journal homepage: http://kont.zsf.jcu.cz
DOI: 10.32725/kont.2021.002

KONTAKT / Journal of nursing and social sciences related to health and illness

Original research article

Inter-rater reliability of delirium measuring instruments in patients with locomotive apparatus trauma

Blažena Ševčíková [1] *, Hana Matějovská Kubešová [1], Lenka Šáteková [2], Elena Gurková [2]

[1] *University of Ostrava, Faculty of Medicine, Department of Nursing and Midwifery, Ostrava, Czech Republic*
[2] *Palacký University Olomouc, Faculty of Health Sciences, Department of Nursing, Olomouc, Czech Republic*

Abstract
Objective: Determine the results for inter-rater reliability for items in two delirium screening measuring instruments (Delirium Observation Scale, Nursing Delirium Screening Scale).
Design: Prospective comparative study of measuring instruments.
Methods: Data were collected between August 2018 and August 2019 at the Trauma Clinic, Teaching Hospital, Olomouc, Czech Republic. Study sample included 50 patients with locomotive apparatus trauma, assessed by two raters – general nurses. To assess the inter-rater reliability the study used Cohen's kappa (κ) method.
Results: For the Delirium Observation Scale, the agreement between the two raters for two of the items was 98% (κ = 0.790); for all other items the agreement was 100% (κ = 1.0). The percentage rate of agreement between both raters using the Nursing Delirium Screening Scale was 100% (κ = 1.0) for all five items.
Conclusions: For both screening measuring instruments a high level of reliability has been established. Nevertheless, the highest agreement between the raters has been achieved for the Nursing Delirium Screening Scale. The recommendation is to further test the screening delirium measuring instruments in a Czech clinical setting.

Keywords: Delirium; Inter-rater reliability; Measuring instrument; Nurse; Screening

Introduction

Delirium is described as an acute disorder of consciousness and changes in cognition that mostly manifest in disorganized thinking and lack of attention, which develop over a very short period of time (Maybrier et al., 2018). Acute brain dysfunction or delirium affects up to 80% of patients with severe conditions. In 57% to 75% of these patients it remains undetected. This situation may be affected by various aspects: lack of knowledge about delirium among the healthcare staff, short experience in the field, and primarily a non-existent standardized and efficient measuring instrument for delirium assessment. Patients who develop delirium are hospitalized for much longer (Ghaeli et al., 2018; Leung et al., 2008). Cognitive disorders resulting from the delirium may last as long as a year after the patient has been discharged from the hospital. These disorders often trigger a cascade of events causing a loss of independence, increased number of falls and last but not least death. The delirium patient mortality is up to three times higher than in patients without delirium (Vasilevskis et al., 2011). Repeated testing and outcomes of research studies investigating delirium are becoming more common knowledge for healthcare staff. Delirium prevention and monitoring have also significantly improved owing to the use of standardized screening measuring instruments to detect delirium. The use of screening measuring instruments to detect delirium by a general nurse is a common everyday nursing activity and is primarily important for the prevention of delirium states (Mueller et al., 2017). The most frequently psychometrically tested screening measuring instruments abroad include the Delirium Screening Scale (DOS), Nursing Delirium Screening Scale (Nu-DESC), and Neelon and Champagne Confusion Scale (NEECHAM) (Gavinski et al., 2016). The key indicators for using the screening measuring instruments do not include only validity – as the resulting values of the inter-rater reliability are important as well. It is defined by the degree of agreement between two raters who, independently of each other, assign obtained scores. To what degree the research measuring tool is reliable is based on the agreement between the results of repeated data collections. Repeatability or also accuracy (reproducibility) of a measuring method are basically just different terms for reliability. Therefore, it is necessary that the measuring instruments for delirium risk assessment are not only valid but also reliable. With low inter-rater reliability values, one general nurse may indicate the patient as at high risk while the second nurse may assign the patient to the no risk group. In the psychometric testing the inter-rater reliability

* **Corresponding author:** Blažena Ševčíková, University of Ostrava, Faculty of Medicine, Department of Nursing and Midwifery, Syllabova 19, 703 00 Ostrava, Czech Republic; e-mail: blazena.sevcikova@upol.cz
http://doi.org/10.32725/kont.2021.002
Submitted: 2020-11-02 • Accepted: 2021-01-12 • Prepublished online: 2021-02-01
KONTAKT 23/1: 20–24 • EISSN 1804-7122 • ISSN 1212-4117

of the total measuring instrument score, as well as their individual items, may be evaluated (Šáteková and Žiaková, 2016; Urbánek et al., 2011). Foreign research studies have demonstrated the inter-rater reliability values for Nu-DESC to be κ = 0.47 – 0.83 (κ, ICC). The DOS instrument inter-rater reliability values show significant agreement (κ = 0.73) (Luetz et al., 2010; Numann et al., 2017; Poikajärvi et al., 2017; Radtke et al., 2010). No delirium screening measuring instrument has been tested for the inter-rater reliability in the Czech Republic.

