Mining Important Symptoms of Adult Depression

Subhagata Chattopadhyay
Bankura Unnayani Institute of Engineering, India

Fethi A. Rabhi
School of CSE, The University of New South Wales, Sydney NSW, Australia

Sangeeta Bhanjachaudhuri
National Institute of Science and Technology, India

Sarat Kumar Jena
Indian Institute of Technology, Kharagpur India

ABSTRACT

Screening and diagnoses of psychiatric diseases are conducted by interviews, where each question aims to capture the information of the symptoms to assess the severity of the illness. The objectives of this paper are to mine important questions (Q) of a depression questionnaire (Qs) as well as significant symptoms, captured through the questions. It proposes that, in this way, the number of questions could be reduced. To examine this hypothesis, one hundred and twenty six subjects suffering from depression are interviewed. Answers are quantified using a 3-point scale: 'Symptom absent'=0, 'Not sure about the symptom'=0.5, and 'Symptom present'=1. Factor analysis is then considered to mine significant questions (Q), followed by a regression study to note the goodness of the proposed model. Results show that the questionnaire is internally consistent and reliable (α=0.79, average '*r*'=0.94 minute, and '*rc*'=51 minute) and factor analysis yields ten significant Components (C), among which C1 is found as the most significant, which captures Q15, 25, 16, 24 and 23 capturing the symptoms such as 'loss of libido', 'self blame and self criticism', 'loss of appetite', 'Individualism', 'distortion of body image', respectively. Regression analysis shows that the model is of a good fit (training R=0.95, test and validations R=~1.0, each). The paper concludes that mining important symptoms through 'Q' is possible and it reduces the complexity of interpretations. Such information could be helpful for depression diagnoses as during the treatment medical doctors may monitor these symptoms as the markers of improvement.

General Terms

Pattern recognition, mental health, depression, questionnaire, significant symptoms.

Keywords

Depression; Questionnaire; Diagnosis; Statistics; Model fit; Factor analysis

1. INTRODUCTION

Depression is a psychological disorder affecting a large number of populations, globally [1]. Demographic studies on the lifetime prevalence of diagnosable depression show that women suffer more than men [2]. However, in both sexes, persistent 'mood fluctuation' is the principal symptom [3]. Studies have revealed several key causes of depression, such as poverty [4], broken homes [5], substance abuse [6], chronic illnesses [7], and so forth and hence, it is considered to be the most common psychological disorder in the global population.
Screening depression is an important task; because it might lead to suicide [8-11]. Due to prevailing issues, it is also a complex task. One issue is that, the symptoms are inconsistent in nature and varies across populations. The other issue lies in the variations of doctors' perceptions in handling those symptoms. It may lead to 'under' or 'over' diagnoses, despite of the use of several questionnaire-based screening tools [12-15], available. These prevailing issues yield ample scope to apply informatics in depression research; however, till date the number of attempts does not score much [16][17].
Data mining techniques are very popular for analyzing large, non-linear and complex clinical data, which is also encountered in the clinical medicine [18]. Hence, mining such data would be challenging and it is the principal focus of our paper.
Precisely, the aim of this study is to mine important questions (Q) from a given questionnaire (Qs) and at the same time the significant symptoms, embedded within the Q. Authors propose that, in this way, the complexity of the symptoms as well as the Q could be better handled in mental health study.

Rest of the paper is organized as follows. Section 2 describes the methodology. Results are shown and discussed in section 3. Finally, section 4 concludes the paper and directs future scope of extension of this study.

2. EARLIER WORK

This paper utilizes a questionnaire, which would be able to capture the symptoms and mine the significant symptoms efficiently [19][20]. At the *first step*, fifteen symptoms have been identified from the available literature. The symptoms are categorized under four major constructs, stated below:

1. '*Emotional*' [21]: The construct 'Emotional' consists of 'Dejected mood' [22], 'Negative feelings about self' [23], 'Reduction in gratification' [24] and 'Loss of emotional attachment' [25].
2. '*Cognitive*' [26]: 'Cognitive' consists of four indicators, such as 'Negative expression' [27], 'Self-blame and self-criticism' [28], 'Distortion of body image' [29] and 'Indecisiveness' [30].
3. '*Motivational*' [31]: 'Motivational' consists of 'Avoidance' [31], 'Suicidal Wishes' [32], and 'Paralysis of Will' [33].
4. '*Vegetative*' [34]: Lastly, under the 'Vegetative' construct, there are three symptoms - 'Loss of appetite' [35], 'Loss of libido' [36], 'Fatigability' [37], and 'Sleep disturbances' [38].

