HBScience (2016) 7:117–127
DOI 10.1007/s16024-016-0278-3

Wie valide und reliabel kann das Delirrisiko eingeschätzt werden?

Systematischer Review über die Qualität der Delirium Observation Screening Scale (DOS-Skala)

Gerhard Müller[1] · Jutta Wetzlmair[2] · Petra Schumacher[1] · Monika Lechleithner[3]

Eingegangen: 19. Februar 2016 / Angenommen: 18. Mai 2016 / Online publiziert: 21. Juni 2016

Zusammenfassung

Hintergrund Um die Diagnose Delir zu behandeln, ist die Aufmerksamkeit aller medizinischen Disziplinen gefordert. Pflegepersonen nehmen in der Früherkennung eine Schlüsselrolle ein, weil sie Screening-Skalen zur Risikoeinschätzung, wie die Delirium Observation Screening Scale (DOS-Skala), dafür einsetzen können. Studienergebnisse zu psychometrischen Eigenschaften der Skala wurden bis dato in keiner deutschsprachigen Übersichtsarbeit zusammengefasst.

Ziel der Arbeit Beschreibung des Aufbaus sowie Vorstellung instrumentenbezogener Güte- und Nebengütekriterien der DOS-Skala.

Methode Zwei unabhängige Personen suchten getrennt von September 2014 bis Februar 2015 mit definierten Suchbegriffen (delirium [MeSH] AND delirium observation screening scale; delirium [MeSH] AND psychometric properties) in den Datenbanken PubMed via MEDLINE, CINAHL, Academic Search Elite und Cochrane Library, PubPsych, Psyndex und DIMDI.

Ergebnisse Neben der gegebenen internen Konsistenz (α = 0,77–0,96) liegt eine Konstruktvalidität zum IQCODE, zu bereits bestehenden psychiatrischen Diagnosen und zum Barthel-Index vor. Die Übereinstimmungsvalidität wurde an mehreren Instrumenten erfolgreich geprüft. Die Sensitivität (81,8–100 %) und Spezifität (76,6–96,6 %) der Skala zeigen hohe Werte. Die Resultate zur AUC (0,93–0,98) weisen auf eine gute Diskriminierung der Skala hin. Die Praktikabilität der Skala wird als anwenderfreundlich beschrieben.

Schlussfolgerung Die Skala verfügt über eine ausreichende Homogenität und Kriteriumsvalidität. Sie ist ein praktikables Instrument für die tägliche Pflegepraxis.

Schlüsselwörter DOS-Skala · Delir · Psychometrische Gütekriterien · Screening · Instrument

How valid and reliable is estimation of the risk of delirium?

Systematic review on the quality of the delirium observation screening scale (DOS scale)

Abstract

Background The diagnosis of delirium requires the attention of all medical disciplines. Nursing personnel play an essential key role in early recognition because they can employ risk assessment scales, such as the delirium observation screening scale (DOS scale) for this purpose. Study results on the psychometric properties of the DOS scale have not yet been reported in a German language systematic review.

Aim The aim of this review article is to describe the structure as well as the characteristics of the psychometric qualities of the DOS scale.

Method Between September 2014 and February 2015 a literature search was independently conducted by 2 people with the defined search terms delirium (MeSH) and

✉ Gerhard Müller
gerhard.mueller@umit.at

[1] Institut für Pflegewissenschaft, Department Pflegewissenschaft und Gerontologie, Universität für Gesundheitswissenschaften, Medizinische Informatik und Technik (UMIT), Eduard-Wallnöfer Zentrum 1, 6060 Hall in Tirol, Österreich

[2] Pflegemanagement, Tirol Kliniken GmbH, Anichstraße 35, 6020 Innsbruck, Österreich

[3] Landeskrankenhaus Hochzirl-Natters, Standort Hochzirl, Tirol Kliniken GmbH, Hochzirl 1, 6170 Zirl, Österreich

delirium observation screening scale, delirium (MeSH) and psychometric properties in the databases PubMed via MEDLINE, CINAHL, Academic Search Elite and the Cochrane Library, PubPsych, Psyndex and DIMDI.

Results The DOS scale showed high internal consistency (α = 0.77–0.96). Construct validity has been established for the DOS scale with IQCODE, pre-existing psychiatric diagnoses and the Barthel index. The concurrent validity was successfully tested with several instruments and demonstrated high values for the sensitivity (81.8–100 %) and specificity (76.6–96.6 %) of the scale. Results for the AUC (0.93–0.98) indicated a good discrimination of the scale. The practicability of the instrument is described as user friendly.

Conclusion The DOS scale shows high internal consistency and satisfactory validity of criteria. It is a practical instrument for use in daily nursing practice.

Keywords DOS scale · Delirium · Psychometric properties · Screening · Instrument

Das Delir ist eine klinische Diagnose und beschreibt einen akuten Verwirrtheitszustand, der sich über einen Zeitraum von einigen Stunden bis zu einigen Tagen entwickeln kann [24]. Für die Praxis und Forschung stehen als Klassifikationssysteme das Diagnostische und Statistische Manual Psychischer Störungen (DSM) und die Internationalen Klassifikationen der Krankheiten (ICD) zur Diagnostik eines Delirs zur Verfügung.

Die Prävalenz von als verwirrt eingeschätzten Patienten variiert weltweit zwischen 7 und 52 % [11]. Mehr als 50 % der im Krankenhaus befindlichen älteren Patienten sind von dieser oft tödlichen Diagnose (Österreichischen Gesellschaft für Geriatrie und Gerontologie, ÖGGG, [19]) betroffen. Die höchste Prävalenzzahl von 63 % zeigt sich auf onkologischen Stationen bzw. Palliativstationen [25]. In 18 europäischen Ländern fielen 2011 diesbezüglich Mehrbelastungen in Höhe von mehr als 182 Billionen Dollar an [13]. Die Kosten für Österreich sind bis heute kaum valide zu beziffern [19].

Die Ätiologie eines Delirsyndroms resultiert aus einer Interaktion mehrerer Prozesse, bevorzugt durch somatische Erkrankungen, durch Wirkung und Nebenwirkung von Pharmaka und durch störende Umgebungsfaktoren [19]. Der Zusammenhang zwischen somatischen und psychischen Aspekten ist beim Delir besonders evident, wie der Bericht der Österreichischen Gesellschaft für Geriatrie und Gerontologie aufzeigt [19]. Für die Entstehung spielen das Verhältnis von Vulnerabilität und Noxe eine wesentliche Rolle (Abb. 1). Wenn die Vulnerabilität hoch ist, reicht eine geringfügige Noxe aus und umgekehrt [12].

Aus Sicht der Pathophysiologie nehmen die Störung des Gleichgewichts der Neurotransmitter Acetylcholin und Dopamin eine zentrale Rolle ein [5, 11]. Sie haben eine hohe Bedeutung für kognitive Funktionen, Vigilanz und Schlaf-Wach-Rhythmus. Bereits eine Hypoxie führt zur reduzierten Synthese von Acetylcholin oder einer vermehrten Freisetzung von Dopamin und kann so zur Entstehung eines Delirs beitragen [5]. Ältere Menschen sind durch den natürlichen Verlust des Acetylcholins sowie Menschen, die an einem pathologischen Verlust leiden (z. B. demenzielle Erkrankung des Typs Alzheimer) besonders betroffen [11].

