EAPAS
East African Journal of Pure and Applied Science

East African Journal of Pure and Applied Science Vol. 2 (2009) 66-73

Comparison of Estimates of False Negative Fraction (FNF) and Predictions when different Non-informative Priors are used in a two-stage Screening Test.

O. R., Argwings

Mathematics and Computer Science Department

Chepkoilel University College

P.O Box 1125-30100 Eldoret

Abstract

Summary measures of the performance of a diagnostic kit require all study subjects to be verified via a gold standard procedure. However the subjection of all subjects to such a procedure may not be possible due to associated risks, invasiveness and cost. In normal practice only those who register at least one positive test result undergo the confirmatory procedure. Over the recent past different models have been proposed to estimate the false negative fraction (FNF) in this partial verification scenario using Maximum Likelihood and Bayesian techniques. In the Bayesian framework different priors have been proposed for the parameter of a Bernoulli distribution. In this work we compared the estimates of FNF obtained when three different non-informative priors are assigned to the probability of an individual testing positive and further did some validation by comparing the predictions with the actual observed data. Results show that the estimates of the FNF under three selected non-informative priors are largely similar. We conclude that though different forms of non-informative priors for the parameter of the Bernoulli distribution are in existence they do lead to similar results. The choice of the non-informative prior to use does not really matter. However, it was found that the predictions based on each of the three selected non- informative priors did not fit the observed data quite well.

KEY WORDS: Non-informative Prior, Bayesian techniques, Partial Verification.

†Corresponding author's E-mail address: ranyot2001@yahoo.com,

ISSN 2070 0903 /2009

1. Introduction

Accurate classification of individuals as either diseased or non-diseased is part of the critical routine processs of providing health care. Binary classification kits are commonly used in such excercises. The kit classifies an individual as either being positive or negative. Two important summary measures of the performance of a binary classifier are sensitivity and specificity. Sensitivity is the proportion of true-diseased correctly identified by the test while specificity is the proportion of the non-cases who are are correctly classified as negatives by the test. The performance of the binary kit can also be expressed in terms of its error rates. These are simply 1- sensitivity, also refered to as the false negative fraction (FNF) and 1-specificity, which is also called the false positive fraction (FPF) (Lloyd and Frommer, 2004). All the above characteristics of a binary classifier can be easily obtained provided all the study subjects undergo a confirmatory gold standard procedure. In certain cases not all the subjects tested using the binary kit go to through the confirmatory procedure. Such factors as cost, associated risks and invasiveness of the gold standard procedure may render such a process to be inapplicable. It is also possible that only those who have been identified as positives by the binary classifier are subjected to the confirmatory procedure. Because of this partial verification the summary measures of the testing kit cannot be obtained directly.

The data used in this work is taken from Lloyd and Frommer (2004) where 38000 subjects volunteered to be screened for bowel cancer. In the primary phase each study subject was required to test for blood in stool on 6 consecutive days using a self- administered kit. About 3000 subjects tested positive at least on one occasion. Those who had at least one positive result had their true disease status verified using physical examination, sigmoidoscopy and colonoscopy. Of all the subjects with a minimum of one positive test 196 were confirmed to be true cases. Individuals who tested negative in all the 6 days were not verified. In the secondary phase further screening was conducted on 122 of the initial 196 verified cases. Each of these 122 individuals volunteered to take further 6 tests using the same self-administered kit as in the primary phase. These tests were conducted about one week after the primary phase. The tabulation of the number of study subjects based on the positive counts in the primary and the secondary phases is given in Table 1. The column marked “missed“are those who did not volunteer to be tested in the second part of the study. Since there is partial verification direct estimation of the summary measures of the test kit is not possible.

Lloyd and Frommer (2004) use maximum likelihood approch while Held and Ranyimbo (2004) propose Bayesian techniques to addressing the estimation problem under partial verification scenario. The Bayesian approach has the advantage that use is made of the prior information regarding the parameter and in addition the uncertainty about the parameter can be captured by specification of the prior distribution. Numerous priors derived under different paradigms are in existence. The selection of the suitable prior is a critical exercise. The first objective of this work is to compare the estimates of the FNF under the partial verification setup when we use three differeent non-informative priors. The second objective is to validate the models.

obtained under the different non-informative priors by making use of the out-of-sample predictions. The paper is organized as follows: In Section 2 we briefly discuss our method which entails Bayesian inference, derivation

of the beta-binomial model and validation. We give the obtained results in Section 3 and our concluding remarks in Section 4.

