Brit. J. Psychiat. (1982), **140**, 174–180

Community Screening for Mental Illness: A Validity Study of the General Health Questionnaire

SIDNEY BENJAMIN, PETER DECALMER and DAVE HARAN

Summary: The 60-item GHQ was validated in a community population by comparison with the CIS. The GHQ failed to identify nearly half of the psychiatric 'cases' in this population. Those missed were similar to those detected except for greater chronicity of illness and more frequent social and interpersonal problems. The GHQ appears to be unsuitable as a screening instrument for mental illness in the community and the possible reasons are discussed. Principal components analysis resulted in a 15-item GHQ factor which, when used with Likert scoring, resulted in considerable improvement and failed to identify only 4 per cent of 'cases'. It is suggested that this may provide a more satisfactory screening instrument.

The General Health Questionnaire (GHQ) (Goldberg, 1972; 1978) was initially developed as a first stage screening instrument for psychiatric illness in order to identify potential 'cases' which could then be verified and the nature of which could be determined by using a second stage instrument such as the Clinical Interview Schedule (CIS) (Goldberg *et al*, 1970). Used in this way the GHQ was found to be an effective means of case identification when validated in a number of studies based on general practice or clinic attenders and administered to the patient as part of the medical consultation. Subsequently it has been widely administered for the purpose of case identification in other settings, including community studies, and to patients admitted to hospital following self-poisoning or for obstetric care. It has also been administered by other means, for example, by post and separately from any other medical assessment. Although it is commonly assumed that the GHQ is equally effective as a screening instrument for mental illness in this wide range of conditions, there is little evidence available to support this. It is questionable as to whether responders to such a questionnaire will behave in an identical manner regardless of how they are identified or the circumstances in which the questionnaire is presented.

A community-based study of the health of women in the fifth decade of life (the collaborative menopause study, to be reported separately) included the GHQ as a first stage screening instrument for mental illness. In view of the reservations stated above, it was considered desirable to validate the GHQ in this population by comparison with the CIS. The purpose of this paper is to present the results of this validation, to identify the limitations of the GHQ in this setting and to examine the means by which these may be overcome.

Method

Ninety-two subjects were involved in the validation study of the GHQ. These were all women aged between 40 and 49 years at the inception of the study. The base population consisted of 2,502 women in this age group who were registered with any one of 26 general practitioners in the Greater Manchester area. A random sample of approximately 1 in 10 included 228 women whose date of birth fell on the 5th, 15th or 25th of the month. It was necessary to exclude many of these as the collaborative menopause study required subjects who were still able to pass through a 'natural' menopause and who could cooperate with multiple investigations of physical, mental and social state. Only those of caucasian origin were included. Of 100 women remaining, eight were either unable or unwilling to participate leaving a total of 92. The distribution of marital status and social class of these subjects is shown in Table I.

The preliminary contact with subjects was made by their general practitioners using a standard form of letter explaining the purpose of the study. Each subject was then seen by the general practitioner who completed a standardized questionnaire about the patient's physical health, took a blood sample and then administered the 60-item GHQ followed by other

https://doi.org/10.1192/bjp.140.2.174 Published online by Cambridge University Press

TABLE I
Social class and marital state (n = 92)

Social class:	
I	10
II	25
III non-manual	17
III manual	26
IV	7
V	6
Unclassified	1
Marital state:	
Married	75
Single	3
Widowed	4
Separated	3
Divorced	7

questionnaires dealing with aspects of mental health. Arrangements were then made for the subject to be seen by one of the two research psychiatrists (P.D. or S.B.) who completed the CIS. This was administered to all subjects, usually within one week of the GHQ, although occasionally a few days later. The interviewer had no knowledge of the GHQ score, nor of the results of other psychological, social and physical assessments being carried out by other members of the research team.