Problem- und Zieldarstellung

Für eine Früherkennung eines Delirs bedarf es der Aufmerksamkeit von allen medizinischen Disziplinen. Dabei übernehmen Pflegepersonen eine zentrale Rolle bereits bei der Aufnahme als auch während des Klinikaufenthalts. Pflegerische Screening- und Assessment-Instrumente unterstützen einerseits die Risikoeinschätzung und andererseits

Prädisposition	Exogener Einfluss
Hohe Vulnerabilität	Schwache Noxe
• Hohes Lebensalter • Kognitive Einschränkung • Frailty • Hohe somatische Komorbidität • Schwere Grunderkrankung • Seh- und Hörbehinderung • Anämie • Malnutrition (niedriges Serum-Albumin) • Depression • Ängstlichkeit • Alkoholismus • Benzodiazepin-Einnahme • Schmerz • Leichte kognitive Störung • Einsamkeit	• Fremde Umgebung • Körperliche Einschränkung • Immobilisation • Störung des Biorhythmus • Psychoaktive Medikamente • Entzugssyndrom (Alkohol, Sedativa) • Respiratorische Insuffizienz (Hypoxie) • Exsikkose • Elektrolytentgleisung • Akute Infektion • Hypo-, Hyperglykämie • Organversagen (Leber, Niere) • Intensivbehandlung • Anticholinergika • Chirurgische Eingriffe
Niedrige Vulnerabilität	Hohe Noxe

Abb. 1 Entstehungsfaktoren eines Deliriums. (mod. nach [19], S. 7)

eine frühzeitige Diagnostik durch klare Verfahrensweisen [1]. Dem Klinikalltag stehen zahlreiche Instrumente zum Screening und zur Diagnostik von Delir zur Verfügung [10]. Ein solches Screening-Instrument ist die von Schuurmans et al. [24] entwickelte Delirium Observation Screening Scale (DOS-Skala). Damit wird den Pflegepersonen ermöglicht, frühzeitig das Risiko eines Delirs zu erkennen [15]. Zur Entwicklung möglicher Entscheidungsstrategien in der Praxis muss ein Gesamtüberblick über das vorhandene Wissen zum Instrument vorgestellt werden. Dies ist bisher nur in Ansätzen erfolgt, so dass es keinen Gesamtüberblick zur Screening-Skala gibt und die methodische Qualität des Instruments nicht eindeutig beurteilt werden kann.

Mit der Literaturanalyse soll einerseits die DOS-Skala dargestellt und andererseits deren psychometrische Eigenschaften inklusive der Nebengütekriterien des Instruments bewertet werden.

Methode

Um der Zielsetzung nachzugehen, wurde zur Identifikation von Reliabilität, Validität und den Nebengütekriterien der DOS-Skala eine systematische Literaturrecherche durchgeführt. Die Durchführung erfolgte nach den Vorgaben der PRISMA-Leitlinie (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) zur Darstellung systematischer Übersichtsarbeiten (Moher et al. [17]).

Die in den Monaten von September 2014 bis Februar 2015 durchgeführte mehrstufige, iterative und systematische Literaturrecherche [16] erfolgte von zwei Personen. Dabei wurden einerseits die Datenbanken MEDLINE via PubMed, CINAHL und Academic Search Elite via EBSCO-Host und andererseits Psyndex via OVID, DIMDI, Cochrane Library und PubPsych unabhängig voneinander durchsucht. Zusätzlich erfolgte eine Handsuche in einschlägigen Online-Katalogen wissenschaftlicher Bibliotheken (z. B. Universität Wien, Graz und Innsbruck). Für eine zielgerichtete Recherche wurden aus den Themenfeldern der Zielsetzung folgende Suchbegriffe identifiziert: *delirium [Mesh], DOS-scale, delirum obersavation screening scale, psychometric properties, validity, sensitivity, specificity, screening, assessement, scale*. Mithilfe der Boolschen Operatoren AND und/oder OR wurden die Suchbegriffe miteinander verknüpft und Trunkierungen gesetzt.

Die Literaturauswahl erfolgte anhand der a priori definierten Ein- und Ausschlusskriterien (Tab. 1). Die Auswahl der Artikel basierte auf der Grundlage des Screenings von Publikationstiteln und Zusammenfassungen.

Auf der Grundlage einer anschließenden Volltextzusammenfassung relevanter Artikel erfolgte die Datenextraktion. Beide Personen führten unabhängig voneinander den Auswahlprozess durch. Für die Bewertung der Qualität einzelner Studien wurden für die systematische Übersichtsarbeit die Bewertungshilfe nach Behrens und Langer [2] und für die Diagnosestudien die *Standards for Reporting of Diagnostic Accuracy Checklist* (STARD-Checkliste; [4]) gewählt. Im Fall von differenzierten Bewertungen unter den beiden Wissenschaftlern erfolgte ein Konsens nach einer Diskussion oder durch Hinzuziehen einer dritten beratenden Person. Der Identifikations- und Selektionsschritt der Literaturrecherche sind im Flussdiagramm (Abb. 2) dargestellt.

Ergebnisse

Die systematische Literaturrecherche ergab sechs relevante Publikationen, die über die Entwicklung und/oder Qualitätsgüte der DOS-Skala berichteten. Insgesamt mussten 93 Publikationen aus unterschiedlichen Gründen ausgeschlossen werden.

Entwicklung, Aufbau und Modifizierung der DOS-Skala

Die im Jahr 2001 von Marieke Schuurmans entwickelte DOS-Skala ist ein für Pflegepersonen entwickeltes Screening-Instrument, das auf Beobachtungen im Rahmen der Pflegetätigkeit beruht und welches auf den Delirkriterien der DSM-IV basiert. Sie zielt darauf ab, Patienten mit Verhaltensauffälligkeiten, die typisch für ein Delir sind, in kur-

Tab. 1 Ein- und Ausschlusskriterien der Literaturrecherche

	Einschlusskriterien	Ausschlusskriterien
Publikationszeitraum	2002 bis 2014	Vor 2002
Sprache	Deutsch- und englischsprachige Studien	Andere Sprachen
Studiendesign und Publikationsart	(Nicht-)Experimentelle Studiendesigns (retro- und prospektive Designs), Reviews	Qualitative Studien
Setting	Prähospitale und klinische Einrichtungen	–
Verwendungszweck	Arbeiten, welche die testtheoretischen Eigenschaften sowie die Nebengütekriterien der DOS-Skala untersuchten	Arbeiten, die die DOS-Skala als Erhebungsinstrument im Rahmen einer klinischen Untersuchung verwendeten