Based on the chosen symptoms, simple, straight forward and closed-ended 26 questions (Qs) were used in the *second step*. The questions are framed with the help of ten experts with mean experience of 7 years and 4 months. It aims to capture the information of each symptom (see table 1). To assess the quantified load of each symptom, a three-point scoring has been used, e.g., the symptom is 'absent' (0), the subject is 'not sure' of its presence (0.5), and the symptom is 'Present' (1.0).

Table 1. Symptoms captured by the Qs [19][20]

No.	Symptoms	Qs.
1	Dejected mood (DM)	Q1
2	Negative expression (NE)	Q2, 9
3	Reduction in gratification (RG)	Q3, 4, 5
4	Loss of emotional attachment (LEA)	Q6, 7
5	Negative feeling about self (NF)	Q8, 10, 26
6	Sleep disturbances (SD)	Q11, 12
7	Fatigability (F)	Q13, 14
8	Loss of libido (LL)	Q15
9	Loss of appetite (LA)	Q16, 17
10	Paralysis of will (PW)	Q18, 19
11	Suicidal intents (SI)	Q20, 21
12	Avoidance (A)	Q22
13	Distortion of body image (DBI)	Q23
14	Individualism (I)	Q24
15	Self-blame and self-criticism (SBC)	Q25

3. METHODOLOGY

The study had been conducted during 2010 (full one year) in two Indian hospitals. Using the questionnaire (Qs), 126 anonymous subjects were interviewed taking appropriate ethical measures. It is important to note that the Qs contain 26 questions. Three representative interviews were conducted each by three experienced psychologists. It is also important to note that those interviewers had not participated in Qs generation. Subjects were randomly chosen which was consisted of 40% males and 60% females with ages between 19 to 30 years (mean age 22 years and 5 months). Each interview took between 10 to 25 minutes (average 12 minutes) to complete (mean 10 minutes for males and 14 minutes for females). It is also worth noting that, two consecutive interviews were conducted at a gap of 30 minutes to avoid (i) repetitions of the answers and (ii) the effect of medications on the subjects. The result of each interview was undisclosed to the peer psychologist to prevent induced biasness and personal influences on each other.

The Qs was tested for internal consistency using *Cronbach's alpha* (α) [39]. It is expressed with equation 1.

$$\alpha = \frac{k(\mathrm{cov}/\mathrm{var})}{1+\{(k-1)(\mathrm{cov}/\mathrm{var})\}} \ldots(1)$$

Here, '*k*' denotes 126 cases, while, '*cov*' and '*var*' denote covariance and variance of the answers to the Qs as [0, 0.5, 1]. *Spearman's rank correlation* [40] has also been measured (see equation 2) between interviews 1 and 2, 1 and 3 and 2 and 3 to measure the cross correlations (see equation 2).

$$r = 1 - \frac{6\sum D^2}{n(n^2-1)} \ldots(2)$$

In this equation, '*r*' denotes the Spearman's rank correlation, '*D*' represents the difference between pair of estimates in any two corresponding rank characteristics, and '*n*' refers to the number of paired observations. Together (i.e., α and r), the reliability of the questionnaire has been assessed.

In the next phase, factor analysis was performed with Kaiser-Meyer-Olkin (KMO) and Bartlett's Test [41] to engineer significant information (i.e., Qs and the relevant symptoms), which is the key focus of this work. Another usefulness of this technique is that, we can reduce the data dimension and thus the involved complexity with the clinical data [41]. Simultaneously, a Factor analysis has been conducted, which is a collection of methods, used to examine how underlying symptoms influence the responses on a number of measured variables. Generally factor analyses are performed through identifying and capturing the pattern of covariance (or correlations) between the observed measures. Symptoms, which are highly correlated with each other (whether positively or negatively) are pruned to reduce the dimension. In this study, 126 responses were used to examine the validity and reliability of the scale to obtain a quantitative and statistically proven identification of the responses. The proof for quantitative variable was conducted by factor analysis on 26 proposed items (i.e., the answers to 26 Qs) using the 'Varimax rotation' (*V*) [42]. In this context, it is worth noting that the value of KMO measures the sampling adequacy, and the 'Varimax rotation' is an orthogonal rotation that minimizes the number of symptoms that have high loadings on each symptom. This method simplifies the interpretation of the factor. 'V' is calculated using equation (3) as follows,

$$V = \sum_{k=1}^{M} \left\{ \frac{\sum_{j=1} \left(\frac{l_{jk}}{h_j}\right)^4}{N} - \left[\frac{\sum_{j=1} \left(\frac{l_{jk}}{h_j}\right)^2}{N} \right]^2 \right\} \ldots(3)$$

