

STATISTICAL ASSESSMENT OF JOINTLY OBSERVED SCREENING TESTS

MARLOS A. G. VIANA AND CARLOS A. DE B. PEREIRA

ABSTRACT. In this article, Lindley and Novick criteria of screening usefulness is applied to the statistical assessment of jointly observed screening test. Posterior probabilities comparing screening sensitivities and specificities, and posterior probability bounds to comparing screening predictive values are obtained.

1. INTRODUCTION

Screening tests are aimed to identify individuals who are carriers of certain physiologic conditions such as the presence of a genetic biomarker that may identify subjects at increased cancer risk or the presence of an enzyme associated with acute myocardial infarction. The clinical relevance of a screening algorithm is a consequence of its ability to properly alter the management of the patient and of its usefulness in anticipating the progress of the physiologic process.

In this article we consider the experimental design in which individuals are jointly screened by two competing binary tests (generally called T_1 and T_2) and by a reference or gold standard assessment rule D which determines (by definition) the presence or absence of the physiologic condition of interest. For example, when screening for liver metastasis the reference test may be an autopsy or any other invasive test such as a liver biopsy, whereas a liver scan and bilirubin assay are typical screening (non-invasive) tests (Lind and Singer 1986). When screening for myocardial infarction, the reference test may be the serial creatine kinase-MB enzyme assay, whereas a single enzyme assay and a specific electrocardiographic response are competing screening tests. The question of interest is the statistical comparison of the two competing screening tests T_1 and T_2 . In the next section we introduce the necessary notation and assumptions, followed by the a descriptive assessment of the competing tests. Section 3 describes the inferential methods required to statistically compare the two screening tests. In Section 4 we apply these methods to assess the relative performance of a rapid optical immunoassay and a non-selective medium for the detection of lower genital tract colonization by group B streptococcus (GBS) using a selective broth enhanced culture as the reference or gold standard test.

Statistical methods to assess the usefulness (Lindley and Novick criteria) of independently observed screening tests have been described previously by Viana and Farewell (1990, 1994). Recent applications to assess the screening usefulness of specific genotypes for breast and lung cancer are described in Rebbeck, Viana, Jordan, Weber and Rogatko (1996) and Rebbeck, Rogatko and Viana (1998).

2. MODELS AND ASSUMPTIONS

Let

$$p_{ijk} = P(D = i, T_1 = j, T_2 = k), \quad i, j, k \in \{0, 1\},$$

indicate the underlying multinomial probabilities associated with the random joint outcome ($D = i, T_1 = j, T_2 = k$). A screening outcome $T = 1$ suggests the presence of the condition of interest and the outcome $T = 0$ suggests its absence. In contrast, the reference outcome $D = 1$ determines the presence of the

Date: Electronic posting of August 2, 2004.

1991 Mathematics Subject Classification. 62F15, 62P10.

Key words and phrases. Posterior probabilities, Dirichlet distributions, Discrete data.

This is a reprint from the original article published in Biometrical Journal up to editorial changes.

condition of interest and the outcome $D = 0$ determines its absence. The corresponding observed frequencies are denoted by x_{ijk} and the total number of joint observations is equal to N . Following standard notation, let

$$\pi = p_{1..} = p_{100} + p_{101} + p_{110} + p_{111} = 1 - p_{0..}$$

indicate the marginal probability of condition D present. The marginal probability of T_1 and T_2 suggesting the presence ($T=1$) of condition D are indicated by $p_{.1.}$ and $p_{.1.}$, respectively. The corresponding screening sensitivities are the conditional probabilities

$$\eta_1 = P(T_1 = 1 \mid D = 1) = \frac{p_{11.}}{p_{1..}}, \quad \eta_2 = P(T_2 = 1 \mid D = 1) = \frac{p_{1.1}}{p_{1..}}, \quad (2.1)$$

whereas the corresponding specificities are denoted by

$$\theta_1 = P(T_1 = 0 \mid D = 0) = \frac{p_{00.}}{p_{0..}}, \quad \theta_2 = P(T_2 = 0 \mid D = 0) = \frac{p_{0.0}}{p_{0..}}$$

