Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Reliability and validity of the Global Assessment of Functioning Scale in clinical outpatients with depressive disorders

Esther M. V. Grootenboer MD,[1] Erik J. Giltay MD PhD,[2] Rosalind van der Lem MD,[3] Tineke van Veen MD PhD,[4] Nic J. A. van der Wee MD PhD[5] and Frans G. Zitman MD PhD[6]

[1]Psychiatrist, Rivierduinen, Leiden, The Netherlands
[2]Psychiatrist, [4]Epidemiologist, Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands
[3]Psychiatrist, Rivierduinen, Leiden, The Netherlands Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands
[5]Psychiatrist, Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands Leiden Institute for Brain and Cognition, Leiden, The Netherlands
[6]Professor in Psychiatry, Rivierduinen, Leiden, The Netherlands Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands

Keywords
Diagnostic and Statistical Manual of Mental Disorders Axis V, Global Assessment of Functioning Scale, inter-rater reliability, mood disorder, psychometric analysis, routine outcome monitoring

Correspondence
Ms Esther M.V. Grootenboer
Rivierduinen
afdeling ambulante volwassenenzorg
Albinusdreef 7
2300 AT Leiden
The Netherlands
E-mail: emv_grootenboer@email.com

Accepted for publication: 10 November 2010

doi:10.1111/j.1365-2753.2010.01614.x

Abstract

Rationale, aims and objectives The Global Assessment of Functioning Scale (GAF) is widely used to assess psychological, social and occupational functioning. The validity and reliability of the GAF in clinical practice have only scarcely been studied in naturalistic samples.

Methods A total of 432 outpatients with a current major depressive disorder (MDD) were evaluated with routine outcome monitoring (ROM). At baseline the GAF score was assessed by the treating clinician and at ROM baseline and follow-up sessions also by a trained test nurse. Sociodemographic data, the Mini International Neuropsychiatric Interview Plus and scores on the Montgomery-Äsberg Depression Rating Scale, Beck Depression Inventory-revised, Brief Symptom Inventory and Short Form-36 were assessed.

Results At baseline, the mean GAF score by the clinician was 54.8 (range 35–85), and this was systematically lower than the mean GAF score by the test nurse of 57.5 (range 31–88). GAF scores by the clinician and test nurse correlated weakly ($r = 0.26$). The GAF scores of the clinicians correlated strongly with disease severity, and social and physical functioning.

Conclusion The GAF showed rather poor inter-rater reliability as well as poor discriminant validity with disease severity and physical limitations in a large naturalistic sample of outpatients with MDD.

Introduction

The Global Assessment of Functioning Scale (GAF) is designed as an easy and brief measure that integrates within a single score three different dimensions of the individual's level of functioning: psychological, social and occupational. The purpose of the GAF is to represent the patient's current (i.e. for the past week) level of functioning and to track the clinical process of that patient with little demand on the clinician's time [1]. The GAF score ranges from 1 to 100, with lower levels reflecting more disablement within the dimension that is most adversely affected. The GAF is widely used in clinical practice as an integral part of the standard multi-axial psychiatric diagnostic system: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV-TR) [1].

After the introduction of the GAF, a number of studies have been published that claim the GAF to be a reliable scale [2–11], but these studies have all been conducted in research settings. Only Vatnaland and colleagues [12] studied the reliability of the GAF used by untrained doctors in a routine clinical setting of patients admitted at an acute setting of a psychiatric hospital. They found a rather poor reliability of the GAF.

The validity of the GAF has also been studied, which showed it to be relatively independent of sociodemographic factors [13], but closely inversely associated with the number of symptoms [13] and the duration of hospitalization [14]. Furthermore, an inverse association with the number and complexity of Axis I diagnoses was found [10,15–19]. A revised version of the GAF, which considers only social and occupational functioning and not symptom severity, was associated with ratings of clinical symptoms, but not

with ratings of functioning [20]. In the version of the Mental Illness Research, Education, and Clinical Center (MIRECC), occupational functioning, social functioning and severity of symptoms are scored on three subscales. All three were shown to be reliable and valid. However, all these studies have been carried out in a research setting, and raters had received extensive training on how to score the GAF [11].

