Thailand Statistician
July 2010; 8(2) : 157-166
http://statassoc.or.th
Contributed paper

Adjusted Estimator of the Sum of Misclassification Errors of Youden’s Index in Sparse Data of a Diagnostic Study

Kamonwan Soonklang [a], Chukiat Viwatwongkasem* [a], Pratana Satitvipawee [a], Rujirek Busarawong [b], Ramidha Srihera [c]

[a] Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok 10400, Thailand.

[b] Department of Applied Statistics, Faculty of Science, King Mongkut’s Institute of Technology, Ladkrabang, Bangkok 10520, Thailand.

[c] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Klongluang, Pathum Thani 12121, Thailand.

***Author for correspondence**; e-mail: phcvw@mahidol.ac.th

Received: 14 February 2010
Accepted: 11 June 2010

Abstract

Youden’s index as a common measure of the accuracy of diagnostic test is defined by $\mathit{sensitivity} + \mathit{specificity} - 1$. In estimating the sum of two misclassification errors of Youden’s index, the conventional estimator, defined by $\hat{\lambda} = \hat{\alpha} + \hat{\beta} = (x_D / n_D) + (x_H / n_H)$ where $\hat{\alpha}$ is an error estimate of false negative, $\hat{\beta}$ is a false positive error estimate, x_D is the frequency of (falsely) negatively classified persons out of n_D diseased groups, and x_H is the frequency of (falsely) positively classified persons out of n_H healthy ones, may have a considerable problem of zero variance in sparse data. The simple way to solve this problem is to add the constants c_D and c_H in the form of $\hat{\lambda}_c = \hat{\alpha}_{c_D} + \hat{\beta}_{c_H} = (x_D + c_D)/(n_D + 2c_D) + (x_H + c_H)/(n_H + 2c_H)$. The minimum Bayes risk approach is proposed in order to find the optimum points of c_D and c_H. Under each arm of prior errors ranged between 0 to 0.25, the optimal value of c_D and c_H equals 5/14. The simulation techniques are provided to confirm that the simple adjusted estimator, $\hat{\lambda}_c$, has the best performance with the smallest average mean square errors.

Keywords: Diagnostic test, misclassification errors, Youden's Index, zero variance correction.

1. Introduction

Diagnostic tests are vital in medical care and play a significant role in health care costs [1]. A diagnostic test has two purposes, i.e. to give reliable information about a patient's condition and to help the health care providers plan on how to manage the patients [2]. Diagnostic accuracy is usually characterized by the sensitivity ($1-\alpha =$ probability of positive tests given diseased persons) and the specificity ($1-\beta =$ probability of negative tests given non-diseased persons) [3]. They are closely related to the concepts of type I error ($\alpha =$ false negative rate) and type II error ($\beta =$ false positive rate). Sensitivity or specificity alone doesn't tell us how well the test predicts. It is therefore useful to summarize the incorporation of sensitivity and specificity into a single index, for examples, odd ratio, receiver operating characteristic (ROC), and Youden's index. The Youden's index is defined as $sensitivity + specificity - 1$ with a maximum value of 1 and a minimum value of 0 [4]. Böhning et al. [5] are interested to use the sum of sensitivity and specificity; $(1-\alpha)+(1-\beta)$, or, equivalently, the sum of the misclassification errors α and β. We note that $(1-\alpha)+(1-\beta)=1+J$, where J is Youden's index. One of their motivations is to find the best cut-off value in the sense of maximizing the sum of misclassification errors under the meta-analysis studies that the cut-off values themselves are frequently not reported and often varies between studies. Figure 1 shows that the sum of sensitivity and specificity is suggested to diagnose since it can diminish the cut-off value problem; furthermore, it is fairly constant.

