Validation of the Risk Model for Delirium in hip fracture patients

Sophie Moerman, M.D., Wim E. Tuinebreijer, M.D., Ph.D., Maarten de Boo, M.D., Peter Pilot, M.Sc. Ph.D., Rob G.H.H. Nelissen, M.D., Ph.D., Anne J.H. Vochteloo, M.D.

⁎ Corresponding author. Department of Orthopaedics LUMC, P.O. Box 9600, 2300 RC Leiden, The Netherlands. Tel.: +31 71 5263606; fax: +31 71 5266743.
E-mail address: ajvochteloo@hotmail.com (A.J.H. Vochteloo).

Abstract

Objective: The Risk Model for Delirium (RD) score is a 10-item questionnaire that allocates hip fracture patients after admission to hospital to be either at high or at low risk for delirium. This allows targeted preventive actions. Clinical reliability, validity and feasibility of the RD score are discussed.

Methods: Demographic data, RD score and delirium incidence of all consecutive admissions for hip fractures in patients 65 years and older were collected. In 102 patients, the RD score was repeated. Interobserver reliability and validity were determined. The correlation between delirium and items both included and not included in the RD score was calculated.

Results: A total of 378 patients were included; 102 (27%) were diagnosed with a delirium. The intraclass correlation coefficient of the RD score was 0.77 (confidence interval (CI) 0.68–0.84). Sensitivity was 80.4% (71.4–87.6), and specificity was 56.2% (50.1–62.1). Area under the receiver operating characteristic curve was 0.73 (CI 0.68–0.77). A multivariable logistic regression analysis showed that besides the RD score, a trochanteric fracture and male gender were independent risk factors for delirium.

Conclusions: The RD score is a reliable, feasible and valid instrument for predicting delirium in hip fracture patients.

Keywords: Delirium; Prediction; Risk model; Score; Validation; Hip fracture patients

1. Introduction

Delirium in hip fracture patients is a serious complication, leading to higher morbidity and mortality [1–4]. A recent Cochrane review states that proactive geriatric consultation can reduce delirium incidence [5]. To optimize patient care, it is important to perform a preoperative risk assessment in order to target preventive interventions. This risk assessment should be simple and brief to increase participation of both patients and medical professionals.

Several authors described a model that identifies patients at high-risk for delirium. These models were applied in cohorts of vascular surgery patients [6], elective (non-cardiac) surgery patients [7] and four cohorts of cardiac surgery patients [8–11]. All these models contained items that are not applicable to hip fracture patients. On the contrary, these models are not only patient group specific, but are also designed for elective surgery patients. Kalisvaart et al. published the outcome of a risk score for delirium in a population that contained both elective hip replacement surgery and hip fracture patients [12]. They scored visual impairment, disease severity [using the Acute Physiology and Chronic Health Evaluation (APACHE) II score] [13], mental impairment [scored with the Mini Mental State Examination (MMSE)] [14] and dehydration (expressed by blood urea nitrogen–creatinine ratio). We chose to develop a simpler model that is easy to use in an acute admission setting to achieve maximum use in daily practice: the Risk Model for Delirium (RD) score. Recently, we published the results of the use of the RD score in daily practice [15].
This study evaluates the clinical reliability, validity and feasibility of the RD score in hip fracture patients.

2. Methods

2.1. Patients

All consecutive admissions for a hip fracture to a 450-bed Dutch teaching hospital between January 2008 and December 2009 were registered prospectively. This was part of daily care and monitoring according to our local hip fracture protocol [16]. Patients of this cohort aged 65 years (n=445) and older were included in the current study.

Sixty-seven patients were excluded because the RD score was not completed sufficiently. Final analysis was therefore performed in 378 patients with a mean (S.D.; range) age of 83.8 (7.3; 65–101) years; 279 (73.8%) were female.

Failure to complete the RD score was mainly because of lack of time of medical personnel at the emergency department (ED) or refusal of patients to cooperate. The excluded patients had a delirium incidence of 28.4%, not significantly different from the included study cohort (27%, P=.82). Furthermore, there was no difference in mean (S.D.) age between these groups [82.4 (7.4) vs. 83.8 (7.3) years; P=.17].