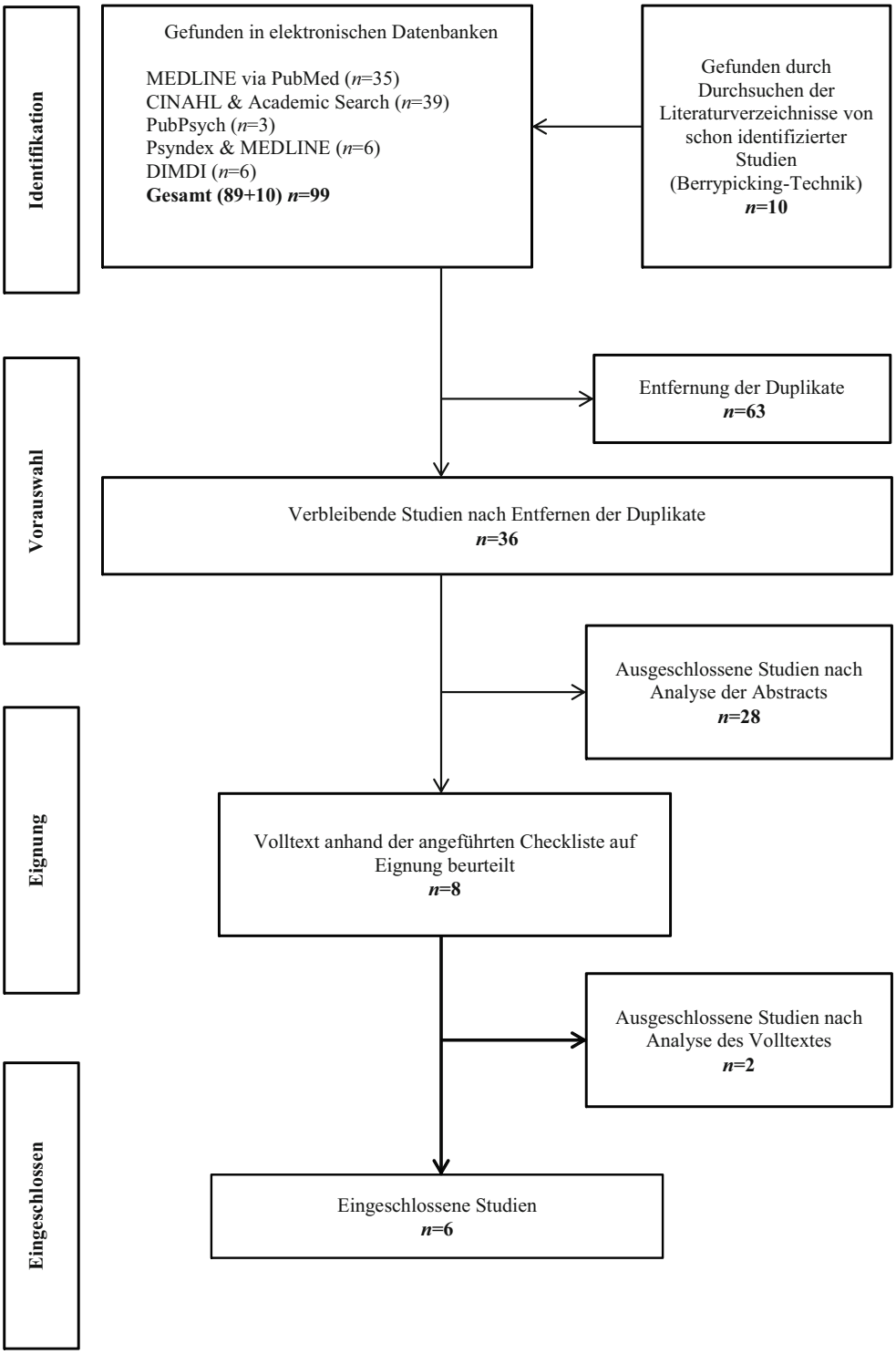

Abb. 2 Flussdiagramm zur Visualisierung des Literatursuch- und Einschlussprozesses

zer Zeit und ohne zusätzliche Belastung durch das Pflegepersonal zu identifizieren [11, 24]. Die Originalversion [24] beinhaltete 25 Items, welche die diagnostischen Delirsymptome über acht 8 Dimensionen darstellt: Bewusstseinsstörungen (3 Items), Aufmerksamkeit und Konzentration (3 Items), Denken (5 Items), Erinnerung – Orientierung (3 Items), psychomotorische Aktivität (4 Items), Schlaf-Wach-Rhythmus (3 Items), Laune (2 Items) und Wahrnehmung (2 Items). Zunächst wurde die Einschätzung anhand einer 5-stufigen Rating-Skala (1 = nie [never] bis 5 = immer [always]) durchgeführt [24].

Aufgrund der Schwierigkeiten beim Einschätzen wurde die Skala auf eine 4-stufige Rating-Skala reduziert (ebd.). Ein „weiß nicht“ kann gegeben werden, wenn ein Item nicht

bewertet werden kann (z. B. Patient kommuniziert nicht, Patient schläft) oder das Pflegepersonal sieht sich nicht im Stande, eine fachliche Einschätzung durchzuführen (z. B. Wissensdefizit; [11]).

Wong et al. [25] beschrieben erstmalig eine Modifizierung der Skala. Dabei erfolgte eine Reduktion von 25 auf 13 Items. Seither besteht die Skala aus 13 Verhaltensweisen, die den Symptomen eines Delirs entsprechen. Die Verhaltensbeobachtungen werden in jeder Arbeitsschicht schon während der üblichen Pflegemaßnahmen festgehalten. Ein „nie" beschreibt, dass der Patient das beschriebene Verhalten nie aufgezeigt hat. Die Antwortmöglichkeit „manchmal–immer" gibt an, jenes zu beobachtende Verhalten zumindest einmal gesehen zu haben. Kann ein Item nicht bewertet werden, erfolgt die Antwortmöglichkeit „weiß nicht" gemäß der oben angeführten 4-stufigen Rating-Skala. Die jeweilige Zahl (0–1) in der Spalte wird umkreist und am Ende jeder Schicht zusammengezählt. Die totale Punkteanzahl beträgt 0 bis maximal 13 Punkte pro Schicht. Am Ende des Tages wird der Durchschnitt aus allen Schichten berechnet. Ergibt die Endsumme eine Anzahl von weniger als 3, hat der Patient vermutlich kein Delir, bei einer Summe von ≥3 liegt mit hoher Wahrscheinlichkeit ein Delir vor.

Untersuchungsdesigns eingeschlossener Studien

Schuurmans et al. [24] veröffentlichten in einem Artikel die Ergebnisse von zwei prospektiven Studien. Sie testeten die DOS-Skala (25 Items) an 82 geriatrischen Patienten (Studie 1) und an 92 chirurgischen Patienten (Studie 2). Die Teilnehmer der Studie 1 waren durchschnittlich 83 Jahre alt (SD ± 6,17; R = 70–96). Ausgeschlossen wurden nur Patienten, die bereits bei Aufnahme ein Delir hatten oder die Entlassung bzw. ein Transfer innerhalb der ersten Woche geplant war. Die Ermittlung der Daten erfolgte unmittelbar nach Aufnahme jeweils am Ende jeder Arbeitsschicht über einen Zeitraum von 7 Tagen. In Studie 2 erfolgte die Datensammlung bei den Teilnehmern (MW = 82,3; SD ± 6,65; R = 70–98) an den ersten 6 Tagen unter Einbezug der DOS-Skala und der *Confusion Assessment Method* (CAM), einer Skala, basierend auf den DSM-IV-Kriterien, welche eine standardisierte Erfassung von Delirsymptomen erlaubt. Eingeschlossen wurden Patienten, bei denen kein Delir erwartet wurde und auch kein Transfer oder eine Entlassung innerhalb der nächsten 5 postoperativen Tage geplant war. In beiden Studien wurde einerseits die Delirdiagnose unabhängig von den Einschätzungen der Pflegepersonen („staff nurse" und „research nurse") durch einen Geriater anhand der DSM-IV-Kriterien gestellt. Andererseits wurden die *Mini Mental State Examination* (MMSE), der *Informant Questionaire on Cognitive Decline in the Elderly* (IQCODE), bereits bestehende psychiatrische Diagnosen und der Barthel-Index verwendet. In einer Literaturübersicht (1986–2002) von Schuurmans et al. [23] wurden die Charakteristika von 13 Instrumenten zum Einschätzen von einem Delir und deren psychometrische Gütekriterien beschrieben. Bei den 13 Instrumenten wurden der Verwendungszweck, die Diagnosekriterien auf welches das jeweilige Instrument basiert, die Anzeichen und Symptome die das Instrument zu messen vorgibt, die Methoden der Datenerhebung, die berufsgruppenspezifischen Anwendungen, die Anzahl der Items und die Dauer der Einschätzung bestimmt.

