BRIEF COMMUNICATION

Stratum-specific likelihood ratios of the General Health Questionnaire in the community: help-seeking and physical co-morbidity affect the test characteristics

T. A. FURUKAWA, G. ANDREWS AND D. P. GOLDBERG

From the Department of Psychiatry, Nagoya City Medical School, Nagoya, Japan; School of Psychiatry, University of New South Wales, St Vincent’s Hospital, Sydney, NSW, Australia; and Institute of Psychiatry, King’s College, London

ABSTRACT

Background. In evidence-based medicine, stratum-specific likelihood ratios (SSLRs) are now being increasingly recognized as a more convenient and generalizable method to interpret diagnostic information than an optimal cut-off and its associated sensitivity and specificity. We previously examined the SSLRs of the General Health Questionnaire (GHQ) in primary care settings. The present paper aims to examine if these SSLRs are generalizable to the community settings.

Methods. The Composite International Diagnostic Interview (CIDI) and the GHQ were administered on a representative sample of the Australian population in the Australian National Survey of Mental Health and Well-Being. We first compared the SSLRs of GHQ in urban Australia with the estimates that we had previously obtained from the developed urban centres in the WHO Psychological Problems in General Health Care study. If the SSLRs in the community were found to differ significantly from those in the primary care, we sought for explanatory variables.

Results. The SSLRs in urban Australia and in the urban centres in the WHO study were significantly different for three out of the six strata. When we limited the sample to those with physical problems who visited a health professional, however, the SSLRs in the Australian study were strikingly close to those observed for primary care settings.

Conclusions. Different sets of SSLRs apply to primary care and general population samples. For general population surveys in developed countries, the results of the Australian National Survey represent the currently available best estimates. For developing countries or rural areas, the results are less definitive and an investigator may wish to conduct a pilot study.

INTRODUCTION

In 1995 the World Health Organization completed the Psychological Problems in General Health Care study to investigate the psychological disorders commonly seen among primary care patients in 15 centres around the world (Üstün & Sartorius, 1995). They administered the GHQ along with the Primary Care Version of the Composite International Diagnostic Interview (CIDI-PC). Although this study took great care in standardizing both the GHQ and the CIDI-PC in uniform manners across the 15 centres, there remained great variation in the best cut-off score for the GHQ-12, ranging from 1/2 to 6/7 out of the possible total score of 12 (Goldberg et al. 1997).

This posed a particular concern because the
GHQ is the most widely used screening questionnaire for mental disorders in general medical and community settings. An investigation followed which showed that the threshold was mainly determined by the variation in prevalence (Goldberg et al. 1998). Because mean GHQ score correlated highly with prevalence, this report gave the former as a rough guide to the best threshold. The variation in the best recommended threshold remained.

In evidence-based medicine, stratum-specific likelihood ratios (SSLRs) are now being increasingly recognized as a more convenient and informative method to interpret diagnostic information than an optimal cut-off and its associated sensitivity and specificity (Sackett et al. 2000; Guyatt & Rennie, 2001). For an explanation of SSLR, please see the Appendix. The SSLR is independent of prevalence and is less subject to, although not totally free from, spectrum bias (change in test characteristics due to a different mix of disease severity and comorbidity) than the optimum threshold and are hence expected to be more generalizable across situations. In fact, when test characteristics of the GHQ were expressed in SSLRs, much of the apparent variation in the performance of GHQ disappeared (Furukawa & Goldberg, 1999). In particular, for primary care centres in urban and developed areas, the estimates of the SSLRs were very homogeneous (Furukawa et al. 2001).

In 1997 the Australian Bureau of Statistics conducted the National Survey of Mental Health and Well-Being in all of its States and Territories. The survey used the GHQ-12 and the CIDI and gave us the unique opportunity to examine the generalizability of the SSLRs of the GHQ in the general population. The present work was undertaken to discover whether the SSLRs that apply to consulting settings also apply to respondents in the community.

METHOD

Subjects and procedure
The Australian National Survey of Mental Health and Well-Being was a nationwide mental health survey of a representative sample of residents of private dwellings to determine the prevalence of ICD-10 and DSM-IV disorders and their associated co-morbidity, disability and service utilization. The response rate was 78%.

Detailed description of the study procedures can be found elsewhere (Andrews et al. 2001).