2.1 Bayesian Infe

2.2 rence

The Bayesian approach consists of two parts: specification of the likelihood $f(y/\vartheta)$ and choice of a suitable prior $\pi(\vartheta)$ where $\underline{y}$ and ϑ are the vector of observations and the unknown parameter respectively. If the functional form of the prior is known then using Bayes' theorem the posterior distibution of ϑ is given by

$$f(\vartheta/\underline{y}) = \frac{f(\underline{y}/\vartheta)\pi(\vartheta)}{\int f(\underline{y}/\vartheta)\pi(\vartheta)d\vartheta} \text{ .... (1)}$$

The denominator in expression (1) above is called the marginal density of $\underline{y}$. The evaluation of the integral is usually difficult if not impossible hence need for approximations. With the fast development of Monte Carlo computing methods the

integrals of the above type can now be accurately estimated and hence advanced

Bayesian data analysis can be conducted (Carlin and Louis, 1996). Since the denominator is not a function of ϑ we can summarize the posterior as

$$Posterior \propto Likelihood \times \Pr ior$$

In certain conditional analysis there may be no or very little information regarding the unknown parameters. In order to employ the Bayesian techniques in such a situation one may choose to work with non-informative priors. Berger (1985) has given an explanation of a non-informarive prior to be one which has virtually no information regarding the unknown parameter. Various approaches have been proposed for the derivation of non-informative priors. The commonly used approach is that due to Jeffreys (1961) which involves choosing th non-informative prior $\pi(\vartheta)$ for the unknown parameter ϑ as

2. Materials and Methods

$$I(\vartheta) = -E\left[\frac{}{\partial\vartheta^2}\right] \text{...............(3)}$$

is the expected Fisher information function. Other methods which have been put forward in deriving non-informative priors include Novick and Hall (1965), Zellner (1971, 1977), Akaike (1978) and Bernado (1979) among others. Berger (1985) has argued that the approach by Bernado does better even where the others fail. In this work we look at the following three priors for

parameter $\vartheta \in (0,1)$:

Prior I: $\pi(\vartheta) \propto \vartheta^{-\frac{1}{2}}(1-\vartheta)^{-\frac{1}{2}}$

Prior II: $\pi(\vartheta) \propto 1$

Prior III: $\pi(\vartheta) \propto \vartheta^{-1}(1-\vartheta)^{-1}$

All the prior listed above belong to the *beta*(α , β) distribution with the density as given below:

$$f(\vartheta|\alpha,\beta) = \frac{\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}}{B(\alpha,\beta)} \quad \alpha>0,\beta>0,0<\vartheta<1 \text{ .....(4)}$$

Where $B(\alpha,\beta)$ is the complete beta function.

2.3 Beta-Binomial model

The beta-binomial model that is used to model the bowel cancer study described above has been discussed in Held and Ranyimbo (2004). First we consider the 196 individuals who tested

Let $X_i^{(j)} = 1$ if the ith subject tests positive on the jth test.

$X_i^{(j)} = 0$ if the ith subject tests negative on the jth test.

And suppose that the probability that an individual tests positive is ϑ then clearly $X_i^{(j)}$ follows a Bernoulli distribution with parameter ϑ . It then follows that the count of the positives tests for the *i*th individual is given by $V_i = \sum_{k=1}^{6} X_i^{(k)}$. The distribution of V_i is binomial with parameters (6, ϑ). If ϑ is assumed to follow the $Beta(\alpha,\beta)$ distribution then the marginal density of V_i is obtainable from the following integral:

$$f(v_i/\alpha,\beta) = \int_0^1 \binom{6}{v_i} \vartheta^{v_i}(1-\vartheta)^{6-v_i} \frac{\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}}{B(\alpha,\beta)} d\vartheta \quad \ldots. (5)$$