Van Gemert und Schuurmans [9] untersuchten in einer monozentrischen prospektiven Studie (n = 87) an einer internen und 3 chirurgischen Stationen die unterschiedlichen Fähigkeiten der DOS-Skala (Kurzversion: 13 Items) mit einem anderen Einschätzungs-Instrument für Delir, der *NEECHAM Confusion Scale*. Eingeschlossen wurden Patienten ab dem 70. Lebensjahr mit 3 oder mehreren Begleiterkrankungen. Das Durchschnittsalter lag bei 79 Jahren (R = 70–96). In einem Zeitraum von 5 Monaten wurden maximal 4 Patienten an einem Tag pro Station eingebunden, um die zusätzliche Belastung für die Pflege möglichst gering zu halten. Die Befragung fand dreimal pro 24 h statt. Am Ende jeder Arbeitsschicht wurde jeder teilnehmende Patient anhand beider Instrumente (DOS-Skala und NEECHAM Confusion Scale) eingeschätzt. Nachdem alle Daten gesammelt waren, wurde ein Geriater hinzugezogen, um anhand der DSM-IV-Kriterien das Delir zu diagnostizieren. Die Prognose erfolgte verblindet gegenüber den pflegerischen Bewertungsergebnissen. Um die Praktikabilität der beiden Skalen zu bestimmen, wurden die Krankenpflegepersonen anhand eines validierten Fragebogens befragt.

In der prospektiven Kohortenstudie (n = 112) von Koster et al. [15] wurden von November 2006 bis Juni 2009 Patienten ab 45 Jahren (MW = 70; SD ± 7,3) auf einer herzchirurgischen Station mittels der DOS-Skala (Kurzversion: 13 Items) eingeschätzt. Die Datenerhebung fand präund postoperativ statt, von der Aufnahme bis zum fünften postoperativen Tag. Personen, die bis dahin kein Delir aufwiesen, wurden nicht weiter befragt. Alle anderen wurden weiterhin eingeschätzt bis der Patient über 2 Tage kein Delir aufwies. Patienten, die bereits bei Aufnahme ein Delir aufwiesen, wurden ausgeschlossen. Ebenso Patienten, die keine präoperative Untersuchung hatten und Patienten mit Demenz. Postoperativ wurden die Patienten am Ende jeder Schicht mit der DOS-Skala eingeschätzt. Ergab die Einschätzung mit der DOS-Skala einen Wert von ≥2 wurde ein Psychiater hinzugezogen, der anhand der DSM-IV-Kriterien die Diagnose Delir bestätigte oder entkräftete.

Im Untersuchungszeitraum von Mai 2005 bis Juli 2008 führten Scheffer et al. [21] ihre Kohortenstudie (n = 97) zur Kurzversion der DOS-Skala (13 Items) durch. Dabei wurden bereits delirante Patienten ab dem 65. Lebensjahr eingeschlossen, die sich kürzlich einer Hüftoperation un-

terzogen (n = 41) haben, mit einem Durchschnittsalter von 82,1 Jahren (SD ± 7,4) bzw. Patienten, die akut auf einer internen Abteilung (n = 56) aufgenommen wurden (MW = 86,7; SD ± 6,4). Das Vorhandensein oder Fehlen von Delir wurde innerhalb der ersten 48 h nach Aufnahme anhand der CAM erhoben. Geriater verwendeten auch die *Delirium Rating Scale-Revised-98* (DRS-R-98). Um die kognitive Funktion durch einen Arzt erfassen zu können, wurde der IQCODE in der kurzen Form (IQCODE-SF) verwendet. Eine körperliche Beeinträchtigung wurde mittels der Katz-ADL ermittelt. Zwei Drittel der Gesamtstichprobe (n = 97) wurden dreimal in 24 h befragt (= einmal in jeder Arbeitsschicht). Das andere Drittel konnte aus verschiedenen Gründen nur zweimal eingeschätzt werden.

In der nichtexperimentellen, deskriptiven, prospektiven Studie von Detroyer et al. [7] schätzte das Krankenpflegepersonal 48 Patienten einer Palliative Care Unit mittels Kurzversion der DOS-Skala (13 Items) ein. Im Vergleich dazu fand eine Einschätzung von Forschern mittels der CAM und dem Delirium-Index statt. Beide Gruppen waren verblindet gegenüber den Resultaten der anderen Messungen. Im Vorfeld wurden sowohl das Krankenpflegepersonal als auch die Forschungsgruppe von zwei speziell auf Delir trainierten klinischen Experten auf die jeweiligen Skalen eingeschult. Eingeschlossen wurden Patienten ab einem Alter von 18 Jahren (Median = 72 Jahre, Q1 = 67,25; Q3 = 78). Patienten, die sich im terminalen Stadium ihres Lebens befanden, wurden ausgeschlossen. Die Einschätzung mit der DOS-Skala fand einmal pro Schicht statt, dreimal in 24 h und über einen Einschätzungszeitraum von 10 Tagen. Die Gruppe der Forscher erhob die Daten maximal dreimal pro Patient an 3 unterschiedlichen Tagen. Um die Praktikabilität der DOS-Skala aufzuzeigen, wurden 10 Pflegepersonen mittels eines Fragebogens mit 25 Antwortmöglichkeiten befragt.

Psychometrische Eigenschaft der DOS-Skala

Bei der Entwicklung von Messinstrumenten ist die Bestimmung der psychometrischen Eigenschaften eines Instruments wichtig [6]. Dabei spielt die klassische als auch die probabilistische Testtheorie eine wesentliche Rolle. Die Vertreter der erstgenannten Theorie sind die Objektivität, Reliabilität und Validität. Zusätzlich werden in der Literatur die anwendungsbezogenen Gütekriterien, wie beispielsweise die Praktikabilität [20], genannt. In allen der identifizierten Publikationen (Tab. 2) wird der Ansatz der klassischen Testtheorie gewählt und zur Reliabilität, Validität als auch zur prädiktiven Validität berichtet.

