Soc Psychiatry Psychiatr Epidemiol (2000) 35: 375–379

ORIGINAL PAPER

N. Schmitz · J. Kruse · W. Tress

Application of stratum-specific likelihood ratios in mental health screening

Accepted: 8 May 2000

Abstract *Background*: The accuracy of a diagnostic procedure is commonly assessed by measuring sensitivity, specificity and positive and negative predictive values. Likelihood ratios provide an alternative method for describing these results, though they are typically reported only for dichotomized outcomes. However, likelihood ratios can also be applied to ordinal or continuous results. *Methods*: The present paper discusses the application of stratum-specific likelihood ratios in a primary care setting using the General Health Questionnaire (GHQ-12) and the Symptom Check List 90-R (SCL-90-R). A randomly selected sample (n = 408) of adult outpatients from primary care offices in Düsseldorf was screened using the German versions of the GHQ-12 and the SCL-90-R. *Results*: Logistic regression analysis indicated that stratum-specific or multilevel likelihood ratios preserve more information than a fixed threshold approach with a single cutoff point. For each test, five clinically useful strata with monotonically increasing stratum-specific likelihood ratios were selected. *Conclusions*: Stratum-specific likelihood ratios have enormous practical value, and they are becoming an important way of expressing and comparing the usefulness of different tests. Stratum-specific likelihood ratios reduce the spectrum bias that might arise if only two categories (cases and non-cases) are chosen. Additionally, multilevel likelihood ratios can be used as bedside information to obtain the post-test probability from the pre-test probability of the disorder.

N. Schmitz (✉) · J. Kruse · W. Tress
Clinic for Psychosomatic Medicine and Psychotherapy,
Heinrich-Heine University,
Bergische Landstrasse 2, H19,
D-40605 Düsseldorf, Germany
e-mail: schmitzn@uni-duesseldorf.de,
Tel.: +49-211-9224723
Fax: +49-211-9224709

Introduction

Screening tests like the General Health Questionnaire are used as predictors for mental disorders. Common use includes screening a specific population for evidence of disease and confirming or ruling out a tentative diagnosis in an individual patient. The interpretation of a diagnostic test result depends on both the ability of the test to distinguish diseased from non-diseased subjects and the particular characteristics of the patient and setting in which the test is being used.

In addition to sensitivity and specificity, likelihood ratios are also used to characterize the behavior of diagnostic tests (Sackett et al. 1991). Sensitivity (the proportion of diseased patients with positive results) and specificity (the proportion of non-diseased patients with negative results) describe the behavior of tests given patients' disease status. The likelihood ratio expresses the likelihood that a given finding will occur in a patient with the target disorder. The key advantage of likelihood ratios is that they are less sensitive to changes in pre-test probabilities of disorders. In addition, likelihood ratios can be calculated for several levels of performance on diagnostic tests, and thus the test result (e.g., GHQ-12) of a given subject can be linked to its likelihood ratio to yield the odds that the subject has a mental disorder. For example, when the pre-test probability (i.e., the probability of the target disorder before a diagnostic test result is known) lies between 30 and 70%, test results with a high likelihood ratio (e.g., above 10) rule in disease. A low likelihood ratio (e.g., below 0.1) virtually rules out the chance that the patient has the disease, and a likelihood ratio of around 1 means that no useful information has been obtained from the clinical finding (Sackett et al. 1997).Thus, likelihood ratios offer the diagnostician powerful evidence concerning an individual patient's status with respect to a disorder.

However, the best cutoff score to achieve optimum sensitivity and specificity may vary considerably from one setting to another, as it would depend on the base