Paper objective

Determine the degree of inter-rater reliability for delirium screening measuring instruments, namely Nu-DESC and DOS.

Materials and methods

Prospective comparative study. Data were collected between August 2018 and August 2019 at a standard Trauma Clinic ward, Teaching Hospital, Olomouc. Prior to the actual administration, both raters were trained directly at the ward where the research was carried out. The training focused on delirium occurrence risk assessment. Rater A: general nurse with a specialized master's degree, 20 years of experience at a standard surgical inpatient ward. Rater B: specialized general nurse, 20 years of experience at a standard surgical inpatient ward. The exclusion criteria were: experience shorter than 2 years, position of practical nurse. The data set used to test inter-rater reliability included 50 patients admitted to the trauma department of standard type. The inclusion criteria were: age 18 years and older, speaking Czech or Slovak, patient with locomotive system trauma, signed consent to participate in the research. The exclusion criteria were: patients with a history of or current dementia, patient with head or brain trauma, patient with delirium condition caused by intoxication.

The first assessment by rater A was performed upon the patient's admission to the ward, using the two screening measuring instruments. Rater B assessed the delirium occurrence risk in the same way as rater A within the following 24 hours. The assessment process was independent, meaning that none of the raters knew about the other's assessment results. To assess the delirium occurrence risk, the following screening measuring instruments were used: Nursing Delirium Screening Scale (Nu-DESC) and Delirium Screening Scale (DOS), which may be administered by a general nurse.

The Nu-DESC measuring instrument includes five items for evaluation: (1) Disorientation; (2) Inappropriate behaviour; (3) Inappropriate communication; (4) Hallucinations; (5) Psychomotor retardation. All items are scored with numbers 1–2. All these items are measured on a three-point scale (0, 1, 2). The limit for delirium is 2 bodies. Evaluation takes place 3 times a day (morning shift, afternoon shift, night shift). The total score for a shift is between 0 and 10, where zero means no initial value and a score ≥2 indicates delirium. A score greater than 2 identifies 86% of patients at risk of delirium.

DOS is a screening measuring instrument based on the Diagnostic and Statistical Manual of Mental Disorders criteria, version IV. It includes 13 items. It is a modified version of this screening measuring instrument. The instrument assesses the following areas: (1) Dozes off during conversation or activities; (2) Is easily distracted by stimuli from the environment; (3) Maintains attention on conversation or action; (4) Does not finish question or answer; (5) Gives answers that do not fit the question; (6) Reacts slowly to instructions; (7) Thinks he/she is somewhere else; (8) Knows which part of the day it is; (9) Remembers recent events; (10) Is picking, disorderly, restless; (11) Pulls IV tubing, feeding tubes, catheters, etc.; (12) Is easily or suddenly emotional; (13) Sees/hears things which are not there. The resulting score for one shift could be up to 13 points.

To assess the inter-rater reliability, the Cohen's kappa (κ) statistical method was used. This statistical method is one of the most frequently applied statistical methods to determine the inter-rater reliability in assessing the delirium occurrence risk in patients at surgical wards (Detroyer et al., 2014; Luetz et al., 2010; Numan et al., 2017; Radtke et al., 2010). Cohen's kappa is designed to determine the raters agreement in the classification of effects to some nominal variable. The kappa calculation is based on the total number of agreements and disagreements (Urbánek et al., 2011). Cohen's kappa values have been classified by Landis and Koch (1977) as follows: 0.00–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, 0.81–1.00 almost perfect agreement. Prior to launching the actual research, consent was obtained from the authors of the used delirium screening measuring instruments, just like the consent with carrying out the research that was given by the Ethics Committee at the Teaching Hospital, Olomouc.

All the participants took part voluntarily and the data were processed anonymously. The research was funded by the Ostrava University SGS project, ref. No. 07/LF/2018–2019.