Similarly, the predictive values of a screening suggestive of condition D are represented by

$$pvp_1 = P(D = 1 \mid T_1 = 1) = \frac{p_{11.}}{p_{.1.}}, \quad pvp_2 = P(D = 1 \mid T_2 = 1) = \frac{p_{1.1}}{p_{.1.}}$$

whereas the predictive values of non-suggestive screening results are

$$pvn_1 = P(D = 0 \mid T_1 = 0) = \frac{p_{00.}}{p_{.0.}}, \quad pvn_2 = P(D = 0 \mid T_2 = 0) = \frac{p_{0.0}}{p_{..0}}$$

2.1. Screening Usefulness. The equations relating sensitivity and specificity to pvp and pvn results from the usual Bayes formula [e.g., Ingelfinger, Mosteller, Thibodeau and Ware (1987) , Press (1989), Lee (1989)], and for that purpose it is convenient to express the resulting probabilities in terms of the corresponding odds. The posterior odds \mathcal{O}_1 on the presence of condition D when a suggestive ($T=1$) screening result is obtained is

$$\mathcal{O}_1 = \frac{\eta}{1 - \theta} \frac{\pi}{1 - \pi},$$

which shows the dependence of \mathcal{O}_1 on the (prior) odds $\frac{\pi}{1 - \pi}$ on the presence of condition D . Its relation to the predictive value of a suggestive test result is given by the equation $pvp = \mathcal{O}_1 / (1 + \mathcal{O}_1)$. Similarly, the posterior odds \mathcal{O}_0 on the presence of condition D when a non-suggestive screening result is observed is

$$\mathcal{O}_0 = \frac{1 - \eta}{\theta} \frac{\pi}{1 - \pi}.$$

The predictive value of a negative result is obtained from the relation $pvn = 1 / (1 + \mathcal{O}_0)$.

Lindley and Novick (1981) argue that a screening test is useless until it is likely to properly change the actions associated with the odds \mathcal{O}_1 on the presence of condition D when a suggestive screening result is observed, relative to the odds \mathcal{O}_0 on $D = 1$ when a non-suggestive test result is observed. A comparison is therefore, made between \mathcal{O}_1 and \mathcal{O}_0 at any given value of the marginal probability for condition $D = 1$, or prevalence π . Since the odds of 1 to 1 is equivalent to 50% in probability, a value of $\mathcal{O}_1 > 1$ corresponds to more than 50% in probability and is necessary to suggest a sequence of interventions seeking to confirm the presence of condition D ; conversely, a value of $\mathcal{O}_0 < 1$ is necessary to suggest an opposing course of action. As pointed out by Sox (1990), p.28, a screening test should be obtained only when its outcome is likely to properly alter the management of the patient. In this sense, a screening process with given sensitivity and specificity is properly useful at those prevalence levels π such that

$$\mathcal{O}_0 \leq f_1 < 1 < f_2 \leq \mathcal{O}_1 \quad (2.2)$$

holds true. Values of f_1 and f_2 may be selected to obtain more stringent criteria of usefulness (Viana and Farewell 1994). For descriptive purposes, it is graphically convenient to express the criterion in terms of the (natural) logarithm of the posterior odds. The interval R of prevalence levels for which (2.2) obtains is called the proper range of the screening test. To express R , let

$$\lambda_0 = \frac{\theta}{1 - \eta}, \quad \lambda_1 = \frac{1 - \theta}{\eta}$$

indicate the likelihoods of condition absent ($D=0$) relative to condition present ($D=1$) at each test outcome ($T=0, T=1$), and define

$$U = \min \{ \lambda_0, \lambda_1 \}, \quad V = \max \{ \lambda_0, \lambda_1 \}.$$