In two prospective analyses, the GAF was found to have good predictive power for prognosis among inpatients [21] and patients with schizophreniform disorder [22]. There are no prior studies on the GAF in adult patients with depressive disorders. Previous studies on the GAF included children [2–4] or adult patients with schizophrenia [10,22], substance abuse disorders [13] and delirium [16], and groups of patients with different disorders, including mood disorders, anxiety disorders and personality disorders [5–7,12].

In view of the lack of studies on the GAF in depressed patients, the objective of this paper was to assess the reliability and convergent and discriminant validity of the GAF in a non-research, naturalistic sample of outpatients with major depressive disorders (MDDs) in a secondary mental health care setting where the GAF is part of the routine outcome monitoring (ROM) procedure [23]. We aimed to discover whether the GAF meets its purposes, when used in this population and setting.

Methods

Patients

In total 4157 patients, who sought treatment at the Regional Mental Health Provider (RMHP) Rivierduinen in the Netherlands (service area with 1.1 million inhabitants) between January 2002 and January 2007, were as part of normal clinical practice, evaluated with ROM. Of these patients, 1654 had a DSM-IV diagnosis of a current MDD as established by the Mini International Neuropsychiatric Interview Plus (MINI-plus) [24]. We excluded the patients without a follow-up assessment in ROM ($n = 901$), to select patients who had received treatment. This resulted in a group of 753 subjects. We examined possible selection bias by comparing the characteristics of these patients ($n = 753$) with the group without a follow-up assessment ($n = 901$). These groups did not differ statistically significant with respect to co-morbidity of mood disorders, co-morbid drug or alcohol abuse/dependency, disease severity and co-morbid eating disorder. The groups did differ, with respect to co-morbid generalized anxiety disorder (7.3% follow-up vs. 4.6% lost-to-follow-up, $P = 0.02$) and psychotic features (0.9% follow-up vs. 2.7% lost-to-follow-up, $P = 0.01$).

Subsequently, we excluded patients who did not have complete ROM assessment ($n = 41$), patients with a bipolar I or II disorder ($n = 20$), patients over 65 years of age ($n = 1$), patients treated (time between first and second GAF assessment) shorter than 4 weeks or longer than 52 weeks ($n = 47$), patients with a depressive disorder in remission at intake [Montgomery-Äsberg Depression Rating Scale (MADRS) score of 15 or lower] ($n = 15$), and patients with a depressive disorder with psychotic symptoms ($n = 3$). Of the remaining 626 patients, 194 were excluded because there was no GAF assessment by a clinician available. This resulted in a research sample of 432 MDD patients (26%).

Instruments/routine outcome monitoring

In ROM, all patients referred to the RMHP Rivierduinen in the Netherlands for treatment of a mood, anxiety or somatoform disorder received an extensive baseline assessment. The assessment comprised a standardized diagnostic interview [Mini-International Neuropsychiatric Interview Plus (MINI-plus)] [24,25], the assessment of sociodemographic and socioeconomic data, the administration of observer-rated scales and self-report questionnaires, and general measures of health and quality of life, including the MADRS, the Beck Depression Inventory-revised (BDI-II) [26], the Brief Symptom Inventory (BSI), which is a short version of the Symptom Checklist (SCL-90) [27], and the Short Form-36 (SF-36) [28,29]. The assessments were performed by specially trained test nurses in the outpatient clinics. Dedicated software has been developed to assist in this task [23]. All patients with sufficient mastery of the Dutch language who were able to complete computerized and written questionnaires were eligible for ROM. The data were made available in an anonymous form, as dictated by the Psychiatric Academic Registration Leiden (PAREL) regulation. This procedure has been approved by the Medical Ethical Committee of the University Medical Hospital Leiden.

