Reliability Study on the Japanese Version of the Clinician's Interview-Based Impression of Change

Akira Homma^a Yu Nakamura^b Shinichi Kobune^c Hirofumi Haraguchi^d
Nobuyoshi Kodani^e Isao Takami^c Joe Matsuoka^f Hiroaki Matsuda^e
Tadashi Kusunoki^g the SKETCH Study Group

^aTokyo Metropolitan Institute of Gerontology, ^bNara Medical University, ^cNovartis Pharma K.K.,
^dAventis Pharma Ltd, ^eNippon Boehringer Ingelheim Co. Ltd, ^fQuintiles Transnational Japan and
^gJapanese Society for Pharmacoepidemiology, Tokyo, Japan

Key Words

CIBIC-plus J · Clinician's Global Impression of Change ·
Dementia · Alzheimer's disease

Abstract

The Japanese version of the Clinician's Interview-Based Impression of Change (CIBIC-plus J) is a semistructured interview format including a 7-point Likert type scale for the Clinician's Global Impression of Change. It comprises subscales of DAD for ADL, Behave-AD for behavioral and psychiatric symptoms and MENFIS for cognitive and emotional impairment. DAD and Behave-AD are used for caregivers and MENFIS for both caregivers and patients. The objective of this study was to examine the interrater reliability of the CIBIC-plus J by the videotape method. Twenty videotaped patients including 13 real and 7 virtual patients were independently assessed by 11 raters. The kappa coefficient from the full set of data of 20 patients and 11 raters was 0.453, and from the limited data of 13 real patients, 0.383. By permitting one-point disagreements on the 7-point scale or by collapsing it into the 3-point scale (improved, no change, or worsened), kappa improved to 0.894 and 0.731, or 0.734 and 0.504,

respectively. From these results, it was concluded that the reliability of a 7-point Clinician's Global Impression of Change was halfway, i.e., neither sufficient nor insufficient. However, if one-point disagreements or collapsing into 3-point scale are allowable, the CIBIC-plus J will be sufficiently reliable.

Copyright © 2006 S. Karger AG, Basel

Introduction

The Japanese version of the Clinician's Interview-Based Impression of Change (CIBIC-plus J) was developed in 1997 [1]. Since then, it has been widely used in Japan as a primary outcome measure for clinical trials in Alzheimer's disease (AD).

Although global measures came into use in the early history of clinical trials in dementia, only a global measure on severity and/or change was used without any special instruments and only observation on patients was requested. In the meantime, the global change measure became to be called the Clinician's Global Impression of Change (CGIC) [2]. However, both its reliability and validity were questioned [1].

In an attempt to seek improved reliability, countermeasures were adopted to use a specifically designed interview format that included subscale items and caregiver interviews in addition to patient interviews. This type of a reformed instrument is called the CIBIC [3, 4], and instruments which include patient interviews 'plus' caregiver interviews are called the CIBIC-plus [5].

The Alzheimer's Disease Cooperative Study has developed a semistructured version of the CIBIC-plus called the ADCS-CGIC [6]. The domains of the ADCS-CGIC are cognition, behavior and social and daily functioning. The rater interviews the subject and informant at baseline and at the follow-up to assess the changes from baseline by the 7-point scales of the CGIC.

The NYU CIBIC-plus [7, 8] has structured interview formats including the Brief Cognitive Rating Scale [9] for patients, Functional Assessment Staging [10] for caregivers, Empirical Behave-AD for patients and Behave-AD [11] for caregivers.

The CIBIC-plus J has been designed with the intention of producing a Japanese version of the CIBIC-plus. It has semistructured interview formats for both caregivers and patients, including subscale items asking about the ADL of demented patients, behavioral and psychiatric symptoms, impaired cognitive functions, motivation activities and emotional functions of AD patients.

The common characteristic of global measures, regardless of whether the format is structured or not, is that the raters (clinicians) are requested to assess each patient's CGIC by their 'interview-based impression'. Because of this subjective nature of global measures, reliability has been an issue of concern, and some studies have examined this issue. Dahlke et al. [3], Boothby et al. [6] and Quinn et al. [12] studied interrater reliability. However, satisfactory results were not always obtained with respect to the reliability of global change measures, and the interrater reliability of the CIBIC-plus J has not yet been studied. The objective of this study was to examine whether the semistructured CIBIC-plus J and its CGIC have a practicable level of reliability.