Then, the test's proper region is given by

$$R = \left(\frac{U}{U+1}, \frac{V}{V+1} \right). \quad (2.3)$$

Note that the $\eta + \theta > 1$ is sufficient to determine $U = \lambda_1$ and $V = \lambda_1$, and consequently,

$$R = \left(\frac{\lambda_1}{\lambda_1+1}, \frac{\lambda_0}{\lambda_0+1} \right). \quad (2.4)$$

The additional notation has the advantage of suggesting the corresponding definition of the proper region when more than two test responses are necessary to represent the screening method (Viana, Rogatko and Rebbeck 1998). Also note that the condition $\eta + \theta > 1$ imposes a natural restriction to the underlying probability describing the association between test outcome and disease condition (further discussed in Section 5). When $\eta + \theta > 1$ we say that the test has positive dependence.

In addition to describing the range R of prevalence levels in which the process is expected to properly screen the individuals, it is of interest to assess the test's log relative likelihood ratio H ,

$$H = \ln \frac{V}{U}, \quad (2.5)$$

which equals the height between the two (parallel) utility curves $\ln \mathcal{O}_1(\pi)$ and $\ln \mathcal{O}_0(\pi)$. Under positive dependence, however, $H = \ln \frac{\lambda_0}{\lambda_1}$. Moreover, H does not depend on the prevalence π and expresses the relative weighting of evidence intrinsic to the screening device (see also Kass and Raftery (1995)).

Any two tests' log relative likelihood ratios H_1 and H_2 are monotonically related with their proper regions R_1 and R_2 in the sense that

$$R_2 \subseteq R_1 \Rightarrow H_2 \leq H_1.$$

The converse of (2.1) is not true. In fact, take any two proper regions R_1 and R_2 such that neither $R_1 \subseteq R_2$ or $R_2 \subseteq R_1$. Then, because either $H_1 \leq H_2$ or $H_2 \leq H_1$ (the real line together with \leq is a completely ordered set), the converse fails for all such regions R_1 and R_2 .

When $R_2 \subseteq R_1$ we say that tests T_1 and T_2 are comparable and that T_1 is uniformly better than T_2 . It follows directly from definition (2.3) that a sufficient condition for $R_2 \subseteq R_1$ under positive dependence ($\eta_2 + \theta_2 > 1$) of Test 2 is $\eta_1 \geq \eta_2, \theta_1 \geq \theta_2$.

Lindley and Novick's criterion is also applicable to the sensitivity and specificity of a clinical trial [e.g., Sacks, Chalmers and Smith (1983)] and to the analysis of case-control studies [e.g., Marshall (1988), Zelen and Parker (1986), Schlesselman (1982), Hallstrom and Trobaugh (1985)].

3. STATISTICAL ASSESSMENT

Let $D(y)$ indicate the posterior Dirichlet probabilities given the data x and prior Dirichlet probabilities $D(\alpha)$. We recognize that the components of y , x and α are related by $y_{ijk} = x_{ijk} + \alpha_{ijk}$.

Proposition 3.1. Given the data,

$$\begin{aligned} P(\eta_1 \geq \eta_2) &= P(\text{Be}(y_{110}, y_{101}) \geq \frac{1}{2}), \\ P(\theta_1 \geq \theta_2) &= P(\text{Be}(y_{010}, y_{001}) \leq \frac{1}{2}), \end{aligned}$$

where $\text{Be}(e, f)$ indicates a Beta random variable with parameters e and f . Moreover, the events $\eta_1 \geq \eta_2$ and $\theta_1 \geq \theta_2$ are independent.

