Agreement on Web-based Diagnoses and Severity of Mental Health Problems in Norwegian Child and Adolescent Mental Health Services

Håkan Brøndbo[1,*], Børge Mathiassen[1,2], Monica Martinussen[3], Einar Heiervang[4], Mads Eriksen[5] and Siv Kvernmo[1,2]

[1]*Department of Child and Adolescent Psychiatry, Divisions of Child and Adolescent Health, University Hospital of North-Norway, Tromsø, P.O. Box 19, 9038 Tromsø, Norway*

[2]*Department of Clinical Medicine, Faculty of Health Sciences, University of Tromsø, 9037 Tromsø, Norway*

[3]*Regional Knowledge Centre for Children and Youth Mental Health and Child Protection, University of Tromsø, 9037 Tromsø, Norway*

[4]*Institute of Clinical Medicine, University of Oslo, 0372 Oslo, Norway*

[5]*Alta Child and Adolescent Mental Health Service, Finnmark Hospital Trust, P.O. Box 1294, 9505 Alta, Norway*

Abstract: *Objective:* This study examined the agreement between diagnoses and severity ratings assigned by clinicians using a structured web-based interview within a child and adolescent mental health outpatient setting.

Method: Information on 100 youths was obtained from multiple informants through a web-based Development and Well-Being Assessment (DAWBA). Based on this information, four experienced clinicians independently diagnosed (according to the International Classification of Diseases Revision 10) and rated the severity of mental health problems according to the Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) and the Children's Global Assessment Scale (C-GAS).

Results: Agreement for diagnosis was κ=0.69-0.82. Intra-class correlation for single measures was 0.78 for HoNOSCA and 0.74 for C-GAS, and 0.93 and 0.92, respectively for average measures.

Conclusions: Agreement was good to excellent for all diagnostic categories. Agreement for severity was moderate, but improved to substantial when the average of the ratings given by all clinicians was considered. Therefore, we conclude that experienced clinicians can assign reliable diagnoses and assess severity based on DAWBA data collected online.

Keywords: Web-based, telepsychiatry, DAWBA, HoNOSCA, C-GAS.

INTRODUCTION

Use of telepsychiatry is related to avoidance of travelling, and therefore widespread in northern Norway with long distances and less access to professionals. Most users find telepsychiatry useful, but lack of equipment may inhibit the use in Norwegian Child and Adolescent Mental Health Service (CAMHS) [1]. Computer-based assessment procedures, which require no special equipment, are becoming more commonplace in CAMHS in order to save time and clinical resources [2]. Use of Development and Well-Being Assessment (DAWBA) information to assign psychiatric diagnoses, collected either by lay interviewers or online, is common in epidemiological research [3-6], but to our knowledge little is known about the reliability for clinical samples when information is collected online. Even less is known about the reliability of severity measures when assigned based on online information.

Mental health diagnoses are based on information generally gleaned from clinical interviews, on developmental history, behavioral observations and reported difficulties in everyday life. The accuracy of any subsequent diagnostic assignment depends on a clinician's ability to integrate information from different sources and perspectives [7, 8]. Inaccuracies and variability in diagnoses are common, and the agreement between clinicians for common mental disorders is fairly poor, ranging from low to moderate [9-11]. Reliability of diagnoses *via* videoconferencing seems as good as face-to-face assignments [12].

Use of structured interviews as opposed to traditional clinical interviews has been shown to significantly improve the accuracy of diagnoses [13, 14]. For example, Ramirez Basco and colleagues [13] added structured procedures to improve diagnostic agreement for adult outpatients with severe mental illness. They found a kappa (κ) value of 0.76-0.87 for the most structured procedure, compared to κ=0.45-0.52 for the standard clinical interview. In a review of clini-

*Address correspondence to this author at the Department of Child and Adolescent Psychiatry, Divisions of Child and Adolescent Health, University Hospital of North-Norway, Tromsø, P.O. Box 19, 9038 Tromsø, Norway; Tel: +47 77755701; Fax: +47 77755761;
E-mail: hakan.brondbo@unn.no

