Published in final edited form as:
Alzheimer Dis Assoc Disord. 2004 ; 18(4): 219–222.

Reliability of Monitoring the Clinical Dementia Rating in Multicenter Clinical Trials

Kimberly A. Schafer[*], **Rochelle E. Tractenberg**[†], **Mary Sano**[‡], **Joan A. Mackell**[§], **Ronald G. Thomas**[*,‖], **Anthony Gamst**[*,‖], **Leon J. Thal**[*], **John C. Morris**[¶], and **for the Alzheimer's Disease Cooperative Study**

[*]Department of Neurosciences, University of California, San Diego, CA

[†]Department of Biomathematics and Biostatistics, Georgetown University School of Medicine, Washington, DC

[‡]Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, NY

[§]Department of Psychiatry, New York University Medical Center, and Pfizer, Inc., New York, NY

[‖]Department of Family and Preventive Medicine, University of California, San Diego, CA

[¶]Department of Neurology, Washington University School of Medicine, St. Louis, MO

Abstract

Context—The Clinical Dementia Rating (CDR) is quickly becoming a criterion standard in multicenter clinical trials in Alzheimer disease. An abbreviated version, with formal monitoring for consistency across sites and raters, is currently used in the Alzheimer's Disease Cooperative Study (ADCS).

Objective—To demonstrate the degree of agreement on CDR scoring of clinical monitors working independently from ADCS-CDR worksheets.

Design—Three members of the ADCS who are experienced and highly trained with respect to the CDR independently reviewed the ADCS-CDR worksheets of 15 subjects, assigning box and global CDR scores according to the prescribed algorithm.

Setting—The ratings were assigned during a single, 3-hour session in a closed room.

Participants—Two clinical monitors and one project director/clinical monitor supervisor.

Main Outcome Measures—Percent agreement, Kendall's tau-b, and Cohen's kappa were used to assess the degree of agreement of the raters with the previously established gold standard assessment on global and box scores for the 15 subjects.

Reprints: Rochelle E. Tractenberg, Department of Biomathematics and Biostatistics, 7 East Main, M7202, Georgetown University Hospital, 3800 Reservoir Rd. NW, Washington, DC 20007 (ret7@georgetown.edu).

Results—Raters, blinded to patient groupings, were in agreement with the Gold Standard global CDR assessment on 87% of ratings. Kappa values indicated good (κ = 0.66, orientation and judgment & problem solving boxes) to excellent (κ = 0.83, global CDR) agreement.

Conclusions—The ADCS-CDR worksheets were reliably and consistently scored by clinical monitors, who may be considered proxy gold standards for CDR assessment.

The Clinical Dementia Rating (CDR) is a global rating instrument developed for staging dementia, and consists of a global score derived from scores of impairment in six individual cognitive domains (memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care).[1] The CDR was developed by clinicians at the Memory and Aging Project at Washington University, and its interrater reliability and validity have been well established.[2,3] For assessment of their subject cohort, the clinicians at Washington University developed a standardized semi-structured interview for both the subject and informant, known as the Initial Subject Protocol (ISP). The ISP requires approximately 90 minutes and involves both physical and neurologic examinations of the patient. Its purpose is to collect information from both the subject and informant regarding the subject’s medical and social histories, as well as current health status.

The CDR has been widely used as both an inclusion criterion and an outcome measure in Alzheimer disease (AD) clinical trials and other studies of dementia.[3–7] At the Alzheimer’s Disease Cooperative Study (ADCS), a multicenter NIA-funded clinical trials consortium, the CDR has been, or will be, used in more than two thirds of study protocols. All ADCS protocols involve multiple sites and are longitudinal in design; the structural elements of the ISP are particularly attractive in this context because they promote consistency across sites. However, practical limitations on subject and staff time within multicenter clinical trials prevent the implementation of the ISP in its entirety.