Bei den in Tab. 2 dargestellten psychometrischen Eigenschaften zeigt sich, dass die Ermittlung der Reliabilität nur durch die interne Konsistenz erfolgte. Die interne Konsistenz der DOS-Skala wurde in 3 Studienergebnissen (Tab. 2) berechnet [7, 24]. Schuurmans et al. [24] publizierten 2 Studienergebnisse in einem Artikel zur Langversion der DOS-Skala (25 Items), die einerseits von geriatrischen (Studie 1) und andererseits von chirurgischen Patienten (Studie 2) entstammen. Die interne Konsistenz in Studie 1 (n = 82) zeigte Werte von Cronbach's Alpha (α) zwischen 0,74 und 0,98. Der durchschnittliche Cronbach's α lag bei 0,96; 12 von 21 Arbeitsschichten wiesen einen α-Wert über 0,90 auf. In Studie 2 variierten die Cronbach's α zwischen 0,78 und 0,98. Im Durchschnitt wies dieser einen α-Wert von 0,97 auf; in 13 von 18 Arbeitsschichten lag der α-Wert über 0,90. Bei einer differenzierten Betrachtung der Studienteilnehmer zeigte sich bei den deliranten geriatrischen Patienten (n = 4) ein durchschnittlicher α-Wert von 0,93 und 0,96 bei den deliranten chirurgischen Patienten (n = 18). Die interne Konsistenz über die Kurzversion der DOS-Skala (13 Items) ermittelten Detroyer et al. [7]. Sie berechneten einen Cronbach's α von 0,77. Bei der Trennschärfenberechnung mit der Summe von allen Items der Skala konnte gezeigt werden, dass 9 von 13 Items eine mittlere Korrelation nach Pearson (r = 0,57–0,40), 3 Items eine ausreichende Korrelation (r = 0,39–0,25) und ein Item eine schwache Korrelation von r = 0,12 aufwiesen. Bis auf den Hinweis zur Interrater-Reliabilität in Schuurmans et al. [24], wonach in 3 Studien zufriedenstellende Resultate dahingehend erzielt wurden [22], aber nicht näher definiert (keine Angaben), sind weitere Ergebnisse zu anderen Reliabilitätsarten nicht bekannt.

Die Ermittlung der Validität der DOS-Skala erfolgte durch die Bestimmung der Inhalts-, Konstrukt- und Kriteriumsvalidität. Die Inhaltsvalidität der Langversion der DOS-Skala wurde von insgesamt 7 Mitgliedern einer holländischen multidisziplinären Arbeitsgruppe überprüft. Von 26 Items wurden 21 als inhaltlich gültig beurteilt und 5 ausgeschlossen. Auf Empfehlung dieses Gremiums benötigten 6 der 21 Items eine geringfügige Adaptierung in der Formulierung. Zusätzlich wurden 4 neue Items (n = 25) hinzugefügt [24], was zur Inhaltsvalidität der DOS-Skala gemäß den Fachleuten beigetragen hat. Die Konstruktvalidität wurde gestützt durch die Korrelationen zwischen den Werten der DOS-Skala (25 Items) und dem IQCODE, den bereits bestehenden psychiatrischen Diagnosen sowie dem Barthel-Index [24]. In Studie 1 betrugen die Korrelationswerte vom Spearmen-Rangkorrelationkoeffizient (r_s) zum IQCODE r_s = 0,33 ($p \leq 0,05$), in Studie 2 r_s = 0,74 ($p \leq 0,001$). Zu bereits bestehenden psychiatrischen Diagnosen korrelierte die DOS-Skala in Studie 1 mit r_s = 0,42 ($p \leq 0,001$), in Studie 2 mit r_s = 0,43 ($p \leq 0,001$). Die Gesamtwerte zwischen dem Barthel-Index und der DOS-Skala korrelierten negativ ausreichend in Studie 1 (r_s = –0,26; $p \leq 0,05$) sowie negativ moderat in Studie 2 (r_s = –0,55; $p \leq 0,001$). Die Kriteriumsvalidität der DOS-Skala wurde mehrmals in Bezug auf die Übereinstimmungsvalidität

Tab. 2 Studienüberblick über psychometrische Eigenschaften der DOS-Skala

Autoren, Jahr, Land	Ort, Stichprobe, (Anzahl der Items) [Anmerkung]	Reliabilität		Validität			Prädiktive Validität					Nebengüte
		Interne Konsistenz	Interrater-Reliabilität	Inhalts-validität	Konstrukt-validität	Kriteriums-validität	Sensitivität	Spezifität	AUC	PPV	PNV	Praktikabel
Schuurmans et al. 2003 [24] Niederlande	Geriatrie (Study 1) $n = 82$	●	k. A.	●	●	●						
	Chirurgie (Study 2) $n = 92$ (25 Items)	●			●	●						
Schuurmans et al. 2003 [23] Niederlande USA	(13 Items) [Literaturarbeit]						●	●		●	●	
Van Gemert, Schuurmans 2007 [9] Niederlande	Interne & Chirurgie n = 87 (13 Items)						●	●		●	●	●
Koster et al. 2009 [15] Niederlande	Herzchirurgie $n = 112$ (13 Items)						●	●	●			
Scheffer et al. 2011 [21] Niederlande	Chirurgie (n = 41) Interne (n = 56) $\sum n = 97$ (13 Items)					●						
Detroyer et al. 2014 [7] Belgien	Palliativ Care Unit $n = 48$ (13 Items)	●				●	●	●	●	●	●	●

● Qualitätskriterium erfüllt, *AUC* „area under the curve“, *PPV* „positive predictive value“ – positiver Vorhersagewert, *NPV* „negative predictive value“ – negativer Vorhersagewert, *k. A.* keine Angaben in der Publikation ausgewiesen

geprüft. Schuurmans et al. [24] überprüften die Übereinstimmungsvalidität der DOS-Items (n = 25 Items), indem sie einerseits den statistischen Zusammenhang zwischen „staff nurses“ und „research nurses“ berechneten. Dabei konnte ein Korrelationskoeffizient von nur ausreichend bis mittelmäßig (r_s = –0,06–0,60) bestimmt werden, wobei 3 Korrelationen nichtsignifikant und 20 signifikant waren ($p \leq 0{,}05$–$p \leq 0{,}001$). Bei 2 Items wurden keine Korrelationsberechnungen durchgeführt, da Vergleichswerte fehlten [24]. Bei Betrachtung der Summenwerte dieser beiden Berufsgruppen zeigte sich eine moderate Korrelation (r_s = 0,54; $p \leq 0{,}001$). Weiters wurde die Übereinstimmungsvalidität auch zwischen der DOS-Skala und der CAM geschätzt. Sie betrug r_s = 0,63 ($p \leq 0{,}001$), was einer mittelmäßigen Korrelation entspricht. Zwischen den Gesamtpunkten der DOS-Skala und dem MMSE zeigten sich negative Korrelationen. Die Korrelationswerte in Studie 2 waren höher (n = 68; r_s = –0,79; $p \leq 0{,}001$) als in Studie 1 (n = 28; r_s = –0,66; $p \leq 0{,}001$). In den Studienergebnissen von Scheffer et al. [21], in die nur delirante Personen aufgenommen wurden, lag zur Übereinstimmungsvalidität zwischen dem DRS-R-98-Score und dem der Kurzversion der DOS-Skala (13 Items) ein signifikanter Pearson-Korrelationskoeffizient von r = 0,67 (p = 0,01) vor. Ebenso zeigte sich eine mittlere Korrelation (r = 0,61; p = 0,01) zwischen diesen beiden Skalen bei kognitiv eingeschränkten Personen (n = 62). In der Gruppe der kognitiv Nicht-Eingeschränkten konnte ein ähnliches Resultat (r = 0,67; p = 0,01) erreicht werden. Zur Bestimmung der Übereinstimmungsvalidität zwischen der DOS-Skala (13 Items)

und dem Delirium-Index wiesen Detroyer et al. [7] gerade noch eine mittlere Korrelation (r_s = 0,53; p = 0,001) nach. Dennoch zeigte sich bei Personen mit diagnostiziertem Delir eine höhere Übereinstimmung zwischen den beiden Skalen (r_s = 0,73; p < 0,01). Personen mit einem hohen Delirrisiko (DOS-Wert ≥3) zeigten beim Delirium-Index einen signifikant höheren Mittelwert (MW = 10,08; SD ± 3,48; p < 0,001) als in der Gruppe, bei der kein Risiko (DOS-Wert ≤2) vorlag (MW = 3,16; SD ± 2,90).