1745-0179/12 2012 Bentham Open

cal diagnoses of depression [14], agreement among mental health care professionals ranged from κ=0.64 to *κ*=0.93 when the diagnoses were aided by semi-structured interviews. For attention deficit hyperactivity disorder (ADHD) in clinically referred youths, [15] high levels of agreement (κ=0.57 to κ=0.76) were observed for diagnoses assigned based on information collected online from the DAWBA, and diagnoses based upon full clinical examination in addition to the DAWBA. The authors concluded that a trained clinician scoring the DAWBA without meeting the patient can be as accurate as an ordinary clinical assessment. Another study by Foreman and Ford [16] showed agreement between ADHD diagnoses made by ordinary English CAMHS teams and independent DAWBA diagnoses for 98% of the cases.

If good accuracy can be established through web-based procedures, there is a huge potential for saving time and clinical resources in the assessment phase and thereby improve accessibility to treatment. High agreement between clinicians is a first step towards valid procedures for assignment of both diagnoses and severity of mental health problems.

The purpose of this study was to examine the agreement between CAMHS clinicians in Norway when assigning diagnosis and severity of mental health problems based only on DAWBA information collected online.

METHODS

Participants

A sample of 100 patients, 58 boys and 42 girls (mean age 11.11 years, SD 3.35), was randomly selected from the 286 patients participating in the CAMHS North study. This is a multicenter study in the northern part of Norway, where clinical procedures, structures and treatment paths were evaluated, with the aim to bridge the gap between clinical practice and academic research. More specific aims included the investigation of factors related to the waiting list, duration of assessment and treatment, implementation and validation of structured instruments, and user satisfaction with services.

Four independent clinicians diagnosed and rated the severity of mental illnesses for the 100 CAMHS North study participants included in the present report. Diagnoses were assigned according to the International Classification of Diseases Revision 10 (ICD-10) and severity was rated according to the Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) and the Children's Global Assessment Scale (C-GAS) (Table **1**).

Written informed consent was obtained from all participants before inclusion in the CAMHS North study. Parents gave the consent for participants younger than 12 years of age. For participants between 12 and 16 years of age, written consent was obtained from both the parents and the adolescents. Participants older than 16 years of age gave consent themselves, according to Norwegian legislation. The Regional Committee for Medical Research Ethics and the Norwegian Social Science Data Services approved the study.

Procedure

From October 2006 to December 2008, children and adolescents referred to the CAMHS Outpatient Clinic at the University Hospital of Northern Norway were invited to participate in the CAMHS North study. For those who accepted to participate, parents, teachers and children 11 years of age or older completed the relevant version of the DAWBA using a web-based interface that they accessed from home or school after receiving a request with the unique web link for that child's case. Written information concerning how to log on, as well as contact information in case of problems, was distributed along with the unique web ID and passwords. For participants younger than 16 years of age, requests were distributed by mail to the parents, who in turn distributed the requests to their children (if aged 11-15 years) and the teachers. For the participants older than 16 years of age, requests to both parents and teachers were distributed *via* the participants themselves. Among the 100 CAMHS North study participants included in the present report, the DAWBA was completed online by at least one parent for 93% of the participants, and by 84% of participants 11 years of age or older. A total of 57% of the participants were 11 years of age or older.

Of the four rating clinicians, three were clinical specialists in neuropsychology with a minimum of 9 years of experience in the field, and one was a specialist in child and adolescent psychiatry with 15 years of experience in the field. All clinicians completed the online training for DAWBA [17]. They also completed a 1-day training session on the categories of severity in the HoNOSCA and the C-GAS, including scoring of vignettes [18, 19]. In addition, all clinicians participated in two separate 2-day training sessions in preparation for this study, including diagnostic assessment and severity ratings of clinical cases. The clinician who lead the 2-day training sessions was trained by Robert Goodman, who developed the DAWBA interview.