Results

The criteria for 'case' identification resulting from the CIS included total score (the sum of the scores for individual symptoms plus twice the sum of the scores for manifest abnormalities), 'overall severity rating' and clinical diagnosis according to the International Classification of Disease (Eighth Revision, 1968). The number of subjects identified at clinical interview by these different criteria as having mental illness is shown in Table II. An overall severity rating of two and above is highly concordant with clinical diagnosis and these two constitute the indices that conform most closely with 'case' identification in clinical psychiatric practice. A total score of 13 and above provided the closest approximation to the overall severity rating and clinical diagnosis. With few exceptions these three criteria identified the same subjects, the majority of whom had mild to moderate neurotic illnesses

TABLE II
'Cases' identified using different criteria (n = 92)

GHQ	Threshold 9/10	25
	10/11	24
	11/12	23
CIS total score	Threshold 10/11	37
	11/12	34
	12/13	33
	13/14	27
CIS overall severity rate	2 and above	32
Clinical diagnosis (other than personality disorder)		31

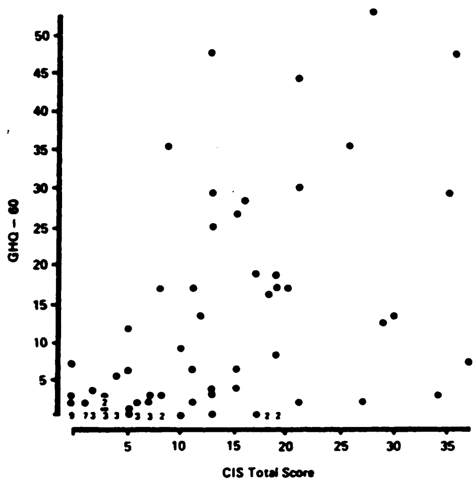

FIG 1a

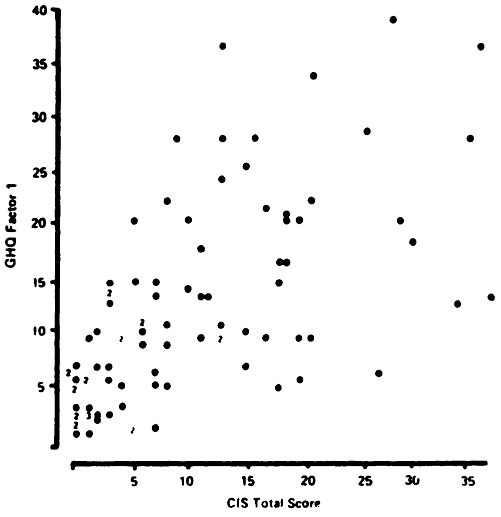

FIG 1b

https://doi.org/10.1192/bjp.140.2.174 Published online by Cambridge University Press

characterized by a mixture of symptoms of anxiety and depression. Analyses for overall severity rating and total score were carried out separately on all data but results were similar and only total score results are presented. The latter provides the better opportunity for examining the effects of adjusting the threshold for 'case' identification.

The correlation of GHQ with CIS total score is +0.63 (Spearman) and with overall severity rating is +0.55. Although they are statistically highly significant ($P < 0.001$), the scattergram (Fig 1a) illustrates the considerable disagreement between these indices of morbidity, and the identification of subjects as 'cases' depends on whether GHQ or clinical interview scores are used. This becomes evident on examining the distribution of high and low scores (Table III). With threshold scores for GHQ of 11/12 and total interview score of 12/13, 23 'cases' are identified by the GHQ of which five are not confirmed on clinical interview ('false positives'). Of 69 'non-cases' identified by the GHQ, no less than 15 are identified as 'cases' on psychiatric interview ('false negatives'). It appears that the GHQ is failing to identify nearly half of the total psychiatric 'cases' in this population. The sensitivity of the test, i.e., its ability to correctly identify 'cases', is only 54.5 per cent and lower than that reported in any other validity study of the 60-item GHQ. By contrast the specificity, i.e., the ability of the test to correctly identify 'non-cases', is 91.5 per cent and compares favourably with other validation studies. The overall misclassification rate is 21.7 per cent.

It might be argued that the 'false negative' group had little wrong with them and should not be considered as 'true cases'. Indeed, of these 15 subjects (Table IV), two received an overall severity rating of only one, indicating that there was no clinically significant disorder. However, the remainder all received an overall severity rating of at least two, indicating a clinically significant 'case'. Three of the four subjects with an overall severity rating of three suffered from chronic and untreated agoraphobia which resulted in considerable disability. The absence of other major symptoms and manifest abnormalities in two of these is indicated by their relatively low total scores (cases 41 and 47).