Materials and Methods

Instruments and Report Format of the CIBIC-Plus J

The format of the CIBIC-plus J consists of the following and an assessment manual is provided. Subscales are compiled in the order of the Disability Assessment of Dementia (DAD) [13], Behavioral Pathology in Alzheimer's Disease (Behave-AD) [11, 14] and Mental Function Impairment Scale (MENFIS) [15, 16]. It is recommended the interviews be used in this order.

The Japanese versions of the DAD and Behave-AD have been translated from their respective originals and verified by back translation. MENFIS has been developed in Japan by Homma et al. [15] and Homma and Hasegawa [16]. Different MENFIS questionnaires are used for caregivers and patients, with the caregivers being interviewed first. Both forms include the same 13 items about impaired cognitive functions, motivation activities and emotional functions. Answers from caregivers and from patients are combined to produce a unified rating for each item using a 7-point scale where 0 signifies not impaired and 6 completely impaired.

Raters (clinicians) are requested to compare symptoms and conditions observed by the DAD, Behave-AD and MENFIS at the follow-up interview to the baseline interview and to assess the CGIC by their 'interview-based impression.' The observation window in the interviews should be the past 2 weeks and the CGIC is assessed by a 7-point scale of 'markedly improved' (score 1), 'moderately improved' (2), 'minimally improved' (3), 'no change' (4), 'minimally worsened' (5), 'moderately worsened' (6) and 'markedly worsened' (7).

Subjects

The subjects were patients with mild to moderate AD with a severity of CDR 1 or CDR 2. The diagnosis was confirmed by NINCDS-ADRDA criteria [17] as possible AD. From August 1998 to July 2001, 21 patients were selected at 15 hospitals by an attending geriatric psychiatrist. Informed written consent was obtained from the patient and/or caregiver.

Videotapes for Reliability Study

Interview and Videotape Recording

The patients and caregivers were interviewed by the attending geriatric psychiatrists twice, at baseline soon after entry into the study and after followup periods of 3–14 months. These interviews were performed using the CIBIC-plus J and were videotaped by the geriatric psychiatrists themselves.

Videotape Editing

The videos of all 21 patients were examined regarding compliance with the CIBIC-plus J and other deficits in recording. The videos of 8 patients were excluded while the remaining 13 were edited to remove unnecessary information for the reliability study and to shorten the viewing time. However, the CGIC of the remaining 13 patients was somewhat limited and ranged from 'minimally improved' (3) to 'moderately worsened' (6). In other words, scores of 'markedly improved' (1), 'moderately improved' (2) and 'markedly worsened' (7) were lacking or scarce. To compensate for this lack and to attain an even distribution over the 7-point scale of the CGIC, it was decided that the videos of 7 'virtual patients' would be added to the videos of the 13 'real patients.' Interview scenarios for the 'virtual patients' were provided and actors played the roles of the patients, caregivers and interviewers. Thus, a total of 20 videotapes of 20 patients were used in the reliability study.

Reliability Study

Eleven experienced geriatric psychiatrists were invited to participate as raters in the reliability study. They replayed the videos of the 20 patients and filled out the CIBIC-plus J forms. The video patients were arranged in a random order which was independently determined for each rater, and two videos for each patient were viewed, starting with the baseline first. The raters were not aware that some of the 20 patients were 'virtual'.

Analysis

The entire CGIC data set was arranged in a two-way table with 11 columns (raters) and 20 rows (patients) to calculate the pairwise proportions of agreement between the raters and the corresponding pairwise kappa using Cohen's method [18]. Schouten's overall kappa [19] was obtained from the mean proportion of agreement and regarded as the index of overall interrater reliability. The same analysis was also performed using the data for the 13 'real patients'. In these analyses, an overall kappa ≥ 0.8 was considered to represent the 'almost perfect' and ≥ 0.6 the 'substantial' level of reliability [19].

Kappa [20] was also calculated by regarding disagreement in adjacent points in the 7-point scale as 'agreed'.