Each clinician individually diagnosed the 100 participants included in this report according to the ICD-10, based on information from the DAWBA, without face-to-face contact with the families. To ensure enough cases for agreement analysis, the diagnoses were categorized as emotional diagnosis (diagnoses related to separation anxiety, specific phobias, social phobia, panic attacks and agoraphobia, post-traumatic stress disorder, generalized anxiety, compulsions and obsession, depression, and deliberate self-harm), ADHD/hyperkinetic diagnosis (diagnoses related to attention and activity), conduct diagnosis (diagnoses related to awkward and troublesome behavior), and other diagnosis (diagnoses related to developmental disorders, eating difficulties, and less common problems). Co-morbidity was documented when diagnoses from at least two categories where assigned, without taking the exclusion rules of the ICD-10 into consideration. The clinicians also rated clinical severity according to the HoNOSCA and the C-GAS.

According to the instruction for the DAWBA, all raters, even experienced ones, are to attend regular consensus meetings to discuss difficult cases [17]. Of course the proportion of difficult cases will be larger in a clinical population, such as ours. Twenty-five of the 100 cases included in the present report had diagnostic disagreement between two or more

raters, and thus were discussed among all four clinicians until a consensus was met. Previous studies, like the British Child and Adolescent Mental Health Survey 1999 [3, 4], the Bergen Child Study [5], and the Italian preadolescent mental health project [6], have used similar procedures.

Measures

Information contained in the DAWBA was used by the clinicians to assign ICD-10 diagnoses and C-GAS and HoNOSCA severity ratings. The DAWBA is a package of measures of child and adolescent psychopathology for administration to multiple informants. It is designed to generate common child psychiatric diagnoses according to the ICD-10 and DSM-IV without neglecting severe but less common diagnoses. The Norwegian web-based version that was used in the CAMHS North study contains modules for diagnoses related to separation anxiety, specific phobias, social phobia, panic attacks and agoraphobia, post-traumatic stress disorder, generalized anxiety, compulsions and obsession, depression, deliberate self-harm, attention and activity, awkward and troublesome behavior, developmental disorders, eating difficulties, and less common problems, as well as modules for background information and strengths. For each module there are both closed questions with fixed response categories and open-ended questions where the informant is asked to give detailed descriptions in his/her own words in text-boxes. Each module has initial screening questions with skip rules, and if problems are reported informants are also asked about their functional impact. Three different versions are available: 1) a detailed psychiatric interview for parents of approximately 50 minutes in length, 2) a youth interview of approximately 30 minutes and 3) a briefer questionnaire for teachers of approximately 10 minutes. The information from all informants is presented to the clinician in a separate program, where all closed questions are used to generate predictions of likelihood for a diagnosis [17]. The predictions can be used as rough prevalence estimates for research purposes [20], but mostly as a convenient starting point for clinicians evaluating all information, including the open-ended questions, in order to determine correct diagnoses to the child. The DAWBA has shown good discriminative properties both between population-based and clinical samples, and between different diagnoses [21]. Both in Norway and Great Britain, the DAWBA generates realistic estimates of prevalence for psychiatric illness as well as a high predictive validity when used in public health services [4, 5]. Good to excellent inter-rater reliability has been reported in both British and Norwegian studies, with κ=0.86-0.91 for any diagnoses κ=0.57-0.93 for emotional diagnoses, and κ=0.93-1.0 for hyperkinetic or conduct diagnoses [3, 22]. Good to excellent agreement has also been reported between diagnoses from clinical practice and those based solely on the DAWBA, with kappa values ranging from κ=0.57-0.76 [15, 16].

The C-GAS was used to rate severity of mental health problems. It is frequently used for this purpose and has several areas of application, such as to quantify impairment levels, as an outcome measure, or as an indicator of prognosis [23, 24]. C-GAS is a single-factor measure of the overall severity of psychiatric disturbance, with a summary score ranging from 1 to 100 that allows for a clinically meaningful index of global psychopathology. Green, Shirk, Hanze and Wanstrath [25] found that C-GAS used in clinical practice measures functional strengths. C-GAS has also been found to better measure change and outcome prediction than diagnosis and multidimensional scales [24]. Several studies have revealed good inter-rater reliability, especially among raters that have experience with C-GAS [25-27].