Results

This research study enrolled 50 patients, hospitalized at the standard trauma ward following a locomotive apparatus trauma. In the text below, the results for inter-rater reliability of the Nu-DESC screening measuring instrument will be presented first, followed by the inter-rater reliability results for the DOS screening measuring instrument. The level of agreement for the Nu-DESC screening measuring instrument was the same for all items (1,000). The agreement was perfect (100%) for all items in this screening measuring instrument. The total inter-rater reliability score for this instrument was 100% (Table 1).

For item 5 (gives answers that do not fit the question) in DOS the level of agreement was 0.790, indicating a substantial agreement, and for item 6 (reacts slowly to instructions) it achieved almost a perfect agreement – 0.960. For the other eleven items the level of agreement was perfect – 1.000 (100%). The total DOS inter-rater reliability reached 100% (Table 2).

Table 1. Inter-rater reliability of the Nu-DESC screening measuring instrument

Nu-DESC Item	Cohen's kappa (κ)	Agreement %	Agreement
1. Disorientation	1.000	100	perfect
2. Inappropriate behaviour	1.000	100	perfect
3. Inappropriate communication	1.000	100	perfect
4. Illusions/Hallucination	–	100	perfect
5. Psychomotor retardation	1.000	100	perfect
Total score	1.000	100	perfect

Nu-DESC – Nursing Delirium Screening Scale.

Table 2. Inter-rater reliability of the DOS screening measuring instrument

Nu-DESC Item	Cohen's kappa (κ)	Agreement %	Agreement
1. Dozes off during conversation or activities	1.000	100	perfect
2. Is easy distracted by stimuli from the environment	1.000	100	perfect
3. Maintains attention to conversation or action	1.000	100	perfect
4. Does not finish question or answer	–	100	perfect
5. Gives answers that do not fit the question	0.790	98	substantial
6. Reacts slowly to instructions	0.960	98	almost perfect
7. Thinks to be somewhere else	1.000	100	perfect
8. Knows which part of the day it is	1.000	100	perfect
9. Remembers recent event	1.000	100	perfect
10. Is picking, disorderly, restless	1.000	100	perfect
11. Pulls IV tubes, feeding tubes, catheters etc.	–	100	perfect
12. Is easily or suddenly emotional	1.000	100	perfect
13. Sees/hears things which are not there	1.000	100	perfect
Total score	1.000	100	perfect

DOS – Delirium Observation Screening Scale.

Discussion

It is known so far that the inter-rater reliability of Nu-DESC and DOS screening measuring instruments in a Czech clinical setting has not been tested. The main aim of this research was to assess how reliable the screening measuring instruments used in a Czech clinical setting in patients with locomotive apparatus trauma are. The Nu-DESC and DOS screening measuring instruments are valid instruments frequently used in patients hospitalized at surgical wards. However, these screening measuring instruments also provide significant results in other clinical settings (internal medicine, palliative care). The screening measuring instruments should be reliable as they are used by general nurses with variable level of skills and experience. Therefore, it is necessary to further investigate the accuracy of the independent measurement between the two raters. The reliability indicators were used to determine the agreement between the two raters in repeated measuring in patients with locomotive apparatus trauma, applied to the Nu-DESC and DOS screening measuring instruments using the Cohen's kappa method.

For Nu-DESC, all items in our research study were assessed as 100%, meaning perfect agreement. High values (0.97) of reliability were also reported by the authors Cinar and Aslan (2019). Almost perfect agreement (0.83) was demonstrated by Radtke et al. (2010), and substantial agreement (0.79) by Luetz et al. (2010). Both these research studies were carried out in Germany. Nu-DESC items were all high (>0.96), even in research studies by Abelha et al. (2013). Weighted kappa ranged from 0.79 to 0.93. On the contrary, only a moderate agreement (0.50) was detected by the authors of a research study conducted in the U.S. – Gavinski et al. (2016), which could have been affected by the low number of patients in the research sample. A moderate agreement (0.47) was also demonstrated by the Finnish authors, Poikajärvi et al. (2017), and this was almost identical to the results of a research study by Leung et al. (2008) from China. Nevertheless, the reliability of the screening measuring instrument is, considering the research study results, good. However, none of the above studies provided specific results for the individual items. They published only the total inter-rater reliability value of the measuring instrument. It should be added that the five-item screening measuring instrument is not considered the most suitable one, despite its feasibility, because the item characteristics are not clearly defined.