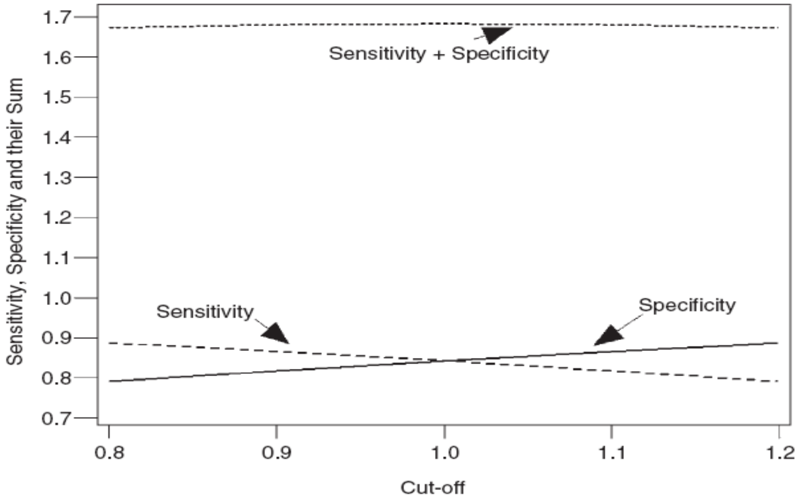

Figure 1. Sensitivity, specificity and their sum as a function of the cut-off value.

In sparse data, Agresti and Coull [6], Agresti and Caffo [7], Ghosh [8], Newcombe [9, 10], Böhning and Viwatwongkasem [11], and Viwatwongkasem et al. [12] indicated a problem of zero variance of the conventional estimators $\hat{\alpha}$ and $\hat{\beta}$. The $\hat{\alpha}$ is an error estimate of false negative, $\hat{\beta}$ is a false positive error estimate, where $\hat{\alpha} = x_D / n_D$ and $\hat{\beta} = x_H / n_H$, x_D is the frequency of (falsely) negatively classified persons out of n_D diseased ones, and x_H is the frequency of (falsely) positively classified persons out of n_H healthy ones. Both $\hat{\alpha}$ and $\hat{\beta}$ have a considerable problem of zero variance since the variance of $\hat{\alpha}$, obtaining by $\alpha(1-\alpha)/n_D$ which is estimated by $\hat{\alpha}(1-\hat{\alpha})/n_D$, equals 0 if $x_D = 0$ or $x_D = n_D$. Similarly, the zero variance can be occurred with $\hat{\beta} = x_H / n_H$. In order to solve the problem, the continuity correction constants, c_D and c_H, are often added to each cell of a 2×2 table. Table 1 shows the 2×2 table of observations with continuity corrections. In each true condition group, the class of parametric forms, namely $\hat{\alpha}_{cD} = \frac{x_D + c_D}{n_D + 2c_D}$ and $\hat{\beta}_{cH} = \frac{x_H + c_H}{n_H + 2c_H}$, is suggested in estimation for binomial parameter α and β respectively. Various choices of c_D and c_H are possible, leading to the main question of this paper to find the best value of c_D and c_H

to minimize the bias and/or the mean square error for the sum of misclassification errors of Youden’s statistics.

Table 1. The 2×2 table adds with continuity corrections.

True Condition	Test outcome		Total
	Positive	Negative	
Present (Diseased)	$n_D - x_D + c_D$	$x_D + c_D$	$n_D + 2c_D$
Absent (Healthy)	$x_H + c_H$	$n_H - x_H + c_H$	$n_H + 2c_H$
Total	$m_1 + c_D + c_H$	$m_0 + c_D + c_H$	$n_+ + 2c_D + 2c_H$