HoNOSCA was also used as a measure of severity of mental health problems in this study. HoNOSCA is a broad measure of behavioral, symptomatic, social, and impairment domains in children and adolescents. A total of 13 clinical features are rated on a five-point severity scale and added into a summary score, ranging from 0 (no problems) to 52 (severe problems in relation to all clinical features). Several studies have found good inter-rater reliability for the total score, as well as for the majority of individual items [28-32].

Statistical Analyses

All statistical analyses were performed using STATA version 11.0. For the exact proportion of cases where all four clinicians agreed on the diagnoses, raw agreement was calculated. To examine the agreement for diagnoses between the four clinicians, Fleiss' kappa was calculated. Fleiss' kappa measures the overall agreement for all raters, without any reference to the consensus diagnoses [33]. Interpretations of kappa values followed the guidelines suggested by Cicchetti and Sparrow [34]. Agreement in the range κ=0.75 to κ=1.00 was interpreted as excellent, κ=0.60 to κ=0.74 as good, κ=0.40 to κ=0.59 as fair, and κ <0.40 as poor.

Intra-class correlation (ICC) between clinicians was computed to assess agreement for HoNOSCA and C-GAS severity ratings. The preferred model for ICC was an alpha model for dichotomous data, two-way mixed type for consistency data [35, 36]. The ICC was calculated as a "single-measure ICC" and an "average-measure ICC", where the single-measure ICC is the reliability of the ratings of one clinician, and the average-measure ICC is the reliability of the ratings of all four clinicians averaged together.

The interpretations of the ICC values were done according to the guidelines suggested by Shrout [37]. Agreement in the range of 0.81 to 1.00 was interpreted as substantial, 0.61 to 0.80 as moderate, 0.41 to 0.60 as fair, 0.11 to 0.40 as slight and 0.00 to 0.10 as virtually none.

RESULTS

Raw Agreement for Diagnoses

Raw agreement between the clinicians was calculated both for agreement on any diagnosis versus no diagnosis, and for agreement on the type of clustered diagnoses. For any diagnosis the raw agreement was 75%, for emotional diagnosis 77%, for ADHD/hyperkinetic diagnosis 84%, and for conduct diagnosis 84% (data not shown).

Agreement for Diagnoses

Fleiss' kappa was used to examine the agreement for diagnoses between the four clinicians. We found that the agreement was good, both for any diagnosis/no diagnosis and for the diagnostic categories emotional diagnosis, ADHD/hyperkinetic diagnosis and co-morbidity. For the

Table 1. Agreement Between Clinician-Assigned Diagnoses and Severity Ratings for 100 Participants of the CAMHS North Study

	Agreement for Diagnoses		Clinician-Rated Severity[2]		ICC for Severity			
	N	**κ (CI)**	**C-GAS *Mean* (*SD*)**	**HoNOSCA *Mean* (*SD*)**	**C-GAS Single (CI)**	**C-GAS Average (CI)**	**HoNOSCA Single (CI)**	**HoNOSCA Average (CI)**
Total sample	100[1]		56.11 (10.56)	11.09 (5.27)	0.74 (0.66-0.81)	0.92 (0.89-0.94)	0.78 (0.71-0.84)	0.93 (0.91-0.95)
Any diagnosis	70	0.69 (0.66-0.73)	51.26 (7.21)	13.20 (4.54)				
Emotional diagnosis[3]	20	0.70 (0.68-0.75)	53.05 (8.24)	13.24 (4.97)				
ADHD/Hyperkinetic diagnosis[3]	6	0.72 (0.68-0.76)	54.88 (6.29)	10.71 (3.39)				
Conduct diagnosis[3]	19	0.82 (0.76-0.87)	54.47 (5.23)	10.57 (3.32)				
Co-morbidity	24[4]	0.70 (0.60-0.82)	46.27 (5.40)	15.89 (3.93)				
Other diagnosis[3]	1		52.75 (-)	12.75 (-)				
No diagnosis	30		67.41 (8.27)	6.17 (3.18)				