Evaluating the integral in (5) leads to the following beta-binomial distribution

$$f(v_i/\alpha,\beta) = \binom{6}{v_i} \frac{B(\alpha+v_i,\beta+6-v_i)}{B(\alpha,\beta)} ,v_i=0,1,2,3,4,5,6 \quad ..(6)$$

The obtained beta-binomial is such that

$$E\left(\frac{V_i}{6}\right) = \frac{\alpha}{\alpha+\beta}$$ and

$$Var\left(\frac{V_i}{6}\right) = \frac{\alpha\beta}{6(\alpha+\beta)^2}\left(1+\frac{5}{\alpha+\beta+1}\right)$$

The last bracket in the expression for variance is known as the variance inflation or the extra-

$$f(v_i/\alpha,\beta,V_i>0) = \frac{f(v_i/\alpha,\beta)}{1-f(0/\alpha,\beta)} ,v_i=1,2,3,4,5,6\ldots.(7)$$

Also if the functional form of the beta-binomial distribution is known then the false negative fraction (FNF) is simpby given by $FNF = f(0/\alpha,\beta)$. It can be shown that the likelihood for the data from the primary phase is

$$L(\alpha,\beta) = \frac{\prod_{l=1}^{6}[f(v_i/\alpha,\beta)]^{m_j}}{[1-f(0/\alpha,\beta)]^{196}} \ldots\ldots\ldots..(8)$$

where m_j is the count of study subjects with *j* positive tests. Instead of working with the parameters (α,β) we adopt the reparameterization $\mu = \frac{\alpha}{\alpha+\beta}$ and $\rho = \frac{□1}{\alpha+\beta+1}$ where is the prior mean, $\alpha+\beta$ is an indicator of the prior precision and is the measure of correlation between an individual's test results. We assign independent priors (using Prior I, II and III) to (μ,ρ). Using Bayes' theorem the posterior distribution of (μ,ρ) given V_i

binomial variation (Carlin and Louis, 1996). Because of this term the beta- binomial model is commonly used to model overdispersed data. Since in the bowel cancer data the individuals who tested negative in all the 6 tests

$(V_i = 0)$ were never verified it follows that the distribution of the V_i is the zero-truncated beta-binomial distribution as derived in Lloyd and Frommer (2004), and given as

Since V_i follows a beta-binomial

distribution, we can use Bayes' theorem to obtain the posterior distribution of ϑ given v_i .Thus

$$\pi(\vartheta / v_i) \propto \vartheta^{\alpha+v_i-1}(1-\vartheta)^{\beta+5-v_i} \quad (9)$$

which is another beta distribution. If v_i^s denotes the count of positives for the *i*th individual who volunteered to participate in the secondary phase after registering v_i positives in the primary phase then $f(v_i^s / v_i, \alpha, \beta)$ can be shown to have a beta-

binomial distribution

$$f(v_i^s/v_i,\alpha,\beta)=\binom{6}{v_i^s}\frac{B(\alpha+v_i+v_i^s,\beta+12-v_i-v_i^s)}{B(\alpha+v_i,\beta+6-v_i)},\ v_i^s=0,1,2,3,4,5,6.(10)$$

Using the above probablity we get the predicted counts of individuals having a given number of positives in the secondary phase given their history of primary counts.

3. Results

Using the primary data alone we applied the MCMC approach to the bowel cancer data assuming each of the three priors. We ran three

(i=1,...,196) can be obtained. However this posterior density is not in a well-known closed form hence the need to use Markov Chain Monte Carlo methods.We used a Metropolis Hastings algorithm where

and are updated separately. From the posterior samples of and we can obtain the posterior estimates of FNF.

2.4 Model Validation

Since we have three beta-binomial models derived using three different priors it is of interest to determine which model would be suitable for the bowel cancer data. To make this assessment we rely on the out-of-sample

prediction approach assuming the secondary data is representative of the primary results.the estimates are quite similar. However the 95% credible interval for the model under

Prior III is relatively larger.

Tables 4, 5 and 6 report the predicted counts in the secondary phase given the particular primary history under the three different models. Generally the predictions based on all the three models are again similar though the predicted values do not appear to match the observed counts.