In collaboration with the researchers at Washington University, the ADCS developed a modified CDR protocol (ADCS CDR Worksheets), which requires only 30 minutes to complete while retaining the critical components of the original (ISP) instrument, including structured interviews with both subject and informant. In both the original ISP and the ADCS modified version, the rater fills out worksheets documenting responses to all questions for later consideration and rating. In addition to saving 60 minutes per interview, the modified CDR protocol results in fewer pages of worksheets, thus further streamlining the procedure and minimizing data collection errors.

Using this abbreviated protocol, the ADCS applied the CDR as an inclusion criterion and an outcome measure in a recent clinical trial.[5] Neither the Washington University ISP nor the ADCS CDR Worksheets had been previously used to obtain the information necessary to derive the CDR in a multicenter clinical trial. To ensure the desired level of consistency across clinical raters at the 23 sites participating in this trial, and for the 9 CDR assessments required for each patient during the 2-year study, clinical monitors were used to act as “proxy gold standards” for the on-site raters. The term “gold standard” refers to the “expert” CDR rater against whose decisions those of other raters’ are compared.[1–3]

In the time since the aforementioned trial was initiated in 1992, the methodology of using the ADCS CDR Worksheets and monitoring of the CDR ratings has also been adopted and used in industry-sponsored clinical trials,[6] and the CDR remains an important instrument in ADCS research.[5,8] Given the frequency of use of the CDR in clinical trials, it is important to demonstrate that its use in a consistent, standardized manner is valid and reliable. In a multicenter clinical trial, the most practical way to monitor the quality of the CDR is to review the worksheets completed by on-site personnel. This provides the rater and the monitor with all of the relevant information for scoring the CDR. The present study was designed to measure the agreement of CDR scoring among ADCS clinical monitors, and their agreement with a gold standard, when scores were derived solely from information recorded on ADCS CDR Worksheets.

METHODS

Materials

Completed ADCS-CDR Worksheets were selected for 15 patients. Expert raters at Washington University, where the CDR was developed, who conducted the in-person interviews and completed the original interview worksheets, represent the *de facto* gold standard ratings for the global CDR and each of the six individual categories. Each of the five possible global CDR scores was represented as follows: CDR = 0 (no dementia): *N* = 3; CDR = 0.5 (questionable dementia): *N* = 3; CDR=1 (mild dementia): *N* = 3; CDR = 2 (moderate dementia): *N* = 4; CDR = 3 (severe dementia): *N* = 2. When dementia was present, the clinical diagnosis for each individual was dementia of the Alzheimer type, in accordance with validated diagnostic criteria.[9] The distributions of subjects and CDR levels were unknown to the raters. The worksheets were randomized, and the order of presentation was fixed for all raters.

Participants

Two clinical monitors and one Project Director, all located at different ADCS sites, participated in this study. All had previously received extensive training from the team at Washington University on use of the CDR and the ADCS CDR worksheets. Each of the two monitors had more than 2 years of experience in monitoring the CDR in ADCS clinical trials. The third rater had extensive clinical experience using the CDR to stage dementia. Additionally, as Project Director of a multicenter clinical trial,[5] she participated in ensuring monitor consistency in reviewing the CDR worksheets at 23 study sites.

Design and Procedure

The experiment took place in a closed conference room in a single 3-hour session. Cases were selected from research participants in the Memory and Aging project at Washington University in accordance with the following criteria: cases should represent each of the five CDR stages to allow reliability to be evaluated across the full CDR spectrum; for individuals with dementia (ie, CDR of 0.5 or greater), the clinical diagnosis should be uncomplicated dementia of the Alzheimer type to permit the focus to be on staging of dementia severity rather than differential diagnosis of dementia; and both the participant and his/her informant must consent to professional-quality videotaping of the CDR assessments for teaching purposes. The raters reviewed the CDR worksheets for each of the 15 subjects and independently completed the CDR scoring.[1] The experimental setting represented the only deviation from monitoring in the field.