In our research, the inter-rater reliability of the second screening measuring instrument, DOS, achieved a perfect agreement (100%), except for item 5 (gives answers that do not fit the question) the score was 0.790 (98%) as a substantial agreement and for item 6 (reacts slowly to instructions) the score was 0.960 (98%), meaning almost perfect agreement. The total inter-rater reliability score for DOS was 100%. Detroyer et al. (2014) also obtained a substantial agreement result of 0.739 for item 5 in their inter-rater reliability research. For item 6 they received a lower value (0.763) than in our research. The value of the other items demonstrated a substantial agreement (0.743–0.774). The highest agreement was found in item 12 (is easily or suddenly emotional) with a score of 0.774, which is in line with a substantial agreement. Numan et al. (2017) published a substantial agreement (0.61) in the DOS total score for inter-rater reliability. However, the authors included a third rater for final assessment. For DOS, only one of the research studies above presented specific values for the individual items. The other studies presented only the total DOS inter-rater reliability score. In our opinion, the DOS screening measuring instrument presents a high quality measuring instrument also because the items have clear explanation of score (0 – without sign, 1 – with sign). This screening measuring instrument has thirteen assessment areas that are clear, brief and easily identifiable for the nurse. The time needed for administration is less than 3 minutes, which does not at all affect the process of daily nursing care provided.

Most screening measuring instruments to assess delirium are based on observation, short tests or a combination of both. It is though necessary to further use and test the standardized measuring instruments for the assessment of delirium and improve the knowledge and skills of the nurses providing nursing care to patients who are at a higher risk of delirium occurrence.

The limitations of our study may include not studying the differences in terms of sex and age, and also that the research was conducted in one clinical setting. Another limitation of the study may have been that measurements were performed 24 hours apart. The patient's clinical condition could have changed during this time frame. A final limitation is the small number of researched overseas studies dealing with this issue.

Conclusions

This research study focuses on the inter-rater reliability for a total score of two screening measuring instruments, namely Nu-DESC and DOS. The study indicates a high agreement between the individual item measurements conducted by two raters in patients with locomotive apparatus trauma. Both screening measuring instruments used in clinical practice systematically make a nurse's job easier and ensure high quality nursing care to patients at risk of delirium occurrence. The Nu-DESC and DOS screening measuring instruments are suitable measuring instruments for clinical settings addressing the problem of patients with locomotive apparatus trauma.

Authorship contributions
BŠ: designed the study, collected and analyzed the data, prepared the manuscript. **HMK:** designed the study, analyzed the data, and revised the manuscript, approved the final version. **LŠ:** analyzed the data and revised the manuscript. **EG:** analyzed the data and revised the manuscript. All authors approved the final version for submission.

Funding statement
07/LF/2018-2019 from the Faculty of Medicine, University of Ostrava, Czech Republic, funded this study.

Conflict of interests
The authors have no conflict of interests to declare.

Acknowledgements
I would like to thank the healthcare facility where the research was conducted, as well as all the participants who voluntarily took part in this study. I also highly appreciate the cooperation with Hana Matějovská Kubešová and with my colleagues Elena Gurková and Lenka Šáteková. Thank you for their support and invaluable advice.

Inter-rater reliabilita meřicích nástrojů pro určení deliria u pacientů po traumatu pohybového aparátu

Souhrn
Cíl: Zjistit výsledky inter-rater reliability položek dvou screeningových měřicích nástrojů pro určení deliria (Delirium Observation Scale, Nursing Delirium Screening Scale).
Design: Prospektivní srovnávací studie měřicích nástrojů.
Metody: Sběr dat probíhal od srpna 2018 do srpna 2019 na oddělení Traumatologické kliniky Fakultní nemocnice Olomouc. Výzkumný vzorek tvořilo 50 pacientů s traumatem pohybového aparátu, posuzovaný dvěma posuzovateli – všeobecnými sestrami. Pro hodnocení inter-rater reliability byla použita statistická metoda Cohenova kappa (κ).
Výsledky: V případě screeningového měřicího nástroje Delirium Observation Scale byla u dvou položek shoda mezi oběma posuzovateli 98 % (κ = 0,790), u ostatních položek byla shoda 100 % (κ = 1.0). Procentuální shoda obou posuzovatelů v případě screeningového měřicího nástroje Nursing Delirium Screening Scale byla 100 % (κ = 1.0) ve všech pěti položkách.
Závěr: U obou screeningových měřicích nástrojů byla zjištěna vysoká míra reliability. Nejvyšší shodu mezi posuzovateli však dosahoval screeningový měřicí nástroj Nursing Delirium Screening Scale. Doporučujeme další testování screeningových měřicích nástrojů pro určení deliria v českém klinickém prostředí.