[1]Consensus diagnoses. [2]Mean of four clinicians. [3]Single diagnosis without co-morbidity. [4]Emotional diagnosis (*N*=14), ADHD/hyperkinetic diagnosis (*N*=16), conduct diagnosis (*N*=20) and other diagnosis (*N*=4)

ICC= Intra-class correlation, C-GAS= the Children's Global Assessment Scale, HoNOSCA= Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) SD=standard deviation, CI=confidence interval.

category of conduct diagnosis, agreement was excellent (Table **1**).

Agreement for Severity

The single-measure ICC for both C-GAS and HoNOSCA was moderate, and average-measure ICCs were substantial (Table **1**).

DISCUSSION

The first aim of this study was to examine agreement between clinicians for diagnoses assigned based on information collected online with the DAWBA, in a Norwegian child and adolescent mental health outpatient population. Our results indicated that agreement for mental health diagnoses can be good to excellent when aided by the DAWBA, and are consistent with the findings of other studies in which diagnostic agreement in mental health populations was examined [13, 14]. Ramirez Basco *et al.* [13] and Williams *et al.* [14] reported agreement between mental health professionals using structured or semi-structured interviews in adult populations on par with our results. Despite differences in population and clinical setting, our results strengthen the claim that when aided by structured or semi-structured instruments, agreement for mental health diagnoses can be good to excellent, even when information is collected online. Previously good to excellent diagnostic agreement is reported for diagnoses assigned *via* videoconferencing [13]. Our results suggest that also an online procedure for collecting information can be sufficient for reliable diagnostic assignments. Compared to other medical disciplines, results of diagnostic agreement for mental health problems are on par or better [38]. When major psychiatric diagnoses were compared to medical/neurological diagnoses, the conclusion was that "there is as much objective science in psychiatry as there is in most other medical specialties, which is to say an impressive but not overwhelming amount" [38, p. 22].

The second aim of this study was to examine agreement between clinician-assigned severity of mental health problems, as measured with C-GAS and HoNOSCA. The DAWBA was not originally designed to measure severity and to our knowledge this is the first study that used the web-based DAWBA as the source of information on which severity ratings were based. The use of the DAWBA as the source of information, instead of written vignettes as most other studies have used, increases the complexity and amount of information available, and thereby lessens the focus on themes that are directly relevant when rating by HoNOSCA and C-GAS. These differences improved the ecological validity of our results, which show that agreement for severity ratings based on DAWBA information collected online can be fair to moderate for a single clinician, and moderate to substantial when an averaged rating from multiple clinicians is used.

Hanssen-Bauer, Aalen, Ruud, and Heyerdahl [18] used both written vignettes and clinical interviews. A total of 169 clinicians rated 10 single-page written vignettes each, based on clinical descriptions from the CAMHS. Three clinicians also rated 20 patients each as part of the hospital admission procedure. They found the ICC for C-GAS to be 0.61, and 0.81 for HoNOSCA. They did not detect any difference in ICC between vignettes and clinical interviews. Even with strengthened ecological validity, our results are on par with the HoNOSCA ICC, and better than the C-GAS ICC. In a study using [39] five single-page written vignettes to obtain

C-GAS ratings in a naturalistic clinical setting, five experts' ratings were compared to ratings by 703 untrained health-care professionals. The vignettes were based on chart information from patients' first visit at outpatient units or emergency rooms. The ICC for the experts was 0.92, compared to 0.73 for the untrained health-care professionals [39]. Our single-measure ratings are thus comparable with those assigned by untrained health-care professionals from the aforementioned study. We believe that it is the increased complexity and amount of information available in our study and also the diminished focus on questions that directly affect HoNOSCA and C-GAS scores, which may explain this phenomenon. Our single-measure ICC was moderate for both C-GAS and HoNOSCA, and the average-measure ICCs were substantial for both instruments. The correct measure to use depends on the clinical or research situation. If you rely on the rating of one clinician, the single-measure ICC is appropriate. If you have multiple ratings available, it is more appropriate to use the average-measure ICC. Multiple ratings generally enhance reliability [40]. It is noteworthy that, by using multiple clinicians, we compensated for the complexity of the DAWBA information and showed an ICC on par with the expert group of Lundh *et al.* [39], who rated less complex vignettes.