In this equation, the term '*k*' denotes answers to the 26 Qs and '*j*' is the number of cases (i.e., 126). Now the new matrix becomes 126×26 and square of h_j is the commonality measure, which mathematically represents the sum of the square of factor loading. It is important to note that in each 'C', Qs are embedded. At the same time each Q captures the corresponding symptoms (see Fig.1). This concept has been discussed in detail in the next section.

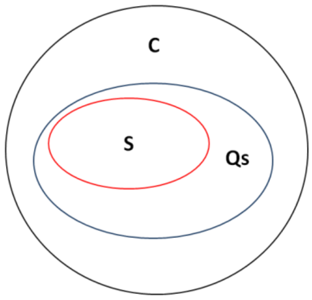

FIG. 1 THE RELATIONSHIPS AMONG COMPONENTS (C), QUESTIONS (QS), AND SYMPTOMS (S) AFTER DATA MINING.

At the end, multiple regressions have been performed to validate the fitness of the model.

4. RESULTS AND DISCUSSIONS

The internal consistency of the Qs has been checked by measuring *Cronbach's alpha* (α), *Spearman's correlation coefficient* (r), and the *repeatability coefficient* (rc). The study reveals that α is 0.79, which is well above the threshold 0.7 [43]. The mean difference between a pair of interviews i.e., '*r*' = 7.75 minutes, with an '*rc*' = 9.5 minutes for the whole sample (for all CI 95%; $p < 0.001$).

Factor analysis was carried out to find the most significant components (C) of depression so that the redundant Qs could be pruned and the data dimension is reduced. In order to accomplish the task, the weighted answers of 26 Qs were analysed by IBM SPSS19.0 software. Among the 26 Qs, 22 Qs were loaded more than 0.5 and taken into consideration for further study, while the rest of the items are pruned (e.g., Q6, Q14, Q13 and Q18). These 22 Qs were then categorized under 10 'C's by the software itself. The value of *KMO* had been used for measuring sampling adequacy. We found that the KMO value is 0.643, indicating that the factor analysis test proceeded correctly and the sample size used was adequate. It is important to note that the minimum acceptable value of KMO is 0.5 [43]. Therefore, it can be argued that the matrix neither suffer from the multi-co-linearity nor singularity. From Bartlett test of specificity, we have seen that the factor analysis process is correct and suitable for multidimensionality control. Table 2 presents ten 'C's and its corresponding factor loadings are given for each question. Here, it may be noted that Q15, Q25, Q16, Q24, and Q23 are dealt with C1, which is the most significant 'C'. It is also important to note that the internal consistency might be reduced after pruning of items (i.e., Qs). To investigate this possibility, Cronbach's alpha (α) has been repeated and the value of α is found as 0.72, which indicates that there is not disturbance in the internal consistency [43].

Table 2. Factor loading scores components & respective Qs.

	COMPONENTS (C)									
Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Q15	.714									
Q25	.664									
Q16	.588									
Q24	.531									
Q23	.506									
Q4		.787								
Q17		.633								
Q5		.628								
Q21			.701							
Q20			.689							
Q8			.551							
Q22			.506							
Q7				.742						
Q26				.669						
Q1					.740					
Q2					.612					
Q11						.832				
Q10						.553				
Q3							.792			
Q19								.777		
Q12									.832	
Q9										.837

Table 3 shows the percentage variation (*V%*) explained by factor analysis with '*V*'.

Table 3. % of variations (V%) explained by factor analysis

C	V%	Rank
1	17.689	1
2	7.389	2
3	6.819	3
4	5.971	4
5	5.467	5
6	5.001	6
7	4.597	7
8	4.427	8
9	4.115	9
10	3.967	10

To note the model fit and the correlation among the components, multiple regressions have been conducted. Here, the independent factors are the answers to the Qs and the dependent parameter is the probable state of depression (given by the domain experts). A sample data (first instance of the 126 instances) has been shown below to understand its structure. Here, Qs are the independent factors and the DS is the dependent factor, which denotes the state of severity. For this case, the severity is 0.7. The shaded Qs are the significant Qs, obtained indirectly by factor analysis.