The following baseline characteristics were collected at admission: age, gender, presence of a partner, presence of dementia, prefracture living situation, American Society of Anaesthesiologists (ASA) physical status classification, psychotropic drug use (antidepressant, antipsychotic or any form of tranquillizer), RD score, type of fracture, hip fracture treatment, type of anesthesia and occurrence of a delirium during admission [17]. Presence of dementia was determined upon history taking from patients, families and caretakers. Ward nurses observe patients three times a day for clinical signs of a delirium as described in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria [18]. Symptoms of a delirium were registered in both the medical and nursing staff records. Directly after discharge, both records were examined independently by two authors (S.M. and A.V.) to register prevalence of a delirium. In case of doubt (36 patients), a psychiatrist was consulted, who diagnosed a delirium in 27 patients.

2.2. RD score

The RD score was developed in 2004 by the Department of Psychiatry of the Reinier de Graaf Hospital, Delft, the Netherlands [15]. At that moment, the most important known risk factors for a delirium were used in the RD score [7,19–24]. Based on the clinical experience, patients with a score of 5 or more points were considered to be high-risk patients. At admission at the ED, nurses (or doctors) filled out the RD score with the patient and his or her family or caretakers. According to the hospital’s delirium protocol, high-risk patients (with a score of 5 or more points) received two times daily 1 mg of haloperidol as delirium prevention (in case of absence of contraindications). Independent of the RD score, all patients were monitored for delirium during hospitalization as described above. When patients developed a delirium, they were fully clinically assessed to exclude a somatic cause and treated in collaboration with the psychiatric department.

In 102 patients, the ward nurse at the orthopaedic department performed the RD score a second time when the patient and his or her family arrived on the ward. These 102 patients together with the raters [both ED and ward nurses (doctors)] were selected randomly. Ward and ED nurses (or doctors) completed the score independent of each other. We used these independently executed RD scores to calculate the interobserver variability.

All RD score sheets of all patients were checked for possible errors and, if applicable, were corrected.

2.3. Statistical analysis

All items of the questionnaires were imported into IBM SPSS Statistics 19. (IBM Corporation, Somers, NY, USA)

2.3.1. Clinical reliability

The clinical reliability of the RD score was analyzed using the intraclass correlation coefficient (ICC) and the kappa coefficient. The ICC and the kappa coefficient are measures of the interobserver reliability, which assesses the degree in which observers assign the same ratings. Values
of 0–0.20 were regarded as slight agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial and 0.81–1 as almost perfect agreement [25]. As a measure of test–retest agreement, the standard error of measurement was calculated by dividing the mean difference in score between the initial test and the retest by the square root of 2 [26]. The standard error of measurement must be interpreted in relation to the mean. To account for this relationship, the coefficient of variation was calculated. In order to test the validity of the results of this retrospective analysis of the RD score, a post hoc power analysis was performed. A sample size of 70 subjects with two observations per subject achieved an 84% power to detect an ICC of 0.70 under the alternative hypothesis when the ICC under the null hypothesis is 0.50 using an F-test with a significance level of .05 (calculated with PASS 2008, version 08.05).

2.3.2. Validity

A receiver operating characteristics curve (ROC) was created by plotting the sensitivity (true-positive rate) versus the 1–specificity (false-positive rate). The actual area under the ROC (AUROC) measures the ability of the instrument to classify correctly the patients with and without a high risk for delirium to identify the best cutoff point.

The percentages of scores below 5 and above 14 for the RD score were calculated to assess floor and ceiling effects.

2.3.3. Test items statistics

Reliability of the individual items of the RD score was expressed using the ICC of these items. Spearman correlation coefficient was calculated between items of the RD score and the occurrence of delirium to determine the convergent validity. Furthermore, the odds ratio (OR) of each score item for the prevalence of a delirium was calculated using logistic regression analysis with the score items separately as independent variables. To test whether risk factors other than the RD score items would improve the risk model, a multivariable logistic regression was repeated with the RD score dichotomized in high- (score ≥4) and low-risk (score <4) patients. The items age, gender, ASA score, fracture type, psychotropic drug use, presence of a partner and institutional residence were used as possible predictors for delirium [27,28]. Based on the ROC curve analysis, the optimal cutoff point of the RD score was considered to be 4 instead of the clinical cutoff of 5 points. The likelihood ratio backward test was conducted to find the best-fit model by selecting the variables one by one. The probability for entry was set at .05, and the probability for removal at .10.