Klíčová slova: derlirium; inter-rater reliabilita; měřicí nástroj; sestra; screening

References

1. Abelha F, Veiga D, Norton M, Santos C, Gaudreau J-D (2013). Delirium assessment in postoperative patients: Validation of the Portuguese version of the Nursing Delirium Screening Scale in critical care. Braz J Anesteziol 63(6): 450–455. DOI: 10.1016/j.bjane.2012.09.003.
2. Cinar F, Aslan FE (2019). Evaluation of Postoperative Delirium: Validity and Reliability of the Nursing Delirium Screening Scale in the Turkish Language. Dement Geriatr Cogn Dis Extra 9(3): 362–373. DOI: 10.1159/000501903.
3. Detroyer E, Clement PM, Baeten N, Pennemans M, Decruyenaere M, Vandenberghe J, et al. (2014). Detection of delirium in palliative care unit patients: A prospective descriptive study of the Delirium Observation Screening Scale administered by bedside nurses. Palliat Med 28(1): 79–86. DOI: 10.1177/0269216313492187.
4. Gavinski K, Carnahan R, Weckmann M (2016). Validation of the delirium observation screening scale in a hospitalized older population. J Hosp Med 11(7): 494–497. DOI: 10.1002/jhm.2580.
5. Ghaeli P, Shahhatami F, Mojtahed Zade M, Mohammadi M, Arbabi M (2018). Preventive Intervention to Prevent Delirium in Patients Hospitalized in Intensive Care Unit. Iran J Psychiatry 13(2): 142–147.
6. Landis JR, Koch GG (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33(2): 363–74. DOI: 10.2307/2529786.
7. Leung JI, Leung VC, Leung CM, Pan PC (2008). Clinical utility and validation of two instruments (the Confusion Assessment Method Algorithm and the Chinese version of Nursing Delirium Screening Scale) to detect delirium in geriatric inpatients. Gen Hosp Psychiatry 30(2): 171–176. DOI: 10.1016/j.genhosppsych.2007.12.007.
8. Luetz A, Heymann A, Radtke FM, Chenitir Ch, Neuhaus U, Nachtigall I, et al. (2010). Different assessment tools for intensive care unit delirium: Which score to use? Crit Care Med 38(2): 409–418. DOI: 10.1097/CCM.0b013e3181cabb42.
9. Maybrier HR, Mickle AM, Escallier KE, Lin N, Schmitt EM, Upadhyayula RT, et al. (2018). Reliability and accuracy of delirium assessments among investigators at multiple international centres. BMJ Open 8(11): 1–15. DOI: 10.1136/bmjopen-2018-023137.

10. Mueller G, Schumacher P, Wetzlmair J, Lechleitner M, Schulc E (2017). Inter-Rater Reliability and User-Friendliness of the Delirium Observation Screening Scale. J Nurs Meas 25(3): 504–518. DOI: 10.1891/1061-3749.25.3.504.
11. Numan T, van den Boogaard M, Kamper AM, Rood PJT, Peelen LM, Slooter AJC (2017). Recognition of Delirium in Postoperative Elderly Patients: A Multicenter Study. J Am Geriatr Soc 12(9): 1932–1938. DOI: 10.1111/jgs.14933.
12. Poikajärvi S, Salanterä S, Katajisto J, Junttila K (2017). Validation of Finnish Neecham Confusion Scale and Nursing Delirium Screening Scale using Confusion Assessment Method algorithm as a comparison scale. BMC Nurs 7(16): 3–10. DOI: 10.1186/s12912-016-0199-6.
13. Radtke MF, Franck M, Schust S, Boehme L, Pascher A, Bail HJ, et al. (2010). A Comparison of Theree Scores to Screen for Delirium on the Surgical Ward. World J Surg 5(34): 487–494. DOI: 10.1111/j.1365-2702.2012.04102.x.
14. Šáteková L, Žiaková K (2016). Inter-rater reliabilita položiek Bradenovej škály, Nortonovej škály, Waterlowej škály. Profese on-line 9(2): 10–15. DOI: 10.5507/pol.2016.007.
15. Urbánek T, Danglerová D, Širůček J (2011). Psychometrika. Měření v psychologii. Praha: Portál, 320 p.
16. Vasilevskis EE, Morandi A, Boehm L, Pandharipande PP, Girard TD, Jackson JC, et al. (2011). Delirium and sedation recognition using validated instruments: reliability of bedside intensive care unit nursing assessments from 2007 to 2010. J Am Geriatr Soc 29(2): 249–255. DOI: 10.1111/j.1532-5415.2011.03673.x