Kappa in each category of the 7-point scale was obtained by the method of Davies and Fleiss [21]. Since it is known that the weighted mean of kappa in each category is equal to Schouten's overall kappa [22], one can examine which category is better or worse with respect to reliability. The adjacent categories with poor reliability were then collapsed into scales having 3 points; 'improved', 'no change' and 'worsened', and the reliability of the collapsed scales was examined.

As another index of reliability, the intraclass correlation coefficient (ICC) was calculated from the two-way table using 3 different methods (Fisher [23], Ebel [24] and Fleiss [25]).

Results

Subjects

Table 1 shows the characteristics of the 13 real and 7 virtual patients separately. The real patients included 6 men and 7 women, with a mean age (SD) of 70.5 (8.8) years. The CDR score was 1 in 8 and 2 in 5 patients. The mean score (SD) was 16.7 (5.3) in HDS-R, 14.4 (4.6) in MMSE and 28.4 (11.9) in ADAS-cog, and the mean duration of AD was 40.6 (21.5) months. A CGIC score of 'markedly improved/worsened' was lacking and 'moderately improved' was only seen in one case in the 13 real patients.

Therefore, scenarios for virtual patients were provided assuming a CGIC score of 'markedly improved' in 2, 'moderately improved' in 2, 'minimally improved' in 1 and 'markedly worsened' in 2. The scenarios for 7 virtual patients included 3 men and 4 women with a mean age (SD) of 72.3 (4.6) years, and the other characteristics were not assumed.

CGIC Data

Table 2 is a two-way table of the CGIC scores assessed by the 11 video raters for 20 patients. In many patients, CGICs by 11 raters agreed in certain two scores next to each other. For patient No. 11, the CGIC scores given by all 11 raters were identical (5, 'minimally worsened').

Table 1. Summary of baseline demographic characteristics of patients

	Real patients ^a	Virtual patients ^b	Total
Patients	13	7	20
Sex, male/female	6/7	3/4	9/11
Age, years	70.5 (8.8)	72.3 (4.6)	71.1 (7.5)
Duration of AD, months	40.6 (21.5)	–	–
CDR, 1/2	8/5	–	–
HDS-R	16.7 (5.3)	–	–
MMSE	14.4 (4.6)	–	–
ADAS-J cog	28.4 (11.9)	–	–
CGIC			
Markedly improved	0	2	2
Moderately improved	1	2	3
Minimally improved	2	1	3
No change	3	0	3
Minimally worsened	4	0	4
Moderately worsened	3	0	3
Markedly worsened	0	2	2

Figures in parentheses indicate SDs. CDR = Clinical Dementia Rating; HDS-R = The Revised Hasegawa Dementia Scale; MMSE = Mini-Mental State Examination; ADAS-J cog = Japanese version of Alzheimer's Disease Assessment Scale.

^a Real patients were interviewed and videotaped with their caregivers by attending physicians. However, patients in CGIC categories of markedly/moderately improved and markedly worsened were lacking or scarce.

^b Virtual patients were added to real patients to fill up the CGIC categories that were lacking. Scenarios were provided for the virtual patients and actors played the roles of patients, caregivers and physicians.

Reliability of the CGIC

Pairwise proportions of agreement for all 55 pairs of raters for 20 patients ranged from 30 to 80%, and Cohen's kappa ranged from 0.193 to 0.758. As shown in table 3, the mean proportion of agreement for all 20 patients was 53.8% and the corresponding Schouten's overall kappa was 0.453. The mean proportion and the overall kappa for the 13 real patients were 53.3% and 0.383, respectively.

Thus, the criterion for a 'substantial' level of reliability, kappa > 0.6 , was satisfied neither by the overall kappa from the 20 patients, nor the 13 real patients. However, by regarding disagreement between adjacent points as 'agreed', the mean proportion (range) for all 20 patients reached a high level of 94% (85–100%) and the criteria for 'almost perfect' reliability, kappa > 0.8 , was attained