Statistical Methods

Validity and reliability were demonstrated with measures of agreement of the raters with the gold standard (percent agreement, Kendall's tau-b, and kappa), while interrater reliability (agreement among the three raters themselves) was demonstrated by percent agreement. We constructed 95% confidence intervals for these statistics via the bootstrap algorithms in S-Plus for Window 98 (1000 replications), except for those intervals for values of Kendall's tau, which were calculated as t ± 2(standard error).

RESULTS

Table 1 presents the measures of agreement of the three raters with the gold standard. The percent agreement, Kendall Tau-b, and kappa were computed for the global CDR and each individual domain, as were their 95% confidence intervals. Agreement with the gold standard on the global CDR was 87% [76%, 96%], t = 0.929 [0.869, 0.989], with kappa = 0.83 [0.68, 0.94], representing excellent agreement after chance agreement is taken into account. For each of the six individual domains, agreement with the gold standard ranged from 73% [60%, 84%] to 87% [78%, 96%], r = 0.870 [0.800, 0.940] to 0.924 [0.864, 0.984], with kappa values ranging from 0.66 [0.49, 0.80] (very good) to 0.83 [0.65, 0.93] (excellent) (Kappa values between 0.40 and 0.75 are classified as ranging from fair to good, with values above 0.75 classified as excellent.[10]). Interrater agreement on the global CDR was 87% [76%, 96%], with interrater agreement for each of the six individual domains ranging from 73% [64%, 87%] to 87% [78%, 96%]. The lowest interrater agreement was in judgment & problem-solving (73% [64%, 87%]).

DISCUSSION

One of the principal reasons for the use of the CDR in clinical studies is the high degree of reliability across physicians and trained non-physicians which has been reported for the original Washington University ISP. Among five physicians reviewing 10 videotapes each, the rate of agreement on the global CDR score was 80%, with a correlation of 0.91.[2] Within this group (n = 5), the agreement with the gold standard on the individual domains ranged from 68% to 88%, with correlations ranging from 0.85 to 0.95. A similar study with three clinical nurse specialists found 81% agreement with a gold standard on the global score, with agreement on the individual domains ranging from 73% to 81% (no correlations were reported for this group).[3]

Both of these studies involved small groups of specialists at a single site (Washington University) rating the CDR based on videotaped interviews with subjects and informants. Although these are small sample sizes, they reflect the circumstances of clinical trials and other situations where the agreement of a small group of specialists is relevant. In the present study, the three ADCS clinical monitors (from different study sites) achieved high levels of agreement, equivalent to or greater than those established levels. This study thus independently confirms the previous reports of high interrater reliability on the CDR among trained raters and establishes the reliability of clinical monitoring of the CDR. While all of these studies involved very small groups of participants, the demonstration of agreement in ratings is clinically realistic and suggests that oversight by a small group of highly trained CDR raters can be reliable.

Reliability for the judgment & problem solving subscore has been more difficult to achieve than for the other subscores (see, eg, Burke et al[2]), likely because this domain samples broad areas of dementia-related dysfunction. For example, judgment & problem solving includes the large cognitive domain of “executive functions,” which encompass problem-solving, abstract reasoning, attention, sequencing, motor control, and response inhibition, as well as personality/behavior changes considered as “social judgment.” The large variety of dysfunctions to be evaluated in judgment & problem solving stand in contrast to the relatively limited domains of orientation and personal care, for which the discrete number of well-characterized behaviors to be evaluated enhance interrater reliability. The excellent agreement reported here and in previous studies[2,3] for the CDR global rating suggests that the algorithm adequately accounts for the different complexities among the box scores.

It is difficult to ensure consistency in CDR ratings in a multicenter clinical trial, as ratings are made by clinicians with varying amounts of experience and training. Since multicenter clinical trials are standard for evaluating the efficacy of pharmaceutical agents, the systematic and reliable collection of data that can be confidently compared across these sites is of utmost importance. This study demonstrates that clinical monitors who undergo extensive training can achieve high levels of consistency and interrater agreement on the CDR. These monitors can therefore be used as proxy gold standards to evaluate CDR ratings by multicenter site personnel and to ensure consistency and comparability in ratings across sites.