rate of the target disorder in the population at hand. Much information is indeed lost when screening tests are evaluated in relation to a single cutoff value of continuous or categorical (with more than two categories) variables (Sox 1986). A way to avoid this shortcomings has recently been recommended by several authors (Furukawa and Goldberg 1999; Peirce and Cornell 1993; Simel et al. 1993). Multilevel, or stratum-specific, likelihood ratios (SSLR), have been advocated as a more informative alternative to the fixed threshold approach. Two properties of SSLR make them more uniformly applicable in screening studies and in clinical practice. First, they do not depend on the prevalences of target disorders, although these must be taken into account when using them. Second, they are less subject to spectrum bias than a single cutoff, because serious and less serious cases tend to show up in their corresponding strata, and the misclassification is reduced (spectrum bias is the phenomenon of the sensitivity and/or specificity of a test varying with different populations tested, Ransohoff and Feinstein 1978). Although SSLR are a powerful tool in diagnosing mental disorders, only a few studies have used this approach so far (e.g., Furukawa et al. 1997). The aim of the present paper is to discuss the application of SSLR in a primary care setting using the General Health Questionnaire (GHQ-12, Goldberg and Williams 1991) and the Symptom Check List (SCL-90-R, Derogatis 1977).

Subjects and methods

Statistics

The conventional method of reporting data from screening studies that describe diagnostic test characteristics is to incorporate results into the standard four-cell table. These tables display the results of diagnostic tests versus the presence or absence of a disease as assessed by an accepted gold standard (Table 1). Although sensitivity and specificity are fundamental values calculated from the 2×2 tables, in actual practice the clinician is more interested in the question "What is the probability that my patient has a mental disorder given a positive test result?" rather than being concerned with the question "What proportion of patients with a mental disorder have a positive test result?" While it might seem that the positive predictive value is the most appropriate value to answer the first question, it is not generalizable, since it depends on the disorder's prevalence. Alternatively, likelihood ratios can be used in clinical decision making, since they are independent of prevalence rates of disorders. The likelihood ratio (LR) is defined as the ratio between the probability of a defined test result given the presence of a disease and the probability of the same test result given the absence of the disease (e.g., Choi 1998):

$$\mathrm{LR} = \frac{\text{probability of a test result among the diseased subjects}}{\text{probability of the same test result among the non-diseased subjects}}$$

In addition, a further advantage of the likelihood ratio is the expression as the ratio of the post-test odds of a disease (odds of disease among subjects with a given test result) to the pre-test odds of a disease (odds of disease among all subjects).

If a test generates dichotomous results (e.g., cases and non-cases), then two likelihood ratios can be defined: a likelihood ratio for a positive result (LR+) and a likelihood ratio for a negative result (LR−). The positive likelihood ratio (LR+) expresses the change in odds favoring disease given a positive test result, whereas the negative likelihood ratio (LR−) expresses the change in odds favoring disease given a negative test result:

$$\mathrm{LR}+ = \frac{\dfrac{\text{Probability of disease among test-positive subjects}}{\text{Probability of no disease among test-positive subjects}}}{\dfrac{\text{Probability of disease among all subjects}}{\text{Probability of no disease among all subjects}}}$$

It has been shown that LR+ and LR− are related to the sensitivity and the specificity of a dichotomous test (Simel et al. 1991): LR+ = sensitivity/(1−specificity) and LR− = (1−sensitivity)/specificity.

However, problems for researchers and clinicians arise when possible test results include non-positive and non-negative results, frequently described as intermediate or uninterpretable results. Tests that are neither positive nor negative may have an important impact on clinical decision making. Important information is lost with the single forced cutoff point, since all subjects are divided into two groups (e.g., cases and non-cases). Therefore, a closely related alternative approach is to use stratum-specific likelihood ratios (SSLR). The stratum-specific or multilevel likelihood ratio is calculated by describing the proportion of diseased subjects with a test result in a given range divided by the proportion of non-diseased subjects with a test result in the same given range (Table 2). It has been shown mathematically that SSLR are related to the sensitivity and specificity defined by the strata (Black and Armstrong 1986). The likelihood ratio for a stratum with the upper and lower bounds (scores) x and y corresponds to the change in sensitivity divided by the change in specificity over the defined interval:

$$\mathrm{LR}(x,y) = [\text{sensitivity}(x) - \text{sensitivity}(y)]/[\text{specificity}(y) - \text{specificity}(x)] \ .$$

Therefore, LR(x, y) corresponds to the slope between two points x and y on the receiver operating characteristic curve (ROC, Choi 1998). Confidence intervals can be computed as a measure of