Table 2. CGICs of 20 patients assessed by 11 raters

Patient No. ^a	Video rater No. ^b										
	1	2	3	4	5	6	7	8	9	10	11
1 ^c	2	1	1	2	1	2	1	2	1	2	1
2 ^c	2	1	1	2	1	1	1	2	1	2	2
3 ^c	1	1	2	2	2	2	2	1	2	2	1
4 ^c	3	2	2	2	2	3	2	3	2	2	2
5 ^c	3	3	2	3	3	3	3	3	3	2	2
6	5	4	3	3	3	3	3	4	3	3	3
7	4	5	5	4	3	3	4	4	4	4	3
8	4	4	3	4	4	4	4	4	4	4	4
9	4	5	5	5	5	5	2	4	3	3	3
10	4	6	5	4	4	4	5	4	5	5	4
11	5	5	5	5	5	5	5	5	5	5	5
12	5	5	5	5	6	6	6	5	5	5	5
13	6	6	6	6	6	5	6	5	4	5	5
14	6	6	6	5	6	5	5	5	7	4	5
15	5	6	6	6	6	6	6	6	5	5	5
16	6	6	6	6	6	6	6	4	5	6	5
17	6	6	6	5	6	6	6	6	6	5	6
18 ^c	7	7	7	6	6	6	6	7	7	7	7
19	7	6	7	7	7	7	7	7	7	6	7
20 ^c	7	7	7	7	7	7	7	6	7	7	7

1 = 'Markedly improved'; 2 = 'moderately improved'; 3 = 'minimally improved'; 4 = 'no change'; 5 = 'minimally worsened'; 6 = 'moderately worsened'; 7 = 'markedly worsened'.

^a Rank order of 20 patients by mean scores. When ties were noted, the rank orders of raters and patients were determined arbitrarily.

^b Rank order of 11 raters by mean scores on 20 patients, from severe to lenient.

^c Virtual patients.

Table 3. Summary of pairwise proportions of agreements and overall kappa coefficients (CGIC)

	All 20 patients		13 real patients	
	proportion of agreement, %	overall kappa	proportion of agreement, %	overall kappa
Complete agreement	53.8 (30.0–80.0)	0.453	53.3 (30.8–84.6)	0.383
Allowance for one-point discrepancy	94.0 (85.0–100)	0.894	90.8 (76.9–100)	0.731

Values for proportion of agreement are expressed as mean (range).

by the corresponding overall kappa of 0.894. The mean proportion and the overall kappa for the real patients were 90.8% and 0.731, respectively (table 3).

The kappa values in each category and in collapsed categories of the CGIC are presented in table 4. A distinctly large kappa of 0.722 was noted in 'markedly worsened', followed by 0.462 in 'no change'. It was smaller

still in the other categories of the original 7-point scale. As the last line shows, their weighted mean of 0.453 is identical to the overall kappa from the full set of data of the 20 patients shown in table 3.

Table 4 also shows that kappa became larger in collapsed categories and the overall kappa for the two kinds of 3-point scale (improved, no change, or worsened) be-

came 0.734 and 0.682, respectively, which exceeded the ‘substantial’ level of 0.6.

The CGIC data (table 2) were also analyzed as numerical data to calculate the ICC. Although three different methods were used, the results were almost the same. The ICCs (95% CI) using the Fisher, Ebel and Fleiss methods were 0.903 (0.837–0.944), 0.905 (0.840–0.945) and 0.903 (0.837–0.944) for all 20 patients, respectively, and 0.702 (0.525–0.831), 0.720 (0.546–0.843) and 0.703 (0.526–0.832) for the 13 real patients, respectively.

Discussion

The results obtained for all 20 patients show the overall kappa in the 7-point scale was 0.453 and the ‘substantial’ level of reliability, kappa 0.6, was not satisfied. Therefore, it was concluded the results were inconclusive since the GCIC interrater reliability was neither sufficient nor insufficient. However, if we permit one level of disparity in the 7-point scale, the overall kappa was 0.89, which reached the ‘almost perfect’ level (kappa > 0.8), and the ‘substantial’ level (kappa > 0.6) was reached by collapsing the 7-point scale into a 3-point scale (improved, no change, worsened) and the overall kappa was 0.734 or 0.682 by collapsing.

If we determine kappa according to the CGIC category, ‘markedly worsened’ was the highest at 0.722, followed by ‘no change’ at 0.462, which were higher than the other categories that had low kappa values. Since the 13 real patients do not include any ‘markedly worsened’ patients, it is natural that the overall kappa of the real patients, at 0.383, was lower than the 0.453 for all 20 patients. Also the ICC was lower for the 13 patients (0.702–0.720) compared with that for all 20 patients (0.903–0.905).