Proof. From the definitions in Section 2 it follows that $\eta_1 \geq \eta_2$ is equivalent to $p_{110} \geq p_{101}$. However,

$$p_{110} \geq p_{101} \leftrightarrow \frac{p_{110}}{p_{101}} \geq 1. \quad (3.1)$$

Given the data, the posterior distribution of

$$\frac{p_{110}}{p_{110} + p_{101}}$$

is Beta with parameters y_{110} and y_{101} . Consequently, the probability of $\eta_1 \geq \eta_2$ is the probability of $\text{Be}(y_{110}, y_{101}) \geq \frac{1}{2}$, as proposed. Similarly, $\theta_1 \geq \theta_2$ is equivalent to $p_{010} \leq p_{001}$, and the proposed result follows from the fact that

$$p_{010} \leq p_{001} \leftrightarrow \frac{p_{010}}{p_{001}} \leq 1. \quad (3.2)$$

The independence of the events $\eta_1 \geq \eta_2$ and $\theta_1 \geq \theta_2$ is obtained from the additional fact that the ratios

$$\frac{p_{110}}{p_{101}}, \quad \frac{p_{010}}{p_{001}}$$

in equations (3.1) and (2.2) above are ratios of independent gamma random variables (representing the underlying Dirichlet model) and consequently are also independent. \square

Proposition 3.2. Given the data,

$$P(\text{pvp}_1 \geq \text{pvp}_2) \geq P(\eta_1 \geq \eta_2)P(\theta_1 \geq \theta_2).$$

Proof. Let $\mathcal{O}(t) = t/(1-t)$ indicate the odds transformation, so that

$$\mathcal{O}(\text{pvp}_1) = \frac{p_{110} + p_{111}}{p_{010} + p_{011}}, \quad \mathcal{O}(\text{pvp}_2) = \frac{p_{101} + p_{111}}{p_{001} + p_{011}}.$$

Therefore, because

$$\eta_1 \geq \eta_2 \leftrightarrow p_{010} \leq p_{001}, \quad \theta_1 \geq \theta_2 \leftrightarrow p_{110} \geq p_{101},$$

it follows that

$$\eta_1 \geq \eta_2 \text{ and } \theta_1 \geq \theta_2 \rightarrow \mathcal{O}(\text{pvp}_1) \geq \mathcal{O}(\text{pvp}_2),$$

so that the joint event $E = \{\eta_1 \geq \eta_2, \theta_1 \geq \theta_2\}$ implies the event $F = \{\text{pvp}_1 \geq \text{pvp}_2\}$. Therefore $P(E) \leq P(F)$ and from the independence part of Proposition 3.1, the proposed result obtains. \square

Note that the events E and F are not equivalent. In fact, take $p_{110} = 0.1$, $p_{111} = p_{010} = 0.2$, $p_{011} = 0.01$, $p_{101} = 0.3$ and $p_{001} = 0.4$. Then $\mathcal{O}(\text{pvp}_1) = 1.42 > 1.21 = \mathcal{O}(\text{pvp}_2)$, whereas $\theta_1 > \theta_2$ and $\eta_1 < \eta_2$.

Proposition 3.3. Given the data,

$$P(\text{pvn}_1 \geq \text{pvn}_2) \geq P(\eta_1 \geq \eta_2)P(\theta_1 \geq \theta_2).$$

Proof. This is similar to the proof of Proposition 3.2 and follows from the fact that

$$\mathcal{O}(\text{pvn}_1) = \frac{p_{001} + p_{000}}{p_{100} + p_{101}}, \quad \mathcal{O}(\text{pvn}_2) = \frac{p_{010} + p_{000}}{p_{100} + p_{110}}.$$

\square

As a consequence of Propositions 3.2 and 3.3, we obtain

$$\min \{P(\text{pvp}_1 \geq \text{pvp}_2), P(\text{pvn}_1 \geq \text{pvn}_2)\} \geq P(\eta_1 \geq \eta_2)P(\theta_1 \geq \theta_2).$$