2.3.4. Feasibility

To assess the clinical feasibility, all the RD score sheets were evaluated for errors in interpreting or skipping items and summation of the individual item scores.

3. Results

3.1. Patients

Final analysis was performed in 378 patients; 110 (29.1%) had a partner at admission, 83 (22%) suffered from dementia, 250 (66.1%) lived noninstitutionalized, 252 (66.7%) had an ASA classification of I/II, and 124 (32.8%) used psychotropic drugs.

The mean (S.D.) RD score was 4.9 (3.7), 221 (58.5%) patients were classified as low risk (<5 points), and 157 (41.5%) patients were classified as high risk (≥5 points) based on the clinical cutoff point. Delirium was diagnosed in 102 (27%) patients, 29 (14.1%) in the low-risk and 73 (42.4%) in the high-risk group.

3.1.1. Clinical reliability

In 102 patients (26.9%), an independent nurse performed the RD score for the second time. The ICC for a single measure was 0.77 (90% CI 0.68–0.84). ICC for an average measure of the two observers was 0.87 (90% CI 0.81–0.91). The standard error of measurement of these 102 duplicate tests was 1.73. The coefficient of variation of the standard error of measurement was 29.4%. RD scores ranged from 0 to 17 points. The RD score was 5 points or less in 56.6% of all patients (i.e., clear floor effect); 2.4% of the cohort scored 15 points or more (i.e., no ceiling effect).

3.1.2. Validity

Fig. 2 shows the ROC; the AUROC was 0.73 (95% CI 0.68–0.77), and the best cutoff point for balancing sensitivity and specificity was 4 points. This was different from the cutoff point of 5 points that was used in daily practice to define high-risk patients.

The ability to predict a delirium for cutoff points of 4 and 5 is shown in Table 1. The likelihood ratio of an RD score ≥4 was 1.86, which means that the probability of a score ≥4 being associated with delirium is 1.86 times higher than the

Fig. 2. ROC curve of the RD score with 95% CIs. The diagonal indicate results no better than by chance.
probability of this outcome to be associated with no delirium. The likelihood ratio of an RD score ≤ 4 was 0.33, meaning that the probability of having < 4 points and a delirium is 0.33 less than the probability of having < 4 points and no delirium.

### 3.1.3. Test items included in the RD

The RD score of all patients was analyzed per item as scored on the form. The prevalence of the following risk factors for a delirium was as follows: a delirium during a previous hospitalization was found in 50 patients (13.2%), 87 patients (23%) suffered from dementia, 89 patients (28.6%) made small mistakes, and 99 patients (33.7%) made big mistakes during clock-drawing. An impaired hearing was scored for in 100 (26.5%) patients, an impaired vision in 66 (17.5%), 214 patients (56.6%) needed help with the preparation of their meals or help for domestic work, and 158 patients (41.8%) received help with physical care. Nine patients (2.4%) had a daily consumption of more than four alcoholic beverages, and only five (1.3%) used heroin, methadone or morphine.

The reliability (ICC and kappa) of the items of the RD score is displayed in Table 2. The reliability of dementia was qualified as low-risk instead of high-risk patients, and five vice versa.

### 3.2. Feasibility

In 378 out of 445 patients (84.9%), the RD score was completed sufficiently. In some of these scores, mistakes were made in interpreting or skipping items and summation of the individual item scores. As a result of this, nine patients were categorized in the wrong group; four should have been qualified as low-risk instead of high-risk patients, and five vice versa.

#### 3.2.1. Interpretation of the form and specific items

In nine patients who lived in a nursing home, no points were assigned for help with meal preparation or domestic help, or for help with physical care. None of these patients would be treated as high-risk patients instead of low-risk patients.