Table 4. 1st sample data and the significant Qs.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
1	1	1	1	1	1	1	0	1
Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18
1	0.5	0.5	1	1	0.5	0.5	1	1
Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	**DS**
1	0	0	1	1	1	0	1	**0.7**

Multiple regressions have been performed with the significant Qs (shaded in Table 4) to note the correlations among the computed and target outputs. Table 5 presents the significant symptoms, obtained by regressions. In Fig. 2, plots show high correlations in training, validation, and test cases. For the training sets, the R (i.e. correlation coefficient) value is 0.95, while for the test and validation data, there are perfect fits (R is close to 1.0). Overall correlation coefficient is 0.68621, which is substantial to validate the performance of the developed model.

Table 5. Significant symptoms, thus obtained

Qs	**Symptom**
15	Loss of libido (LL)
16	Loss of appetite (LA)
23	Distortion of body image (DBI)
24	Individualism (I)
25	Self-blame and self-criticism (SBC)

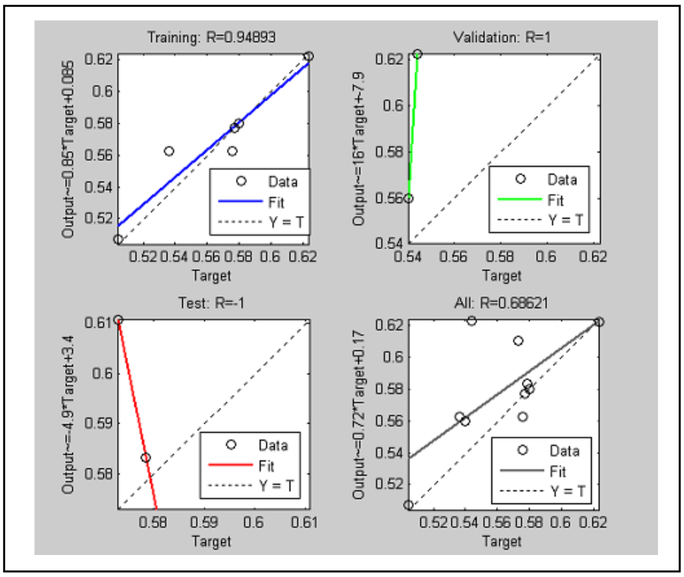

FIG 2. REGRESSION PLOTS OF THE MODEL.

5. CONCLUSIONS AND SCOPE OF FUTURE WORK

The motivation of this paper is to investigate the utility of a depression questionnaire to capture a set of important symptoms. The 'α' (0.79 and 0.72), '*r*' (7.75 minutes) and '*rc*' (9.5 minutes) values substantiate that the questionnaire is reliable and consistent.

Factor analysis is able to extract important components (e.g., C1 for this case), which point towards the related Qs (e.g., 15, 25, 16, 24, 23), embedded into C1. These could be treated as significant Qs and rest can be pruned. It helps reducing the data dimension without losing much of the internal consistency. The authors mare making such a statement after repeating the α test; the result of which is 0.72, i.e., confirmatory. This is certainly an advantage to deal with high dimensional mental health data.

Once the important Qs are identified, it is then easy to get the important symptoms, which might be considered with greater significance to note the effectiveness of the treatment.

Moreover, such a Qs-mining approach could be pioneering for the construction of the clinical decision support systems for the diagnosis and grading of mental illnesses.

6. ACKNOWLEDGEMENT

The authors gratefully acknowledge the contributions of the fellow psychologists and psychiatrists in the construction of the clinical model of depression through interview-based data collection.

7. REFERENCES

[1] Chen, R., Copeland, J. Epidemiology of Depression: Prevalence and Incidence, in *Principles and Practice of Geriatric Psychiatry*, Third Edition (eds. M. T. Abou-Saleh, C. Katona and A. Kumar), John Wiley Sons, Ltd, Chichester, UK.2010. doi: 10.1002/9780470669600.ch77.

[2] Essau C. A., Lewinsohn P. M., Seeley J. R., Sasagawa S. Gender differences in the developmental course of depression. *Journal of Affective Disorders*, 2010; 127:185-190.