Table 1 Generalized contingency table for sensitivity, specificity and predictive values of a screening test

	"Gold standard"	
	Positive	Negative
Screening test		
Positive	a	b
Negative	c	d
Prevalence = (a+c)/(a+b+c+d)		
Sensitivity = a/(a+c)		
Specificity = d/(b+d)		
Positive predictive value = a/(a+b)		
Negative predictive value = d/(c+d)		
Likelihood ratio positive = sensitivity/(1−specificity)		
Likelihood ratio negative = (1−sensitivity)/specificity		

Table 2 The 5×2 table when obtaining multilevel test results

Screening test result	"Gold Standard"	
	Positive	Negative
Level 1	a	b
Level 2	c	d
Level 3	e	f
Level 4	g	h
Level 5	i	j
	n_1	n_2

Example: likelihood ratio for level 2: $\mathrm{LR}_2 = (c/n_1)/(d/n_2)$
$\mathrm{LR}_{\leq 2} = [(c+a)/n_1]/[(d+b)/n_2]$

precision for the estimated likelihood ratios (Simel et al. 1991). An example for the interpretation of SSLR is shown in Table 3.

The number of strata should be chosen carefully, because with too many strata the likelihood ratios become unstable and degenerate. Following Furukawa et al. (1997) and Peirce and Cornell (1993), the following rules of thumb are recommended:

1. Provide sufficient abnormal and normal cases in each stratum to allow the SSLR to be monotonically related, and
2. Collapse those strata where the SSLR are close to one another and their 95% confidence intervals easily overlap.

Population sample

The present study was an epidemiological study designed to evaluate the prevalence of psychological disorders in primary care. A sample of 572 German consecutive attenders of primary care offices was assessed. The sample comprised 179 men (31.3%) and 393 women (68.7%), mainly aged between 23 and 65 (mean = 42.7, SD = 15.7). Subjects were informed about the general purposes of the study, and asked to give their informed consent. After the general practitioner's consultation, patients filled in symptom checklists (GHQ-12, SCL-90-R) and were examined and diagnosed by a mental health professional. For case identification, the Impairment Score (Schepank 1995) and the SCID interview (Wittchen et al. 1990) served as the gold standard. The prevalence rate of mental disorder was 36.8%. Detailed study methodology is reported by Tress et al. (1997). The application of the GHQ-12 and the SCL-90-R as screening tools for mental disorders was reported in a previous issue of this journal (Schmitz et al. 1999).

Instruments

The GHQ-12 (Goldberg and Williams 1991) is a 12-item self-report instrument for the detection of mental disorders in the community and non-psychiatric clinical settings. The GHQ-12 asks the respondents to report how they have been feeling over the past few weeks, using a four-point scale ("not at all" to "much more than usual"). It can be scored in a bimodal fashion (0-0-1-1). Alternatively, it can be scored by using a four-point response format (Likert-Scale: 0-1-2-3), resulting in a scale range of 0 to 36, with higher scores representing higher distress.

The participants also completed the SCL-90-R (Derogatis 1977; Franke 1995), which is a 90-item self-report symptom inventory, multi-dimensional in nature and oriented toward the measurement of psychopathology. Each of the 90 items is rated on a five-point scale of distress, ranging from "not at all" (0) to "extremely" (4). The SCL-90-R is scored on nine primary symptom dimensions: Somatization, Obsessive-compulsive, Interpersonal sensitivity, Anger-hostility, Depression, Anxiety, Paranoid ideation, Phobic anxiety and Psychoticism. In addition, three global indices provide measures of overall psychological distress: the General Symptom Index, the Positive Symptom Total and the Positive Symptom Distress Index.