Since reliability is especially important for global measures such as the CGIC, several groups have already reported on the issue. The work of Knopman et al. [4] involved the test-retest reliability of CIBI, while the Alzheimer’s Disease Cooperative Study by Schneider et al. [7] examined the stability of repeated GCIC assessments over a short period of time. However, neither of these studies looked at interrater reliability. Interrater reliability has been studied by Dahlke et al. [3], Boothby et al. [6] and Quinn et al. [12]. These studies used various different indicators, such as the Pearson correlation coefficient, ICC and kappa coefficient to express reliability. Since the indicators were not the same, their comparison

Table 4. Overall kappas according to original and combined categories of CGIC

Category	Original 7-point scale	Modified	
		3-point scale, type 1	3-point scale, type 2
Markedly improved	0.409		0.838
Moderately improved	0.432	0.821	
Slightly improved	0.434		
No change	0.462	0.462	0.596
Slightly worsened	0.383		
Moderately worsened	0.395	0.784	
Markedly worsened	0.722		0.663
Overall ^a	0.453	0.734	0.682
95% CI	0.404–0.501		

^a Overall kappa by Schouten [19].

is difficult. Interrater reliability results reported include 0.66 (Pearson correlation coefficient) by Dahlke et al. [3] and 0.702 (ICC) by Boothby et al. [6], values which are intermediate between the 0.903–0.905 for all 20 patients and 0.555–0.576 for the 13 real patients, and not greatly different to those obtained in the present study. However, in the study by Dahlke et al. [3] comparisons were made with psychometric data, while in that of Boothby et al. [6] there was no baseline interview and 6-month retrospective artificial data on the conditions of the patients was prepared, which are methods offered to global assessment raters. The only interrater reliability research using video for both is believed to be that of Quinn et al. [12] and the present study.

Quinn et al. [12] reported a poor kappa coefficient of 0.18 for interrater reliability when 10 raters assessed videos for 12 patients using the 7-point CGIC. Compared to this result, our overall kappa coefficient was 0.453, which is higher than that of Quinn et al. [12]; however, one reason for the difference may be that the CIBIC they used did not include an interview of the caregivers, whereas our CIBIC-plus J did. Furthermore, the raters in the study by Quinn et al. [12] included 5 neurologists with experience in clinical trials for AD, as well as 5 raters without adequate experience. Since the interrater reliability was poor, Quinn et al. [12] collapsed the 7-point CGIC into 3 points (better, worse, unchanged) before analyzing, and reported the kappa coefficient increased to 0.76 (0.51 for correct video sequence and 0.12 for reverse sequence) if the raters were limited to 5 neurologists only. One other reason why our overall kappa coefficient was higher than

that of Quinn et al. [12] may be that our 11 raters were all experienced psychogeriatrists. An increase in the kappa coefficient after collapsing the 7-point scale to a 3-point scale was also observed in our study, and the kappa coefficient of 0.76 reported by Quinn et al. [12] with the 3-point scale was not greatly different to our results of 0.743 and 0.682.

The interrater reliability of global change measure, while not the same for each study, cannot necessarily be considered to be adequately high enough. However, if our results are examined together with theirs, various factors are thought to influence the interrater reliability, such as whether or not interviews to caregivers are taken into consideration, whether or not the experience of the raters is adequate, and so on. Based on these results, points that should be considered when conducting a clinical trial include a caregiver interview and the participation of experienced raters. If a clinical trial includes inexperienced raters, it is important to provide sufficient education and training. Another potential measure is to increase the number of subjects enrolled in the study.

The CGIC of the CIBIC-plus J is a 7-point scale ranging from 'markedly improved' to 'markedly worsened', and the overall kappa coefficient of 0.453 obtained in the 20 patients using the 7-point scale indicates the 7-point scale is reliable. Meanwhile, the 13 real patients did not include any patients rated as 'markedly improved' or 'markedly worsened'. In actual clinical trials, it is believed there are virtually no subjects who are rated as 'markedly improved' or 'markedly worsened'. In this respect, therefore, it is thought the results obtained in these 13 subjects represent the reliability of the CGIC scale when used in an actual clinical trial, and because the high value of 0.731 was obtained if a discrepancy of one point was allowed, it is believed the CGIC will have clinically sufficient reliability even when used in an actual clinical trial.