In an earlier report,[11] we studied agreement on CDR performance within a large sample (82 raters). That study quantified the extent, and qualified the nature, of disagreements on CDR global and box scores to identify areas in need of further training. Because the instrument involves semi-objective ratings, some disagreement is always possible. In that report, we found that the ADCS monitors were fundamentally different from clinicians administering the CDR at the sites of a multicenter clinical trial, although the levels of agreement with the gold standard in terms of global CDR for site raters and monitors were good (kappa = 0.74 in persons newly trained to administer the CDR) or excellent (kappa = 0.78 in experienced raters; kappa = 0.83 for the same three ADCS monitors described here). Taken together, these results suggest that the CDR can be reliably used by trained clinical personnel with oversight from highly trained monitors in multicenter studies.

ACKNOWLEDGMENT

The authors thank Chris Ernesto who participated in the development of this protocol.

Supported by the ADCS NIA grant AG10483.

REFERENCES

1. Morris JC. The Clinical Dementia Rating current version and scoring rules. Neurology. 1993; 43:2412–2414. [PubMed: 8232972]
2. Burke WJ, Miller JP, Rubin, et al. Reliability of the Washington University Clinical Dementia Rating. Arch Neurol. 1988; 45:31–32. [PubMed: 3337672]
3. McCulla MM, Coates M, Van Fleet N, et al. Reliability of clinical nurse specialists in the staging of dementia. Arch Neurol. 1989; 46:1210–1211. [PubMed: 2818255]
4. Thal LJ, Carta A, Clarke WR, et al. A 1-year multicenter, placebo-controlled study of acetyl-L-carnitine in patients with Alzheimer's disease. Neurology. 1996; 47:705–711. [PubMed: 8797468]
5. Sano M, Ernesto C, Thomas RG, et al. A controlled trial of selegiline, alpha-tocopherol, or both as treatment for Alzheimer's disease. N Engl J Med. 1997; 336:1216–1222. [PubMed: 9110909]
6. Rogers SL, Farlow MR, Doody RS, et al. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Neurology. 1998; 50:136–145. [PubMed: 9443470]
7. Morris JC, Heyman A, Mohs RC, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology. 1989; 39:1159–1165. [PubMed: 2771064]
8. Morris JC, Ernesto C, Schafer K, et al. Clinical Dementia Rating training and reliability protocol: the Alzheimer's Disease Cooperative Study Experience. Neurology. 1997; 48:1508–1510. [PubMed: 9191756]
9. Morris JC. Validation of clinical diagnostic criteria for Alzheimer's disease. Ann Neurol. 1988; 24:17–22. [PubMed: 3415196]
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:671–679.
11. Tractenberg RE, Schafer K, Morris JC. Interobserver disagreements on Clinical Dementia Rating assessment: interpretation and implications for training. Alzheimer Dis Assoc Disord. 2001; 15:155–161. [PubMed: 11522933]

TABLE 1

Agreement on CDR Global and Box Scores (95% Confidence Interval)

Domain	% Agreement with Gold Standard	Kendall Tau	Kappa	% Agreement Among Raters
Global CDR	87 (76, 96)	0.93 (0.87, 0.99)	0.83 (0.68, 0.94)	87 (76, 96)
Memory	82 (71, 91)	0.92 (0.86, 0.97)	0.78 (0.63, 0.91)	82 (73, 93)
Orientation	73 (62, 84)	0.87 (0.82, 0.93)	0.66 (0.49, 0.80)	87 (78, 96)
Judgment and problem-solving	73 (60, 84)	0.87 (0.80, 0.94)	0.66 (0.48, 0.83)	73 (64, 87)
Community affairs	78 (67, 89)	0.87 (0.81, 0.93)	0.69 (0.53, 0.84)	78 (67, 89)
Home & hobbies	80 (67, 89)	0.91 (0.85, 0.96)	0.74 (0.58, 0.88)	80 (69, 91)
Personal care	87 (78, 96)	0.92 (0.86, 0.98)	0.81 (0.65, 0.93)	87 (78, 96)