The study was a one-phase survey, administered face-to-face by trained interviewers of the Australian Bureau of Statistics using a laptop computer. The Composite International Diagnostic Interview (CIDI version 2.1) (Andrews & Peters, 1998) was used to determine the 12-month and current diagnoses according to ICD-10 and DSM-IV. The survey also included the 12-item GHQ (Goldberg & Williams, 1988). The GHQ-12 was administered independently of the CIDI diagnoses.

Analyses
In order to make the study comparable to our previous analyses of GHQ in the WHO Psychological Problems in General Health Care study (Furukawa et al. 2001), we defined our ‘gold standard’ in the present analyses as current ICD-10 diagnosis of depression, dysthymia, agoraphobia, panic disorder, generalized anxiety disorder, social phobia, obsessive–compulsive disorder, post-traumatic stress disorder and neurasthenia. We also limited the sample to age range of 18 to 64 in order to make the sample as comparable as possible to those recruited in the WHO study which surveyed consecutive patients between the ages of 18 and 65 years.

In the WHO study, homogeneous estimates of SSLRs were obtained for primary care settings in urban and developed areas. We therefore first compared the SSLRs of the GHQ-12 in the urban Australia with those that we had previously reported for seven primary care centres in developed countries around the world in the WHO study (Berlin, Mainz, Manchester, Nagasaki, Paris, Seattle and Verona).

If the average SSLRs of the GHQ-12 in the community settings were found to differ significantly and substantively from those in the primary care settings, we then sought for variables which might influence the SSLRs of the GHQ-12. Candidate variables included demographic ones (age, sex) as well as ones which were thought to characterize primary care visitors vis-à-vis general community samples, such as concurrent physical condition and help-seeking behaviour. Physical morbidity was determined by asking for asthma, chronic bronchitis, anemia, high blood pressure, heart trouble, arthritis, kidney disease, diabetes, cancer, stomach or
duodenal ulcer, chronic gall bladder or liver trouble, and hernia or rupture. Help-seeking behaviour was defined as seeing a doctor or other health professional because of one’s health within the past 4 weeks.

We used SPSS 10.0 (SPSS Inc., 1999) and Microsoft Excel for data analyses. Because this was a one-phase survey in which both the questionnaire in question and the gold standard diagnostic instrument were administered to all the participating subjects, no weighting was required. We divided the GHQ scores into the same six strata as in our previous analyses for the primary care centres, applying the procedures detailed by Peirce & Cornell (1993) to calculate variances and 95% confidence intervals of the SSLRs. The statistical significance of the difference between two likelihood ratios (LR₁ and LR₂) was determined by calculating (Breslau & Day, 1980):

\[
\chi^2_{df=1} = \frac{(\ln LR_1 - \ln LR_2)^2}{\text{VAR}(\ln LR_1) + \text{VAR}(\ln LR_2)}
\]

RESULTS

Of the 10641 respondents in the Australian Survey, 8849 were aged between 18 and 64 (3974 men and 4875 women). Of these, 937 subjects (10.6%) were diagnosed as suffering from one or more of the depressive or anxiety disorders listed in the Method section within the past 4 weeks preceding the interview.

Table 1 shows the SSLRs of the GHQ-12 in urban Australian areas in comparison with those of the developed and urban primary care centres in the WHO study. The SSLRs were statistically significantly different at the conventional \( P \) value of 0.05 for three out of the six strata.

We therefore looked for variables that may modify SSLR and explain the observed differences. Sex or age did not affect the SSLRs, which were not significantly different for the two sexes or for the younger versus older generations for all the strata. Physical co-morbidity and recent visit to a health professional both appeared to affect the SSLRs; in both cases, people with physical co-morbidity and people who had visited a health professional within 4 weeks presented with SSLRs closer to those observed in primary care settings. We therefore made a subgroup of people with physical co-morbidity and seeking help within the last month. This subgroup is the most similar, as far as the study variables in the Australian survey permit, to primary care patients in the WHO study. The SSLRs of the GHQ for this subgroup were strikingly close to those observed for the developed and urban primary care centres in the WHO study (Table 1).