[3] Querques, J. Kontos, N. An approach to the patient with dysregulated mood: major depression and bipolar disorder. *Medical Clinics of North America*, 2010; 94:1117-1126.

[4] Simmons L. A., Braun B., Charnigo R., Havens J. R., Wright D. W. Depression and poverty among rural women: a relationship of social causation or social selection? *Journal of Rural Health*, 2008, 24(3):292-298.

[5] Farzaneh E., Mehrpour O., Alfred S., Moghaddam H. H., Behnoush B., Seghatoleslam T. Self-poisoning suicide attempts among students in Tehran, Iran. *Psychiatria Danubina*, 2010, 22:34-38.

[6] Olfson M., Tobin J. N., Cassells A., Weissman M. Improving the detection of drug abuse, alcohol abuse, and depression in community health centers. *J Health Care Poor Underserved*, 2003, 14(3):386-402.

[7] Katon W. J. Epidemiology and treatment of depression in patients with chronic medical illness. *Dialogues in Clinical Neuroscience*, 2011; 13:7-23.

[8] Chattopadhyay S., Daneshgar F. –"A Study on Suicidal Risks in Psychiatric Adults". *International Journal of Biomedical Engineering and Technology* (2011)*;* 5(4): pp. 390-408.

[9] Chattopadhyay S., Ray P., Chen H. S., Lee M. B., Chiang H. C. 'Suicidal Risk Evaluation using a Similarity-based Classifier'. Proceedings of *International Conference on Advanced Data Mining and Applications (ADMA)* Chengdu, China in book format *Lecture Notes in Artificial Intelligence (LNAI)* (2008), Springer, 5139: pp 51-61.

[10] Chattopadhyay S., Ray P, Lee M. B, Chen H. S. – 'Towards the Design of an e-Health System for Suicide Prevention'. In proceedings of *11th IASTED Conference on Artificial Intelligence and Soft Computing (ASC 2007), ACTA Press*, pp. 191-196, Palma de Malorca, Spain (2007)

[11] Chattopadhyay S. – 'A Study on Suicidal Risk Analysis'. In proceedings of *9th IEEE Intl. Conf. on e-Health Networking, Applications and Service (HEALTHCOM 2007)*, pp. 74-79, Taipei, Taiwan (2007).

[12] Hamilton, M. A. Rating Scale for Depression. *Journal of Neurology and Neurosurgery in Psychiatry*, 1960, 23:56-62.

[13] Zung, W. W. K. The depression status inventory: An adjunct to the self-rating depression scale. *Journal of Clinical Psychology*, 1972, 28:539 – 543.

[14] Beck, A. T., Alford, B. A.. Depression: Causes and Treatment, 2nd Edition; University of Pennsylvania Press.Depression Guideline Panel (1993). 'Depression in primary care: Detection and diagnosis'. Volume 1; AHCPR publication No. 93-0550, *National Library of Medicine DOCLINE Information: CAT/9411422*. 2008.

[15] Lindstrom, E., Lewander, T., Malm, U., Malt, U. F, Lublin H., Ahlfors, U. G. Patient-rated versus clinician-rated side effects of drug treatment in schizophrenia - Clinical validation of a self-rating version of the UKU Side Effect Rating Scale (UKU-SERS-Pat) *Nordic Journal of Psychiatry, 2001;* 55: 5-69.

[16] Chattopadhyay S. Neurofuzzy models to automate the grades of old-age depression, *Expert Systems: the Journal of Knowledge Engineering*, accepted in press 2012.

[17] Chattopadhyay S. A prototype depression tool for rural healthcare: a step towards e-health informatics, *Journal of Medical Imaging and Health Informatics*, 2(3): 244-249 (2012).

[18] Satapathy S., Chattopadhyay S. Observation-Prevention of Cardiac Risk Factors: an Indian Study, *Journal of Medical Imaging and Health Informatics* (2012); 2(2):102-113.

[19] Chattopadhyay S., Kaur P., Rabhi F. A., Acharya U. R. Neural Network Approaches to Grade Adult Depression, *Journal of Medical Systems* (2012), 36(5): 2803-2815.

[20] Chattopadhyay S., Kaur P., Rabhi F. A, Acharya U.R. An Automated System to Diagnose the Severity of Adult Depression". In the *proceedings* of *2nd International Conference on Emerging Applications of Information Technology (CSI* EAIT-2011), 19-20th February Kolkata India, (Editors - Jana D., and Pal P.), pp. 121-124. IEEE Computer Society and Conference Publishing Services

[21] Ball, H. A., McGuffin, P., Farmer, A. E. Attributional style and depression. *British Journal of Psychiatry,* 2008; 192: 275 – 278.

