VARIATION OF SENSITIVITY, SPECIFICITY, LIKELIHOOD RATIOS AND PREDICTIVE VALUES WITH DISEASE PREVALENCE

by H. Brenner and O. Gefeller, *Statistics in Medicine*, **16**, 981–991 (1997)

From: *Ralf Bender*,[1,2] *Stefan Lange*,[2] *Gudrun Freitag*[2] *and Hans Joachim Trampisch*[2]
[1] *Department of Metabolic Diseases and Nutrition* (*WHO-Collaborating Centre for Diabetes*), *Heinrich-Heine-University of Düsseldorf, D-40001 Düsseldorf, Germany*
[2] *Department of Medical Informatics*, *Biometry and Epidemiology*, *Ruhr-University of Bochum*, *D-44780 Bochum, Germany*

Recently, Brenner and Gefeller presented calculations in which the basic efficacy measures of diagnostic tests, sensitivity, specificity and likelihood ratios, vary strongly with disease prevalence. The calculations are based upon the situation where the disease classification is made by means of a dichotomization of a continuous trait subject to measurement error. In short, X denotes the true value of the continuous trait and is assumed to follow a Normal distribution with mean μ and variance 1. The disease is defined as present if $X \geqslant C$. An approximate measurement of X is given by $Z = X + e$, where e is the measurement error. It is assumed that X and e are independent and that e is Normally distributed with mean 0 and variance σ_e^2. Hence, Z is Normally distributed with mean μ and variance $1 + \sigma_e^2$. Without loss of generality the diagnostic cutpoint is chosen as $C = 0$. To investigate the association between the efficacy measures of the diagnostic test based on the variable Z and the disease prevalence, the mean μ is varied between -3 and $+3$, producing populations with a range of disease prevalence from 0·1 per cent to 99·9 per cent. Within this model, the expected sensitivity and specificity as well as the expected predictive values can be expressed as direct functions of μ and σ_e^2. Under the above assumptions it was shown that sensitivity, specificity, and likelihood ratios strongly vary with disease prevalence, and that the association of the predictive values with disease prevalence is lower than one would expect if sensitivity and specificity were constant between populations. These results are interpreted as a 'methodological framework for quantitative assessment of the importance of disease prevalence'.

Brenner and Gefeller address an important issue. Defining and diagnosing a disease by means of the same continuous trait observed with measurement error involves some problems. However, we think that the results of their calculations are misinterpreted in so far as the simple model used has no practical relevance. To investigate the accuracy of a diagnostic test, two classifications are required: one ideally based on the truth (for example, variable X of the paper) or – if X is not observable – based on the gold standard (for example, variable Z of the paper) and one classification based on a diagnostic variable Y. In the model described above either the diagnostic variable Y or the gold standard is missing. Hence, an evaluation study of a diagnostic test as described in the paper will never be performed in practice. The authors do not use the terms 'sensitivity" and 'specificity' as it is usually the case, that is, with respect to a diagnostic test as an indicator for a known ('truth') or assumed (gold standard) disease status; they rather use these terms to describe the impact of measurement error of a quantitative test on its precision at the tails of the underlying distribution.

The fact that a diagnostic classification of patients depends on whether a continuous trait is above or below some defined cutpoint does not automatically mean that the disease itself is defined by means of a dichotomized continuum. For example, diabetes is often diagnosed by means of the 2 hour plasma glucose value (oral glucose tolerance test). Nevertheless, there is the idea that diabetes is inherently binary although the true state is not directly observable. For example, Engelau *et al.* define the true diabetes states theoretically by means of two components of a bimodal distribution fitted to 2 hour glucose values.[1] The historical background for defining cutpoints of 2 hour plasma glucose values was to predict the development of microvascular complications associated with diabetes, not to define diabetes itself.[2]

Most importantly, the formulation used by the authors that sensitivity and specificity strongly vary with the disease prevalence is misleading. At first, the event A|B is stochastically independent of B by definition. This means that the event '*test positive|disease present*' is stochastically independent of the event '*disease present*'. As sensitivity only refers to the subgroup of ill subjects one can get an unbiased estimate of sensitivity from a random sample of ill subjects, although the disease prevalence itself cannot be assessed. In other words, sensitivity is independent of the disease prevalence. This fact naturally holds only within the same probability space. If one changes the diagnostic criterion or the population, then the efficacy measures of the diagnostic test may not be constant. The efficacy of diagnostic tests is dependent on the population characteristics. If the sample used for estimation does not contain the whole spectrum of ill and healthy patients, respectively, estimation of test efficacy is subject to spectrum bias. This was described by Ransohoff and Feinstein nearly 20 years ago[3] and was repeatedly reported in a number of papers thereafter. If the population characteristics change, sensitivity and specificity as well as disease prevalence may also change. Hence, the observed variation of sensitivity and specificity in the model of Brenner and Gefeller is trivial, but this observation does not mean that sensitivity and specificity vary in dependence on disease prevalence. The main point is the variation of the population characteristics not that of the disease prevalence. For example, it is possible that the sensitivity of a diagnostic test is different between two populations with equal disease prevalence but with more seriously ill persons in one population. Hence, it is misleading to say that sensitivity and specificity vary with disease prevalence.

It is to the credit of the authors that they underline the impact of measurement error on the validity of diagnostic tests. However, the calculations based on their simple model are by no means a theoretical framework for the explanation or quantification of the dependence of sensitivity and specificity on general population characteristics.

REFERENCES

1. Engelau, M. M., Thompson, T. J., Herman, W. H., Boyle, J. P., Aubert, R. E., Kenny, S. J., Badran, A., Sous, E. S. and Ali, M. A. 'Comparison of fasting and 2-hour glucose and HbA_{1c} levels for diagnosing diabetes', *Diabetes Care*, **20**, 785–791 (1997).
2. Davidson, M. B., Peters, A. L. and Schriger, D. L. 'An alternative approach to the diagnosis of diabetes with a review of the literature', *Diabetes Care*, **18**, 1065–1071 (1995).
3. Ransohoff, D. F. and Feinstein, A. R. 'Problems of spectrum bias in evaluating the efficacy of diagnostic tests', *New England Journal of Medicine*, **299**, 926–930 (1978).

AUTHORS' REPLY

We thank Dr. Bender and his colleagues for their interest in our recently published paper. We anticipated this type of reaction given the provocative title and content of our paper. We also understand this type of reaction given that our paper challenges a widely accepted dogma in medical statistics. In a previous personal communication, Dr. Bender expressed his concerns on the validity of our calculations (which are outlined in the Appendix of our paper). He later found that these concerns were unjustified, and he and his colleagues now appear to accept the statistical validity of the type of variation of sensitivity and specificity illustrated in our paper. They now express some conceptual concerns about the adequacy of our approach and the interpretation of the illustrated patterns. As we will outline below, some of the issues they address have been carefully discussed in our paper and seem to have been overlooked, ignored or misunderstood by the authors. We will further show that other arguments are subject to serious logical flaws.

First, Bender *et al.* interpret our model as reflecting mere measurement error and ignore the fact, explicitly expressed in our paper, that the 'measurement error' in this model 'may also reflect other factors, such as intra-individual variability of the underlying trait or the influence of uncontrolled covariates' (p. 983). This misconception has important implications. For example, it is clearly incorrect to interpret our model as