Conditions: $m_1 = n_D - x_D + x_H$, $m_0 = x_D + n_H - x_H$, $n_+ = n_D + n_H$

1. Estimating an Error of Misclassifications

Youden’s statistic is usually defined as

$$\hat{J} = \left(1-\hat{\alpha}\right)+\left(1-\hat{\beta}\right)-1 = \frac{\left(x_D - n_D\right)\left(x_H - n_H\right) - x_D x_H}{\left(x_D + \left(n_D - x_D\right)\right)\left(x_H + \left(n_H - x_H\right)\right)},$$

and its estimated variance is

$$\hat{V}\left(J\right) = \frac{x_D\left(n_D - x_D\right)}{\left(x_D + \left(n_D - x_D\right)\right)^3} + \frac{x_H\left(n_H - x_H\right)}{\left(x_H + \left(n_H - x_H\right)\right)^3}.$$

However, in this section, we are interested in estimating a misclassification error α (or β) of Youden’s index under the sparse data coping with continuity correcting terms. For a diseased group, the simple adjusted estimate defined by $\hat{\alpha}_{c_D} = \frac{x_D + c_D}{n_D + 2c_D}$ with correction term c_D is proposed for estimating a binomial parameter α . The expectation, bias, variance, and mean square error of $\hat{\alpha}_{c_D}$ can be found in the following formulae:

1. $E\left(\hat{\alpha}_{c_D}\right) = \frac{n_D\alpha + c_D}{n_D + 2c_D}$

2. $Bias\left(\hat{\alpha}_{c_D}\right) = \frac{c_D\left(1-2\alpha\right)}{n_D + 2c_D}$

3. $V\left(\hat{\alpha}_{c_D}\right) = \dfrac{n_D\alpha\left(1-\alpha\right)}{\left(n_D + 2c_D\right)^2}$

4. $MSE\left(\hat{\alpha}_{c_D}\right) = \dfrac{n_D\alpha\left(1-\alpha\right)}{\left(n_D + 2c_D\right)^2} + \left(\dfrac{c_D\left(1-2\alpha\right)}{n_D + 2c_D}\right)^2$

5. $V\left(\hat{\alpha}_{c_D}\right) \leq V\left(\dfrac{x_D}{n_D}\right) = \dfrac{\alpha\left(1-\alpha\right)}{n_D}$ if $c_D \geq 0$ **.**

Unfortunately, it is impossible to find the optimal point c_D such that $\hat{\alpha}_{c_D}$ has the smallest mean square error (MSE) for all values of α. The minimum point c_D is not a unique solution. The solution of c_D depends inversely on the values of α which is not practical with real situations. Therefore, an alternative method in which we are considered is the average MSE or the Bayes risk with respect to a uniform prior on $\left[0,a\right]$ where a is a maximum value of α. We suppose that the square error loss function is given by $Loss = \left(\hat{\alpha}_{c_D} - \alpha\right)^2$. The average squared error loss (or risk, or MSE of $\hat{\alpha}_{c_D}$ in this case) is given as $Risk = Var\left(\hat{\alpha}_{c_D}\right) + \left(Bias\left(\hat{\alpha}_{c_D}\right)\right)^2$. Given the prior uniform density, $g(\alpha) = 1/a$, over $\left[0,a\right]$; consequently, the Bayes risk of $\hat{\alpha}_{c_D}$ denoted by $m(c_D)$ with respect to the Euclidean loss function is

$$m(c_D) = \int_0^a MSE\left(\hat{\alpha}_{c_D}\right)g\left(\alpha\right)d\alpha = \frac{1}{a}\int_0^a \frac{n_D\alpha(1-\alpha) + c_D^2(1-2\alpha)^2}{(n_D + 2c_D)^2}\,d\alpha\,.$$

A straight computation of Bayes risk shows that

$$m\left(c_D\right) = \frac{2c_D^2\left(3-6a+4a^2\right) + an_D\left(3-2a\right)}{6\left(2c_D + n_D\right)^2}\,.$$

The first derivative of $m(c_D)$ is

$$\frac{d}{dc_D}m\left(c_D\right) = \frac{2c_D\left(3-6a+4a^2\right)}{3\left(2c_D+n_D\right)^2} - \frac{2\left(2c_D^2\left(3-6a+4a^2\right) + an_D\left(3-2a\right)\right)}{3\left(2c_D+n_D\right)^3}.$$

Setting $\dfrac{d}{dc_D}m\left(c_D\right) = 0$, we have $c_D = \dfrac{3a-2a^2}{3-6a+4a^2}$ as the solution of $m(c_D)$. We note that the c_D is a globally concave function of a with a maximum point at $a = 0.75$.