[22] Wollershiem, J. P. Depression, women, and the workplace, *Occupational Medicine*. 1993; 8: 787-95.

[23] Showers, C., Ruben, C. Distinguishing Defensive Pessimism from Depression: Coping Mechanisms', *Cognitive Therapy and Research.* 1990;14: 385-399.

[24] Hirschfeld, R. M., Klerman, G. L., Chodoff, P., Korchin, S., Barrett, J. Dependency: Self-Esteem—Clinical

Depression. *Journal of American Academy of Psychoanalysis*, 1976; 4: 373-388.

[25] Ainsworth, M. D. S., Eichberg G. C. Effects on infant-mother attachment of mother's experience related to loss of an attachment figure. In: C. M. Parkes, J. Stevenson-Hinde, and P. Marris (eds.), Attachment across the life cycle. Routledge., *New York*, 1991, pp. 160-183.

[26] Austin, M. P., Mitchell, P., Goodwin, G. M. Cognitive deficits in depression: Possible implications for functional neuropathology. *British Journal of Psychiatry*, 2001; 178: 200 – 206.

[27] Plumb, M. M., Holland, J. Comparative studies of psychological function in patients with advanced cancer--I. Self-reported depressive symptoms, *Psychosomatic Medicine.* 1977; 39: 264-276.

[28] Coyne, J. C., Fechner-Bates, S., Schwenk, T. L. Prevalence, nature, and comorbidity of depressive disorders in primary care. *General Hospital Psychiatry,* 1994; 16, 267-76.

[29] McCaulay, M., Mintz, L., Glenn, A. A. Body image, self-esteem, and depression-proneness: Closing the gender gap, *Springer Netherlands.* 1992; 18:381-391.

[30] Forsell, Y., Jorm, A. F., Winblad, B. Association of age, sex, cognitive dysfunction, and disability with major depressive symptoms in an elderly sample. *American Journal of Psychiatry,* 1994; 151:1600-1604.

[31] Kemp, A., Rawlingsm, E. I., Green, G. L. Post-traumatic stress disorder (PTSD) in battered women: A shelter sample. *Journal of Traumatic Stress,* 1991; 4: 137-148.

[32] Minkoff, K., Bergman, E., Beck, A. T., Beck, R. Hopelessness, Depression, and Attempted Suicide, *American Journal of Psychiatry.* 1973; 130: 455-459.

[33] Aldrich M. S., *Beck*, R. W., Alber, J. W. Familial recurrent Bell's palsy with with ocular motor palsies. *Neurology,* 1987; 37(8) :1369-71.

[34] Aneshensel, C. S., Frerichs, R. R., Huba, G. J. Depression and Physical illness: A multiwave, no recursive causal model. *Journal of Health and Social Behavior.* 1984; 25: 350-371.

[35] Casper, R. C., Redmond, D. E., Katz, M. M., Schaffer, C. B., Davis, J. M., Koslow, S. H. Somatic Symptoms in Primary Affective Disorder; Presence and Relationship to the Classification of Depression, *Archives of General Psychiatry.* 1985; 42: 1098-1104.

[36] Mathew, R. J., Weinman, M.L. Sexual dysfunctions in depression. *Archives of Sexual Behavior.* 1982; 11: 323-328.

[37] Walford, G. A., Nelson, W. M., McCluskey, D. R. Fatigue depression, and social adjustment in chronic fatigue syndrome, *Archives of Disease in Childhood,* 1993; 68: 384-388.

[38] Livingston, G., Blizard B., Mann A. Does sleep disturbance predict depression in elderly people? A study in inner London. *British Journal of General Practice,* 1993; 43: 445-448.

[39] Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika,* 1951; 16:297-334.

[40] Spearmann C. 'The proof and measurement of association between two things', *American Journal of Psychology,* 1902; 15:72-101.

[41] Kim, J. O., Mueller, C. W. Introduction to factor analysis: What it is and how to do it. *Beverly Hills, CA: Sage,* 1978.

[42] Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis, *Psychometrika,* 1958, 23: 187–200.

[43] Nunnally, J. C. Psychometric theory (2nd ed.). *McGrawHill., New York,* 1978.