DISCUSSION

Our previous analyses of the GHQ have demonstrated that the fixed best threshold approach to the GHQ can be very misleading and introduce unnecessary variations across centres with different prevalences of mental disorder. On the other hand, the SSLR approach was more generalizable across various settings differing in prevalence and spectrum of the target disorders, because SSLRs are independent of prevalence of the target disorders and they are less subject to

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4-6</th>
<th>7-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seven urban primary care centres in developed countries in WHO study (( N = 2323 ))</td>
<td>0.23</td>
<td>0.55</td>
<td>0.90</td>
<td>1.7</td>
<td>2.6</td>
</tr>
<tr>
<td>Urban Australian sample (( N = 6005 ))</td>
<td>(0.15 to 0.35)</td>
<td>(0.36 to 0.85)</td>
<td>(0.39 to 2.1)</td>
<td>(1.0 to 2.6)</td>
<td>(2.0 to 3.4)</td>
</tr>
<tr>
<td>(0.30)</td>
<td>1.2</td>
<td>1.4</td>
<td>2.3</td>
<td>4.1</td>
<td>17</td>
</tr>
<tr>
<td>Statistical significance</td>
<td>( \chi^2 = 1.43 )</td>
<td>( \chi^2 = 1.11 )</td>
<td>( \chi^2 = 1.11 )</td>
<td>( \chi^2 = 1.31 )</td>
<td>( \chi^2 = 7.51 )</td>
</tr>
<tr>
<td>( P = 0.23 )</td>
<td>( P = 0.001 )</td>
<td>( P = 0.29 )</td>
<td>( P = 0.25 )</td>
<td>( P = 0.01 )</td>
<td>( P = 0.001 )</td>
</tr>
<tr>
<td>People with physical co-morbidity and help seeking in urban Australia (( N = 1093 ))</td>
<td>0.25</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>36</td>
</tr>
<tr>
<td>(0.19 to 0.35)</td>
<td>(0.75 to 1.4)</td>
<td>(0.64 to 1.5)</td>
<td>(1.2 to 3.2)</td>
<td>(2.6 to 5.1)</td>
<td>(5.8 to 14.5)</td>
</tr>
<tr>
<td>Statistical significance</td>
<td>( \chi^2 = 0.15 )</td>
<td>( \chi^2 = 5.28 )</td>
<td>( \chi^2 = 0.04 )</td>
<td>( \chi^2 = 0.24 )</td>
<td>( \chi^2 = 2.44 )</td>
</tr>
<tr>
<td>( P = 0.71 )</td>
<td>( P = 0.02 )</td>
<td>( P = 0.84 )</td>
<td>( P = 0.63 )</td>
<td>( P = 0.12 )</td>
<td>( P = 0.95 )</td>
</tr>
</tbody>
</table>

95% confidence intervals are shown in parentheses.

For all \( \chi^2 \) tests, \( df = 1 \).
spectrum bias than the dichotomous best threshold approach (Furukawa et al. 2001). The present study further sought to examine the generalizability of the SSLRs from primary care setting to the community settings.

Reasons why test characteristics of a questionnaire may vary include the difference of the ‘gold standard’ diagnostic instrument used to define a case, the particular disorders included as ‘cases,’ the manner in which the questionnaire is administered, and the health and psychosocial characteristics of the population studied. Our previous analyses of the WHO Primary Care Study showed that, holding the first three factors constant as required in any multi-site study, the SSLRs of the GHQ were reasonably constant for developed urban centres. In other words, for developing and/or non-urban areas, health attitudes and other characteristics of the populations studied may have rendered the SSLRs dissimilar.

In the present comparison between the Australian National Survey and the WHO Primary Care Study, the first two factors were held as constant as possible: both used the CIDI and the definition of cases were very similar, although not identical due to the different versions of the CIDI used. The manners in which the GHQ was administered were also minimally different, because in the Australian Survey the questions were read out to the subjects in accordance with the instructions on the laptop computer at the end of the computerized interview whereas in the WHO Study it was delivered as a self-rating questionnaire before the interview.

Despite these subtle differences, we found that, if we limit the sample to those who have recently sought help for their physical condition in urban Australia, the SSLRs of the GHQ were strikingly similar to those found with primary care attenders in various cities in seven different countries across the world. In contrast, for people who do not have a physical health problem or who do not seek professional help, the test characteristics of the GHQ as a mental health screener were different. It therefore follows that SSLRs obtained from surveys of patients attending primary care cannot be generalized to community surveys in general. Our study represents, on one hand, an independent replication of our findings with the WHO Study and, on the other, increases the applicability of the Australian findings for community surveys elsewhere. Overall, higher scores on the GHQ were more discriminatory and had stronger ruling-in capabilities in the general population than among primary care attenders.