### Table 2

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>ICC-single</th>
<th>ICC-av.</th>
<th>Kappa</th>
<th>OR</th>
<th>CI</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delirium during previous hospitalization</td>
<td>0.59</td>
<td>0.45−0.70</td>
<td>0.74</td>
<td>0.62−0.83</td>
<td>0.59</td>
<td>0.39−0.79</td>
</tr>
<tr>
<td>Dementia</td>
<td>0.86</td>
<td>0.80−0.90</td>
<td>0.92</td>
<td>0.89−0.95</td>
<td>0.86</td>
<td>0.74−0.97</td>
</tr>
<tr>
<td>Clock drawing (displaying 10 past 11)</td>
<td>0.72</td>
<td>0.57−0.83</td>
<td>0.84</td>
<td>0.73−0.91</td>
<td>0.72</td>
<td>0.56−0.88</td>
</tr>
<tr>
<td>Small mistakes in clock-drawing</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Big mistakes in clock-drawing</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Age</td>
<td>0.72</td>
<td>0.61−0.80</td>
<td>0.84</td>
<td>0.76−0.89</td>
<td>0.72</td>
<td>0.54−0.89</td>
</tr>
<tr>
<td>70–85 years</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Older than 85 years</td>
<td>0.59</td>
<td>0.45−0.71</td>
<td>0.74</td>
<td>0.62−0.83</td>
<td>0.59</td>
<td>0.42−0.76</td>
</tr>
<tr>
<td>Impaired hearing (patient is not able to hear speech)</td>
<td>0.49</td>
<td>0.33−0.63</td>
<td>0.66</td>
<td>0.50−0.77</td>
<td>0.49</td>
<td>0.30−0.69</td>
</tr>
<tr>
<td>Impaired vision (vision less than 40%)</td>
<td>0.52</td>
<td>0.36−0.65</td>
<td>0.68</td>
<td>0.53−0.78</td>
<td>0.52</td>
<td>0.34−0.70</td>
</tr>
<tr>
<td>Domestic help or help with meal preparation</td>
<td>0.51</td>
<td>0.35−0.64</td>
<td>0.67</td>
<td>0.51−0.78</td>
<td>0.50</td>
<td>0.34−0.67</td>
</tr>
<tr>
<td>Use of heroin, methadone or morphine</td>
<td>0.01</td>
<td>0.00−0.18</td>
<td>0.03</td>
<td>0.00−0.31</td>
<td>-0.01</td>
<td>-0.03 to 0.005</td>
</tr>
<tr>
<td>Daily consumption of 4 or more alcoholic beverages</td>
<td>0.66</td>
<td>0.53−0.76</td>
<td>0.79</td>
<td>0.70−0.86</td>
<td>0.66</td>
<td>0.21−1.0</td>
</tr>
</tbody>
</table>

Av. # = average (after repeated measurement).
of low-risk patients if the age points had been assigned correctly. One patient would shift to the low-risk group.

3.2.2. Skipping items
In 84 (22.2%) patients, the clock-drawing part of the RD score was not completed, and no points were assigned to this item. In 47 of the 84 patients, an RD score of more than 5 points was already achieved based on the other items. In eight patients with no points on any other items on the RD score, clock-drawing was not performed; therefore, it was without consequences for shifting to the high-risk group. In the other patients that did not perform the clock-drawing, nine would have shifted to the high-risk group if they made big mistakes in this task.

3.2.3. Summation errors
In six patients, errors were made in summation of the points of the individual score items; two patients were treated as low-risk patients while they were actually high-risk patients. In the other four patients, the summation error had no consequence. Three other patients were assigned 1 point for, respectively, delirium or dementia (instead of 5 points). This had no consequence for the risk group that they were assigned to originally.

4. Discussion
In this paper, we analyze the validity of the RD score, a valuable new risk model for delirium in hip fracture patients. The RD score showed good reliability and validity. Furthermore, its clinical feasibility was reasonable, with an acceptable participation rate in daily use. The validity of the RD score was improved by adding the items “male gender” and “type of fracture” and removing the items “daily consumption of more than four alcoholic beverages” and “use of heroin, methadone or morphine.”

The RD score had a good interobserver agreement, which improved when a second observer was added. However, in daily practice, the distinction between two groups (i.e., low- or high-risk patients) with a single observer will suffice. In this population of hip fracture patients, the RD score had a clear floor effect; thus, it is not very sensitive to detect a delirium risk in patients with a low score (i.e., low risk). An explanation might be that the delirium incidence in our cohort (27%) was too low to differentiate between patients at risk or not at risk for delirium in this low-risk group. Another explanation might be that the current RD score items are not sensitive enough to differentiate for the risk of developing delirium in low-risk patients. The RD score could be improved by adding more relevant risk factors in order to diminish this floor effect, as is discussed below.