In the global change measure using the 7-point scale, as the respective operational criteria of the changes are not determined, it is natural that there is some degree of variation between raters; therefore, it is impossible to attain complete agreement with the 7-point assessment. Rather, high reliability is seen with the 3-point scale consisting of improved, unchanged and worsened. In other words, if we consider that whether or not all the raters can assess the pharmacological action in a right direction is important, we should be satisfied with the present results. However, future studies are needed to determine whether or not the 7-point assessment, which is currently used as a global assessment method, is a truly appropriate method with which to assess pharmacological action.

Appendix

The SKETCH study group at the time: H. Morimoto (Fujisawa Pharmaceutical Co., Ltd), T. Otani, T. Seita, N. Sekine (Takeda Chemical Industries, Ltd), M. Fujimoto, F. Kimura (Daiichi Pharmaceutical Co., Ltd), H. Kawaguchi (Mitsubishi Pharma Co., Ltd), K. Matsubara, H. Oi, N. Yoshida (Shionogi & Co., Ltd), T. Shimizu, T. Tsuyuki, H. Kasahara, S. Yoshida (Meiji Seika Kaisha, Ltd), K. Sha, M. Uchida (Nippon Boehringer Ingelheim Co., Ltd), H. Konno, T. Kihara, S. Ueno (Suntory Ltd), K. Shimada, J. Shimizu, S. Ishizaki (Nikken Chemicals Co., Ltd), T. Higuchi, H. Ishimado, H. Hashimoto (Janssen Pharmaceutical K.K.), M. Okamoto, K. Tsukahara (Bayer Yakuhin Co., Ltd), A. Chiba, T. Negishi, Y. Matsuda (Toyama Chemical Co., Ltd).

Acknowledgements

We would like to thank the following investigators for their patient recruiting, interviews and video-recording: S. Kawakatsu, Department of Neuropsychiatry, Yamagata University School of Medicine; Y. Ichimiya, Department of Neuropsychiatry, Juntendo University Urayasu Hospital; M. Shigeta, Department of Psychiatry, The Jikei University School of Medicine; K. Koyama, Department of Psychiatry, Tokyo Metropolitan Geriatric Hospital; S. Kitamura, Department of Internal Medicine, Nippon Medical University 2nd Hospital; Y. Imai, Graduate School of Social Services, Japan College of Social Work; H. Endo, Division of Geriatric Medicine, Chubu National Hospital; K. Kono, Division of Geriatrics, Aichi Koseiren Kainan Hospital; Y. Nakamura, Department of Psychiatry, Nara Medical University; A. Ueki, Department of Neuropsychiatry, Hyogo College of Medicine; Y. Kishikawa, Department of Psychiatry, Kobe Central Hospital; Y. Mimori, Third Department of Internal Medicine, Hiroshima University School of Medicine; N. Kamimura, Department of Neuropsychiatry, Kochi Medical School; J. Sanada, Psychiatry, Nanohana Clinic.

We also wish to thank following raters who participated in the study: N. Okamura, Department of Geriatric and Respiratory Medicine, Tohoku University School of Medicine; H. Tago, Department of Neuropsychiatry, Fukushima Medical University School of Medicine; M. Amari, Department of Neurology, Geriatrics Research Institute and Hospital; T. Asada, Department of Neuropsychiatry, Institute of Clinical Medicine, University of Tsukuba; Y. Sugai, Department of Psychiatry, Yokufukai Geriatric Hospital; K. Miyazu, Department of Psychiatry, Kanazawa University School of Medicine; K. Okada, Vice-president of Ohda Municipal Hospital; K. Urakami, Department of Biological Regulation School of Health Science, Faculty of Medicine, Tottori University; M. Takita, Department of Psychiatry, Imazu Red Cross Hospital; K. Isse, Department of Neuropsychiatry, Tokyo Metropolitan Toshima General Hospital; M. Usui, Psychiatry, Jiyu Clinic.

We are especially thankful to our advisor, Dr. Y. Mitsuyama: Vice-president of Miyazaki Medical College.