Die in Tab. 2 dargestellten Ergebnisse zur prädiktiven Validität der DOS-Skala (13 Items) wurden als erster von Schuurmans et al. [23] untersucht. Dabei zeigte sich bei einem Cut-off-Punkt von ≥3 eine Sensitivität von 94,4 % (CI 95 % 69–99) und eine Spezifität von 76,7 % (CI 95 % 65–84). Der PPV beträgt 50 % und der NPV 98,2 %. Die positive Likelihood-Ratio (LR^+) beträgt 3,9 (CI 95 % 2,6–5,9) und die negative Likelihood Ratio (LR^-) ist mit einem Wert von 0,07 (CI 95 % 0,01–0,50) hoch. Van Gemert und Schuurmans [9] berechneten bei einer Prävalenz von 10,3 % (n = 87) eine Sensitivität von 89 % (CI 95 % 50–98) und eine Spezifität von 88 % (CI 95 % 79–94; PPV = 47 %, NPV = 98,5 %). Die LR^+ ist mit 7,6 (CI 95 % 4,0–15) tendenziell höher und besser als bei Schuurmans et al. [23], wobei die LR^- mit 0,13 (CI 95 % 0,02–0,80) im Vergleich dazu einen schlechteren Wert ergab. In der Studie von Koster et al. [15] konnten hohe Werte von Sensitivität und Spezifität erzielt werden (Sensitivität = 100 %, Spezifität = 96,6 %). Sowohl die NPV von 100 % als auch die Ergebnisse der AUC der Receiver-Operating-Characteristics(ROC)-Kurve von 0,98 (CI 95 % 0,96–1,00; p < 0,001) zeigten beachtliche Werte. In den Studienergebnissen von Detroyer et al. [7] sind Berechnungen zur Genauigkeit der DOS-Skala und der CAM an 48 Personen durchgeführt worden. Sie bestätigen den derzeitigen Cut-off-Punkt von 3. Die Fläche unter der ROC-Kurve (AUC) zeigte Werte, die auf eine hervorragende Diskriminierung zwischen Sensitivität und Spezifität hinweisen (AUC = 0,93; CI 95 % 0,82–1,00). Bei einer Sensitivität von 81,8 % (CI 95 % 52–59) und Spezifität von 96,1 % (CI 95 % 90–98) lag der PPV bei 69,2 % (CI 95 % 42–87) sowie der NPV bei 98 % (CI 95 % 93–99). Die diagnostische Übereinstimmung beider Skalen kann als überdurchschnittlich hoch bezeichnet werden (prozentuelle Übereinstimmung = 94,7 %; κ = 0,72; CI 95 % 0,51–0,93; p < 0,001). Zur DOS-Skala mit 25 Items liegen keine Studienergebnisse zur Vorhersagevalidität vor.

Die Ermittlung der anwenderbezogenen Gütekriterien der DOS-Skala erfolgte durch die Bestimmung der Praktikabilität. Die Nutzerfreundlichkeit der Kurzversion der DOS-Skala wurde in 2 Studien überprüft. Van Gemert und Schuurmans [9] untersuchten an 37 Pflegepersonen (MW = 34 Jahre) die Benutzerfreundlichkeit der DOS-Skala. Sie waren durchschnittlich 13 Jahre im Beruf tätig, von den 37 hatten 33 Teilnehmer (88 %) einen Bachelor-Abschluss. Im Durchschnitt benötigten die Pflegepersonen 5 min (R = 1–15 min, Median = 5 min) zur Einschätzung. Pflegende bezeichneten die DOS-Skala als ein leicht anwendbares und praxistaugliches Instrument. In den Studienergebnissen von Detroyer et al. [7] stimmten alle 10 Teilnehmer darüber ein, dass die Konzepte der DOS-Items klar, kompatibel mit der pflegerischen Alltagssprache sowie wert- und urteilsfrei formuliert sind. Neun der 10 Pflegenden beurteilten die DOS-Skala als handliches und ein für die Praxis wertvolles Instrument. Im Gegensatz zu den Ergebnissen von van Gemert und Schuurmans [9] betrug die mittlere Einschätzungszeit 1 min.

Diskussion

Die DOS-Skala wurde auf Basis der DSM-IV-Diagnosekriterien entwickelt [24], damit Pflegepersonen ein vorliegendes Delirrisiko rechtzeitig erfassen können. Vor dem Hintergrund der Tatsache, dass in der gesamten (nicht)wissenschaftlichen Literatur zur DOS-Skala kein Hinweis zur Begründung für die Kürzung der Skala von 25 auf 13 Items zu finden ist, existieren einige Untersuchungen zur psychometrischen Qualität der Skala im Bereich der Palliativpflege, Geriatrie, internen Medizin und Chirurgie. Als unzufrieden ist die Dokumentation der Interrater-Reliabilität in der Literatur zu erwähnen. Obwohl Schuurmans et al. [24] in ihrer Publikation darauf hinweisen und auf Schuurmans et al. [22] verweisen, lassen sich keine Daten zu diesem Ergebnis ausforschen. Die Ergebnisse zur internen Konsistenz (α = 0,74 und α = 0,98) der DOS-Skala (Langversion) [24] zeigte im Vergleich mit der Kurzversion der Skala einen ähnlichen Cronbach's α von 0,77 [7]. Je mehr Items eine Skala besitzt, desto höhere Cronbach's α-Werte können erreicht werden [3]. Trotz der Itemreduktion von 25 auf 13 Items ist der α-Wert von 0,77 [7] als ausreichend hoch anzusehen [3, 8]. Gründe für diesen durchaus hohen Wert können einerseits die gute Objektivität der Skala sein, welche bekannterweise einen positiven Einfluss auf die Reliabilität (geringe Fehlervarianz) hat, andererseits die hohe Item-Interkorrelation der Skala [3]. Insgesamt können die Werte zur prädiktiven Validität der Skala in den unterschiedlichsten Untersuchungsfeldern als ideal bezeichnet werden (Sensitivität: 82–100 %; Spezifität: 77–97 %). Dennoch ist zu beachten, dass es in keiner der angeführten Studien einen tatsächlichen objektiven Goldstandard gab. Oftmals wurde die Diagnose eines Arztes anhand der DSM-IV-Kriterien als Goldstandard verwendet. Jedoch hat dieser Goldstandard eine eingeschränkte Reliabilität, die Vergleichbarkeit ist fragwürdig und die Interpretation dieses Standards ist anfällig für Subjektivität [23]. Die Tatsache, dass verschiedene Assessment-Instrumente von anderen Berufsgruppen auch als Goldstandard