As for validity, the RD score had a high sensitivity and a moderate specificity using the optimum cutoff level of 4 points. The AUC was moderate, and the negative predictive value was high; thus, the RD score is suitable as a screening tool for evaluation of the risk for delirium in this patient group. The moderate specificity is of less clinical importance since this high-risk group is probably more closely monitored for delirium in daily practice.

All items of the RD score, as well as potential new items (like gender and type of fracture), were evaluated for their individual effect (i.e., multivariable regression analysis) on the presence of delirium, as was their correlation with each
other. The item “dementia” of the RD score showed high correlations with the prevalence of delirium and with the other RD score items and thus severely influenced the RD score. Two items of the RD score could be removed (“use of heroin, methadone or morphine” and “daily consumption of four or more alcoholic beverages”). These two items were only scored positive in a very low number of hip fracture patients and had no significant correlation with the occurrence of a delirium. Two new items (“male gender” and “trochanteric fracture”) were added to improve the model. This is in contrast to two meta-analyses on delirium which report no or “a non-convincing” correlation between male gender and presence of delirium [27,28]. Two other studies did find a negative correlation between male gender and delirium [29,30]. However, the individual effect of these two items on occurrence of a delirium was smaller than the compound of a high RD score (i.e., 4 and more points). As for the effect of “fracture type,” none of the meta-analyses did include this risk factor [27,28].

The items of the RD score can easily determined at the ED in an interview with both patients and family or caregivers. Adherence to the RD score protocol was good (85%). Of the 15% of all RD scores which were not completed, clock-drawing was the most frequently not-completed item. Since painful hip fracture patients are in a supine position on a stretcher and often have an additional injury to the dominant arm, drawing might be difficult. Furthermore, half of the noncompleted clock-drawings were found in patients that already scored into the high-risk group (4 and more points) due to the other RD score risk items. Nevertheless, clock-drawing had a reasonable reliability and correlates with delirium. Therefore, better scoring of this item will be facilitated with the addition of a short manual explaining the necessity to have the clock-drawing performed by each patient.

Several risk models for delirium have previously been published [6–12,31]. However, they contained items that were not applicable to hip fracture patients since they were either specific for a certain patient group [medical ward, (cardio)vascular surgery] or designed for elective admissions. Furthermore, some items of these models take time and skills to obtain, like the APACHE II score or the MMSE used in the model of Kalisvaart [12–14]. The RD score is a relatively simple model that works in daily practice and has now been validated for acute admissions of hip fracture patients.

The prospective study design, large sample size and the use of a predefined risk stratification model are important issues for interpretation of our results. A substantial amount of the RD scores was repeated; thus, reliability could be calculated. However, two limitations remain. The main limitation is the diagnosis of presence of a delirium and mental impairment, this despite the fact that in this study delirium was diagnosed based on the DSM-IV classification [18]. We did not use a specific instrument like the Confusion Assessment Method to establish delirium [32]. Furthermore, presence of dementia was based on history taking; a cognitive impairment score like the MMSE was not used [14].

High-risk patients were treated with prophylactic low-dose haloperidol. A recent Cochrane review, a randomized controlled trial of Kalisvaart et al. and our clinical series demonstrated that haloperidol did not have a diminishing effect on delirium incidence [5,15,33]. Therefore, prophylactic haloperidol most probably did not influence the results of the present study. However, labeling patients as “high-risk” might bias the nursing staff; they might be triggered to observe patients more closely for presence of a delirium. Discarding patients because of an incomplete RD score could potentially have influenced our outcome. However, this is highly unlikely since the incidence of delirium and the age of these patients were comparable to the evaluated cohort. Finally, we decided to correct errors in summation and errors as a consequence of not assigning the right amount of points to a specific item. This was only necessary in a limited number of cases, which emphasizes the clinical usefulness of the RD score.

5. Conclusions

The RD score is a recently introduced score that determines the risk for a delirium in hip fracture patients. It has a reasonable clinical feasibility and a good reliability and validity. It would be of additional value to adjust the model by adding “male gender” and ‘trochanteric fracture’ and removing “daily consumption of more than four alcoholic beverages” and “use of heroin, methadone or morphine” as score items.

References