Usually, a false negative error α should not be greater than $a = 0.25$. This statement of the boundary of a ranged from 0 to 0.25 is supported by searching through the online biomedical literature database, PubMed, using Sensitivity and Specificity as keywords in Thailand 2009. It is revealed that most of studies (more than 70% out of 404 studies) have an upper limitation of $a = 0.25$. Hence, under the prior criterion for $\alpha \in [0, 0.25]$, the minimum point $c_D = 5/14$ can meet the Bayes risk function verifying minima with the condition of $\frac{d^2 m(5/14)}{dc_D^2} > 0$. Figure 2 shows that the average mean square errors, $m(c_D)$, have a locally minimum point at $c_D = 5/14$ for various values of n_D.

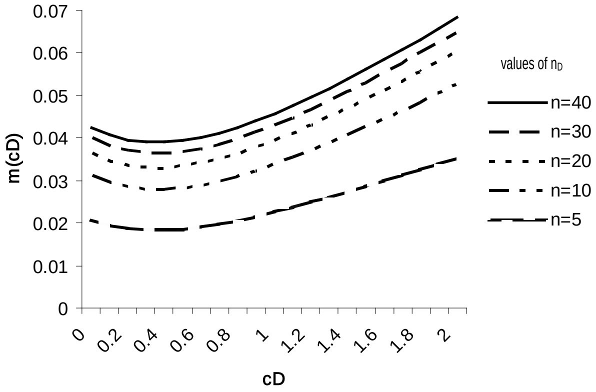

Figure 2. $m(c_D)$ as a function of c_D for values of n_D = 5, 10, 20, 30, 40, respectively.

2. Estimators of the Sum of Misclassification Errors

According to the assumption that diseased and healthy groups are independence, one can write the conventional estimate and its variance estimate for estimating the sum of misclassification errors $\lambda = \alpha + \beta$ as follows:

$$\hat{\lambda} = \hat{\alpha} + \hat{\beta} = \frac{x_D}{n_D} + \frac{x_H}{n_H}$$

$$\hat{V}(\hat{\lambda}) = \hat{V}\left(\hat{\alpha} + \hat{\beta}\right) = \frac{\hat{\alpha}(1-\hat{\alpha})}{n_D} + \frac{\hat{\beta}\left(1-\hat{\beta}\right)}{n_H}.$$

The proposed estimate which minimizes the Bayes risk with respect to the prior of $\alpha \in [0, 0.25]$ is

$$\hat{\lambda}_c = \hat{\alpha}_{c_D} + \hat{\beta}_{c_H} = \frac{x_D + c_D}{n_D + 2c_D} + \frac{x_H + c_H}{n_H + 2c_H} = \frac{x_D + 5/14}{n_D + 5/7} + \frac{x_H + 5/14}{n_H + 5/7}$$

and the variance estimate for the sum of misclassification errors is

$$\hat{V}(\hat{\lambda}_c) = \hat{V}\left(\hat{\alpha}_{c_D} + \hat{\beta}_{c_H}\right) = \frac{n_D \hat{\alpha}_{c_D}\left(1 - \hat{\alpha}_{c_D}\right)}{\left(n_D + 5/7\right)^2} + \frac{n_H \hat{\beta}_{c_H}\left(1 - \hat{\beta}_{c_H}\right)}{\left(n_H + 5/7\right)^2}$$