4. Discussion and Conclusions

Prior ellicitation is usually a difficult exercise that requires experience. The choice

of non-informative prior lessens the burden of determining which prior to use. By

choosing to a prior with little or no

separate MCMC algorithms with 10,000 iterations each. Our burn-in was 500 runs. We approximated Prior III with *Beta*(0.0001,0.0001) distribution as an

approximation to avoid computation collapse as the MCMC algorithm is run. Table 2 shows the median posterior estimates of the FNF obtained using the three specified priors.

The estimates are very close and we can say that the

FNF is about 27%. The 95% credible intervals are largely similar except that under the model with Prior III this interval is wider. This is likely due to the approximation of Prior III. In Table 3 we report the posterior median estimate of the number of individuals who could have been misclassified as negatives by the test. Again information about the unknown parameter it means that the likelihood will be dominant in determination of the posterior distribution (Carlin and Louis, 1996). The three priors that are considered in this paper have different functional form since they are derived under different conditions. Though the non-informative priors for the parameter of the Bernoulli distribution considered in this paper are derived under different assumptions, the final posterior estimates of the false negative fraction under the beta- binomial models are largely similar. It therefore means that when one is not certain which prior to use, in a scenario where there in no adequate information, any non- informative prior would do. The results also indicate that the out-of-sample predictions of the secondary data are not well fitting. One probable reason as to why the models do not appear to give good predictions is that the considered priors are not invariant under reparameterization. Secondly the tests from each subject have been assumed to be independent. This may not be the case hence the need to incorporate the dependency in the models. An alternative approach would be to consider the study subjects to be coming from a mixture ot two populations, diseased and non-diseased, with two different prevalence parameters

References

Akaike, H. (1978). A new look at Bayes procedure. *Biometrka.* **65**: 53-59.

Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York. 632 pp.

Bernardo, J. M. (1979). Reference prior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc.* **41**: 113- 147.

Carlin, B. P. & Louis T. A. (1996). Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall, New York. 436 pp.

Held, L. & Ranyimbo, A. O. (2004). A Bayesian approach to estimate and validate the false negative fraction in a two-stage multiple screening test. *Methods of Information in Medicine.* **43**: 461-464.

Jeffreys, H. (1961). Theory of Probability (3rd edn.). Oxford University Press, London. 472 pp.

Lloyd, C. J. & Frommer, D. J. (2004). Estimating the false negative fraction for a multiple screen test for bowel cancer when the negatives are not verified. *Australian and New Zealand Journal of Statistics.* **46**: 531-542.

Novick, M. R. & Hall, W. J. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**: 1104-1117.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. Wiley, New York. 448 pp.

Zellner, A. (1977). Maximal data information prior distributions. *In*: Aykac, A. and Brumat, C. (Eds). New Methods in the Applications of Bayesian Methods. Proceedings of the CEDEP-INSEAD Conference, (Contribution to Economic Analysis Series) Vol. **119**, North-Holland, Amsterdam.pp. 211-232

Table 1. Tabulation of the count of subjects on the basis of the number of positive tests in the primary as well as the secondary phase.

	Secondary								
Primary	0	1	2	3	4	5	6	missed	total
1	10	3	3	2	2	1	2	14	37
2	3	2	2	1	4	1	1	8	22
3	4	1	5	3	4	1	2	5	25
4	1	1	0	3	4	1	4	15	29
5	3	1	1	4	4	3	6	12	34
6	1	0	1	3	3	5	16	20	49
Total	22	8	12	16	21	12	31	74	196

Table 2. Posterior estimates of the false negative fraction (FNF) under different priors.

	FNF estimate	95% Credible Interval
Prior I	0.269	[0.122, 0.717]
Prior II	0.265	[0.123, 0.626]
Prior III	0.272	[0.120, 0.944]

Table 3. Posterior estimate of the missed cases under different priors.

	Missed cases	95% Credible Interval
Prior I	73	[27, 298]
Prior II	71	[27, 329]
Prior III	74	[26, 3334]

Table 4 Observed and predicted number of individuals with counts of secondary positives conditional on primary positives. Predictions based on beta-binomial model with Prior I.