verwendet wurden, kann kritisch gesehen werden [7]. Die Berechnung der Likelihood-Ratio (LR)[1] in der Studie von Schuurmans et al. [23], zeigte eine mäßig LR^+ von 3,9 (CI 95 % 2,6–5,9) und eine sehr gute LR^- von 0,07 (CI 95 % 0,01–0,50) gemäß dem Interpretationsschema von Jaeschke et al. [14]. Im Vergleich dazu berechneten van Gemert und Schuurmans [9] sowohl eine gute $LR^+ = 7,6$ (CI 95 % 4,0–15) als auch eine gute $LR^- = 0,13$ (CI 95 % 0,02–0,80). In nur einem Drittel der eingeschlossenen Studien [7, 9] wurde tatsächlich die Praktikabilität erhoben, obwohl sie mehrmals positiv erwähnt wird. Das teilnehmende Pflegepersonal in der Studie von Detroyer et al. [7] empfiehlt, die Kurzversion der DOS-Skala um eine Antwortkategorie zu erweitern, um so die Möglichkeit zu schaffen, eine Begründung für ein Nichteinschätzen anführen zu können, womit bei Patienten im bevorstehenden terminalen Zustand die Nutzbarkeit der Skala verbessert werden kann. Insgesamt betrachtet, handelt es sich bei der DOS-Skala mit 13 Items um ein praktikables Screening-Instrument [7, 9].

In allen eingeschlossen Studien war die Anzahl der Stichproben gering. So lag die Spannweite der Stichprobengrößen bei R = 64 (min. 48; max. 112) Teilnehmern. Des Weiteren wurde in keiner Studie eine Poweranalyse durchgeführt. Dennoch geben alle Autoren der jeweiligen Studien die Gründe für die niedrige Stichprobenzahl an. Daher kann in allen Resultaten davon ausgegangen werden, dass möglicherweise keine Teilnehmerheterogenität in den Stichproben vorlag und somit die Ergebnisse nicht unkritisch interpretiert werden dürfen [3]. Einen Fehler zweiter Ordnung (β-Fehler) bei statistischen Tests zu begehen, nimmt bei zunehmender Abnahme von Stichprobengrößen wahrscheinlich zu. Aus diesem Grund sind die internen Validitäten der Studien kritisch zu betrachten [3]. So beruhen beispielsweise die Werte zur Validität und Reliabilität in der Studie von Schuurmans et al. [24] auf Zahlen einer mittleren Prozentzahl an Patienten mit Delir. Deshalb kann angenommen werden, dass diese Ergebnisse bei einer größeren Gruppe an deliranten Patienten höher ausfallen hätte können. Neben der Stichprobengröße löst auch die Häufigkeit eines vorliegenden Phänomens in einer Population (= Prävalenz) einen nicht unberücksichtigten Einfluss auf die Bestimmung der Sensitivitäts- und Spezifitätshöhe von Risikoeinschätzungsinstrumenten aus. Beide Werte sind stark durch die Häufigkeit eines vorliegenden Phänomens beeinflussbar [18]. In einer Studie an einer Population mit niedrigen Delirraten könnten hohe Spezifitätswerte errechnet werden, während eine hohe Anzahl an Delirien zu einer hohen Sensitivität führen könnten. Die Tatsache, dass zum Teil vulnerable Patienten ebenso falsch-positiv eingeschätzt werden können, verhindert nicht das Ziel eines Screenings [18]. Die hier präsentierten moderaten bis hohen Werte zur prädiktiven Validität, Übereinstimmungs- und Konstruktvalidität der DOS-Skala lassen sich u. a. darauf zurückführen, dass die Anwender des Instruments geschult wurden oder über eine längere Erfahrungszeit im Umgang [20] mit der DOS-Skala zurückgreifen konnten. Ebenso begegneten Koster et al. [15] dem Umstand der möglichen Beeinflussung bei der Diagnosestellung durch den Arzt, wenn das Risiko eines Delirs bestand, dadurch, dass dieser bereits schon bei einem Summenwert von 2 anstatt von 3 konsultiert wurde. Bei der Interpretation der sehr hohen Sensitivität und Spezifität der DOS-Skala sind diese Bias jedoch als gering zu sehen. Scheffer et al. [21] weisen auf drei wichtige Punkte bei der Verwendung der DOS-Skala im psychiatrischen Setting hin. Zum einen können bei einem demenziell erkrankten Menschen die Symptome eines Delirs nicht eindeutig von den Symptomen der Demenz diskriminiert werden. Zum anderen ist der Schweregrad eines Delirs auch mit der DOS-Skala einschätzbar, auch wenn sie primär nicht dafür entwickelt wurde [21, 9]. Als dritter Punkt führen sie an, dass Pflegepersonen die DOS-Skala bei unruhigen oder hyperaktiven Patienten anwenden und diese dann tendenziell als delirant einschätzen. Im Gegensatz dazu, wird das hypoaktive Delir leicht übersehen [21]. Die vereinzelten vorliegenden Studienergebnisse aus den vier unterschiedlichen Fachdisziplinen der Pflege sind nur begrenzt vergleichbar, da sie aus unterschiedlichen Settings stammen, eine zu stark schwankende Stichprobengröße aufweisen und verschiedene Instrumente als Übereinstimmung oder als Kontrolle der Vorhersage eingesetzt wurden.

Die Limitation der Literaturarbeit liegt darin, dass trotz des beschriebenen Suchvorgangs durch zwei unabhängige Rater nur englisch- und deutschsprachige Literatur einbezogen werden konnte. Somit kann nicht gänzlich ausgeschlossen werden, dass weitere Studienergebnisse in anderen Sprachen vorliegen. Eine weitere Eingrenzung stellt die Tatsache dar, dass keine Studie identifiziert werden konnte, die in Österreich durchgeführt wurde. Daher ist eine Übertragung der publizierten Ergebnisdaten auf die österreichische Population nur begrenzt möglich.

Schlussfolgerungen

Bei der Kurzversion der DOS-Skala handelt es sich um ein Screening-Instrument, das reliabel im Sinne der internen Konsistent ist. Des Weiteren ist sie ein valides Instrument in Bezug auf Inhaltsvalidität, Übereinstimmungsvalidität und prädiktiver Validität. Aufgrund der beiden letztgenannten Validitätskriterien kann die DOS-Skala als ein Screening-

[1] Positive und negative $LR^{(+/-)}$: Ein Ausdruck der Wahrscheinlichkeit eines Testergebnisses für Personen mit tatsächlichem Vorliegen eines Delirs im Verhältnis zur Wahrscheinlichkeit für Personen, die kein Delir aufweisen.

Instrument mit Kriteriumsvalidität bezeichnet werden. Hinsichtlich der Konstruktvalidität liegen derzeit nur Daten zu den 25 Items der DOS-Skala aus 2 Stichproben vor. Zur Kurzversion der DOS-Skala wurde keine Untersuchung dahingehend ausgeforscht. Derzeit kann somit nur von einer hinreichenden Konstruktvalidität ausgegangen werden. Des Weiteren konnten keine Studienergebnisse zur Interrater-Reliabilität der Skala, in Zusammenhang mit den Diagnosekriterien ICD-10 und keine Studien aus dem deutschsprachigen Raum ausfindig gemacht werden.

