

Calculating Accuracy Rates from Multiple Assessors with Limited Information *

Richard TWEEDIE and Kerrie MENGERSEN

February 23, 1996

Abstract

In order to increase the accuracy of a particular diagnosis or similar type of binary decision, a common approach is to use multiple assessors. Through the use of Bayes' Theorem we illustrate how to compute accuracy rates with limited information under a median agreement method and show, perhaps counter-intuitively, that in some circumstances a seemingly greater accuracy among assessors actually implies a greater rate of misclassification. The approach is exemplified through an investigation of the accuracy of radiological classification of opacities of the lung associated with exposure to asbestos. This provides a good example of the need for care in defining the conditioning events involved in discussing "accuracy of diagnosis".

KEYWORDS Bayes Theorem, sensitivity, predictive values, diagnosis accuracy, asbestos, lung cancer, fibrosis, opacities, median agreement

*Richard Tweedie is Professor of Statistics, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA. Kerrie L. Mengersen is Senior Lecturer in Statistics, School of Mathematics, Queensland University of Technology, Brisbane, Australia. The authors are grateful to Richard Travers of Clayton Utz (Solicitors and Attorneys) for first bringing to our attention the medical questions that motivate this analysis, and to Sue Taylor for discussing the epidemiological relevance of the results.

1. INTRODUCTION

In many situations in which an expert diagnosis or other binary evaluation (for example, yes/no, absent/present, selected/unselected) is required, the question of accuracy of the diagnosis is clearly paramount.

Suppose, for example, that the presence of a particular disease in a patient is to be evaluated by a single assessor. Two types of misclassification may occur: a false positive (stating that the disease is present when in fact it is absent) or a false negative (stating that the disease is absent when in fact it is present).

With only one assessor it is of course not possible to estimate the probability of these adverse diagnoses without some independent assessment of the presence or absence of the disease. Moreover such independent evaluation may be difficult or even unethical to obtain: if a disease is diagnosed as absent, say, then it may not be reasonable to operate on patients just to verify clinically that the diagnosis was correct, although false positive rates may be easier to clinically establish.

In such cases, a common practice used to increase confidence is to employ multiple assessors and specify an overall decision criterion, such as requiring unanimous agreement on a diagnosis, or using the median of the assessors' evaluations. Although the arguments below can be carried over to a variety of situations, for focus in this paper we consider a team of three assessors, all equivalent in skill, who employ a median evaluation criterion: that is, if two or three assessors agree, that diagnosis is chosen as the overall or "median-based" diagnosis.

The contribution of this note is to illustrate how, through an application of Bayes' Theorem, information on various measures of accuracy may be retrieved from limited information. In doing so we show the somewhat counter-intuitive result that apparently high levels of accuracy for individual assessors, which on the face of it seem to imply a high degree of concordance, may in fact result in unacceptably low levels of accuracy for the median-based diagnosis.

2. FORMALISATION OF THE PROBLEM

Let F be the event that the disease (fibrosis in our example below) is present, and \bar{F} that it is absent; and let D and \bar{D} be the events of diagnosing presence and absence of the disease, respectively.

We will use p and $1 - p_f$ to denote respectively the "sensitivity" and "specificity" of the individual assessors, defined (Mausner and Bahn (1985) Chapter 7) by:

$$\begin{aligned} p &:= \text{Probability of true positive evaluation for an individual assessor} \\ &= \Pr [D|F] \end{aligned}$$

Note that from

$$\begin{aligned} 1 - p_f &:= \text{Probability of true negative evaluation for an individual assessor} \\ &= \Pr [\bar{D}|\bar{F}] \end{aligned}$$

these we get without other calculation that $1 - p = \text{Probability of false negative evaluation for an individual assessor} = \Pr [\bar{D}|F]$ and $p_f = \text{Probability of false positive evaluation for an individual assessor} = \Pr [D|\bar{F}]$.