Limitations of this study should also be noted. Due to both the impossibility of receiving additional or follow-up information, and the problem of circularity between the clinicians' individual ratings and the consensus diagnoses, the consensus procedure could not be used to validate the diagnoses and severity ratings. Such procedures are imperfect, but nevertheless valuable, as assessment of mental health continues to be based on developmental history, behavioral observations and reported difficulties in every day life [3-6, 41]. The use of one single expert rating may not always be sufficient to achieve reliable diagnoses [42]. The consensus discussion has the benefit of providing intelligent input from several experts in order to refine the final diagnosis. Nevertheless, in some cases additional information or longitudinal course might have refined the diagnosis even more than a consensus discussion. One can also question if the assessment of young children, where the DAWBA interview is not applicable, raise different challenges to the validity of a web-based interview, or if the information from parents and teachers are sufficient. In order to examine the validity and accuracy of the diagnoses and severity ratings, a study design where clinicians examine the patients by traditional methods such as a clinical interview, would have been more appropriate.

Another limitation is the relative homogeneity of the clinicians. All four clinicians in our study were male health-care professionals with long educational and clinical backgrounds, although one is a medical doctor and three are psychologists. It is possible that the relative homogeneity of the raters strengthened the agreement for both the diagnoses and the severity ratings compared to regular staff composition in some routine clinical settings. However, Hanssen-Bauer *et al.* [18], did not find any clinically significant difference in ICC for either the HoNOSCA, or the C-GAS based on a rater's profession or experience. In contrast, Lundh *et al.* [39] found that rater characteristics, such as profession, gender and age did affect the C-GAS ratings. Older raters and men were found to be more likely to be aberrant raters. Likewise, psychologists were found to be more aberrant raters than medical doctors. The characteristics of the clinicians in our study may therefore weaken, rather than strengthen, the agreement we report here [39].

The clinical implication of the results is that a trained, experienced clinician is usually sufficient to assign reliable diagnoses and also rate severity of mental health problems based on information collected online in the DAWBA. However, reliability could be further improved if several independent trained clinicians contribute to the assessment of the same patient. In a clinical setting this will of course be a question of resources, but using even two independent raters is likely to raise the reliability substantially. Another implication may be that even with structured measures and multiple clinicians available, a small amount of cases have diagnostic ambiguity that needs to be handled by either collecting additional information or in a longitudinal course, but if online procedures can save time and clinical resources for most cases, this seems to be a minor expense.

Further research is needed to examine the agreement in even more naturalistic settings, where clinicians may differ more with regard to clinical experience, occupational background and training. Also, further research is needed on the agreement of these instruments when used for less prevalent mental health disorders, such as sub-types of anxiety, autism and psychosis. Finally, although high agreement between raters is important for validity, it does not ensure it. More research is needed into the validity of clinician-assigned diagnoses and severity ratings according to the HoNOSCA and C-GAS, when using the online DAWBA as the main source of information, compared to clinicians evaluating the patients using a clinical interview..

CONCLUSIONS

In conclusion, information obtained with the online DAWBA may be a sound basis on which reliable clinical diagnoses and severity ratings for common mental health disorders in a clinical setting can be established. A clinical practice that includes systematic multiple, independent assignments of diagnosis and severity are preferable due to the resulting improved reliability of the severity ratings.

REFERENCES

[1] Hanssen B, Wangberg SC, Gammon D. Use of videoconferencing in Norwegian psychiatry. J Telemed Telecare [Randomized Controlled Trial] 2007; 13(3): 130-5.