Proposition 3.4. Given the data, test T_1 has positive dependence with posterior probability

$$P(\text{Be}(y_{00.}, y_{01.}) \geq \text{Be}(y_{10.}, y_{11.})),$$

based on independent Beta distributions. Similarly, test T_2 has positive dependence with posterior probability

$$P(\text{Be}(y_{0.0}, y_{0.1}) \geq \text{Be}(y_{1.0}, y_{1.1})).$$

Proof. By definition, test T_1 has positive dependence when $\eta_1 + \theta_1 \geq 1$, that is

$$\frac{\theta_1}{1 - \eta_1} \geq 1 \iff \frac{p_{00.}}{p_{0..}} \geq \frac{p_{10.}}{p_{1..}}.$$

Similarly, T_2 has positive dependence when $\eta_2 + \theta_2 \geq 1$, or

$$\frac{\theta_2}{1 - \eta_2} \geq 1 \iff \frac{p_{0.0}}{p_{0..}} \geq \frac{p_{1.0}}{p_{1..}}.$$

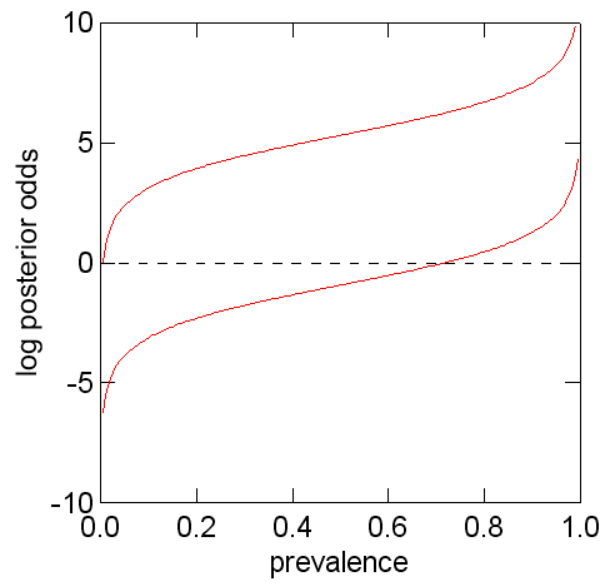
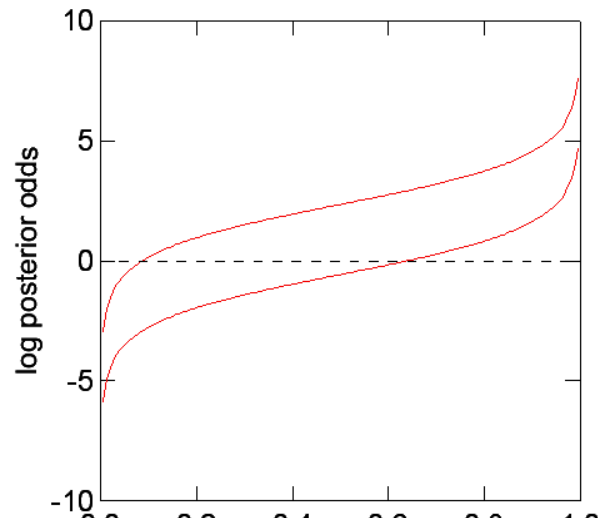
The proposed result follows by obtaining the required posterior marginal Beta distributions and the fact that the corresponding sides of the inequalities above are independently distributed. \square

4. DETECTION OF GROUP B STREPTOCOCCUS - AN EXAMPLE

The study reported by Nguyen, Gauthier, Myles, Viana and Schreckenberger (1998) considered the relative performance of a rapid optical immunoassay (shortly indicated here by quick strep - QS, or Test 1) and a non-selective medium (Trypticase Soy Agar - shortly TSA, or Test 2) for the detection of lower genital tract colonization by group B streptococcus (GBS) using a selective broth (Lim broth) enhanced culture as the gold standard. The assessment of lower genital tract colonization with GBS is indicated, for example, when there is a history of previous child with GBS neonatal sepsis. Infants born to GBS-colonized parturient could develop early-onset GBS disease, a leading cause of neonatal morbidity and mortality. Prompt intrapartum treatment can significantly decrease serious GBS-related sequelae. The observed data consist of 513 women enrolled in the year-long study. Each subject was screened by QS and TSA, in addition to the Lim broth culture. Table 1 summarizes the observed joint data and descriptive estimates of screening sensitivities, specificities and predictive values.