Raters

During intake, each new patient was interviewed by a clinician (i.e. psychiatrist, resident in psychiatry or psychologist-psychotherapist), who subsequently scored the GAF. The clinicians did not receive specific training on the scoring of the GAF, other than the general education for their profession. One week later each patient took part in the ROM. As part of the ROM, the test nurse scored the GAF as well. The test nurses received an extensive instruction on how to score the GAF. They based their GAF score on their own individual interview of the patient, MINI-plus, MADRS and BAS. Neither the test nurses nor the clinicians knew in advance that their GAF scores were being used for a direct comparison as a study aim, and these two mental health professionals were not aware of each other's GAF scores, resulting in independent assessment. At follow-up (after median 5.2 $\pm$ SD 2.2 months), the GAF was scored again as part of the ROM by a test nurse using the same procedure.

Statistical analysis

Of the final sample of 432 depressed patients, sociodemographic characteristics and co-morbid disorders were described according to tertiles, based on the clinician's GAF scores. Differences between study groups were analysed by analysis of variance (ANOVA) or χ^2 (linear-by-linear) test for categorical variables. To examine whether patients were categorized similarly by clinicians and test nurses, the equivalence between the GAF scores of the clinician and test nurse was analysed by Kappa statistics. Correlations between the GAF scores of the clinicians and the test nurses were calculated using the Pearson's correlation coefficient. For other analyses, the GAF score was used as a continuous measure. Univariate and multivariate regression analyses were used to assess the association between the GAF scores and other instruments on social and occupational functioning and symptoms. Multivariate analyses were adjusted for potential confounding by

sex, age, ethnic background and educational level. Employment status was not included as a confounder, as unemployment or disablement may be a consequence of the psychiatric illness, or an alternative measure of poor occupational functioning. Using stepwise multiple regression analysis with the GAF score as the dependent variable and MADRS score, BDI score, BSI total score, SF-36 subscale scores, the independent predictors for the GAF score by the clinician and the test nurse were explored. Again, these analyses were adjusted for sex, age, ethnic background and educational level. Statistical significance was inferred at $P < 0.05$. All statistical analyses were undertaken with SPSS 17.0.

Results

At baseline, the mean GAF score by the clinician was 54.8 (SD: 6.5; median 55) and ranged from 35 to 85. The mean GAF score by the test nurse was 57.5 (SD: 7.6; median 57) and ranged from 31 to 88. The clinicians scored on average 2.3 points lower than the test nurses (95% confidence interval: 1.3–3.2; $P < 0.001$). The clinicians were much more likely to round GAF scores to the nearest 5's or 10's than the test nurses. The sociodemographic characteristics of our patient sample are presented in Table 1, divided in three groups according to the tertiles of the clinician's GAF scores. There were no significant differences in age, marital status or co-morbid anxiety or somatoform disorder. Lower GAF scores were significantly associated with a non-Dutch ethnic background ($P = 0.04$), lower education ($P = 0.008$) and more unemployment ($P < 0.001$).

Reliability of the GAF

Pearson's correlation coefficients between the baseline GAF score by the clinician and the baseline GAF score by the test nurse was 0.26 ($P \leq 0.001$) (Table 2 and Fig. 1), and between the baseline GAF score by the clinician and the follow-up GAF score by the test nurse was 0.19 ($P < 0.001$). In comparison, the correlation between the baseline and follow-up GAF scores by the test nurse was 0.45 ($P < 0.001$). Likewise, the concordance reflected by Cohen's Kappa between the GAF score by the clinician and the baseline GAF score by the test nurse when using categories was statistically significant but of very poor strength (Table 2). When using more extreme cut-offs for the GAF score (group 1: ≤45; group 2: 46–60; group 3: >60) the Cohen's Kappa did not improve in strength (data not shown). Moreover, Cohen's Kappa did not

Table 1 Baseline sociodemographic characteristics and co-morbid disorders according to tertiles of the GAF in outpatients with current MDD