Table 3 Interpretation of stratum-specific likelihood ratios (SSLR)[a]

SSLR for a mental disorder	Interpretation
0.1	Most likely normal
0.5	More likely normal than abnormal
1	Normal/abnormal
3	More likely abnormal than normal
10	Most likely abnormal

[a] The diagnostic decision depends on the prior probability of the target disease. The rules described in the table should be used for pre-test probability between 30 and 70%

Results

In a first step, the recommended SSLR for the GHQ-12 (bimodal scoring) from the World Health Organization Collaborative Study of "Psychological Problems in General Health Care" (Sartorius et al. 1993) were compared with the SSLR obtained in our sample. In addition, SSLR were computed for the Verona center (one center of the WHO study) based on the data reported in Dunn et al. (1999). As shown in Table 4, similar values with overlapping confidence intervals were obtained for the lower GHQ-12 scores (scores 0–3). For the GHQ scores 4–12, somewhat higher stratum-specific likelihood ratios were obtained for the WHO centers.

In a further step, we computed SSLR for the Likert scaling of the GHQ-12, since the SSLR did not discriminate well between the lower GHQ scores (bimodal scoring) in our study. Results for the general score of the SCL-90-R and the global severity index (GSI) of the SCL-90-R are shown in Table 5 and Table 6. We selected five strata for each test with monotonically increasing SSLR. Although the strata were somewhat different than those shown in the example in Table 3, the SSLR can be interpreted in the same way.

Likelihood ratios can be calculated for several strata; the question is whether the benefit is worth the trouble. From a statistical perspective, logistic regression models can be used to test whether the strata provide significantly more information than a single cutoff point (Simel et al. 1993). First, a logistic regression model with the dichotomous predictor is fitted. Then, a second

Table 4 SSLR for the 12-item General Health Questionnaire (GHQ-12) in the Primary Care Study (Düsseldorf) and the WHO study

GHQ-12 score	WHO study (15-center) ($n = 5438$)		Verona (WHO center) ($n = 250$)		Primary Care Study (Düsseldorf) ($n = 421$)	
	SSLR	95% CI	SSLR	95% CI	SSLR	95% CI
0	0.14	0.12–0.17	0.32	0.27–0.38	0.40	0.28–0.58
1	0.54	0.43–0.67	1.11	0.86–1.44	0.75	0.46–1.21
2	0.78	0.61–1.00	1.39	1.05–1.83	0.79	0.40–1.56
3	1.59	1.26–2.02	1.39	1.05–1.83	1.17	0.55–2.53
4–6	3.29	2.83–3.82	2.26	1.80–2.84	1.43	0.90–2.30
7–12	11.45	9.68–13.54	4.33	2.95–6.35	5.31	3.23–8.75

Table 5 SSLR for the GHQ-12 sumscore (Likert scaling) in the Primary Care Study (Düsseldorf)

GHQ-12 scores	SSLR	95% CI
0–5	0.16	0.06–0.43
6–9	0.51	0.36–0.72
10–15	1.05	0.77–1.44
16–23	2.98	1.96–4.53
24–36	5.42	2.61–11.26

Table 6 SSLR for the Global Severity Index (SCL-90-R) in the Primary Care Study (Düsseldorf)

SCL-GSI scores	SSLR	95% CI
0–0.20	0.38	0.24–0.61
0.21–0.45	0.58	0.41–0.82
0.46–0.70	1.23	0.82–1.82
0.71–1.65	2.06	1.49–2.84
1.66–4	16.89	3.97–71.79

model with the same dichotomous predictor and the strata as a categorical predictor is fitted. A statistically significant improvement is reached when the model fit for the second model is better than the fit for the first model, e.g., when the likelihood ratio test yields a statistically significant chi-square statistic. For the GHQ-12, the $-2 \cdot \log$ LR for the model with a single cutoff point (cutoff 11/12) was 490.9. The $-2 \cdot \log$ LR for the model with the strata as a categorical predictor was 468.2. The difference between these two values is 22.7 and is a chi-square statistic with 4*df* under the null hypothesis that the strata predictor does not add more predictive ability than the single cutoff point. The corresponding *P*-value is quite small ($P < 0.001$), indicating a significant improvement using the strata for the GHQ-12. Similar results were obtained for the SCL-90-R (cutoff 0.65; single cutoff point: $-2 \cdot \log \text{LR} = 530.3$, single cutoff point and strata: $-2 \cdot \log \text{LR} = 509.5$, difference = 20.8, $df = 4$, $P < 0.001$).