Tarnopolsky and his colleagues (1979), using the 30-item version of the GHQ, have demonstrated that the sensitivity of the instrument is related to the ratio of high to low scores in the population concerned. Most validation studies of the GHQ have been based on artificial conditions in which the number of high scores has approximated to the number of low scores and in these 'ideal' conditions both specificity and sensitivity are high. They calculated that when the ratio of high to low scores falls from approximately 50 per cent to 22 per cent there is a reduction in sensitivity from 78 per cent to 54 per cent, i.e., approximately the level found in the present study. They suggest that sensitivity can be improved to more acceptable levels by reducing the GHQ threshold for 'case' identification. Thus the threshold needs to be determined independently for each population to which the GHQ is applied. The aim is to improve sensitivity at the expense of specificity as clearly 'false positives' are more acceptable than 'false negatives' in a first stage screening process which is to be followed by a psychiatric interview to confirm potential 'cases'.

We have attempted to apply this procedure to our own data (Table II), but reduction of GHQ threshold to 9/10 makes little difference to the number of 'cases' identified. Table IV shows the GHQ and interview scores for 'false negative' subjects. It becomes clear that the 'false negatives' have not simply arisen as a result of an inappropriate high GHQ threshold or low clinical interview total score threshold as there is a wide-ranging disparity between GHQ and

TABLE III

'Case' identification using GHQ and CIS total score ($n = 92$)

	CIS total score		
GHQ	0–12	13 and above	Total
0–11	'True negative' 54	'False negative' 15	69
12 and above	'False positive' 5	'True positive' 18	23
	Total 'non-cases' 59	Total 'cases' 33	92

https://doi.org/10.1192/bjp.140.2.174 Published online by Cambridge University Press

TABLE IV

GHQ and CIS scores for 15 'false negative' subjects

Subject	GHQ	CIS overall severity rating	CIS total score
2	0	2	20
5	7	2	15
12	0	2	13
34	3	2	27
37	4	3	34
41	4	3	14
47	0	3	17
50	8	3	37
53	1	2	18
54	5	2	15
55	3	1	21
64	0	2	18
66	5	1	13
77	9	2	19
84	10	2	18

interview scores. As many as 12 per cent of the total population are identified as 'cases' on interview but have GHQ scores of five or less.

Attempts to raise the threshold for total interview score are equally unsatisfactory. For example, an increase in total score threshold to 16/17 is required to reduce the 'false negatives' from 15 to 10, but also has the effect of reducing the 'true positives' from 18 to 13. Not only does this leave the same unacceptably low sensitivity (55 per cent), but it also incorrectly identifies as 'false positives' (i.e., 'non-cases') five subjects who are identified as 'cases' on all other available criteria including GHQ, CIS overall severity rating and clinical diagnosis.

We have tried to identify the factors that contribute to 'false negative' status by a comparison of 'false negatives' with both 'true negatives' and 'true positives'. In most respects the 'false negatives' appear similar to the 'true positives' and different from 'true negatives'. This applies to comparisons of age distribution, social class, past history of mental illness and family history of mental illness. It also applies to distribution of scores for individual symptoms and manifest abnormalities and to distribution of total scores for symptoms and manifest abnormalities. Some difference was found in the frequency of social and interpersonal problems acknowledged at interview by subjects in the 'false negative' group as compared with those in the 'true negative' group ($\chi^2 = 15.180$; degrees of freedom = 2; $P < 0.005$). There was also a tendency (which was not statistically significant) for more such problems to be declared by 'false negative' subjects than by 'true positive' subjects ($\chi^2 = 5.10$; degrees of freedom = 2; $P < 0.1$). The differences between 'true positive' and 'true negative' subjects was not significant ($\chi^2 = 2.025$).