	GAF by clinician			
	Low GAF *n* = 152	Intermediate GAF *n* = 147	High GAF *n* = 133	*P*-value
GAF				
Mean ± SD	48.1 ± 3.4	55.0 ± 0.4	62.2 ± 4.4	
Range	35–50	51–55	56–85	
Age – geometric mean; P_5–P_{95}	39.2 ± 10.5	39.2 ± 11.0	37.9 ± 12.2	0.32
Sex (%)				
Male	38.2	32.0	24.8	0.02
Female	61.8	68.0	75.2	
Ethnic background (%)				
Dutch	73.0	79.7	84.0	0.04
Other ethnicity	27.0	20.3	16.0	
Marital status (%)				
Married/living with partner	46.8	53.9	49.6	0.82
Divorced/widow(er)	20.6	11.7	17.6	
Living alone	32.5	34.4	32.8	
Education (%)				
Lower education	21.4	7.8	10.1	0.008
High school (lower)	29.4	39.1	27.7	
High school (higher)	36.5	35.9	42.9	
College/university	12.7	17.2	19.3	
Employment status (%)				
Unemployed	34.9	25.8	23.5	<0.001
Work-related disability/retired	46.8	36.7	29.4	
Employed	18.3	37.5	47.1	
Co-morbid anxiety of somatoform disorder (%)				0.13
Absent	35.5	40.8	44.4	
Present	64.5	59.2	55.6	

P-values by analysis of variance (ANOVA) or χ^2 test for categorical variables.
For education and marital status, *P*-value by χ^2 test, linear-by-linear term.
GAF, Global Assessment of Functioning Scale; MDD, major depressive disorder.

Table 2 Concordance of the GAF between the doctor and the test nurse in outpatients with current MDD

	GAF by clinician, *n* (%)				
GAF by test nurse	Low (≤50)	Intermediate (51–55)	High (>55)	Kappa	Pearson's R
Baseline					
Low GAF (≤ 50)	30 (27.8)	16 (13.9)	11 (9.6)	0.14	0.26
Intermediate GAF (51–55)	33 (30.6)	31 (27.0)	21 (18.3)	($P < 0.001$)	($P < 0.001$)
High GAF (>55)	45 (41.7)	68 (59.1)	83 (72.2)		
Follow-up					
Low GAF (≤ 50)	25 (19.7)	19 (15.0)	6 (5.2)	0.07	0.19
Intermediate GAF (51–55)	20 (15.7)	15 (11.8)	14 (12.1)	($P = 0.02$)	($P < 0.001$)
High GAF (>55)	82 (64.6)	93 (73.2)	96 (82.8)		

GAF, Global Assessment of Functioning Scale; MDD, major depressive disorder.

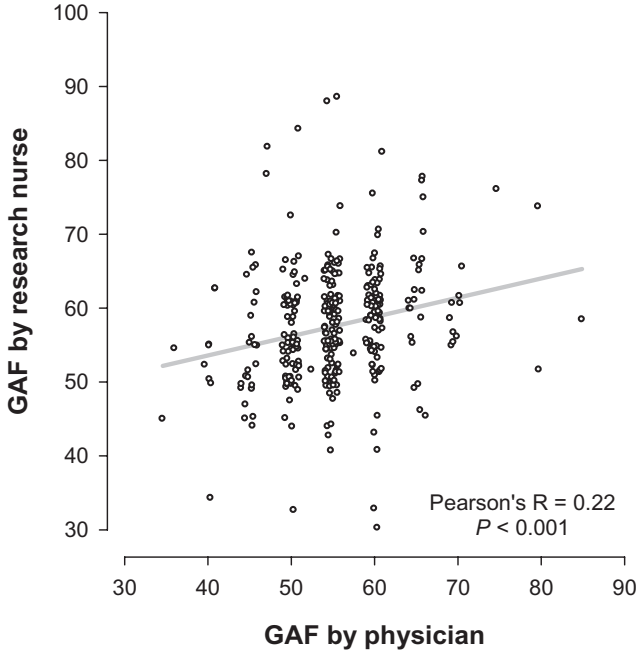

Figure 1 Association between baseline GAF scores by the treating doctor and the trained test nurse in outpatients with current MDD. Pearson's correlation coefficient is given with a univariate regression line. GAF, Global Assessment of Functioning Scale; MDD, major depressive disorder.

increase in strength when the time interval between baseline and follow-up was restricted to 6 months or shorter (data not shown).