Considerable care needs to be taken in distinguishing these (logically and in nomenclature) from the related positive and negative "predictive values" PV_{pos} and PV_{neg} (Mausner and Bahn (1985) Chapter 7), which in this paper focus on the final diagnosis arising from the median evaluation criterion:

$$\begin{aligned} PV_{pos} &= \text{Probability that a positive median evaluation is correct} \\ &= \Pr [F|D] \end{aligned}$$

Note that similar predic-

$$\begin{aligned} PV_{neg} &= \text{Probability that a negative median evaluation is correct} \\ &= \Pr [\bar{F}|\bar{D}] \end{aligned}$$

tive values could be calculated also for individual assessors, but this is not of great interest in this context.

We show in Section 4 that the positive and negative predictive values for the median test, as well as the specificity for the individual raters, may be retrieved with only three types of data that are commonly reported or available in such situations: the sensitivity parameter p for the individuals, estimated independently, perhaps from clinical studies, and

$$\hat{D} = \text{Proportion of observed positive diagnoses based on the median evaluation;}$$

$$p_i = \text{Proportion of positive diagnoses by rater } i \text{ on the declared positives, } \quad i = 1, 2, 3$$

Through this approach we avoid the need to have independent estimates of errors for negative evaluations, which as noted above are commonly unavailable. Moreover, even without the estimate of p it is possible to calculate values of p and p_f which indicate mutually consistent ranges for these misclassification probabilities, and also to indicate the predictive values of the median evaluation corresponding to these consistent values.

Before indicating how these calculations are formed, we motivate the problem through a specific occurrence in radiological assessments of lung fibrosis.

3. MOTIVATION: DETECTION OF RADIOGRAPHICALLY CLASSIFIED OPACITIES

Our interest in this problem arose when considering the role of radiologically-diagnosed fibrosis as an indicator of asbestos related lung cancer.

Fibrosis in the lungs can be diagnosed in a number of ways. Although histopathological (clinical) methods are most accurate, radiological evaluation is more often undertaken since it is less invasive. Hughes and Weill (1991) hypothesise that the presence of fibrosis, as detected radiologically, is a reliable indicator of asbestosis, and that this type of asbestosis is itself an indicator of asbestos-related lung cancer. In their study of workers in asbestos factories, Hughes and Weill found 71% of the "excess" lung cancers (above a population baseline) in subjects with positive radiological diagnoses for fibrosis: this supports their hypothesis, since it implies excess cancers are 2.4 times as likely to occur in the patients with radiologically determined fibrosis than in those without. If this method is valid it could give a very useful screening tool for cancers in those exposed to asbestos.

Of course, one question that must be clarified concerns the accuracy of methods used to detect fibrosis itself. In the extreme case, if *all* subjects were diagnosed as having fibrosis then all "excess" cancers would be in such subjects but little would have been achieved.

In detecting fibrosis radiologically, raters classify the presence and size of "small opacities" in the lung of the subject using, for example, guidelines set out by the International Labour Office (1980).

Fibrosis is considered to be “diagnosed” if the film is classified “ $\geq 1/0$ ” (which we shall call a *positive* diagnosis) and “not diagnosed” if the film is classified “ $< 1/0$ ” (which we shall call a *negative* diagnosis). Because of the difficulties of this radiological classification, Hughes and Weill (1991) employed three raters and based the final diagnosis on the median classification. (Other methods can be used: the decision criterion adopted by Lidell and McDonald (1980), for example, required six readers to agree before a film was accepted as “normal”.)

In the terminology above, the three raters used by Hughes and Weill (1991) had values of

$$p_1 = 0.9, p_2 = 0.86, p_3 = 0.82$$

which seems on the face of it to give a high degree of concordance. Not unreasonably, then, these levels of accuracy were given as a strength of the claims in Hughes and Weill (1991). However, as a consequence of the counter-intuitive situation described in Section 1, we show below that this is in fact rather optimistic: on the basis of the limited reported information the true predictive value of the median test may be poor enough to render doubtful the usefulness of the screening test.

4. PROBABILISTIC ANALYSIS

Suppose that there are N individuals diagnosed and that N_p and N_n represent the number of true positives and true negatives.

