all the individuals had equal probability of missing the secondary tests. Although the beta-binomial model do not appear to fit the data very well it's theoretically and computationally simple to handle and in addition models the heterogeneity and correlation patterns within an individual. The Bayesian approach further provides direct estimates of credible intervals as we are dealing directly with samples from the posterior distribution of the FNF. We therefore avoid the problem of estimation of standard errors based on the delta approximation method. One particular niece feature of the Bayesian approach is the ability to validate whether our model fits the data well.

References

Begg, C. B. (1983) Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 39, 207-215.

Lee, J. C. and Sabavala, D. J. (1987) Bayesian estimation and prediction for the beta-binomial model. *Journal of Business & Economic Statistics* 5, 357-367.

Lloyd, C. J. and Frommer, D. J. (2003) Regression based estimation of the false negative fraction when multiple negatives are unverified.

van der Linde, A. (2003) Personal communication.

Walter, S. D. (1999) Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. *Epidemiology* 10,67-72.

Williams, D. A. (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* 31, 949-952.

Gelman, A., Roberts, G. O. and Gilks, W. R. (1995) Efficient Metropolis jumping rules. In *Bayesian Satistics 5* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith). Oxford University Press, 599-607.

Table 1: Estimates of the FNF.

model	mean	s.e	median	95% CI	Missed cases
1	0.295	0.135	0.264	0.123-0.650	70.3
2	0.249	0.095	0.232	0.122-0.486	59.2
3	0.245	0.087	0.229	0.122-0.461	58.2

Table 2: Logistic models using maximum likelihood method (MLE) and the MCMC based on conditional FNFs (cf. Lloyd(2003))

	MLE		MCMC			
	estimate	s.e	estimate	s.e	median	95%CI
intercept	1.573	0.231	1.567	0.227	1.564	1.122-2.015
x/k	-3.602	0.356	-3.590	0.351	-3.587	-4.285-2.906
x/k^3	1.008	0.294	1.010	0.288	1.013	0.448-1.574

Table 3: Predicted counts based on beta-binomial model and the Bayesian Logistic regression model. [1]based on beta-binomial model with μ and ρ having $beta(0.5, 0.5)$ prior each; [2]based on Bayesian logistic model.

		0	1	2	3	4	5	6
Observed		22	8	12	16	21	12	31
Expected[1]								
	mean	11.7	14.0	14.7	15.1	16.1	18.8	31.8
	p-value	1.000	0.000	0.001	0.783	1.000	0.000	0.431
Expected[2]								
	mean	29.0	18.7	12.1	9.7	12.3	16.5	24.8
	"p-value"	0.189	0.000	0.457	1.000	1.000	0.041	0.925

Table 4: Observed and expected numbers of individuals testing positive y times in the primary test and y^* times on secondary test.

		y^*						
y		0	1	2	3	4	5	6
1		10	3	3	2	2	1	2
	mean	8.2	6.7	4.3	2.3	1.0	0.3	0.1
	"p-value"	0.950	0.000	0.000	0.152	0.999	1.000	1.000
2		3	2	2	1	4	1	1
	mean	2.2	3.3	3.3	2.6	1.6	0.8	0.2
	"p-value"	1.000	0.000	0.000	0.000	1.000	0.997	1.000
3		4	1	5	3	4	1	4
	mean	1.1	2.8	4.1	4.5	3.9	2.6	1.0
	"p-value"	1.000	0.000	1.000	0.000	0.712	0.000	1.000
4		1	1	0	3	4	1	4
	mean	0.2	0.8	1.7	2.7	3.3	3.2	2.0
	"p-value"	1.000	1.000	0.000	1.000	1.000	0.000	1.000
5		3	1	1	4	4	3	6
	mean	0.1	0.4	1.1	2.4	4.3	6.5	7.3
	median	0.1	0.4	1.1	2.4	4.3	6.4	7.3
	"p-value"	1.000	1.000	0.016	1.000	0.001	0.000	0.000
6		1	0	1	3	3	5	16
	mean	0.0	0.1	0.2	0.8	2.1	5.7	21.0
	"p-value"	1.000	0.000	1.000	1.000	0.988	0.146	0.001
total	mean	11.8	14.0	14.8	15.3	16.4	19.0	30.7
	"p-value"	1.000	0.000	0.000	0.8308	1.000	0.000	0.5647