[2] Berger M. Computer assisted clinical assessment. Child Adolescent Mental Health 2006; 11(2): 64-75.

[3] Ford T, Goodman R, Meltzer H. The British Child and Adolescent Mental Health Survey 1999: The Prevalence of DSM-IV Disorders. JAACAP 2003; 42(10): 1203-11.

[4] Meltzer H, Gatward R, Goodman R, *et al.* Mental health of children and adolescents in Great Britain. Int Rev Psychiatr 2003; 15(1-2): 185-7.

[5] Heiervang E, Stormark KM, Lundervold AJ, *et al.* Psychiatric disorders in Norwegian 8- to 10-year-olds: an epidemiological survey of prevalence, risk factors, and service use. JAACAP 2007; 46(4): 438-47.

[6] Frigerio A, Vanzin L, Pastore V, *et al.* The Italian preadolescent mental health project (PrISMA): rationale and methods. Int J Methods Psychiatr Res 2006; 15(1): 22-35.

[7] McClellan JM, Werry JS. Introduction--research psychiatric diagnostic interviews for children and adolescents. J Am Acad Child Adolesc Psychiatry 2000; 39(1): 19-27.

[8] Costello EJ, Egger H, Angold A. 10-Year Research Update Review: The Epidemiology of Child and Adolescent Psychiatric Disorders: I. Methods and Public Health Burden. JAACAP 2005; 44(10): 972-86.

[9] Galanter CA, Patel VL. Medical decision making: a selective review for child psychiatrists and psychologists. J Child Psychol Psychiatry 2005; 46(7): 675-89.

[10] Ezpeleta L, de la Osa N, Domenech JM, *et al*. Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents--DICA-R--in an outpatient sample. J Child Psychol Psychiatry 1997; 38(4): 431-40.

[11] Lauth B, Levy SR, Juliusdottir G, *et al*. Implementing the semi-structured interview Kiddie-SADS-PL into an in-patient adolescent clinical setting: impact on frequency of diagnoses. Child Adolesc Psychiatry Ment Health 2008; 2(1): 14.

[12] Martin-Khan RM, Wootton R, Whited J, *et al*. A systematic review of studies concerning observer agreement during medical specialist diagnosis using videoconferencing. J Telemed Telecare 2011; 17(7): 350-7.

[13] Basco RM, Bostic JQ, Davies D, *et al*. Methods to improve diagnostic accuracy in a community mental health setting. Am J Psychiatry 2000; 157(10): 1599-605.

[14] Williams JW, Jr., Noel PH, Cordes JA, *et al*. Is this patient clinically depressed? JAMA 2002; 287(9): 1160-70.

[15] Foreman D, Morton S, Ford T. Exploring the clinical utility of the Development And Well-Being Assessment (DAWBA) in the detection of hyperkinetic disorders and associated diagnoses in clinical practice. J Child Psychol Psychiatry 2009; 50(4): 460-70.

[16] Foreman DM, Ford T. Assessing the diagnostic accuracy of the identification of hyperkinetic disorders following the introduction of government guidelines in England. Child Adolesc Psychiatry Ment Health 2008; 2(1): 32.

[17] Youthinmind. DAWBA information for researchers and clinicians about the Development and Well-Being Assessment. youthinmind; Available from: http://www.dawba.info [Accessed 17 Nov. 2011].

[18] Hanssen-Bauer K, Aalen OO, Ruud T, *et al*. Inter-rater reliability of clinician-rated outcome measures in child and adolescent mental health services. Adm Policy Ment Health 2007; 34(6): 504-12.

[19] Hanssen-Bauer K, Gowers S, Aalen OO, *et al*. Cross-national reliability of clinician-rated outcome measures in child and adolescent mental health services. Adm Policy Ment Health 2007; 34(6): 513-8.

