STATISTICS IN MEDICINE, VOL. 16, 981–991 (1997)

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

HERMANN BRENNER*[1] AND OLAF GEFELLER[2]

[1]*University of Ulm, Department of Epidemiology, Albert-Einstein-Allee 43, D-89081 Ulm, Germany*
[2]*University of Göttingen, Department of Medical Statistics, Humboldtallee 32, D-37073 Göttingen, Germany*

SUMMARY

The sensitivity, specificity and likelihood ratios of binary diagnostic tests are often thought of as being independent of disease prevalence. Empirical studies, however, have frequently revealed substantial variation of these measures for the same diagnostic test in different populations. One reason for this discrepancy is related to the fact that only few diagnostic tests are inherently dichotomous. The majority of tests are based on categorization of individuals according to one or several underlying continuous traits. For these tests, the magnitude of diagnostic misclassification depends not only on the magnitude of the measurement or perception error of the underlying trait(s), but also on the distribution of the underlying trait(s) in the population relative to the diagnostic cutpoint. Since this distribution also determines prevalence of the disease in the population, diagnostic misclassification and disease prevalence are related for this type of test. We assess the variation of various measures of validity of diagnostic tests with disease prevalence for simple models of the distribution of the underlying trait(s) and the measurement or perception error. We illustrate that variation with disease prevalence is typically strong for sensitivity and specificity, and even more so for the likelihood ratios. Although positive and negative predictive values also strongly vary with disease prevalence, this variation is usually less pronounced than one would expect if sensitivity and specificity were independent of disease prevalence. © 1997 by John Wiley & Sons, Ltd. Stat. Med., Vol. 16, 981–991 (1997).

(No. of Figures: 3   No. of Tables: 1   No. of Refs: 26)

INTRODUCTION

The basic measures of the validity of binary diagnostic tests are sensitivity, which denotes the probability of a positive test in the presence of the disease of interest, and specificity, which denotes the probability of a negative test in the absence of disease.[1] Measures frequently derived from these basic measures are the likelihood ratios. The positive likelihood ratio is the probability of a positive test result in the presence of the disease divided by the probability of a positive test result in the absence of the disease. Analogously, the negative likelihood ratio is the probability of a negative test result in the presence of the disease divided by the probability of a negative test result in the absence of the disease. Measures of validity of particular interest in clinical practice are the predictive values,[2] which denote posterior probabilities of disease given a certain test result: the positive predictive value is the probability of presence of disease given a positive test result, and the negative predictive value is the probability of absence of disease given a negative test result.

* Correspondence to: H. Brenner

CCC 0277–6715/97/090981–11$17.50

Received October 1995
Revised June 1996

A widely recognized limitation of the predictive values is their dependence on disease prevalence. By contrast, sensitivity and specificity and the likelihood ratios are generally thought of as being independent of disease prevalence.[3,4] Consequently, these measures are often regarded as constant benchmarks of test performance which can be used for comparing the diagnostic value of alternative tests. This belief results from simple mathematical reasoning concerning the properties of the measures defined in fourfold tables representing the joint probability distribution of diagnostic classification and true disease status. Such reasoning is justified in situations with a truly dichotomous disease status and a homogeneous probability of diagnostic misclassification within the population of diseased individuals and within the population of non-diseased individuals.

In many practical situations, however, the assumptions underlying this reasoning are likely to be violated; with few exceptions, disease status is not simply black or white (for example, dead versus alive). More typically, there is a continuum of (measurable or unmeasurable) traits on which the classification of disease status is based, varying from clear absence to clear presence of the disease.[5] In the simplest case, in which a binary categorization of disease status is made on the basis of a single continuous trait, the diagnostic classification of patients depends on whether the measurement of this trait is above or below some defined cutpoint. For example, individuals are classified as hypertensive or normotensive on the basis of the continuous trait blood pressure. Ideally, this classification should be based on individuals' usual (average) blood pressure levels. In practice, single or dual measurements are often taken that lead to diagnostic misclassification due to intra-individual variability of blood pressure or less than perfect measurement techniques. In other situations, underlying continuous traits are less readily quantifiable, in which case subjectivity of ratings provides yet another source of measurement error. Common examples include diagnoses based on clinical symptoms or radiological changes. In yet many other situations, inaccuracy of test procedures predominantly reflects the influence of unmeasured covariates on test results which are unrelated to the disease of interest. For example, serum levels of gamma-glutamyltransferase (GGT) are often used as tests for alcohol related liver disorders using various cutpoints.[6,7] Although alcohol is the most important single causal factor for GGT elevation, some variation of GGT levels is due to other factors, such as viral infections.

