
doi: 10.1093/ageing/afn278
Published electronically 15 January 2009

Inter-rater reliability of the DRS-R-98 in detecting delirium in frail elderly patients

SIR—Delirium is a common, but often under-recognised problem in elderly people. Simple instruments to detect and grade delirium have been proposed as a remedy to under-recognition. Such instruments need to be reliable across different raters, as they are likely to be used by clinicians with differing clinical specialisations and levels of experience. Inter-rater reliability (IRR) has not been rigorously studied for most delirium rating scales, and even then typically by developers of the scales [1]. These studies commonly report different IRR statistics, and use IRR and inter-rater agreement interchangeably, even though they are not exactly the same [1]. Pearson’s correlation (r) is a measure of consistency between raters which may be high even when agreement is low. An intraclass correlation coefficient (ICC) is a measure of variance, and hence measures IRR. Cohen’s kappa (κ) is a measure of chance-corrected absolute agreement between raters. Its variation, the weighted κ, weights agreement by degree (agreement is higher if there is only a one-point difference and progressively lower as the difference increases) [1, 2].

Here, we assessed the IRR of DRS-R-98 [3]. This instrument is a revised version of the Delirium Rating Scale (DRS) [4] and allows for assessment of both delirium diagnosis and severity. It has been validated for use by psychiatrists with experience in delirium [3]. We also sought to further explore the construct validity of the DRS-R-98 by investigating whether IRR systematically varies by cognitive diagnosis and level of frailty.

Methods

Sample

We used a convenience sample of geriatric medicine patients at a 1000-bed tertiary care teaching hospital in Halifax, Canada, between November 2003 and November 2006. Using a standard formula [5] and accepting that an intraclass correlation of 0.7 would be the minimum acceptable, we aimed for at least 36 per diagnostic group of no cognitive impairment (NCI), delirium and dementia.

IRR was assessed for pairs of raters, from a pool of three staff geriatricians and six residents in internal, family or geriatric medicine. Each pair included a staff geriatrician with experience in delirium. The raters did not have extensive training in the instrument, except that it had been demonstrated by someone familiar with its use (KR) and they referred to the standard DRS-R-98 instructions at the time of assessment [3]. Each patient was independently and blindly assessed by two raters within 1 h.

Instruments

All patients had a Comprehensive Geriatric Assessment (CGA) [6] and Mini-Mental State Examination (MMSE) [7] as part of usual care prior to administration of the DRS-R-98. These were done during the same clinical encounter (e.g. outpatient visit or hospital admission) but for inpatients, not necessarily on the same day. Dementia and delirium were diagnosed using DSM-IV diagnostic criteria. We used standard criteria for ‘cognitive impairment, no dementia’ (CIND) [8]. The cognitive diagnoses were NCI, CIND, dementia, delirium, delirium superimposed on CIND and delirium superimposed on dementia.

The DRS-R-98 is a clinician-rated instrument. It includes 13 severity ‘symptoms’ [sleep–wake cycle disturbance, perceptual disturbances and hallucinations, delusions, liability of affect, language, thought process abnormalities, motor agitation or retardation, orientation, attention, memory (short- and long-term) and visuospatial ability] and 3 ‘diagnostic items’ (temporal onset, fluctuation and physical disorder) [3]. The ICC between two raters for the DRS-R-98 total score was first reported as 0.98 [3].

We rated frailty using a Frailty Index derived from the patient’s CGA at admission [6, 9, 10]. Impairments were counted in 10 domains: cognition, emotion, communication, mobility, balance, bowel and bladder function, nutrition, activities of daily living and social resources. Each item was scored 0 = no problem, 0.5 = minor problem and 1.0 = major problem. Based on prior distributions of scores, [9, 10] we made two modifications to the FI-CGA to better characterise co-morbidities and medications. First, instead of the original scoring, the co-morbidities were simply counted and the medications were scored as 0 = no medications, 0.5 for each of the first five medications and 1.0 for each medication ≥ 6. The deficit count was divided by 80, the highest possible score if all problems were present and given a maximum of 40 illnesses and medications, to yield an index ranging from 0 to 1.
Research letters

Table 1. Demographic and clinical characteristics of patients assessed for delirium, by cognitive subgroup