The other outstanding feature of the 'false negative' group appeared to be the chronicity of the disorders suffered. Some of these were long-standing but the onset could not be dated for all subjects and a few were acute exacerbations of long-standing disorders. For the 12 whose onset of illness could be approximately dated, the mean duration was 4.8 years (range 1 to 12 years; standard deviation 3.92), and an additional two subjects were noted to have disorders described as 'chronic'. By contrast, 13 of the 'true positives' had a mean duration of illness of 1.4 years (range 1 month to 4 years; standard deviation 1.22 years). Scrutiny of the GHQ responses for the 'false negative' group showed that they frequently checked responses 'about same as usual' or 'no more than usual' indicating no change in their habitual state rather than the absence of symptoms. The standard GHQ scoring fails to differentiate such responses from those indicating total absence of pathology. This may be an important factor in the failure of the GHQ to identify chronic disability in this community sample. Whereas clinic attenders are likely to be characterized by complaint of symptoms, it is likely that there will be others in the community with similar disorders but of longer duration who are non-complainers and have come to accept habitual symptoms as a part of their 'normal' state. Nevertheless, for many purposes it is desirable to identify chronic as well as more acute disorders in community studies and it seems that in its present form the GHQ is an unsatisfactory instrument for this purpose.

Although the standard scoring system of the GHQ has proved to be satisfactory when standardized in clinic attenders, the above findings raise the possibility that alternative score methods which give positive weighting to 'same as usual' responses might result in improved sensitivity in community studies. However, alternative scoring (Table V) makes remarkably little difference to correlations of GHQ and CIS scores and only a moderate improvement in sensitivity, provided that GHQ threshold is maintained at a level which retains reasonably high specificity.

We have also considered whether some form of factor analysis might identify a GHQ factor which is more discriminating in identifying 'cases'. The GHQ data were subjected to principal components analysis. The first eight factors identified accounted for two-thirds of the total variance, and these eight factors were then rotated using the Varimax method. Fifteen items loaded higher than 0.5 on the first rotated factor. Seven of these items belong to Goldberg's 'severity of psychiatric illness' factor (Goldberg, 1972).

https://doi.org/10.1192/bjp.140.2.174 Published online by Cambridge University Press

TABLE V

Effects of alternative scoring for GHQ-60 (n = 92)

GHQ scoring and threshold	Correlation with CIS total score (Spearman)	Overall misclassification rate	Sensitivity	Specificity
Standard GHQ (0–0–1–1) 11/12	+0.61	21.7%	54.6%	91.5%
Modified Likert (0–1–2–2) 42/43	+0.63	25%	63.6%	81.4%
Standard Likert (0–1–2–3) 42/43	+0.63	25%	63.6%	81.4%

TABLE VI

Effects of alternative scoring for GHQ factor 1 (n = 92)

GHQ factor 1 scoring and threshold	Correlation with CIS total score (Spearman)	Overall misclassification rate	Sensitivity	Specificity
Standard GHQ (0–0–1–1) 5/6	+0.61	22.2%	51.5%	93.0%
Modified Likert (0–1–2–2) 9/10	+0.68	30.4%	87.9%	59.3%
Modified Likert (0–1–2–2) 12/13	+0.68	30.4%	69.7%	69.5%
Standard Likert (0–1–2–3) 12/13	+0.68	28.3%	66.7%	74.6%

This first factor resulting from our data appears to represent a general factor of morbidity and total scores resulting from the items in this factor show a remarkably high correlation with the total scores of the GHQ–60 (r = +0.90). Further details of the principal components analysis after Varimax rotation are shown in the Appendix.

Scores resulting from the fifteen items of the first GHQ factor, using both standard and alternative scoring methods, have been correlated with CIS total scores (Table VI). Both standard and modified Likert scoring provided the highest overall correlation (+0.68) of any obtained between GHQ and CIS scores, although the overall misclassification rate was increased. The resultant sensitivity and specificity could then be modified by adjusting the GHQ factor 1 threshold score. With a threshold of 9/10 sensitivity was increased to 88 per cent at the expense of a reduction in specificity to approximately 60 per cent. If the GHQ factor 1 threshold is raised to 12/13, both sensitivity and specificity are approximately 70 per cent. Modified Likert scoring provides slightly higher sensitivity than standard Likert scoring, but the latter is marginally superior with regard to overall misclassification rate and specificity. Thus for this population the 15 items with high loading on factor 1, in conjunction with Likert scoring, provided the best positive case identification at the expense of a moderate reduction in specificity, which could be modified by selecting a suitable GHQ threshold score. Fig 1b shows the scatter of GHQ factor 1 with standard Likert scoring and CIS total score, and in contrast with Fig 1a illustrates the reduction in 'false negatives'.