Diagnostic cutpoints are often established by experts' mutual or explicit consensus. Examples are cutpoints for hypertension or hyperlipedemia. In other situations, cutpoints are imposed by the threshold of clinical manifestation. This would apply to many seemingly clearcut diagnoses, such as myocardial infarction or prostate cancer, which often escape diagnostic verification.

Because individuals with true levels close to the diagnostic test cutpoint are more likely to be misclassified than other individuals in the presence of measurement error or intra-individual variability of the underlying traits, or due to the influence of uncontrolled covariates, the misclassification rates are expected to vary between populations, depending on the distribution of true levels of the underlying traits relative to the diagnostic test cutpoint. This distribution, however, also determines the prevalence of the disease in the population. Diagnostic misclassification and disease prevalence are therefore interrelated.

The variation of sensitivity and specificity with the distribution of the underlying traits has been pointed out by Ransohoff and Feinstein.[8] They coined the term 'spectrum bias' for this phenomenon which has been demonstrated in a variety of empirical studies.[9–16] Published examples include, among others, laboratory tests,[9,12,15] electrocardiography,[10,11,14] scintigrams,[13] and ultrasonography.[16] Despite the repeated recognition of variation of sensitivity and specificity with disease prevalence, quantitatively oriented methodological work on this phenomenon

Table I. Classification of individuals by the true value of the trait (X) and by the approximate measure of the trait (Z)

True value of trait, X	Approximate measure of trait (Z) $\geqslant 0$	< 0
$\geqslant 0$	true positive	false negative
< 0	false positive	true negative

is lacking. In this paper, we illustrate the expected variation of the sensitivity and specificity, the likelihood ratios and the predictive values with disease prevalence for simple models of the distribution of the underlying traits and the measurement error.

METHOD

Let X^* denote the true value of a continuous trait that underlies diagnosis. X^* is assumed Normally distributed with mean μ^* and standard deviation σ within populations (results for other distributional assumptions will be discussed below). The standard deviation of X^* is assumed the same in all populations, but the mean is allowed to vary between populations, which translates to different disease prevalences relative to a diagnostic cutpoint C^*, and to different distributions of severity of disease for diseased and non-diseased individuals. For simplicity, X^* is transformed into a variable $X = (X^* - C^*)/\sigma$ which follows a Normal distribution with mean $\mu = (\mu^* - C^*)/\sigma$ and standard deviation 1 within populations, and C^* is transformed to a cutpoint $C = (C^* - C^*)/\sigma = 0$, accordingly.

Let Z be an approximate measurement of X with $Z = X + e$, where e denotes an additive term which will be referred to as 'measurement error' throughout this paper for simplicity, although e may also reflect other factors, such as intra-individual variability of the underlying trait or the influence of uncontrolled covariates. It is assumed that the measurement error e is independent of the true value X and that e follows a Normal distribution with mean 0 and standard deviation σ_e. Then, Z is Normally distributed with mean μ and variance $1 + \sigma_e^2$ within populations. The measurement error causes some diagnostic misclassification if the measurements are categorized. Table I shows the classification of individuals according to the relation of the values of X and Z to the cutpoint C.

For our numerical illustration, σ_e is varied over 0·125, 0·25, 0·50 and 1·0, that is, the standard deviation of the measurement error is assumed to vary between 12·5 per cent and 100 per cent of the standard deviation of the true trait X^* within the populations, which translates to a wide range from rather low to rather high misclassification rates. This order of magnitude of σ_e is typical of intra-individual variation of many traits that are commonly measured in clinical and epidemiological applications. For example, ratios of within-person standard deviation to between-person standard deviation of about 0·6 and 0·7 have been reported for cholesterol and blood pressure measurements.[17] The population means μ are varied between -3 and $+3$, while the diagnostic cutpoint is kept constant at $C = 0$, reflecting a wide range of disease prevalence from 0·1 per cent to 99·9 per cent. The expected sensitivity and specificity, the expected likelihood ratios and the expected predictive values are derived as a function of μ and σ_e by numerical integration as outlined in the Appendix. These expected values are then plotted against disease prevalences. In addition, the predictive values that are expected from this model, which allows for variation of sensitivity and specificity with disease prevalence, are compared with the predictive values that would be expected with constant levels of sensitivity and specificity. For this comparison, sensitivity and specificity are arbitrarily fixed at the level expected for $\mu = 0$, that is, at the level expected for a population with a disease prevalence of 50 per cent. All analysis were carried out with the software package SAS using the function PROBNORM to integrate the probability density function of the Normal distribution.[18]

The following notation is used throughout this paper, including the Appendix and the figures:

P	disease prevalence
TP (TN)	expected proportion of true positive (negative) diagnoses among all diagnoses
Se (Sp)	expected sensitivity (specificity)
PPV (NPV)	expected positive (negative) predictive value
PPV′ (NPV′)	positive (negative) predictive value that would be expected if the sensitivity and specificity were constant at the level expected for $\mu = 0$
LR + (LR −)	expected positive (negative) likelihood ratio.