<table>
<thead>
<tr>
<th></th>
<th>NCI</th>
<th>CIND</th>
<th>Dementia</th>
<th>Delirium</th>
<th>CIND &amp; delirium</th>
<th>Dementia &amp; delirium</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (%)</td>
<td>33 (23%)</td>
<td>21 (14%)</td>
<td>36 (25%)</td>
<td>23 (16%)</td>
<td>10 (7%)</td>
<td>22 (15%)</td>
</tr>
<tr>
<td>Age (95% CI)</td>
<td>82.5 (79.9–85.2)</td>
<td>81.7 (78.2–85.2)</td>
<td>78.1 (75.7–80.6)</td>
<td>80.7 (77.9–83.4)</td>
<td>83.2 (78.5–87.9)</td>
<td>83.5 (81.3–85.7)</td>
</tr>
<tr>
<td>N (% female)</td>
<td>22 (67%)</td>
<td>16 (76%)</td>
<td>18 (50%)</td>
<td>17 (74%)</td>
<td>8 (80%)</td>
<td>14 (64%)</td>
</tr>
<tr>
<td>From community N (%)</td>
<td>27 (82%)</td>
<td>20 (95%)</td>
<td>32 (89%)</td>
<td>20 (87%)</td>
<td>9 (90%)</td>
<td>21 (95%)</td>
</tr>
<tr>
<td>Frailty index mean (CI)</td>
<td>0.31 (0.27–0.34)</td>
<td>0.32 (0.28–0.37)</td>
<td>0.40 (0.36–0.44)</td>
<td>0.42 (0.37–0.46)</td>
<td>0.40 (0.32–0.47)</td>
<td>0.44 (0.40–0.49)</td>
</tr>
<tr>
<td>MMSE score mean (CI)</td>
<td>26.2 (24.8–27.6)</td>
<td>23.2 (20.4–26.1)</td>
<td>16.3 (13.6–19.1)</td>
<td>18.6 (15.1–22.1)</td>
<td>18.9 (13.8–24.0)</td>
<td>14.0 (10.2–17.7)</td>
</tr>
<tr>
<td>DRS-98 scores mean (CI)</td>
<td>5.4 (3.0–7.9)</td>
<td>7.5 (4.9–10.0)</td>
<td>14.1 (10.7–17.4)</td>
<td>18.3 (15.0–21.6)</td>
<td>18.3 (11.6–24.9)</td>
<td>20.2 (16.1–24.3)</td>
</tr>
</tbody>
</table>

NCI = no cognitive impairment, CIND = cognitive impairment, no dementia, CI = 95% confidence interval.

Analysis

We assessed IRR for each of the 16 DRS-R-98 items and for the total scores using four techniques: Pearson’s correlation, unweighted and weighted kappa, and ICC (two-way random, absolute agreement on single measures model) [2]. A score ≥17.75 points screened positive for delirium [3]. Weighted $\kappa$ values were compared by cognitive diagnostic category, and by grade of frailty [11]. Sensitivity and specificity of the DRS-R-98 compared with clinical diagnosis of delirium was calculated using standard formulae. Less than 1% of the item values were missing; these were replaced with 1.5 (the middle of the 0–3 range). No individual form had more than four missing items.

The study protocol was approved for use by the Research Ethics Committee of the Capital District Health Authority, Halifax, Canada.

Results

The mean age was 81.2 years; 66% were women. Most (106/145, 73%) were inpatients. Outpatients were seen in the emergency department (n = 17), clinics (n = 7) or at home (n = 15) (Table 1). Medications ranged from none to 25 (mean 6.1, SD 3.8); most people (77; 53%) took six or more medications. The co-morbidity count ranged from 1 to 18 (mean 7.1, SD 2.7). The group was frail (mean FI = 0.38 ± 0.12) and frailty was approximately normally distributed. Delirium was clinically diagnosed in 38%, and 40% had dementia. The mean MMSE score was 19.8 (SD 8.2); the DRS-R-98 and MMSE were negatively correlated ($r = -0.70$).

Delirium did not vary by sex or age (Table 1). Patients with delirium had no more medications (mean 6.2 vs. 6.0 medications) or higher co-morbidity counts (mean 7.4 vs. 7.0 illnesses), but they were more frail (mean FI-CGA 0.42 vs. 0.31; P<0.001) and more cognitively impaired (MMSE scores: 16.8 ± 8.2 vs. 21.6 ± 7.7; P > 0.001).

DRS-R-98 scores ranged from 0 to 44. Patients with NCI had the lowest scores (mean 5.4, 95% CI: 3.0–7.9). Those with delirium had the highest scores (mean 19.1, 95% CI: 16.8–21.4) with patients with dementia having intermediate scores (mean 16.4, 95% CI: 13.7–19.0); 11 people with dementia had scores of 17.75 or greater, but did not have a clinical diagnosis of delirium. Pearson’s $r$ and ICC gave almost exactly equal estimates of IRR (0.93 and 0.92). The weighted $\kappa$ was lower at 0.76. The inter-rater agreement (weighted $\kappa$) for individual items varied between 0.52 for ‘physical disorder’ to 0.77 for ‘orientation’ item (Figure 1). The absolute total score agreement was 16.6%, with 95.5% agreement in identifying delirium using the published cut-off score of 17.75. There was no difference in agreement by diagnostic subgroups (weighted $\kappa$ 0.72 subgroup with delirium and dementia; 0.76 with delirium and NCI) or by level of frailty (weighted $\kappa$ 0.70 for least frail vs. 0.75 for most frail). In the undifferentiated clinical sample (N = 145), the DRS-R-98 had a sensitivity of 0.56 and specificity of 0.82 as a screen for delirium in relation to the ‘gold standard’ of clinically diagnosed delirium, and the area under the receiver operating characteristic (ROC) curve was 0.80. In the subset of 58 patients with underlying dementia, its sensitivity and specificity were 0.59 and 0.67, respectively, and the area under the ROC curve fell to 0.67 (Figure 2).