Validity of the GAF

Significant associations were found between the clinicians' GAF score and diverse scales measuring disease severity of depressive symptoms, being the MADRS, BDI-II and BSI total score (Table 3). However, the correlation with the BSI total score disappeared, when adjusting for sex, age, ethnic background and education. Multiple correlations were found between the clinicians' GAF score and subscales of the SF-36, of which the strongest were with poor physical functioning, poor social functioning and poor mental health (Table 3), which persisted after adjustment for sex, age, ethnic background and education.

Using stepwise multiple regression analysis, the clinicians' GAF score correlated independently with the MADRS score and the SF-36 subscale physical functioning, when adjusting for sex, age, ethnic background and education (Table 4). However, there were no independent associations with SF-36 subscales that reflected psychological, social and occupational functioning.

The GAF scores by the test nurses were significantly associated with the MADRS, BDI-II and total score on the BSI, in this case also after adjustment for sex, age, ethnic background and education. All associations between nurse GAF scores and SF-36 subscales were much weaker than the GAF scores by the clinician and the SF-36 subscales, which may indicate that the GAF score by the trained test nurse was less affected by global physical functioning. In multivariate analysis, the MADRS was the only independent correlate of the GAF by the test nurse.

Discussion

In this prospective study in a large secondary care naturalistic sample of outpatients with a MDD, the inter-rater reliability of the GAF was weak, as indicated by the association between the GAF scored by the clinicians and the GAF scored by the test nurses. Also, the GAF showed rather poor discriminant validity.

These results contradict a number of previous studies on the GAF. However, all but one of these previous studies were conducted in research settings. In many of these, the patient information was presented in a standardized way: after watching a videotape [5,6] or reading case vignettes [2–4]. In others the clinicians had a more extensive training than is usual in clinical practice [4,5,7,10,11]. The only study with results more or less comparable to those of our study was also carried out in a naturalistic setting, but with inpatients in an acute setting [12].

We found evidence for a substantial contribution of the patient's psychiatric symptoms and disease severity to the GAF score. The GAF scores by the clinicians as well as by the nurses were strongly associated with disease severity measures from both the MADRS and the BDI. This concurs with findings in other populations of the GAF score being associated with psychiatric symptoms [3,10,13,14,16–19].

Although the GAF should be scored independently from physical functioning [1], the clinicians' GAF score (as opposed to the nurses' GAF score) was most strongly associated with

Table 3 Comparison of GAF scores by the treating doctor and the trained test nurse for associations with rating scales in 432 outpatients with current MDD