For symmetry and simplicity, results are presented for the multiplicative inverse of LR − rather than LR − .

RESULTS

Figure 1 illustrates the relationship of the expected sensitivity and specificity and the expected predictive values with disease prevalence as a function of the magnitude of the error in measuring the underlying trait. The expected sensitivity and specificity vary from 86·9 per cent to 99·97 per cent, and the expected predictive values range from 80·5 per cent to 99·98 per cent for $\sigma_e = 0{\cdot}125$ and the spectrum of disease prevalence considered in these analyses. Obviously, the expected values of these measures decrease with increasing levels of measurement error of Z. For all levels of measurement error, the expected sensitivity and the expected positive predictive value increase with disease prevalence, while the expected specificity and the expected negative predictive value decrease with disease prevalence. The prevalence dependence is of similar magnitude for sensitivity and the positive predictive value for prevalences above 50 per cent. For prevalences below 50 per cent, the expected positive predictive value decreases more strongly with decreasing disease prevalence than sensitivity. Prevalence dependence of both the positive predictive value and sensitivity is most pronounced at very low levels of disease prevalence. Symmetrical results are obtained for the expected specificity and the expected negative predictive value. Their prevalence dependence is of similar magnitude for prevalences below 50 per cent, but the negative predictive value decreases more strongly with increasing disease prevalence for disease prevalences above 0·50. Prevalence dependence of the negative predictive value and of specificity are most pronounced at very high levels of disease prevalence.

If sensitivity and specificity were constant at the levels expected for $\mu = 0$, the prevalence dependence of the predictive values would be even stronger, however (see Figure 2). In other words, the well-known variation of the predictive values with disease prevalence is mitigated to some extent by the fact that sensitivity and specificity also vary with disease prevalence. This particularly applies to situations with small measurement error.

Figure 3 illustrates the relationship between the expected likelihood ratios and disease prevalence. As expected, both the positive likelihood ratio and the inverse of the negative likelihood ratio are negatively associated with the magnitude of the measurement error. In addition, the likelihood ratios are extremely dependent on disease prevalence, with a negative association between disease prevalence and the positive likelihood ratio, and a positive relationship between disease prevalence and the inverse of the negative likelihood ratio. For example, with $\sigma_e = 0{\cdot}25$,