TABLE 1. Observed joint frequencies x_{ijk} of QS (T_1), TSA (T_2) and Lim broth culture (D) and corresponding screening sensitivities, specificities and predictive values.

culture=i	QS=j	TSA=k	x_{ijk}
1	0	0	29
1	0	1	19
1	1	0	6
1	1	1	35
0	0	0	416
0	0	1	0
0	1	0	18
0	1	1	1
sensitivity	0.460	0.606	
specificity	0.956	0.997	
pvp	0.683	0.981	
pvn	0.896	0.925	



Figures 1 and 2 show the utility curves for the two competing tests. The wider proper region of the TSA method fully includes the proper region of the QS method, thus suggesting the superiority of the TSA method. The proper region of TSA (Test 2), following (2.3), ranges from 0.048 to 0.716, whereas the proper region of the QS (Test 1) ranges from 0.087 to 0.639. The corresponding log likelihood ratios are $H_2 = 6.24 \geq H_1 = 2.91$, thus reflecting the fact that $R_1 \subseteq R_2$. The statistical assessment of the two screening test includes the evaluation of the posterior probability associated with the event $R_1 \subseteq R_2$. We adopt a prior Dirichlet model with $\alpha_{ijk} = 1/4$, which weights the same total information as a uniform beta prior model. First note, following Proposition 3.4 with $x_{00.} = 416, x_{01.} = 19, x_{10.} = 48, x_{11.} = 41$, that Test 1 has positive dependence with posterior probability very close to one (about 0.9999, using expression (A.1) in the Appendix). Under positive dependence, from definition (2.3), a sufficient condition for $R_1 \subseteq R_2$ is $\eta_2 \geq \eta_1, \theta_2 \geq \theta_1$. From Table 1 and Proposition 3.1 we obtain (again using expression (A.1) in the Appendix)

$$P(\eta_2 \geq \eta_1) = 0.9963, \quad P(\theta_2 \geq \theta_1) > 0.9999,$$

so that $P(R_1 \subseteq R_2) \geq 0.996$. Consequently, the TSA screening method is uniformly more useful than the QS screening method with posterior probability greater than 0.99. From Proposition 3.2 we conclude that the same lower bound is valid for $P(\text{pvp}_2 \geq \text{pvp}_1)$ and for $P(\text{pvn}_2 \geq \text{pvn}_1)$.

5. DISCUSSION

In this paper we considered the statistical assessment of jointly observed screening tests. The underlying multinomial probability model determines the dependence structure among the observations. We have shown that the comparison between the sensitivity of the two competing test (e.g., Proposition 3.1) depends essentially on assessing the symmetry of the joint distribution of the two competing tests when the condition is present ($D=1$). This is the usual condition (marginal homogeneity) found when assessing dependent proportions. The classical large-sample solution is McNemar's test to assess the sharp hypothesis equivalent to $\eta_1 = \eta_2$ (e.g., Agresti (1990, p. 348)). The same comment applies to the comparison of the specificity parameters of the two tests. A lower probability bound for the relative assessment of the corresponding predictive values is the product of the two posterior probabilities for the sensitivity and specificity parameters (Proposition 3.2). The clinical usefulness of a screening test is determined by its proper region R . When $R_2 \subseteq R_1$ we say that tests T_1 and T_2 are comparable and that T_1 is uniformly more useful than T_2 . We argued that a sufficient condition for $R_2 \subseteq R_1$ under positive dependence ($\eta_2 + \theta_2 > 1$) of Test 2 is $\eta_1 \geq \eta_2, \theta_1 \geq \theta_2$. The statistical assessment of the condition is provided by Proposition 3.1, which also provides for a lower probability bound for the relative screening usefulness.