In a developed country like Australia, the urban versus rural differences did not affect the SSLRs of the GHQ (direct comparison between urban Australia and rural Australia available from the first author on request). However, it is a matter of course that the Australian sample does not provide ultimately reliable estimates of SSLRs in the developing world or in rural communities.

Several limitations of the present study must be noted. First, although SSLRs were statistically significantly different for only one stratum between the Australians with physical co-morbidity who consulted a doctor and the primary care attenders in seven urban centres surveyed in the WHO study, some may attribute this to the decreasing statistical power as we sought subsamples of the Australian National Survey who resembled primary care patients. However, both the Australian Survey and the WHO study were large scale studies and still allowed over 1000 for such subgroups. Moreover, the point estimates of the SSLRs were essentially identical between these two populations. It must also be pointed out that the only significantly different stratum corresponding to a GHQ score of 1 in the Australian National Survey is aberrant in the sense that it is not monotonically increasing from the left to the right cell. Secondly, the validity of the CIDI or other fully structured interviews has recently been under scrutiny (Brugha et al. 1999, 2001). This issue apparently demands further studies and discussions that are beyond the scope of the present paper (Wittchen et al. 1999).

In summary, we make the following recommendations.

1. For primary care samples, the SSLRs as reported in our previous study (Furukawa et al. 2001) apply.
2. For community samples in developed countries, the currently available best evidence suggests that we may apply the results of the Australian National Survey. These findings, however, need independent replication from other community surveys.
(3) For community samples in developing countries or rural areas, the results are less definitive. If an investigator has some doubt that the GHQ would perform differently, then it would be best to do a pilot study to obtain SSLRs of GHQ applicable to the population in question.

APPENDIX
Explanation of SSLR
A likelihood ratio (LR) is a ratio of two likelihoods, one of showing the test result in question among those with the disease, over one of showing the same test result among those without the disease. Suppose we had a diagnostic test resulting in very high scores, high scores, moderate scores and low scores in conjunction with a range of patients with and without the target disorder, as shown in the table below.

<table>
<thead>
<tr>
<th>Test</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very high</td>
<td>$n_{1+}$</td>
</tr>
<tr>
<td>High</td>
<td>$n_{1-}$</td>
</tr>
<tr>
<td>Moderate</td>
<td>$n_{2+}$</td>
</tr>
<tr>
<td>Low</td>
<td>$n_{2-}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N$</td>
</tr>
</tbody>
</table>

By definition, the likelihood ratio for the ‘very high’ test scores in Table 1 is $(n_{1+}/N)/(n_{1-}/N)$. Because this likelihood ratio is specific to a stratum out of a continuum of test results, it is called the stratum-specific likelihood ratio (SSLR).

According to Bayes’ theorem, the relationship between the pre-test probability, SSLR and post-test probability can be described as follows:

Pre-test odds \times SSLR = Post-test odds, where,

\[
\text{Odds} = \frac{\text{Probability}}{(1 - \text{Probability})}.
\]

What this means is that the SSLR of a certain test result modifies the pre-test probability of the disorder in question and yields the post-test probability.

Advantages of SSLR can be summarized as follows. First, the likelihood ratio approach, in conjunction with the Bayes theorem, is intuitive. It nicely incorporates the differences in pre-test probabilities (or prevalences) between populations and illustrates in clearer terms what we should do in diagnostic processes of an individual or a group of individuals.

Secondly, in the case of a continuous scale (which most tests in medicine are), SSLR retains more information than a single cut-off and its associated sensitivity and specificity. We gave some examples on how the SSLR approach is more informative and hence less misleading both for individual-level clinical predictions and for group-level epidemiological estimations in our previous report (Furukawa et al. 2001).

Thirdly, SSLR is expected to be more resilient to, although not totally free from, spectrum bias, because by cutting the total distributions of the test scores of the diseased and non-diseased populations into several strata, we can expect that serious and less serious cases will show up in their corresponding strata and SSLR of each stratum will be less affected by the change in the mix of the serious and less serious cases. These characteristics make the SSLR to be more generalizable than the best cut-off in interpreting test scores, as previously illustrated in our analyses of the GHQ in WHO primary care study.

REFERENCES
Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W. &