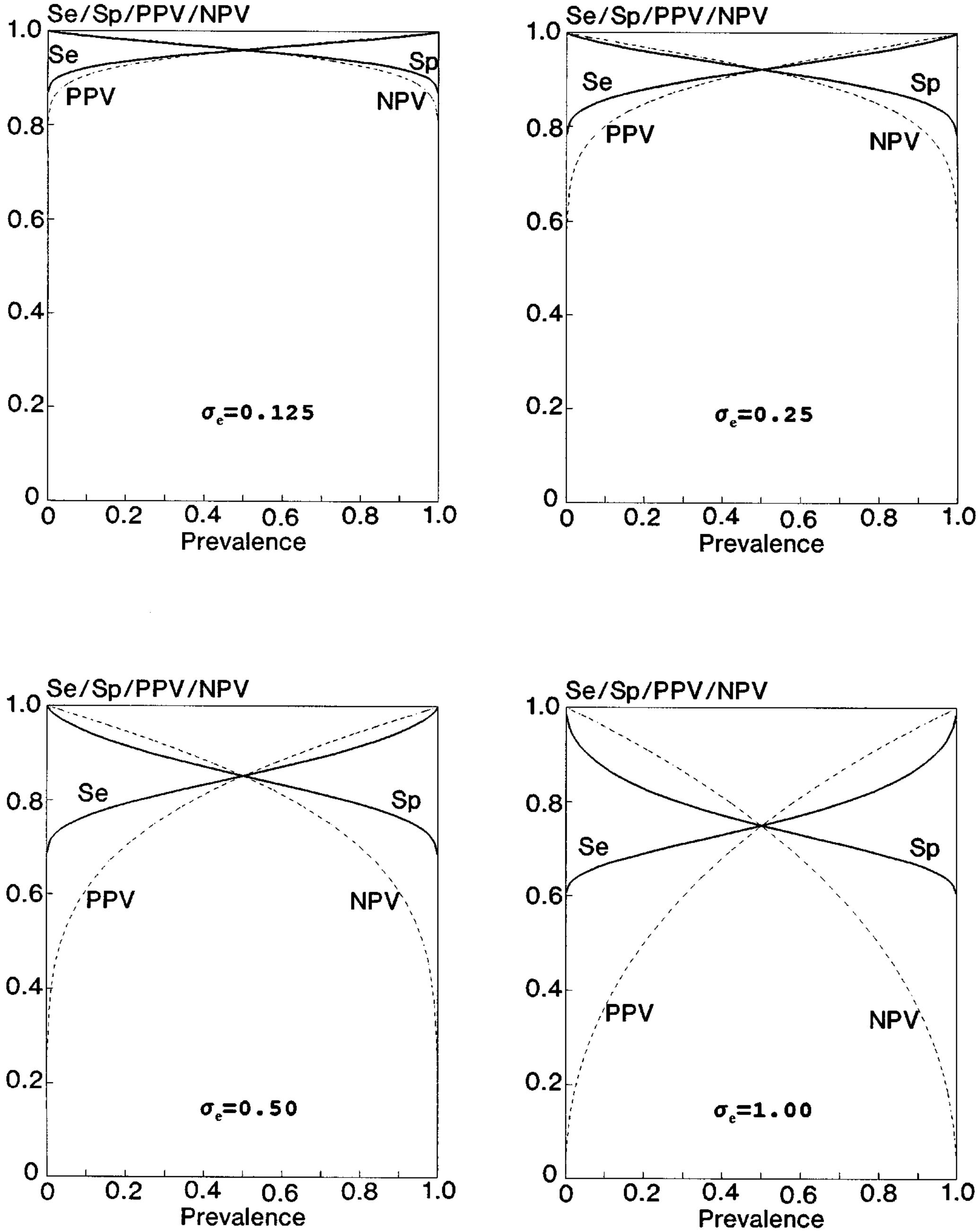

Figure 1. Relationship of sensitivity, specificity, and the predictive values with disease prevalence as a function of the error in measuring the underlying trait

the expected positive likelihood ratio decreases from 108·4 for a disease prevalence of 2·3 per cent to 5·6 for a disease prevalence of 97·7 per cent. Generally, prevalence dependence of the positive likelihood ratio is most pronounced (and can be extreme even for low levels of measurement