[20] Goodman A, Heiervang E, Collishaw S, *et al*. The 'DAWBA bands' as an ordered-categorical measure of child mental health: description and validation in British and Norwegian samples. Soc Psychiatry Psychiatr Epidemiol 2011; 46: 521-32.

[21] Goodman R, Ford T, Richards H, *et al*. The development and well-being assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. J Child Psychol Psychiatr Allied Disc 2000; 41(5): 645-55.

[22] Heiervang E, Goodman A, Goodman R. The Nordic advantage in child mental health: separating health differences from reporting style in a cross-cultural comparison of psychopathology. J Child Psychol Psychiatry 2008; 49(6): 678-85.

[23] Pirkis J, Burgess P, Coombs T, *et al*. Routine measurement of outcomes in Australia's public sector mental health services. Aust N Z Health Policy 2005; 2(1): 8.

[24] Schorre BE, Vandvik IH. Global assessment of psychosocial functioning in child and adolescent psychiatry. A review of three unidimensional scales (CGAS, GAF, GAPD). Eur Child Adolesc Psychiatry 2004; 13(5): 273-86.

[25] Green B, Shirk S, Hanze D, *et al*. The Children's global assessment scale in clinical practice: an empirical evaluation. JAACAP 1994; 33(8): 1158-64.

[26] Dyrborg J, Larsen FW, Nielsen S, *et al*. The Children's Global Assessment Scale (CGAS) and Global Assessment of Psychosocial Disability (GAPD) in clinical practice--substance and reliability as judged by intraclass correlations. Eur Child Adolesc Psychiatry 2000; 9(3): 195-201.

[27] Bird HR, Canino G, Rubio-Stipec M, *et al*. Further measures of the psychometric properties of the Children's Global Assessment Scale. Arch Gen Psychiatry 1987; 44(9): 821-4.

[28] Gowers SG, Harrington RC, Whitton A, *et al*. Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA). Glossary for HoNOSCA score sheet. Br J Psychiatry 1999; 174: 428-31.

[29] Gowers SG, Harrington RC, Whitton A, *et al*. Brief scale for measuring the outcomes of emotional and behavioural disorders in children. Health of the Nation Outcome Scales for children and Adolescents (HoNOSCA). Br J Psychiatry 1999; 174: 413-6.

[30] Pirkis JE, Burgess PM, Kirk PK, *et al*. A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures. Health Qual Life Outcomes 2005; 3: 76.

[31] Garralda ME, Yates P, Higginson I. Child and adolescent mental health service use: HoNOSCA as an outcome measure. Br J Psychiatry, 2000; 177(1): 52-8.

[32] Brann P, Coleman G, Luk E. Routine outcome measurement in a child and adolescent mental health service: an evaluation of HoNOSCA. Aust N Z J Psychiatry 2001; 35(3): 370-6.

[33] Leeflang MM, Deeks JJ, Gatsonis C, *et al*. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149(12): 889-97.

[34] Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic 1981; 86(2): 127-37.

[35] McGraw KO. Forming inferences about some intraclass correlation coefficients. Psychol Med. 1996; 1(1): 30-46.

[36] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86(2): 420-8.

[37] Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res 1999; 7(3): 301-17.

[38] Pies R. How "objective" are psychiatric diagnoses? (guess again). Psychiatry (Edgmont) 2007; 4(10): 18-22.

[39] Lundh A, Kowalski J, Sundberg CJ, *et al*. Children's Global Assessment Scale (CGAS) in a naturalistic clinical setting: Inter-rater reliability and comparison with expert ratings. Psychiatry Res 2010; 177(1-2): 206-10.

[40] Nichols D. Choosing an intraclass correlation coefficient. SPSS keywords 1998; 67.

[41] Miller PR. Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews. Psychiatry Res 2001; 105(3): 265-71.

[42] Noda AM, Kraemer HC, Yesavage JA, *et al*. How many raters are needed for a reliable diagnosis? Int J Methods Psychiatr Res 2001; 10(3): 119-25.

Received: October 21, 2011 Revised: January 24, 2012 Accepted: January 28, 2012