Using the median evaluation criterion, we can calculate the probabilities

$$p^* := \text{Probability of being diagnosed positive given true positive} = p^3 + 3p^2(1-p)$$

$$p_f^* := \text{Probability of being diagnosed positive given true negative} = p_f^3 + 3p_f^2(1-p_f)$$

Thus the overall probability of being diagnosed positive is given by

$$p_D := p^* \frac{N_p}{N} + p_f^* \frac{N_n}{N}.$$

Now consider the situation in which there is a split vote, so that diagnosis is not unanimous among the three assessors. This leads to the probabilities (analogous to p, p_f)

$$p^{**} := 3p^2(1-p); \quad p_f^{**} := 3p_f^2(1-p_f) :$$

we thus have from Bayes' Theorem

$$\begin{aligned} p_S &:= \text{Probability of a split vote given a positive diagnosis} \\ &= \frac{\text{Probability of a split vote and a positive diagnosis}}{\text{Probability of a positive diagnosis}} \\ &= \left\{ p^{**} \frac{N_p}{N} + p_f^{**} \frac{N_n}{N} \right\} / p_D \end{aligned}$$

We can now derive a relation between p and p_f : specifically, after a little manipulation we find that they must satisfy the equation

$$p_D(p^{**} - p_f^{**}) + p_f^{**}p^* = p_S p_D(p^* - p_f^*) + p_f^* p^{**}.$$

If we now equate p_D to the observed proportion \widehat{D} and p_S to the observed proportion of split votes \widehat{S} among the positive diagnoses in this relationship, then for any given values of p (which as we have seen may often be estimated independently) we derive a consistent value for p_f by numerically solving this equation.

We have, of course, not assumed that we have observed \widehat{S} , but rather p_i for each observer. But since, whenever any one reader gave a negative diagnosis, both other readers must have classified it as positive for the median-based diagnosis to be positive, it follows that the proportion of "split votes" on positives must be

$$\widehat{S} = (1 - p_1) + (1 - p_2) + (1 - p_3),$$

and so we can implement this method using the information provided.

We are now able to estimate various other parameters of interest. For example, the true proportion of positives in the population can then be estimated by

$$Pr[F] = \frac{\widehat{N}_p}{N} = \frac{p_D(1 - p_S) - p_f^3}{p^3 - p_f^3}.$$

where \widehat{N}_p is the observed number of positives. Finally, the predictive values of the median diagnosis comes again from Bayes' Theorem

$$PV_{pos} = Pr[D|F]Pr[F]/Pr[D] = p^* \frac{\widehat{N}_p}{N} / p_D;$$

$$PV_{neg} = Pr[\overline{D}|\overline{F}]Pr[\overline{F}]/Pr[\overline{D}] = (1 - p_f^*)(1 - \frac{\widehat{N}_p}{N}) / (1 - p_D).$$

4. APPLICATION: DETECTION OF OPACITIES

We now apply these results to find the predictive value of the median reading in Hughes and Weill (1991) for radiological detection of opacities.

On pp.230–231 of Hughes and Weill (1991), we find that $N = 642$, $N_p = 77$ and so

$$\widehat{D} = 77/642 = 12\%$$

of films were classified $\geq 1/0$ on a median reading. Of those, the three readers classified 82%, 86% and 90% individually as $\geq 1/0$, as noted above. It follows that the proportion of "split votes" on positives must have been

$$\widehat{S} = 42\% (= 18\% + 14\% + 10\%).$$

Using these data in the formulae in Section 3, we give in the first two rows of Table 1 a range of mutually consistent values for both p and $1 - p_f$. The last two rows of Table 1 record the predictive values PV_{pos} and PV_{neg} corresponding to these values.