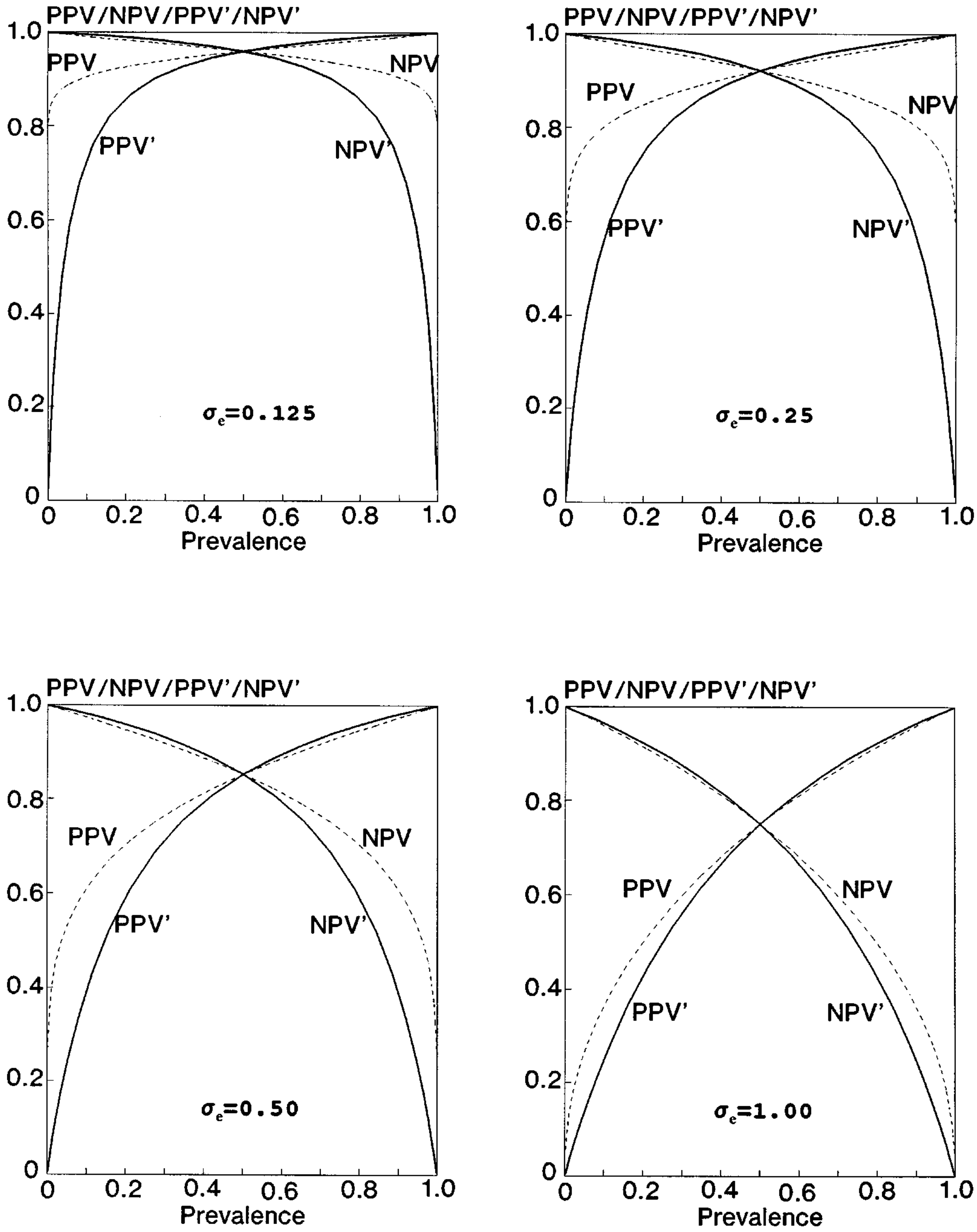

Figure 2. Relationship of the predictive values with disease prevalence as a function of the error in measuring the underlying trait: PPV, NPV = expected predictive values taking prevalence dependence of sensitivity and specificity into account. PPV′, NPV′ = predictive values that would be expected with constant levels of sensitivity and specificity