Discussion

The GHQ–60 correctly identified only about half of the psychiatric 'cases' found at interview in the collaborative menopause study and therefore appears to be of limited value as a screening instrument in this sample. The validity of the GHQ–60 has not previously been determined in community samples. However, validity studies of the GHQ–30 in nonconsultings settings (Mann, 1977; Tarnopolsky *et al*, 1979) show lower sensitivity when compared with studies of clinic attenders, and it seems likely that the low sensitivity in our investigation is related to the community setting. From our findings it appears that the most important factor in determining 'false negative' status was chronicity of illness. Subjects who have become accustomed to their long-standing symptoms are not identified by the GHQ in view of the way in which questions are necessarily worded and scored. Such subjects are likely to be relatively fewer amongst clinic attenders and validity studies of such subjects result in higher sensitivity.

Goldberg (1978) has drawn attention to the tendency for the GHQ to miss those with chronic illness. In a consulting setting the 'false negative' rate is increased from 1.7 per cent in those who have been ill for less than one month to 18.4 per cent in those who have been ill for more than one year. He suggested that in settings in which a high proportion have long-standing disorders (e.g., a psychiatric out-patient 'support' clinic) it may be possible to compensate by lowering the threshold score. Unfortunately, this was not possible in this community setting where there was a mixture of acute and chronic illness and modifications both to threshold and scoring methods were unhelpful. The fact is that those with long-standing illness who attend any clinic are 'consulting' whereas most of those in the community are not, and the latter provide fewer positive responses on the GHQ. It seems likely that this particular combination of chronicity with non-consultation is especially difficult to detect by questionnaire and is an important source of bias in screening community populations.

Our sample is extremely restricted compared with other validation samples, in terms of age and sex. However, validation studies of the GHQ–30 in a consulting setting (Goldberg, 1978) do not appear to be affected by these demographic variables, and there is no evidence to suggest that they might act differently in a community sample. Both the demographic features and the community setting are likely to have been important in determining the different factor structure in this sample compared with those previously published (Goldberg, 1972).

It might be argued that the 'cases' not identified by the GHQ are less relevant to a community study, but apart from chronicity they have been shown to be similar in most other respects to those who were identified as 'cases' and different from those identified as 'non-cases'. The relative excess of social and interpersonal problems identified in these 'false negatives' is concordant with the relatively poorer prognosis for neurotic illness found in those with chronic social problems and personality disorders (Huxley *et al*, 1979).

The total morbidity identified in our sample by the GHQ is remarkably similar to that found in a sample similar with regard to sex, social and marital state and geographical location. Goldberg *et al* (1976), in a community sample of 124 women in South Manchester, found 25.8 per cent had GHQ scores of 12 or more. In our own study the comparable figure is 23 per cent, but prevalence based on the CIS is increased to approximately 32 per cent. Whilst this estimate may seem high, the criteria for exclusion applied to our sample, and also the non-responders, are likely to have reduced the sample prevalence. Morbidity in the population is almost certainly greater than this.

The results of the principal component analysis raises the possibility of using a short (15-item) version of the GHQ with Likert scoring and considerably increased sensitivity, at the expense of a moderate reduction in specificity. On balance this is a more acceptable form of misclassification, particularly if 'cases' identified by the questionnaire are subsequently interviewed. This shortened GHQ, which correlates highly with the 60-item GHQ, would also require less time to administer and would be advantageous in large scale surveys. Validation of the 15 items derived from the GHQ would of course be necessary in an independent community sample, preferably of both sexes and wider age range, before use as a screening instrument. In the present sample this procedure resulted in failure to identify only 4 per cent of 'cases' and this would be highly satisfactory in spite of 26 per cent 'false positives' if used as the first stage of a two stage screening process. An alternative approach might be to use a short form of the GHQ together with a separate brief questionnaire specifically designed to detect chronic illness in 'non-consulters'. It should be recognized that as a screening instrument for mental illness in community populations the GHQ–60 is likely to result in a serious underestimate of prevalence, with a particular bias against the identification of long-standing illness.

Acknowledgements

This research was supported by a grant from the North Western Regional Health Authority. We also wish to thank Dr Gareth Lloyd and Professor David Metcalfe, the coordinators of the collaborative menopause study, and Professor David Goldberg for his valuable comments on the paper.