TABLE 1: Ranges of possible values of sensitivity, specificity, and predictive values

Sensitivity	p	0.82	0.85	0.88	0.9
Specificity	$1 - p_f$	0.958	0.925	0.904	0.894
Positive predictive value	PV_{pos}	0.86	0.80	0.74	0.71
Negative predictive value	PV_{neg}	0.95	0.93	0.93	0.92

In order to decide on appropriate values of p , and hence of $1 - p_f$, PV_{pos} , PV_{neg} , we need a separate estimate. One source is the data on accuracy based on histology on p. 97 of Kipen *et al.* (1987), where we find that out of 138 cases with histopathological fibrosis, 113 (82%) were diagnosed radiologically as $\geq 1/0$; that is, were true positives. This appears to be based on a single reading of the film, and gives an estimate of p . If we take this value of $p = 0.82$ then we find from Table 1 that there are very few false positives (4%), and that the positive predictive value of the median test is 86%.

If we assume that readers are more accurate at identifying true positives than indicated by Kipen *et al.* (1987), then somewhat paradoxically the proportion of false positives must go up: essentially, by being optimistic about positives a reader will guarantee a good rate of true positives but will generate too many false positives also. At this point more of the split votes are in fact wrong diagnoses, representing two optimistic readers and one accurate reader who is outvoted. Not dissimilar results have also been observed by Aberle *et al.* (1988), who reported an even higher rate of true positives (96%) but also a very high rate of false positives (36%) due to overoptimistic diagnosis when using high resolution CT scans as the diagnostic tool.

Note from the values of PV_{pos} and PV_{neg} in Table 1 that if the diagnosis is negative in a study such as that of Hughes and Weill (1991) then there is a fairly constant probability (around 0.92–0.95) that there really is no small opacity present; but that there is much more variation (from 0.86 down to 0.71) in the positive predictive value, the probability that given a positive diagnosis there really are small opacities present. This stems from the fact that if there is a 10% false positive rate as in the last column of Table 1, then the high prevalence of true non- opacity cases will provide many false positives.

Now the need for a reliable estimate of p becomes clear. It may not seem entirely plausible that the true positives should be as low as the 0.82 reported in Kipen *et al.* (1987) since this implies the true negatives are extremely accurate; but it might seem reasonable to conclude that both true positives and true negatives are around 90%, as would happen in radiographic diagnosis if the specificity and the sensitivity of the raters more or less coincide. The implication of Table 1 is that at this point the overall median positive predictive value may be lower than one might initially think ($PV_{pos} = 0.71$).

In this case the purported relationships may be considerably weakened. If, for example, the true value is indeed $p = 0.9$, for a study with the other parameters in Hughes and Weill (1991) we find the true number of positives to be around 62, rather than the 77 actually diagnosed. It then follows that the percentage of excess lung cancers in the $\geq 1/0$ group reduces to 54%, indicating that an excess lung cancer is almost as likely to be associated with a nonopacity as an opacity. The consequence would be that Hughes and Weill's (1991) claimed associations become invalid, and the screening procedure would have almost no value.

REFERENCES

- Aberle, D.R., Gamsu, G., and Ray, C.S. (1988) High-resolution CT of benign asbestos-related diseases: clinical and radiographic correlation. *AJR*, 151:883-891.
- Hughes, J.M. and Weill, H. (1991) Asbestosis as a precursor of asbestos related lung cancer: results of a prospective mortality study. *Brit. J. Ind. Med.*, 48:229-233.
- International Labour Office (1980) *Guidelines for the use of ILO International Classification of Radiographs of Pneumonconioses*. ILO Occupational Safety and Health Series No 22, Geneva.
- Kipen, H.M., Lilis, R., Suzuki, Y., Valciukas, J.A. and Selikoff, I.J. (1987) Pulmonary fibrosis in asbestos insulation workers with lung cancer: a radiological and histopathological evaluation. *Brit. J. Industrial Med.*, 44:96-100.
- Liddell, F.D.K. and McDonald, J.C. (1980) Radiologic findings as predictors of mortality in Quebec asbestos workers. *Brit. J. Industrial Med.*, 37:257-267.
- Mausner, J.S. and Kramer, S. (1985). *Epidemiology - An Introductory Text*. W.B. Saunders Co., Philadelphia.
-