A combination of the SSLR for the GHQ-12 and the SCL-90-R is shown in Table 7. Most of the cases (subjects with a mental disorder) are most likely abnormal, or more likely abnormal than normal, in at least one of the two questionnaires. On the other hand, the non-cases are most likely normal, or more likely normal than abnormal, in at least one of the two questionnaires. However, the tables show that the questionnaires do not identify the same subjects. For example, there are subjects with a low Global Severity Index (GSI of the SCL-90-R) and a high GHQ sumscore.

Discussion

Non-positive, non-negative and uninterpretable test results occur frequently in medical practice and research settings when tests are designed to yield dichotomous results (Simel et al. 1987). It seems reasonable that such indeterminate results should be uniformly reported and assessed. We discussed the application of multilevel likelihood ratios in mental health screening to report diagnostic data. Likelihood ratios offer clear grounds for deciding whether a test is sufficiently informative to warrant its use for identifying mental disorders. They have important properties, that make them useful in epidemiological studies. SSLR, like sensitivity and specificity but unlike the positive predictive value and negative predicative value, do not depend on the prevalence of the target disorder. Additionally, SSLR reduce the spectrum bias that might arise if only two categories (cases and non-cases) are chosen. Serious and less serious cases will tend to show up in their corresponding strata, and the misclassification based on the dichotomous criterion will be reduced (Feinstein 1990). Moreover, SSLR have enormous practical value, and they are becoming an important way of expressing and comparing the usefulness of different tests. For a given pre-test probability (prevalence) and an SSLR, one can estimate

Table 7 SSLR for non-cases (**A**) and cases (**B**)[a] in the Primary Care Study (Düsseldorf)

SSLR GHQ-12	SSLR GSI (SCL-90-R)				
	Most likely normal	More likely normal than abnormal	Normal /abnormal	More likely abnormal than normal	Most likely abnormal
A Non-cases					
Most likely normal	27	18	0	1	0
More likely normal	46	43	12	8	0
Normal/abnormal	9	22	23	19	0
More likely abnormal	0	9	5	14	1
Most likely abnormal	0	0	1	7	1
B Cases					
Most likely normal	2	1	1	0	0
More likely normal	8	8	3	10	0
Normal/abnormal	6	11	12	11	2
More likely abnormal	1	7	10	20	8
Most likely abnormal	0	0	0	14	7

[a] The case criterion is based on the clinical interview by the mental health professional. The strata of the questionnaires are similar to those in Tables 5 and 6

the post-test probability or chance of a mental disorder. For example, if a patient enters a primary office, there is a 0.56 odds ratio of a mental disorder, since one person in three of consecutive attenders of primary care offices has this condition [in the language of diagnostic tests, the pre-test probability or prevalence of mental disorders in primary care is 0.36 in our study (Schmitz et al. 1999), odds ratio = 0.36/(1−0.36)]. Now, if the GHQ-12 is filled in, the result will usually make the diagnosis of a mental disorder more or less likely. A GHQ sumscore between 16 and 23 has a likelihood ratio of nearly 3, so the odds of a patient with this result having a mental disorder is $0.56 \times 3 = 1.68$. This value is known as the post-odds of the screening test. The likelihood ratio of a very low GHQ score (below 6) is 0.16, making the odds of a mental disorder with this result smaller than unity ($0.56 \times 0.16 = 0.09$). Likelihood ratios, coupled to the prior odds, represent an alternative approach that adapts to populations with different prevalences of mental disorders. SSLR are a general methodology for analyzing diagnostic tests, they can be used for ordinal and continuous test results. Additionally, information about covariates can be incorporated in the analysis and allow insight into important clinical variables. The number of strata should be evaluated carefully. Peirce and Cornell (1993) developed a computer program to arrive at the optimal number of strata of test scores with different but overlapping confidence intervals. However, the definition of useful strata should be made on a clinical and statistical basis. Nevertheless, clinicians themselves should understand the trade-offs investigators make when deciding to report SSLR.