Indeed, the conventional estimator is a shrinkage form of the proposed estimator when *c_D* = 0 and *c_H* = 0. An alternative choice of c_D = 1 and c_H = 1, based on the Bayes risk with prior uniform over [0, 1] suggested by Viwatwongkasem et al. [12] in the context of proportion risk, leads to the candidate estimate and its variance estimate in the following:

$$\hat{\lambda}_c = \hat{\alpha}_{c_D} + \hat{\beta}_{c_H} = \frac{(x_D + 1)}{(n_D + 2)} + \frac{(x_H + 1)}{(n_H + 2)}$$

$$\hat{V}(\hat{\lambda}_c) = \hat{V}\left(\hat{\alpha}_{c_D} + \hat{\beta}_{c_H}\right) = \frac{n_D \hat{\alpha}_{c_D}\left(1 - \hat{\alpha}_{c_D}\right)}{\left(n_D + 2\right)^2} + \frac{n_H \hat{\beta}_{c_H}\left(1 - \hat{\beta}_{c_H}\right)}{\left(n_H + 2\right)^2}.$$

3. A Simulation Study

To compare the performance of the proposed estimator (adjusting with c_D = 5/14 and c_H = 5/14) to the conventional estimator (adjusting with c_D = 0 and c_H = 0) and the choice of Viwatwongkasem et al. (adjusting with c_D = 1 and c_H = 1), the simulation plan is requested with the performance criterion of the smallest average mean square error. We proposed a simulation study in the following designs:

Parameters: Let the sum of misclassification error λ be some constants varying from 0.01 to 0.50 steps of 0.01. False positive error β is some constants varying from 0.001 to 0.250 in steps of 0.001. And we calculate α by $\alpha = \lambda - \beta$ where $\lambda > \beta$. The sample size in each arm is fixed and varied as 5, 7, 10, 20, 30, 40.

Statistics: Binomial random variable x_D in the disease group is generated with parameters (n_D, α) and binomial variable x_H in the non-disease group is generated with parameters (n_H, β). The procedure is replicated over 6,000 times.

4. Results

To evaluate the performance of estimators, we concentrate on the smallest average mean square error. Simulation results show that the proposed estimator (adjusting with c_D = 5/14 and c_H = 5/14) yields the best performance with the smallest average mean square error for every sample size. The average mean square error of the conventional estimator (adjusting with c_D = 0 and c_H = 0) is less than those of the choice of Viwatwongkasem et al. (adjusting with c_D = 1 and c_H = 1), especially for small sample sizes ($n_D \le 20$ and $n_H \le 20$). For moderate to large sample sizes ($n_D \ge 30$ and $n_H \ge 30$), all estimators yield the equality of performance with the similar results. This can be clearly demonstrated in Figure 3. The graph of the average mean square errors of the proposed estimator has the lowest line with the best performance for all sample sizes.

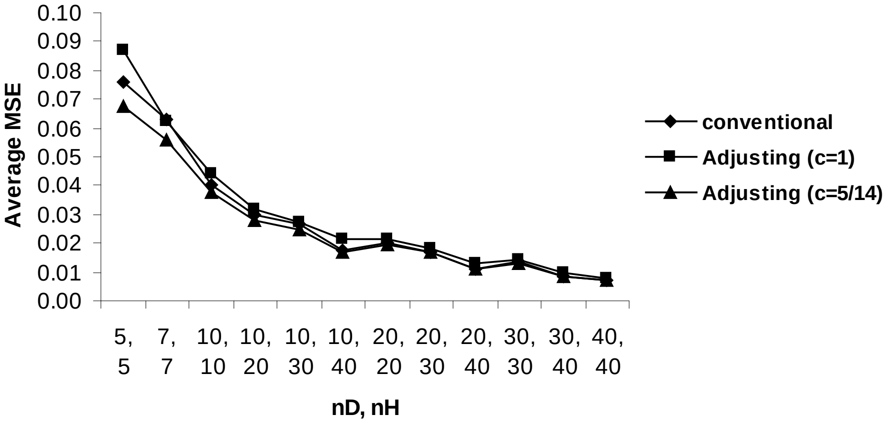

Figure 3. Average MSE of proposed estimator (adjusting with c_D = 5/14 and c_H = 5/14), conventional estimator (adjusting with c_D = 0 and c_H = 0) and the choice of Viwatwongkasem et al. (adjusting with c_D = 1 and c_H = 1) ($\lambda \in [0.00, 0.50]$).