Discussion

In a usual care setting, the DRS-R-98 showed good IRR (ICC = 0.92), moderate inter-rater agreement (weighted $\kappa = 0.76$) and good construct validity. The ICC was lower than 0.98 reported in the original validation paper [3]. There DRS-R-98 performance was assessed by psychiatrists in a sample...
of 68 individuals divided into mutually exclusive diagnostic groupings: delirium (N = 24, mean age 64), dementia (N = 13, mean age 76), depression (N = 12, mean age 58), schizophrenia (N = 9, mean age 41) and other (N = 10, mean age 50) [3].

Higher inter-rater agreement was found for individual items with behavioural anchors, with more specific criteria for rating and those influenced by historical information from the patient’s charts. Hence reliability and agreement of future scales might be increased by providing behavioural anchors and more specific methods of rating, strategies that have improved item reliability for the Hamilton Depression Rating Scale [12]. The disruption of sleep–wake cycle disturbances and motor agitation may lead to better recognition, something also reported for hyperactive delirium [13, 14].

Eleven patients without delirium had scores higher than the suggested cut-off. Of these, three had severe Alzheimer’s disease, three stroke, two dementia with Behavioural and Psychological Symptoms of Dementia (BPSD) and one Lewy body dementia. The overlap in DRS-R-98 scores between delirium and dementia is clinically important. Here, sensitivity was low (<0.6) and specificity suffered for those with underlying dementia in contrast to previous reports of very favourable sensitivity and specificity (both >0.90) [3]. The difference might reflect differences in the study samples: we used a usual-care sample in which delirium and dementia often coexisted, whereas the previous study was limited to mutually exclusive diagnostic groups. In addition, some features of Lewy body dementia (fluctuation, clouding of consciousness), vascular dementia (acute onset, fluctuation) and many BPSD (hallucinations, psychomotor agitation) overlap with delirium; this might also influence IRR.

Our data must be interpreted with caution. The raters had varying expertise and most did not have extensive training in the use of the instrument. On the other hand, we studied the scale in a ‘real-life’ setting to better identify problems associated with delirium detection [15]. We found that the DRS-R-98 has good IRR even between clinicians with varying levels of training. The clinical assessments and DRS-R-98 ratings were not necessarily done on the same day. This is a limitation particularly for the determination of sensitivity and specificity of the DRS-R-98 as a screen for delirium, given that patients with delirium have a fluctuating course.

IRR and agreement are important properties of delirium assessment tools. Our findings support the call for IRR to be assessed whenever an instrument is used [16]. The DRS-R-98 has the advantage of rating severity, and may be appropriate for studies of delirious patients in which quantifying severity is important, although overlap in scores between dementia and delirium is an important consideration. The DRS-R-98 appears to serve as an appropriate adjunct to—but not a substitute for—the clinical diagnosis of delirium.

**Key points**

- Simple, reliable, instruments to detect and grade delirium have been proposed as a remedy to the common problem that delirium is under-recognised.
- Even so, the IRR of many delirium scales is commonly not tested except by the scale’s originators.
- Here, the 1998 revision of the Delirium Symptom Rating Scale showed good IRR, moderate inter-rater agreement and good construct validity. Its use as a diagnostic instrument was compromised in people with dementia, where it showed only moderate sensitivity and specificity.

**Funding**

M.A. is supported by a post-doctoral research fellowship from the Canadian Institutes of Health Research and by a Killam Scholarship. R.B. is supported on academic leave at the Dalhousie University in part by the Fountain Innovation Fund of the Queen Elizabeth II Health Sciences Foundation and in part by Goulburn Valley Health. K.R. receives
Research letters

career support from the Dalhousie Medical Research Foundation as Kathryn Allen Weldon Professor of Alzheimer Research.

Conflict of interest

Each of the authors asserts no conflict of interest in relation to this manuscript.

MELISSA K. ANDREW1, RAVI BHAT2, BARRY CLARKE3, SUSAN H. FRETER1, MICHAEL R. H. ROCKWOOD4, KENNETH ROCKWOOD1,*

1 Division of Geriatric Medicine, Dalhousie University, Halifax, NS, Canada
Email: Kenneth.Rockwood@Dal.ca

2 Goulburn Valley Area Mental Health Service, University of Melbourne, Shepparton, Vic, Australia

3 Department of Family Medicine, Dalhousie University, Halifax, NS, Canada

4 Geriatric Medicine Research Unit, Dalhousie University, Halifax, NS, Canada

* To whom correspondence should be addressed

References


doi: 10.1093/ageing/afn298
Published electronically 28 January 2009