A natural generalization of the SSLR presented here would be to model the likelihood ratio as a continuous monotonic function of the test score itself. From a statistical point of view, continuous likelihood ratios provide more information than is available from SSLR. Following this approach, likelihood ratios are calculated for all test scores, rather than for a range of scores. On the other hand, large samples are needed to have sufficient cases and non-cases for each test score (e.g., to calculate a likelihood ratio for the individual GHQ-12 score 36, a number of healthy and diseased subjects with this extreme score must be available). In addition, as pointed out by Simel et al. (1993), clinicians often break down continuous scales into ordered outcomes, which are used as bedside information in clinical reality. Therefore, pragmatic reasons should drive the choice between continuous and multilevel likelihood ratios.

References

Black WC, Armstrong P (1986) Communicating the significance of radiologic test results: the likelihood ratio. A J R 147: 1313–1318

Choi BCK (1998) Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. Am J Epidemiol 148: 1127–1132

Derogatis LR (1977) SCL-90-R, administration, scoring, and procedures manual for the (Revised) version. Johns Hopkins University, Ann Arbor confirmed

Dunn G, Pickles A, Tansella M, Vazquez-Barquero JL (1999) Two-phase epidemiological surveys in psychiatric research. Br J Psychiatry 174: 95–100

Feinstein AR (1990) The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. J Clin Epidemiol 43: 109–113

Franke GH (1995) SCL-90-R: Die Symptom-Check-Liste von Derogatis – Deutsche Version. Beltz Test Gesellschaft, Göttingen

Furukawa T, Hirai T, Kitamura T, Takahashi K (1997) Application of the Center for Epidemiologic Studies Depression Scale among first-visit psychiatric patients: a new approach to improve its performance. J Affect Disord 46: 1–13

Furukawa T, Goldberg DP (1999) Cultural invariance of likelihood ratios for the General Health Questionnaire. Lancet 353: 561–563

Goldberg D, Williams P (1991) A user's guide to the General Health Questionnaire: NFER-Nelson, Windsor

Peirce JC, Cornell RG (1993) Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med Decis Making 13: 141–151

Ransohoff DF, Feinstein AR (1978) Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299: 926–930

Sackett DL, Haynes RB, Guyatt GH, Tugwell P (1991) Clinical epidemiology: a basic science for clinical medicine, 2nd edn. Little Brown, Boston

Sackett DL, Richardson WS, Rosenberg W, Haynes RB (1997) Evidence-based medicine: how to practice and teach EBM. Churchill Livingstone, London

Sartorius N, Ustun TB, Costa e Silva JA, Goldberg D, Lecrubier Y, Ormel J, Von Korff M, Wittchen HU (1993) An international study of psychological problems in primary care. Preliminary report from the World Health Organization Collaborative Project on 'Psychological Problems in General Health Care'. Arch Gen Psychiatry 50: 819–824

Schepank H (1995) Der Beeinträchtigungsschwerescore (BSS). Ein Instrument zur Bestimmung der Schwere einer psychogenen Erkrankung. Beltz Test Gesellschaft, Göttingen

Schmitz N, Kruse J, Heckrath C, Alberti L, Tress W (1999) Diagnosing mental disorders in primary care: The General Health Questionnaire (GHQ) and the Symptom Check List (SCL-90-R) as screening instruments. Soc Psychiatry Psychiatr Epidemiol 34: 360–366

Simel DL, Feussner JR, DeLong ER, Matchar DB (1987) Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making 7: 107–114

Simel DL, Samsa GP, Matchar DB (1991) Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol 44: 763–770

Simel DL, Samsa GP, Matchar DB (1993) Likelihood ratios for continuous test results – making the clinicians' job easier or harder? J Clin Epidemiol 46: 85–93

Sox HCJ (1986) Probability theory in the use of diagnostic tests. Ann Int Med 104: 60–66

Tress W, Kruse J, Heckrath C, Schmitz N, Alberti L (1997) Psychogene Erkrankungen in hausärztlichen Praxen. Zsch Psychosom. Med. 43: 211–232

Wittchen HU, Schramm E, Zaudig M, Spengler P, Rummler R, Mombour W (1990) SKID-Strukturiertes Klinisches Interview für DSM-III-R. Beltz, Weinheim