	GAF by clinician		GAF by test nurse	
	Unadjusted	Adjusted*	Unadjusted	Adjusted*
MADRS score	–0.22 ($P < 0.001$)	–0.20 ($P < 0.001$)	–0.40 ($P < 0.001$)	–0.39 ($P < 0.001$)
Beck Depression Inventory score	–0.19 ($P < 0.001$)	–0.17 ($P = 0.001$)	–0.21 ($P < 0.001$)	–0.23 ($P < 0.001$)
BSI total score	–0.11 ($P = 0.02$)	–0.08 ($P = 0.12$)	–0.18 ($P = 0.001$)	–0.16 ($P = 0.007$)
SF-36 subscales:				
Physical functioning	0.25 ($P < 0.001$)	0.26 ($P < 0.001$)	0.11 ($P = 0.05$)	0.09 ($P = 0.16$)
Social functioning	0.20 ($P < 0.001$)	0.18 ($P = 0.001$)	0.10 ($P = 0.06$)	0.09 ($P = 0.13$)
Role limitations due to physical problems	0.12 ($P = 0.02$)	0.12 ($P = 0.02$)	0.06 ($P = 0.32$)	0.04 ($P = 0.48$)
Role-limitations due to emotional problems	0.10 ($P = 0.04$)	0.11 ($P = 0.04$)	0.00 ($P = 1.00$)	0.02 ($P = 0.67$)
Mental health	0.20 ($P < 0.001$)	0.16 ($P = 0.002$)	0.18 ($P = 0.001$)	0.17 ($P = 0.003$)
Vitality	0.15 ($P = 0.002$)	0.14 ($P = 0.008$)	0.13 ($P = 0.02$)	0.15 ($P = 0.009$)
Bodily pain	0.13 ($P = 0.009$)	0.12 ($P = 0.03$)	0.03 ($P = 0.63$)	0.03 ($P = 0.58$)
General health perceptions	0.17 ($P = 0.001$)	0.17 ($P = 0.001$)	0.08 ($P = 0.15$)	0.13 ($P = 0.03$)

MADRS, Montgomery-Äsberg Depression Rating Scale; BSI, Brief Symptom Inventory; SF-36, Short Form (36) Health Survey; GAF, Global Assessment of Functioning Scale; MDD, major depressive disorder.
Data are beta (*P*-value between brackets) by regression analysis.
*Adjusted for sex, age, ethnic background and education.

Table 4 Independent predictors of GAF scores by the treating doctor and the trained test nurse in outpatients with current MDD

	GAF by clinician		GAF by test nurse	
	Beta	*P*-value	Beta	*P*-value
SF-36 subscales: Physical functioning	0.23	<0.001		
MADRS score	–0.16	0.003	–0.38	<0.001

MADRS, Montgomery-Äsberg Depression Rating Scale; SF-36, Short Form (36) Health Survey; GAF, Global Assessment of Functioning Scale; MDD, major depressive disorder.
Data are beta by stepwise multiple regression analysis with GAF as dependent variable and MADRS score, Beck Depression Inventory score, Brief Symptom Inventory total score, SF-36 subscale scores as independent variables (adjusted for sex, age, ethnic background and education) with accompanying *P*-values.

physical functioning on the SF-36. However, it should be noted that depressive patients often show physical symptoms [30]. In addition, our study showed an association between the clinicians' GAF score and social functioning on the SF-36, similar to the finding from a previous study with veterans with substance use disorders of whom some had Axis I co-morbidity [13]. In our study the nurses' GAF score showed no independent association with social functioning.

Our study has several limitations and strengths. A limitation of this study is the fact that the test nurses scored the GAF a week after the clinicians. Knowing though, that these patients had a depressive disorder, it can be assumed that the severity of their psychiatric symptoms and their level of functioning did not change importantly within a week. Second, this patient sample only included patients who received follow-up. This sample was chosen explicitly to be able to evaluate the GAF when used before and during treatment of outpatients. The results are therefore generalizable to outpatients who receive treatment. The ethnicity of our study participants was mainly Caucasian. Therefore, the results need to be confirmed in populations of other origin. Finally, patients with a co-morbid Axis I or Axis II disorder were included. It is therefore unclear whether findings are also applicable to patients with a pure MDD without co-morbidity (who, however, are the exception in naturalistic settings). Our study also has several strengths. First, we used a routine clinical setting with a large sample of outpatients. Therefore, we believe our conclusions are of clinical relevance. Second, we included patients with depressive disorders. We are not aware of previous studies on the GAF that evaluated its use in this specific patient population. Previous studies have shown that a (short) training on how to score the GAF improves the reliability of this scale [4,5,7,10,11]. Whether training also improves the reliability of the GAF in depressed out-patients should, however, be subject for future research.