5. Discussion and Conclusion

The problem of zero-variance of the conventional estimator of the sum of misclassification errors of Youden's index is arisen in sparse data of a diagnostic study. We are interested in solving this problem by adding some continuity correction constants (c_D and c_H) since it is easy to implement. Indeed, Sweeting et al. [14] proposed the alternative method of the reciprocal of the opposite error size to solve this problem to avoid the use of continuity corrections; however, it is not popular because of its complicated formulae. A simple way to find the optimum values of c_D and c_H is derived from the Bayes risk with prior uniform over [0, 0.25]. The smallest average mean square error yields the minimum when c_D = 5/14 and c_H= 5/14. The simulation plan is provided to confirm that the proposed estimator has the least average mean square error with the best performance comparing the conventional estimator (adjusting with c_D = 0 and c_H = 0) and the choice of Viwatwongkasem et al. (adjusting with c_D = 1 and c_H = 1). However, for moderate to large sample size ($n_D \geq$ 30 and $n_H \geq$ 30), all estimators are not different regarding the performance equality.

References

[1] Epstein, A.M., Begg, C.B., McNeil, B.J., The use of ambulatory testing in prepaid and free-for-service group practices. *The New England Journal of Medicine*, 1986; 314: 1089-94.

[2] Sox, Jr. HC., Blatt, M.A., Higgins, M.C., Marton, K.I., Medical decision making. Boston: Butterworths-Heinemann, 1989.

[3] Harper, R., Reeves, B., Reporting of precision of estimates for diagnostic accuracy: review. *British Medical Journal*, 1999; 318: 1322-23.

[4] Zhou, X.H., Nancy, A. Obuchowski, Donna, K., McClish, Statistical Methods in Diagnostic Medicine. America: John Wiley & Sons, 2002.

[5] Böhning, D., Böhning, W., Holling, H., Revisiting youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. *Statistical Methods in Medical Research*, 2007;17: 543-54.

[6] Agresti, A., Coull, B.A., Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 1998; 52: 119-26.

[7] Agresti, A., Caffo, B., Simple and Effective Confidence Interval for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician*, 2000; 54: 280-8.

[8] Ghosh, B.K., A Comparison of Some Approximate Confidence Interval for the Binomial Parameter. *Journal of the American Statistical Association*, 1979; 74: 894-900.

[9] Newcombe, R.G., Two-sided Confidence Intervals for a Single Proportion: Comparison of Seven Methods. *Statistics in Medicine,* 1998a; 17: 857-72.

[10] Newcombe, R.G., Interval Estimation forthe Difference between Independent Proportions: Comparison of Seven Methods. *Statistics in Medicine,* 1998a: 17: 873-90.

[11] Böhning, D., Viwatwongkasem, C., Revisiting proportion estimators. *Statistical Methods in Medical Research*, 2007; 17: 543-54.

[12] Viwatwongkasem, C., Bohning, D., Jitthavech, J., Vorapongsathorn, T., Satitvipawee, P., Srihera, P., Research Report of Minimum Mean Square Error Estimator Using the Simple Adjusted Proportion for Testing the Significance of a Common Effect Measure in Sparse Multi-Center Data. National Research Council of Thailand, 2005.

[13] Yates, F., Contingency Tables Involving Small Number and the Chi-squared Test. *Journal of the Royal Statistical Society 1 (Supplement)*, 1934:217-35.

[14] Sweeting M.J., Sutton, A.J., Lambert, P.C., What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine,* 2004; 23: 1351-75.