References

- Homma A, Asada T, Arai H, Isse K, Imai Y, Nishikawa T, Kobune S: Clinician's Interview-Based Impression of Change plus-Japan (CIBIC plus-J) concept and assessment manual (in Japanese). *Jpn J Geriatr Psychiatry* 1997; 8:855-869.
- Guy W: ECDEU Assessment Manual for Psychopharmacology. Washington, US Public Health Service, 1976, pp 534-537.
- Dahlke F, Lohaus A, Gutschmann H: Reliability and clinical concepts underlying global judgments in dementia: implications for clinical research. *Psychopharmacol Bull* 1992;28:425-432.
- Knopman D, Knapp MJ, Gracon SI, Davis CS: The Clinician Based Impression (CIBI): a clinician's global change rating scale in Alzheimer's disease. *Neurology* 1994;44:2315-2321.
- Oremus M, Perrault A, Demers L, Wolfson C: Review of outcome measurement instruments in Alzheimer's disease drug trials: psychometric properties of global scales. *J Geriatr Psychiatry Neurol* 2000;13:197-205.
- Boothby H, Mann AH, Barker A: Factors determining interrater agreement with rating global change in dementia: the CIBIC-plus. *Int J Geriatr Psychiatry* 1995;10:1037-1045.
- Schneider LS, Olin JT, Doody RS, Clark CM, Morris JC, Reisberg B, Schmitt FA, Grundman M, Thomas RG, Ferris SH: Validity and reliability of the Alzheimer's Disease Cooperative Study - Clinical Global Impression of Change. *The Alzheimer's Disease Cooperative Study. Alzheimer Dis Assoc Disord* 1997;11 (suppl 2):S22-S32.
- Corey-Bloom J, Anand R, Veach J: A randomized trial evaluating the efficacy and safety of ENA 713 (rivastigmine tartrate), a new acetylcholinesterase inhibitor, in patients with mild to moderately severe Alzheimer's disease. *Int J Geriatr Psychopharmacol* 1998;1:55-65.
- Reisberg B, Ferris SH: Brief Cognitive Rating Scale (BCRS) findings in primary degenerative dementia (PDD). *Psychopharmacol Bull* 1988; 24:629-636.
- Reisberg B, Ferris SH, Anand R, et al: Functional staging of dementia of the Alzheimer type. *Ann N Y Acad Sci* 1984;435:481-483.
- Reisberg B, Borenstein J, Salob SP, Ferris SH, Franssen E, Georgotas A: Behavioral symptoms in Alzheimer's disease: phenomenology and treatment. *J Clin Psychiatry* 1987;48(suppl):9-15.
- Quinn J, Moore M, Benson DF, Clark CM, Doody R, Jagust W, Knopman D, Kaye JA: A videotaped CIBIC for dementia patients. Validity and reliability in a simulated clinical trial. *Neurology* 2002;58:433-437.
- Gauthier L, Gauthier S, McIntyre M: Functional assessment in Alzheimer's disease (abstract). 16 Annu Meet Can Coll Neuropsychopharmacol Br Assoc Psychopharmacol. Montreal, Canada, June 1993.
- Asada T, Homma A, Kimura M, Uno M: Study on the reliability of the Japanese version of the BEHAVE-AD (in Japanese). *Jpn J Geriatr Psychiatry* 1999;10:825-834.
- Homma A, Niina R, Ishii T, Hasegawa K: Development of a new rating scale for dementia in the elderly: Mental Function Impairment Scale (MENFIS) (in Japanese). *Jpn J Geriatr Psychiatry* 1991;2:1217-1222.
- Homma A, Hasegawa K: Recent clinical trials of cholinomimetics for patients with Alzheimer's Disease in Japan; in Hanin I, et al. (eds): *Alzheimer's and Parkinson's Diseases*. New York, Plenum Press, 1995.
- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984;34:939-944.
- Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46.
- Schouten HJA: Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Stat Neerl* 1982;36:45-61.
- Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
- Davies M, Fleiss JL: Measuring agreement for multinomial data. *Biometrics* 1982;38:1047-1051.
- Cohen J: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213-220.
- Fisher RA: *Statistical Methods for Research Workers*, ed 13. Edinburgh, Oliver & Boyd, 1958.
- Ebel RL: Estimation of the reliability of ratings. *Psychometrika* 1951;16:407-424.
- Fleiss JL: *The Design and Analysis of Clinical Experiments*. New York, Wiley, 1986.