Conclusions

The findings of this study support the idea that the GAF used in a routine clinical practice with depressive outpatients has rather poor inter-rater reliability and discriminant validity. More importantly, in routine clinical practice the GAF is used not only to judge the patient's health, but also to determine the use of health care resources, such as whether a patient should receive welfare benefits [4,5]. Also, in the Netherlands, the GAF is used by the government for the evaluation of the quality of care in Dutch psychiatric hospitals, as part of a benchmarking procedure. The poor reliability and validity we found in clinical practice throws serious doubts on these practices.

Furthermore, the GAF score was associated with such diverse factors, namely disease severity, social functioning and physical complaints, that its clinical value is dubious. Therefore, to our view, it should be seriously questioned whether a place should be reserved for the GAF in its present form in DSM-V. Together with other research teams [5,15,31] we therefore suggest the

introduction of a scale which reflects the level of functioning separate from the severity of the psychiatric symptoms of a patient. In that context, Niv and colleagues have developed the MIRECC version of the GAF [11], which explicitly breaks up the GAF into three subscales for occupational functioning, social functioning and symptom severity. This version was tested in schizophrenic and schizoaffective patients, and future research is needed to examine whether this approach is also reliable and valid for depressed patients and patients with other psychiatric illnesses.

References

1. American Psychiatric Association (2002) Diagnostic and Statistical Manual of Mental Disorders, DSM IV-TR. Washington, DC: American Psychiatric Association.
2. Russel, A. T., Cantwell, D. P., Mattison, R. & Will, L. (1979) A comparison of DSM-II and DSM-III in the diagnosis of childhood psychiatric disorders: multiaxial features. *Archives of General Psychiatry*, 36, 1223–1226.
3. Mezzich, A. C., Mezzich, J. E. & Coffman, G. A. (1985) Reliability of DSM-III vs DSM-II in child psychopathology. *Journal of American Academy of Child and Adolescent Psychiatry*, 24, 273–280.
4. Rey, J. M., Stewart, G. W., Plapp, J. M., Bashir, M. L. & Richards, I. N. (1988) Validity of axis V of DSM-III and other measures of adaptive functioning. *Acta Psychiatrica Scandinavia*, 77, 534–542.
5. Hilsenroth, M. J., Ackerman, S. J., Blagys, M. D., *et al*. (2000) Reliability and validity of DSM-IV Axis V. *American Journal of Psychiatry*, 157, 1858–1863.
6. Söderberg, P., Tungström, S. & Armelius, B. A. (2005) Reliability of global assessment of functioning ratings made by clinical psychiatric staff. *Psychiatric Services*, 56 (4), 434–438.
7. Jones, S. H., Thornicroft, G., Coffrey, M. & Dunn, G. (1995) A brief mental health outcome scale- reliability and validity of the Global Assessment of Functioning (GAF). *British Journal of Psychiatry*, 166, 654–659.
8. Beitchman, J. H., Adlaf, E. M., Douglas, L., Atkinson, L., Young, A., Johnson, C. J., Escobar, M. & Wilson, B. (2001) Comorbidity of psychiatric and substance use disorders in late adolescence: a cluster analytic approach. *American Journal of Drug and Alcohol Abuse*, 27, 421–440.
9. Haro, M., Kamath, S. A., Ochoa, S., *et al*., on behalf of the SOHO Study Group (2003) The Clinical Global Impression-Schizophrenia scale: a simple instrument to measure the diversity of symptoms present in schizophrenia. *Acta Psychiatrica Scandinavia*, 107 (Suppl 416), 16–23.
10. Startup, M., Jackson, M. C. & Bendix, S. (2002) The concurrent validity of the Global Assessment of Functioning (GAF). *British Journal of Clinical Psychology*, 41, 417–422.