Die Relevanz für die Pflegepraxis von der DOS-Skala liegt darin, dass sie sich in der täglichen Praxis als valides und reliables Screening-Instrument erwies. Die 13 Items befähigen Pflegepersonen zwischen Menschen mit und ohne Delirrisiko zu screenen. Die Anwendung der DOS-Skala im klinischen Alltag kann deshalb zu einer raschen Risikoidentifizierung und Diagnose führen, womit ein individuelles Delirmanagement für Patienten ermöglicht wird. Die Anwendung der Skala in der Praxis ist einfach, gering im zeitlichen Aufwand und führt zur Steigerung der Pflegequalität schon nach einer kurzen Schulung, was zur Reduktion zukünftiger Delirien beiträgt.

In der Forschung wurde die DOS-Skala bereits mehrfach in unterschiedlichen Fachdisziplinen auf ihre testtheoretischen Gütekriterien getestet. Die geringe Anzahl vorliegender Studienergebnisse in den vier unterschiedlichen Fachdisziplinen der Pflege sind nur begrenzt vergleichbar. Daher sind weitere Studien zur Validität und Reliabilität in unterschiedlichen Ländern sowie in denselben und anderen Pflegebereichen erstrebenswert, um einerseits die Qualitätskriterien der DOS-Skala zu verdeutlichen und andererseits Vergleichsanalysen durchführen zu können. Ein Forschungsvorhaben zur Testung der Interrater-Reliabilität und zu den ersten Erkenntnissen zur Praktikabilität in Österreich ist angelaufen.

Interessenkonflikt G. Müller, J. Wetzlmair, P. Schumacher und M. Lechleithner geben an, dass kein Interessenkonflikt besteht.

Literatur

1. Bartholomeyczik, S. (2007). Einige kritische Anmerkungen zu standardisierten Assessmentinstrumenten in der Pflege. *Pflege*, *20*(4), 211–217.
2. Behrens, J., & Langer, G. (2010). *Evidence-based Nursing and Caring. Methoden und Ethik der Pflegepraxis und Versorgungsforschung* (3. Aufl.). Bern: Hans Huber.
3. Bortz, J., & Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. Aufl.). Heidelberg: Springer.
4. Bossuyt, P., Reltsma, J., Bruns, D., Gatsonls, C., Glaszlou, P., Irwig, L., Lljmer, J., Moher, D., Rennie, D., & Vet, H. de (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Annals of Internal Medicine*, *138*(1), 40–44.
5. Cerejeira, J., Firmino, H., Vaz-Serra, A., & Mukaetova-Ladinska, E. B. (2010). The Neuroinflammatory hypothesis of delirium. *Acta Neuropath*, *119*(6), 737–754.
6. Clark, L. E., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.
7. Detroyer, E., Clement, P. M., Baeten, N., Pennemans, M., Decruyenaere, M., Vendenberghe, J., Menten, J., Joosten, E., & Milisen, K. (2014). Detection of delirium in palliative care unit patients: A prospective descriptive study of the delirium observation screening scale administered by bedside nurses. *Palliative Medicine*, *28*(1), 79–86.
8. Ewers, A. (2004). Die interne Konsistenz der Confusion Rating Scale zur Messung acuter postoperativer Verwirrtheit: Eine Testung bei kardiochirurgischen Patienten in Deutschand. In E. M. Panfil (Hrsg.), *Fokus: Klinische Pflegeforschung. Beispiele quantitativer Studien* (S. 158–172). Hannover: Schlütersche Verlagsgesellschaft.
9. Gemert, L. A. van, & Schuurmans, M. J. (2007). The Neecham confusion scale and the delirium observation screening scale: capacity to discriminate and ease of use in clinical practice. *BMC Nursing*, *6*(3), 1–6.
10. Grover, S., & Kate, N. (2012). Assessment scales for delirium: A review. *World Journal of Psychiatry*, *2*(4), 58–70.
11. Hasemann, W., Kressig, R. W., Ermini-Fünfschilling, D., Pretto, M., & Spirig, R. (2007). Screening, Assessment und Diagnostik von Delirien. *Pflege*, *20*(4), 191–204.
12. Inouye, S. K. (2006). Delirium in older persons. *N. Engl. J. Med.*, *354*(11), 1157–1165.
13. Inouye, S. K., Westendorp, R. G. J., & Saczynski, J. S. (2014). Delirium in elderly people. *National Institute of Health*, *383*(9920), 911–922.
14. Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). Users guides to the medical literature III. how to use an article about a diagnostic test. B. what are the results and will they help me in caring for my patients? the evidence-based medicine working group. *Journal of the American Medical Association*, *27*(2), 703–707.
15. Koster, S., Hensens, A. G., Oosterveld, F. G. J., Wijma, A., & Palen, J. van der (2009). The delirium observation screening scale recognizes delirium early after cardiac surgery. *European Journal of Cardiovascular Nursing*, *8*(4), 309–314.
16. Kunz, R., Khan, K., Kleijnen, J., & Antes, G. (2009). *Systematische Übersichtsarbeiten und Meta-Analysen. Einführung in Instrumente der evidenzbasierten Medizin für Ärzte, klinische Forscher und Experten im Gesundheitswesen*. Bern: Hans Huber.
17. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. & The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta – Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097.
18. Moosbrugger, H., & Kelava, A. (2008). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger, & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 8–26). Heidelberg: Springer.
19. ÖGGG – Österreichische Gesellschaft für Geriatrie und Gerontologie (2013). *Delir 2013 – Ein häufiges Syndrom im Alter – eine interdisziplinäre Herausforderung*. Wien: ÖGGG – Österreichische Gesellschaft für Geriatrie und Gerontologie.
20. Reuschenbach, B. (2011). Gütekriterien. In B. Reuschenbach, & C. Mahler (Hrsg.), *Pflegebezogenene Assessmentinstrumente – Internationales Handbuch für Pflegeforschung und -praxis* (S. 57–79). Bern: Hans Huber.
21. Scheffer, A. C., Munster, B. C. van, Schuurmans, M. J., & Rooij, S. E. de (2011). Assessing severity of delirium by the delirium observation screening scale. *International Journal of Geriatric Psychiatry*, *26*(3), 284–291.
22. Schuurmans, M. J., Donders, R. T., Shortridge-Baggett, L. M., & Duursma, S. A. (2002). Delirium case finding: pilot testing of a

new screening scale for nurses. *Journal of the American Geriatrics Society*, *50*(4), 3.
23. Schuurmans, M. J., Deschamps, P. I., Markham, S. W., Shortridge-Baggett, L. M., & Duursma, S. A. (2003). The measurement of delirium: review of scales. *Research and Theory for Nursing Practice*, *17*(3), 208–224.
24. Schuurmans, M. J., Shortridge-Baggett, L. M., & Duursma, S. A. (2003). The delirium observation screening scale: A screening instrument for delirium. *Research and Theory for Nursing Practice*, *17*(1), 31–50.
25. Wong, C. L., Holroyd-Leduc, J., Simel, D. L., & Straus, S. E. (2010). Does this patient have delirium? Value of bedside instruments. *Journal of the American Medical Association*, *304*(7), 779–786.