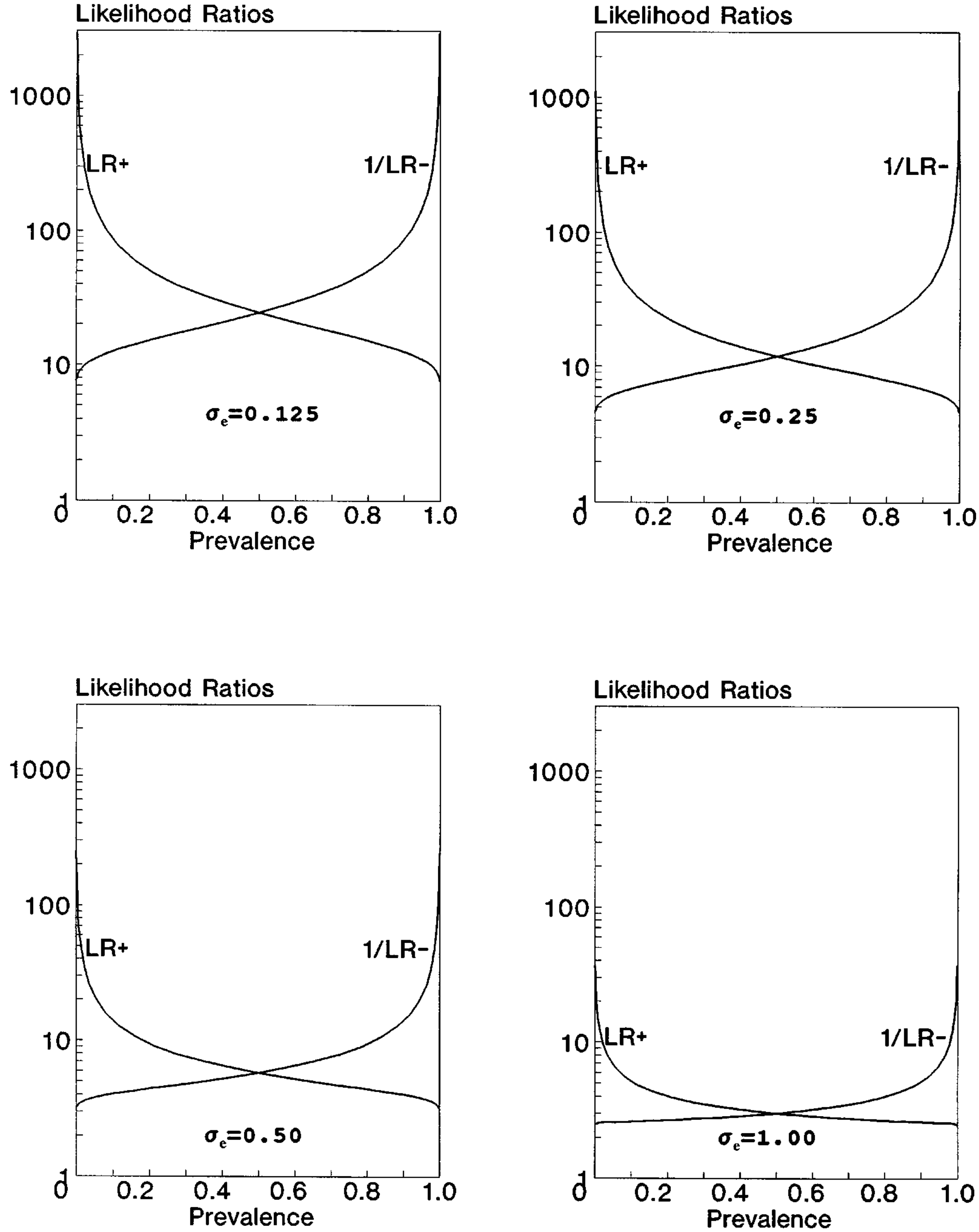

Figure 3. Relationship of the likelihood ratios with disease prevalence as a function of the error in measuring the underlying trait

error) at very low levels of disease prevalence, and prevalence dependence of the inverse of the negative likelihood ratio is most pronounced at very high levels of disease prevalence.

Although the underlying trait can often be assumed approximately normally distributed, other distributional forms are relevant in other situations. In particular, there are situations in which the distribution of the underlying trait is skewed, and in which the magnitude of the measurement error varies with true values of the underlying trait. We therefore carried out additional analyses assuming an exponential distribution of X along with normal distribution of measurement or perception error of $\ln(X)$. Patterns of prevalence dependence were generally similar as in the scenarios shown in Figures 1–3 (increase of Se, PPV, PPV′ and 1/LR − with disease prevalence, decrease of Sp, NPV, NPV′ and LR + with disease prevalence) and are therefore not depicted separately.

DISCUSSION

This paper illustrates that the basic measures of the validity of binary diagnostic tests, sensitivity and specificity, are expected to vary strongly with disease prevalence in many situations of practical relevance. For a wide spectrum of scenarios assessed in this paper, the prevalence dependence of these measures (increase of sensitivity and decrease of specificity with disease prevalence) was almost as strong as the well-known prevalence dependence of the positive and the negative predictive value. Conversely, the prevalence dependence of the predective values was mitigated to some extent by the fact that sensitivity and specificity also vary with disease prevalence. An even stronger dependence on disease prevalence was demonstrated for the likelihood ratios, which are simple functions of sensitivity and specificity.

The results presented in this paper are in contrast to the widespread belief among clinicians that sensitivity and specificity and the likelihood ratios of diagnostic tests are independent of disease prevalence. This common view has repeatedly been challenged by empirical studies demonstrating a substantial degree of heterogeneity in estimates of sensitivity and specificity for the same diagnostic test applied to different populations.[9–16,19] While several authors have discussed a variety of mechanisms that may be responsible for such heterogeneity (which is incompatible with mere random variability),[4,15,16,19–23] our approach provides a methodological framework for quantitative assessment of the importance of disease prevalence.

The results of the present paper underline the warning that the assumption of independence of measures of validity like sensitivity, specificity, and the likelihood ratios from disease prevalence (which continues to be taught in many textbooks) is only justified for the rather unusual situation of a truly binary diagnosis with homogeneous misclassification probabilities given the true disease status. It is extremely hard to think of any diagnosis that would satisfy this criterion. Even such seemingly clearcut diagnosis like presence or absence of certain cancers are often based on a continuum of symptoms or clinical or histological changes. For example, histologic changes often follow to continuum varying from dysplasia to carcinoma *in situ* to invasive cancer. Similarly, while there is a continuum of severity of lesions even for acute events such as myocardial infarction, the diagnosis is typically dichotomized at a cutpoint imposed by the threshold of clinical detectibility. For most diagnoses of chronic diseases, such as osteoarthritis, hypertension, or diabetes, the continuity of the distribution of the underlying traits (for example, radiographic changes, blood pressure, serum glucose levels) is even more obvious. This paper illustrates that both disease prevalence and basic measures of the validity of diagnostic tests, such as the sensitivity and specificity and the likelihood ratios, strongly depend on the distribution of the underlying traits relative to the diagnostic cutpoint. As a result, the validity measures and disease prevalence are closely related. Since the distribution of the traits that underlie clinical diagnoses (and hence the spectrum of severity of disease) is known to show strong variation between populations[24] or between subgroups within a single population (for example, various age groups),[25,26] these results have important practical implications.