11. Niv, N., Cohen, A. N., Sullivan, G. & Young, S. (2007) The MIRECC version of the Global Assessment of Functioning Scale: reliability and validity. *Psychiatric Services*, 58 (4), 529–535.
12. Vatnaland, T., Vatnaland, J., Friis, S. & Opjordsmoen, S. (2007) Are GAF scores reliable in routine clinical use? *Acta Psychiatrica Scandinavia*, 115, 326–330.
13. Moos, R. H., McCoy, L. & Moos, B. S. (2000) Global assessment of functioning (GAF) ratings: determinants and role as predictors of one-year treatment outcomes. *Journal of Clinical Psychology*, 56 (4), 449–461.
14. Gordon, R. E., Jardiolin, P. & Gordon, K. K. (1985) Predicting length of hospital stay of psychiatric patients. *American Journal of Psychiatry*, 142, 235–237.
15. Skodol, A. E., Link, B. G., Shrout, P. E. & Horwath, E. (1988) Toward construct validity for DSM-III axis V. *Psychiatry Research*, 24, 13–23.
16. Trzepacz, P. T., Brenner, R. & Van Thiel, D. H. (1989) A psychiatric study of 247 liver transplantation candidates. *Psychosomatics*, 30, 147–153.
17. Westermeyer, J. & Neider, J. (1988) Social networks and psychopathology among substance abusers. *American Journal of Psychiatry*, 145, 1265–1269.
18. Fabrega, H. Jr, Pilkonis, P., Mezzich, J., Ahn, C. W. & Shea, S. (1990) Explaining diagnostic complexity in an intake setting. *Comprehensive Psychiatry*, 31, 5–14.
19. Tungström, S., Söderberg, P. & Bengt-Ake, A. (2005) Relationship between the Global Assessment of Functioning and other DSM axes in routine clinical work. *Psychiatric Services*, 56 (4), 439–443.
20. Roy-Byrne, P., Dagadakis, C., Unützer, J. & Ries, R. (1996) Evidence for limited validity of the revised global assessment of functioning scale. *Psychiatric Services*, 47, 864–866.
21. Mellsop, G., Peace, K. & Fernando, T. (1987) Pre-admission adaptive functioning as a measure of prognosis in psychiatric inpatients. *Australian New-Zealand Journal of Psychiatry*, 21, 539–544.
22. Beiser, M., Fleming, J. A. E., Iacono, W. G. & Lin, T. Y. (1988) Refining the diagnosis of schizophreniform disorder. *American Journal of Psychiatry*, 145, 695–700.
23. de Beurs, E. & Zitman, F. G. (2007) Routine Outcome Monitoring. Measuring therapy effects in clinical practice with web-based software. *Maandblad Geestelijke Volksgezondheid*, 62 (1), 13–28.
24. Van Vliet, I. M. & de Beurs, E. (2007) The MINI-International Neuropsychiatric Interview. A brief structured diagnostic psychiatric interview for DSM-IV and ICD-10 psychiatric disorders. *Tijdschrift Voor Psychiatrie*, 49 (6), 393–397.
25. Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R. & Dunbar, G. C. (1998) The MINI-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, 59 (Suppl 20), 22–33. quiz 34-57.
26. Beck, A. T., Steer, R. A. & Carbin, M. G. (1988) Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. *Clinical Psychology Review*, 8, 77–100.
27. de Beurs, E. & Zitman, F. G. (2006) The Brief Symptom Inventory (BSI): reliability and validity of a practical alternative to SCL-90. *Maandblad Geestelijke volksgezondheid*, 61, 120–141.
28. Ware, J. E. Jr (1992) The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, 30, 473.
29. Aaronson, N. K., Muller, M., Cohen, P. D., Essink-Bot, M. L., Fekkes, M., Sanderman, R., Sprangers, M. A., te Velde, A. & Verrips, E. (1998) Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *Journal of Clinical Epidemiology*, 51 (11), 1055–1068.
30. Greden, J. F. (2003) Physical symptoms of depression: unmet needs. *Journal of Clinical Psychiatry*, 64 (Suppl 7), 5–11.
31. Goldman, H. H., Skodol, A. E. & Lave, T. R. (1992) Revising axis V for DSM-IV: a review of measures of social functioning. *American Journal of Psychiatry*, 149 (9), 1148–1156.