APPENDIX A

When the probability density function of U is Beta with parameters α, α' and the density of V is Beta with parameters β, β' , independent of U , then

$$P(U \leq V) = \frac{1}{B(\beta, \beta')} \sum_{j=\alpha}^{\alpha+\alpha'-1} \binom{\alpha+\alpha'-1}{j} B(\beta+j, \alpha+\alpha'+\beta'-j-1), \quad (\text{A.1})$$

where $B(c, c') = \Gamma(c)\Gamma(c')/\Gamma(c+c')$ for non-negative real numbers c and c' .

REFERENCES

- Agresti, A. (1990), *Categorical Data Analysis*, Wiley, New York.
- Hallstrom, A. and Trobaugh, G. (1985), 'Specificity, sensitivity, and prevalence in the design of randomized trials: A univariate analysis', *Controlled Clinical Trials* **6**, 128–135.
- Ingelfinger, J., Mosteller, F., Thibodeau, L. and Ware, J. (1987), *Biostatistics in Clinical Medicine*, MacMillan Publishing Co., Inc., New York.
- Kass, R. E. and Raftery, A. (1995), 'Bayes factors', *Journal of the American Statistical Society* **90**(430), 773–795.
- Lee, P. M. (1989), *Bayesian Statistics: An Introduction*, Oxford University Press, New York.
- Lind, S. E. and Singer, D. E. (1986), 'Diagnosing liver metastases: A bayesian analysis', *Journal of Clinical Oncology* **4**(3), 379–88.
- Lindley, D. and Novick, M. (1981), 'The role of exchangeability in inference', *The Annals of Statistics* **9**(1), 45–58.
- Marshall, R. (1988), 'Bayesian analysis of case-control studies', *Statistics in Medicine* **12**(7), 1223–1230.
- Nguyen, T. M., Gauthier, D., Myles, T., Viana, M. and Schreckenberger, P. (1998), 'Detection of group b streptococcus: Comparison of an optical immunoassay with direct plating and broth-enhanced culture methods', *The Journal of Maternal-Fetal Medicine* **7**, 172–176.
- Press, S. J. (1989), *Bayesian Statistics*, John Wiley, New York.
- Rebbeck, T. R., Viana, M., Jordan, H. A., Weber, B. L. and Rogatko, A. (1996), 'Evaluation of susceptibility genes in disease risk assessment', *American Journal of Human Genetics* **59**(4), A28.
- Rebbeck, T., Rogatko, A. and Viana, M. (1998), 'Evaluation of genotype data in clinical risk assessment: Methods and applications to brca1, brca2, and n-acetyl transferase-2 genotypes in breast cancer', *Genetic Testing* (3), 157–164.
- Sacks, H., Chalmers, T. and Smith, H. (1983), 'Sensitivity and specificity of clinical trials', *Archives of Internal Medicine* **143**, 753–755.
- Schlesselman, J. (1982), *Case-Control Studies: Design, Conduct, Analysis*, Oxford University Press, New York.
- Sox, H. C., ed. (1990), *Common Diagnostic Tests Use and Interpretation*, American College of Physicians, Philadelphia.
- Viana, M. A. G. and Farewell, V. (1990), 'A test for diagnostic utility', *Canadian Journal of Statistics* **18**(4), 289–295.
- Viana, M. A. G. and Farewell, V. (1994), 'Assessing the diagnostic utility of a test', *Biometrical Journal* **36**(2), 131,145.
- Viana, M. A. G., Rogatko, A. and Rebbeck, T. (1998), 'Statistical assessment of multi-valued diagnostic tests', *Canadian Journal of Statistics* **26**(4), 657–668.
- Zelen, M. and Parker, R. (1986), 'Case-control studies and Bayesian inference', *Statistics in Medicine* **5**, 261–269.