The validity of diagnostic tests can typically be quantified only in relation to a defined distribution of the underlying traits. As a consequence, estimates of sensitivity and specificity of various diagnostic tests cannot be compared, unless they are derived from populations in which this distribution is comparable. Similarly, estimates of sensitivity and specificity derived from a specific population in a specific age interval in which a new diagnostic test was evaluated cannot generally be assumed relevant for the application of the test in other populations or other age groups.

For example, a diagnostic procedure tested in a clinical environment among patients suspected to have a certain disease will typically have lower sensitivity and higher specificity when applied as a screening tool in the general population (in which disease prevalence is typically lower). According to our results, the prevalence dependence of validity measures is expected to be particularly strong in screening for very rare diseases, such as HIV infection in populations of blood donors. In such situations, sensitivity and positive likelihood ratio may strongly vary by disease prevalence even for very low levels of measurement error of the underlying trait. Conversely, prevalence dependence of the positive predictive value may be strongly overestimated in such situations with traditional approaches that do not take prevalence dependence of sensitivity and specificity into account. These findings therefore require careful consideration by public health professionals in the design of screening programs.

Similarly, formal approaches to diagnostic decision making, which are increasingly used by clinicians, may yield strongly misleading conclusions, unless the prevalence dependence of sensitivity and specificity and the predictive values are correctly specified. Our results imply that such approaches may be considerably improved by taking the variability of performance of diagnostic tests between patient subgroups into account.

APPENDIX

Let $\varphi(\cdot)$ denote the density function and $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution. Then, for the diagnostic cutpoint $C = 0$, the prevalence of disease in a population, in which the underlying trait X follows a normal distribution with mean μ and variance 1, is given as

$$P = P(X \geqslant 0) = \Phi(\mu).$$

For diagnostic classifications based on the approximate measurement Z of the underlying trait (with $Z = X + e$, where e is a normally distributed measurement error with mean 0 and variance σ_e^2), the expected proportions of true positive and true negative diagnoses can be derived as

$$\mathrm{TP} = \int_{-\mu}^{+\infty} (\varphi(x) \times \Phi((\mu + x)/\sigma_e))\,\mathrm{d}x$$

and

$$\mathrm{TN} = \int_{-\infty}^{-\mu} (\varphi(x) \times (1 - \Phi((\mu + x)/\sigma_e)))\,\mathrm{d}x$$

respectively, and the expected proportions of false positive and false negative diagnoses can be derived as

$$\mathrm{FP} = 1 - P - \mathrm{TN} \quad \text{and} \quad \mathrm{FN} = P - \mathrm{TP}$$

respectively. The expected sensitivity and specificity are given as

$$\mathrm{Se} = \mathrm{TP}/P \quad \text{and} \quad \mathrm{Sp} = \mathrm{TN}/(1 - \mathrm{P})$$

respectively. The expected positive and negative predictive value are given as

$$\mathrm{PPV} = \mathrm{Se} \times P/(\mathrm{Se} \times P + (1 - \mathrm{Sp}) \times (1 - P))$$

and

$$\mathrm{NPV} = \mathrm{Sp} \times (1 - P)/(\mathrm{Sp} \times (1 - P) + (1 - \mathrm{Se}) \times P)$$

respectively, and the expected positive and negative likelihood ratio are given as

$$\mathrm{LR}+ = \mathrm{Se}/(1 - \mathrm{Sp}), \quad \text{and} \quad \mathrm{LR}- = (1 - \mathrm{Se})/\mathrm{Sp}.$$