References

GOLDBERG, D. P. (1972) *The Detection of Psychiatric Illness by Questionnaire*. Maudsley Monograph No 21. London: Oxford University Press.

https://doi.org/10.1192/bjp.140.2.174 Published online by Cambridge University Press

—— (1978) *Manual of the General Health Questionnaire.* Windsor: NFER Publishing Company.

—— COOPER, B., EASTWOOD, M., KEDWARD, H. & SHEPHERD, M. (1970) A psychiatric interview suitable for use in community surveys. *British Journal of Social and Preventive Medicine*, **24**, 18–26.

—— KAY, C. & THOMPSON, L. (1976) Psychiatric morbidity in general practice and the community. *Psychological Medicine*, **6**, 565–9.

HUXLEY, P., GOLDBERG, D. P., MAGUIRE, G. P. & KINCEY, V. A. (1979) The prediction of the course of minor psychiatric disorders. *British Journal of Psychiatry*, **135**, 535–43.

MANN, A. H. (1977) The psychological effect of a screening programme and clinical trial for hypertension upon the participants. *Psychological Medicine*, **7**, 431–8.

TARNOPOLSKY, A., HAND, D. J., MCLEAN, E. K., ROBERTS, H. & WIGGINS, R. D. (1979) Validity and uses of a screening questionnaire (GHQ) in the community. *British Journal of Psychiatry*, **134**, 508–15.

Appendix

Factor analysis of General Health Questionnaire

Varimax rotation of first eight factors (items with loadings greater than 0.50) accounting for 66 per cent of total variance.

Factor 1 (21 per cent of variance)
Been feeling in need of a good tonic?
Been feeling run down and out of sorts?
Felt that you are playing a useful part in things?
Felt you're just not able to make a start on anything?
Felt yourself dreading everything that you have to do?
Felt constantly under strain?
Felt you couldn't overcome your difficulties?
Been taking things hard?
Been getting edgy and bad-tempered?
Been getting scared or panicky for no good reason?
Been finding life a struggle all the time?
Found everything getting on top of you?
Been feeling unhappy and depressed?
Been losing confidence in yourself?
Been feeling nervous and strung-up all the time?

Factor 2 (21 per cent of variance)
Been feeling perfectly well and in good health?
Been unable to concentrate on whatever you're doing?
Been feeling mentally alert and wide awake?
Been feeling full of energy?
Been taking longer over the things you do?
Felt on the whole you were doing things well?
Been satisfied with the way you've carried out your task?
Felt capable of making decisions about things?
Been able to enjoy your normal day-to-day activities?
Been able to face up to your problems?
Been feeling reasonably happy, all things considered?

Factor 3 (14 per cent of variance)
Afraid you are going to collapse in public.
Feel people are looking at you (also Factor 5).
Life entirely hopeless.
Hopeful about future.
Life isn't worth living.
Make away with yourself.
Wishing you were dead.
Idea of taking life.

Factor 4 (13 per cent of variance)
Waking early.
Too tired and exhausted to eat.
Difficulty getting to sleep.
Difficulty staying asleep.
Restless disturbed nights.

Factor 5 (12 per cent of variance)
Frightening dreams.
Losing interest in personal appearance.
Late getting to work.
Afraid to say anything.
People looking at you (also Factor 3).

Factor 6 (7 per cent of variance)
Pains in your head.
Tightness or pressure in head.
Couldn't do anything because nerves too bad.

Factor 7 (6 per cent of variance)
Hot or cold spells.
Perspiring a lot.

Factor 8 (6 per cent of variance)
Managing to keep busy.
Managing as well as most people.

Sidney Benjamin, M.D., M.Phil., F.R.C.Psych., *Senior Lecturer in Psychiatry, University of Manchester and Honorary Consultant Psychiatrist, Department of Psychiatry, Manchester Royal Infirmary, Swinton Grove, Manchester M13 0EU*

Peter B. S. Decalmer, M.B., Ch.B., M.R.C.Psych., D.Obst.R.C.O.G., *Senior Registrar in Psychiatry, The University Hospital of South Manchester, West Didsbury, Manchester M20 8LR*

Dave Haran, M.Ed., *Senior Research Officer, Department of Epidemiology and Social Research, Christie Hospital, Manchester M20 9BX*

(*Received 6 July 1981*)

https://doi.org/10.1192/bjp.140.2.174 Published online by Cambridge University Press