Abstract  Background  The finding of variation in the optimal threshold of the General Health Questionnaire (GHQ) across different settings has proved difficult to explain. This analysis aimed to examine the optimal threshold of the GHQ, its variability and relationship with prevalence of psychiatric disorder.  Methods  A cross-sectional two-phase epidemiological survey was used in a study of non-psychotic psychiatric disorder of General Practice consulters. A total of 1670 consecutive patients were screened using the GHQ-28 and the GP encounter form, and 336 were interviewed using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID). The total prevalence of the psychiatric disorders was estimated using three different methods and was calculated in the rural and urban practices and among those attending the General Practitioners and Practice Nurses.  Results  The frequency distribution of the GHQ score for the whole sample of respondents was skewed resulting in the biased mean GHQ score. The mean values for different sample categories were found to be higher than the median values indicating that the median GHQ score may be a better parameter to describe the score distribution than the mean. The median value of the GHQ score showed a strong correlation with the prevalence estimated by the three different methods.  Conclusions  The median of GHQ score gives a better estimate for the cut-off score than the mean. The median score may be used as an estimate for the optimal threshold, in settings where the sensitivity and specificity are not known or where an estimate of the prevalence from a questionnaire survey is required. Alternative methods are preferable when screening for individual cases or in the context of a two-phase design.

Key words  GHQ-28 – optimal threshold – score median – psychometric properties – prevalence

Introduction

The General Health Questionnaire (GHQ) is a self-administered screening questionnaire designed to detect probable psychiatric disorder in primary care settings (Goldberg 1972) which is highly popular and widely used in research (e.g., Lobo et al. 1986; Gureje and Obikoya 1990; Schmitz et al. 1999). There has been some difficulty in interpreting the variation of the threshold adopted for different settings (Goldberg et al. 1998) and recent efforts to understand the discriminatory power of the GHQ have not provided a convincing explanation as to why the GHQ threshold varies from one place to another (Wilmink and Snijders 1989; Van Hemert et al. 1995; Goldberg et al. 1998). Clarke and McKenzie (1994) suggested that when a screening instrument was applied with a cut-off point, measures of validity such as sensitivity and specificity tended to lose their accuracy and generalisability. In fact, the optimal threshold concept may be more useful for estimating prevalence in large population than screening for individual cases. A recent report by Furukawa et al. (2001) moved away from the use of GHQ optimal threshold to the use of stratum-spe-
specific likelihood ratios for the screening test, which may provide a more informative alternative to the fixed threshold approach (Sackett et al. 2000). Goldberg et al. (1998) noted that the mean score of the GHQ provided a rough guide to the best threshold and that there was a relationship between the best threshold and the ability of the GHQ to discriminate between cases and non-cases. There needs to be clarity as to the use of the threshold: as a means of deciding who should get a second-stage interview or as a means of providing an estimate of the prevalence in a given population, and these need to be considered when examining the GHQ cut-off score and the psychometric properties of the GHQ.

This analysis, which used data from the Mid Cheshire and Keele General Practice (MiCK) Study (see Boardman et al. 2003 for further details), aimed to examine the basic GHQ score characteristics and its variability in order to provide more insight and understanding of the GHQ. The analysis centred on the optimal threshold, its variability and possible relationship with other factors such as prevalence.

Subjects and methods

Screening and selection of cases for interview

The data used in this paper were obtained from a two-phase survey of General Practice attenders conducted in Mid Cheshire (Boardman et al. 2003). In summary, a total of 1670 consecutive patients aged 17–65 years attending the routine surgeries of 21 General Practitioners (GPs) and 12 Practice Nurses (PNs) completed the General Health Questionnaire (GHQ-28). The patients’ degree of emotional disturbance was rated by the GP or PN on a six-point scale ranging from 0 (no disturbance) to 5 (severe disturbance) (see Marks et al. 1979; Boardman 1987). A cut-off score of 4/5 was used to define a GHQ case (GHQ+) and all subjects rated as 2 or more by the GP or PN were considered to be a GP (or PN) case (GP+). Subjects were selected for interview on the basis of their GHQ and GP caseness using a modification of the method described by Ormel et al. (1990). All patients in the categories GHQ+/GP+ and GHQ-/GP+ were systematically chosen for interview and random individuals from the categories GHQ+/GP− and GHQ-/GP− were selected (Fig. 1). Further detail on sampling method and sample weights can be found in Boardman et al. (2003).

Measures

A total of 336 individuals (20% of the screened samples) were interviewed using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) (First et al. 1996). The interviewed subjects were classified according to DSM-IV criteria (American Psychiatric Association 1995) and Bedford College criteria (Finlay-Jones et al. 1980). The nature of illness as seen in primary care meant that the study needed to examine the broadest range of non-psychotic disorders including sub-threshold disorders. Thus, the definition of psychiatric cases in our study followed the DSM-IV criteria as the standardised definitions with the inclusion of sub-threshold or ‘borderline’ cases defined by the Bedford College criteria (Finlay-Jones et al. 1980). The range of DSM-IV disorders included Major and Minor Depressive Disorders, Substance Misuse and Dependancy, Generalised Anxiety Disorder, Panic Disorder, Agoraphobia, Social Phobia, Specific Phobia, Obsessive Compulsive Disorder, Post Traumatic Stress Disorder, Somatisation Disorder, and Eating Disorder. No attempt was made to define DSM-IV Axis II disorders. Bedford College cases included Bedford College depression cases and borderline cases, anxiety cases and borderline cases. No hierarchical definition of cases was used and many of the cases defined at interview were comorbid, including the simple phobias and eating disorders which may not otherwise be detected by the screening test. For the purposes of the present analysis, an interview case was defined as a subject who was a case on either of the DSM-IV or Bedford College case criteria used.

In summary, three types of case identification were used in this study: (i) GHQ case identification; (ii) GP case identification; and (iii) DSM-IV or Bedford College case at the time of interview. This was used as a gold standard against which to compare the performance of the GHQ.

Analysis

In order to examine the variability of the GHQ score, the screened samples were categorised according to the practice settings (rural or urban) and the health personnel consulted (GPs or PNs). There are five categories of samples as shown in Table 1. For each of these categories, the summary statistics and frequency distribution of the GHQ score were calculated using SPSS (SPSS Inc. 1999).

Based on the 336 SCID-interviewed samples in the second phase of the survey, the screening prevalence was estimated following the methods of Dunn et al. (1999) and Ormel et al. (1990). Both methods differ in the way the weighting factors were estimated and, thus, have different values for the weights. In Dunn et al. (1999), the sampling weight for interviewed GHQ+ is the ratio between the total number of the screened GHQ+ and the total number of the interviewed GHQ+. Similarly, the sampling weight for the interviewed GHQ− is the ratio between the screened GHQ− and the interviewed GHQ−. The prevalence π is given by:

\[ \pi = \frac{\sum w_i y_i}{\sum w_i} \]

where \( w_i \) is the sampling weight of the \( i \)th subject and \( y_i \) is the gold standard after the interview which is 1 if the subject is a ‘true’ case and 0 otherwise. In Ormel et al. (1990), the sampling weight is the same ratio, but the sampling categories increased from two (GHQ+ and GHQ−) to four (GHQ+/GP+, GHQ−/GP+, GHQ+/GP− and GHQ−/GP−) (see Fig. 1 for sampling fractions). The third method was based on the proportion of those scoring above the GHQ-28 threshold: the probable prevalence (see Goldberg 1978; Goldberg and Hillier 1979; Boardman 1987 for further detail). The probable prevalence represents the proportion of patients who obtain a high score on the GHQ after a correction term has been applied. The formula can be written as