REFERENCES

1. Altman, D. G. and Bland, J. M. 'Diagnostic tests 1: sensitivity and specificity', *British Medical Journal*, **308**, 1552 (1994).
2. Altman, D. G. and Bland, J. M. 'Diagnostic tests 2: Predictive values', *British Medical Journal*, **309**, 102 (1994).
3. Feinstein, A. R. *Clinical Epidemiology: The Architecture of Clinical Research*, Saunders, Philadelphia, 1985.
4. Kramer, M. S. *Clinical Epidemiology and Biostatistics: A Primer for Clinical Investigators and Decision-Makers*, Springer, Berlin, 1988.
5. Altman, D. G. and Bland, J. M. 'Diagnostic tests 3: receiver operating characteristic plots', *British Medical Journal*, **309**, 188 (1994).
6. Penn, R. and Worthington, D. J. 'Is γ-glutamyltransferase a misleading test?', *British Medical Journal*, **286**, 531–535 (1983).
7. Peterson, B., Trell, E., Kristensson, H., Fex, G., Yettra, M. and Hood, B. 'Comparison of gamma-glutamyltransferase and other health screening tests in average middle-aged males, heavy drinkers and alcohol non-users', *Scandinavian Journal of Clinical and Laboratory Investigations*, **43**, 141–149 (1983).
8. Ransohoff, D. F. and Feinstein, A. R. 'Problems of spectrum and bias in evaluating the efficacy of diagnostic tests', *New England Journal of Medicine*, **299**, 926–930 (1978).
9. Richardson, B. and Epstein W. V. 'Utility of the fluorescent antinuclear antibody test in a single patient', *Annals of Internal Medicine*, **95**, 333–338 (1981).
10. Philbrick, J. T., Horwitz, R. I., Feinstein A. R., Langou, R. A. and Chandler, J. P. 'The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg', *Journal of the American Medical Association*, **248**, 2467–2470 (1982).
11. Hlatky, M. A., Pryor, D.B., Harrell, F. E. Jr., Califf, R. M., Mark, D. B. and Rosati, R. A. 'Factors affecting sensitivity and specificity of exercise electrocardiography', *American Journal of Medicine*, **77**, 64–71 (1984).
12. Fletcher, R. H. 'Carcinoembryonic antigen', *Annals of Internal Medicine*, **104**, 66–73 (1986).
13. Detrano, R., Janosi, A., Lyons, K. P., Marcondes, G., Abbassi, N. and Froelicher, V. F. 'Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram', *American Journal of Medicine*, **84**, 699–710 (1988).
14. Levy, D., Labib, S. B., Anderson, K. M., Christiansen, J. C., Kannel, W. B. and Castelli, W. P. 'Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy', *Circulation*, **81**, 815–820 (1990).
15. Lachs, M. S., Nachamkin, I., Edelstein, P. H., Goldman, J., Feinstein, A. R. and Schwartz, J. S. 'Spectrum bias in the evaluation of dignostic tests: lessons from the rapid dipstick test for urinary tract infection', *Annals of Internal Medicine*, **117**, 135–140 (1992).
16. van der Schouw, Y. T., van Dijk, R. and Verbeek, A. L. M. 'Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests', *Journal of Clinical Epidemiology*, **48**, 417–422 (1995).
17. Gardner, M.J. and Heady, J. A. 'Some effects of within-person variability in epidemiological studies', *Journal of Chronic Diseases*, **26**, 781–795 (1973).

18. SAS Institute Inc. *SAS language*: *reference*, version 6, 1st edn, SAS Institute Inc., Cary, NC, 1990.
19. Fahey, M.T., Irwig, L. and Macaskill, P. 'Meta-analysis of pap test accuracy', *American Journal of Epidemiology*, **141**, 680–689 (1995).
20. Harris, J.M. 'The hazards of bedside Bayes', *Journal of American Medical Association*, **246**, 2602–2605 (1981).
21. Begg, C. B. and Greenes, R. A. 'Assessment of diagnostic tests when disease verification is subject to selection bias', *Biometrics*, **39**, 207–215 (1983).
22. Hlatky, M. A., Mark, D. B., Harrell, F. E., Lee, K. L., Califf, R. M. and Pryor, D. B. 'Rethinking sensitivity and specificity', *American Journal of Cardiology*, **59**, 1195–1198 (1987).
23. Begg, C. B. 'Biases in the assessment of diagnostic tests', *Statistics in Medicine*, **6**, 411–423 (1987).
24. Sánchez, R. G., Labarthe, D. R., Forthofer, R. N. and Fernández-Cruz, A. 'National standards of blood pressure for children and adolescents in Spain: international comparisons', *International Journal of Epidemiology*, **21**, 478–487 (1992).
25. The HDFP Cooperative Group. 'Race, education and prevalence of hypertension', *American Journal of Epidemiology*, **106**, 351–361 (1977).
26. Heiss, G., Tamir, I., Davis, C. E., Tyroler, H. A., Rifkind, B. M., Schonfeld, G., Jacobs, D. and Frantz, I. D. 'Lipoprotein-cholesterol distributions in selected North American populations: The Lipid Research Clinics Program Prevalence Study', *Circulation*, **61**, 302–315 (1980).