		Secondary positives						
Primary positives		0	1	2	3	4	5	6
1	Observed	10.00	3.00	3.00	2.00	2.00	1.00	2.00
	predicted	8.18	6.72	4.32	2.32	1.04	0.34	0.07
2	Observed	3.00	2.00	2.00	1.00	4.00	1.00	1.00
	predicted	2.14	3.30	3.33	2.62	1.64	0.76	0.21
3	Observed	4.00	1.00	5.00	3.00	4.00	1.00	4.00
	predicted	1.14	2.75	4.06	4.57	3.94	2.59	1.03
4	Observed	1.00	1.00	0.00	3.00	4.00	1.00	4.00
	predicted	0.23	0.84	1.75	2.71	3.33	3.18	1.97
5	Observed	3.00	1.00	1.00	4.00	4.00	3.00	6.00
	predicted	0.08	0.38	1.10	2.42	4.32	6.43	7.28
6	Observed	1.00	0.00	1.00	3.00	3.00	5.00	16.00
	predicted.	0.01	0.06	0.24	0.75	2.09	5.69	20.16
Total	Observed	22.00	8.00	12.00	16.00	21.00	12.00	33.00
	predicted	11.78	14.05	14.80	15.39	16.36	18.99	30.72

Table 5 Observed and predicted number of individuals with counts of secondary positives conditional on primary positives. Predictions based on beta-binomial model with Prior II.

		Secondary positives						
Primary positives		0	1	2	3	4	5	6
1	Observed	10.00	3.00	3.00	2.00	2.00	1.00	2.00
	predicted	8.15	6.73	4.33	2.33	1.03	0.34	0.07
2	Observed	3.00	2.00	2.00	1.00	4.00	1.00	1.00
	predicted	2.14	3.30	3.33	2.62	1.64	0.76	0.21
3	Observed	4.00	1.00	5.00	3.00	4.00	1.00	4.00
	predicted	1.13	2.75	4.06	4.50	3.94	2.58	1.03
4	Observed	1.00	1.00	0.00	3.00	4.00	1.00	4.00
	predicted	0.24	0.84	1.74	2.71	3.33	3.18	1.97
5	Observed	3.00	1.00	1.00	4.00	4.00	3.00	6.00
	predicted	0.08	0.38	1.10	2.42	4.32	6.43	7.28
6	Observed	1.00	0.00	1.00	3.00	3.00	5.00	16.00
	predicted.	0.01	0.06	0.24	0.75	2.10	5.71	20.13
total	Observed	22.00	8.00	12.00	16.00	21.00	12.00	33.00
	predicted	11.75	14.06	14.80	15.33	16.36	19.00	30.69

Table 6 Observed and predicted number of individuals with counts of secondary positives conditional on primary positives. Predictions based on beta-binomial model with Prior III.

		Secondary positives						
Primary positives		0	1	2	3	4	5	6
1	Observed	10.00	3.00	3.00	2.00	2.00	1.00	2.00
	predicted	8.27	6.68	4.29	2.32	1.03	0.35	0.07
2	Observed	3.00	2.00	2.00	1.00	4.00	1.00	1.00
	predicted	2.17	3.30	3.32	2.61	1.63	0.76	0.21
3	Observed	4.00	1.00	5.00	3.00	4.00	1.00	4.00
	predicted	1.15	2.76	4.06	4.50	3.93	2.58	1.03
4	Observed	1.00	1.00	0.00	3.00	4.00	1.00	4.00
	predicted	0.24	0.84	1.75	2.71	3.33	3.18	1.97
5	Observed	3.00	1.00	1.00	4.00	4.00	3.00	6.00
	predicted	0.08	0.38	1.11	2.42	4.32	6.43	7.27
6	Observed	1.00	0.00	1.00	3.00	3.00	5.00	16.00
	predicted.	0.01	0.06	0.24	0.76	2.10	5.66	20.17
total	Observed	22.00	8.00	12.00	16.00	21.00	12.00	33.00
	predicted	11.92	14.02	14.77	15.35	16.34	18.96	29.72