Pain and its consequences in dementia: Observing the complex relationship between pain, behaviour and ADL in nursing home residents

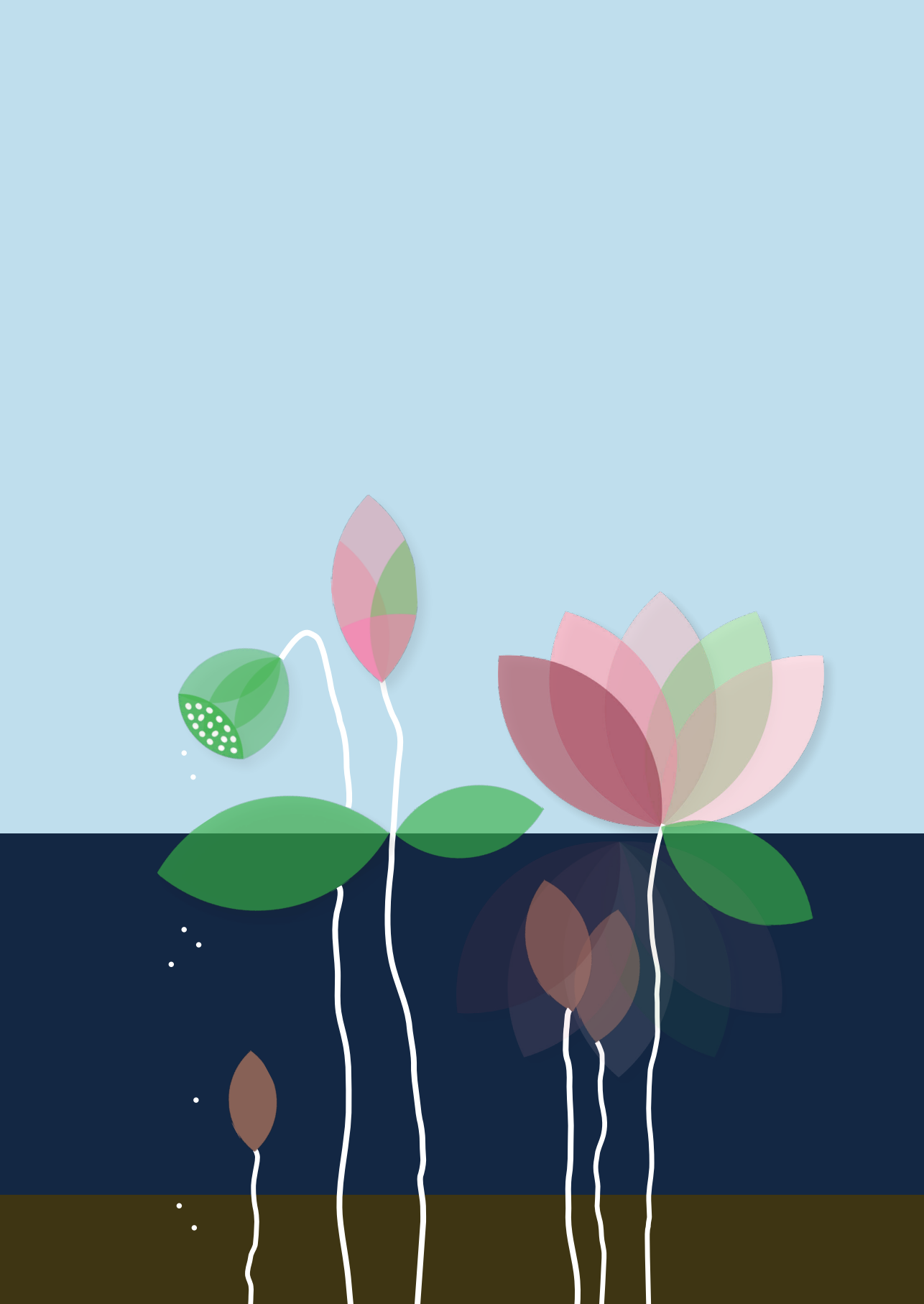
Dalen-Kok, A.H. van

CHAPTER 7

Observational pain assessment in older persons with dementia in four countries: observer agreement and factor structure of the pain assessment in impaired cognition

Margot W.M. de Waal , Annelore H. van Dalen-Kok, Henrica C.W. de Vet, Lydia Gimenez-Llort, Ljubica Konstantinovic, Marina de Tommaso, Thomas Fischer, Albert Lukas, Miriam Kunz, Stefan Lautenbacher, Frank Lobbezoo, Brian E. McGuire, Jenny T. van der Steen, Wilco P. Achterberg.

Eur J Pain. 2020 Feb;24(2):279-296. doi: 10.1002/ejp.1484a



Abstract

Background

Recognition of pain in people with dementia is challenging. Observational scales have been developed, but there is a need to harmonise and improve the assessment process. In EU initiative COST-Action TD1005, 36 promising items were selected from existing scales to be tested further. We aimed to study the observer agreement of each item, and to analyse the factor structure of the complete set.

Methods

One hundred and ninety older persons with dementia were recruited in four different countries (Italy, Serbia, Spain and The Netherlands) from different types of healthcare facilities. Patients represented a convenience sample, with no pre-selection on presence of (suspected) pain. The *Pain Assessment in Impaired Cognition* (PAIC, research version) item pool includes facial expressions of pain (15 items), body movements (10 items), and vocalizations (11 items). Participants were observed by health professionals in two situations, at rest and during movement. Intrarater and interrater reliability was analysed by percentage agreement. The factor structure was examined with principal component analysis with orthogonal rotation.

Results

Health professionals performed observations in 40 to 57 patients in each country. Intrarater and interrater agreement was generally high (≥70%). However, for some facial expression items, agreement was sometimes below 70%. Factor analyses showed a 6-component solution, which were named as follows: Vocal pain expression, Face anatomical descriptors, Protective body movements, Vocal defence, Tension, and Lack of affect.

Conclusions

Observation of PAIC items can be done reliably in healthcare settings. Observer agreement is quite promising already without extensive training.

Significance

In this international project, promising items from existing observational pain scales were identified and evaluated regarding their reliability as an alternative to pain self-report in people with dementia. Analysis on factor structure helped to understand the character of the items. Health professionals from 4 countries using 4 different European languages were able to rate items reliably. The results contributed to an informed reduction of items for a clinical observer scale (*Pain Assessment in Impaired Cognition* scale with 15 items: PAIC15).

Introduction

Recognition of pain in people with impaired cognition and communication problems is challenging because of impairment of self-report capacities [1]. International epidemiological research shows that people with dementia typically receive inadequate pain medication and experience inadequate pain management [2]. This may be because people with cognitive impairment do not reliably report when they have pain. In an effort to find an alternative to self-report, in various countries, scales have been developed that rely on observations, but they often lack sufficient psychometric evaluation. For instance, lack of a gold standard in the clinical setting (as opposed to experimental testing) hinders evaluation of validity. Also reliability and clinical utility is tested in small samples of raters in specific clinical settings, and (international) clinical implementation is hampered [3]. At this moment a considerable number of scales is available. There is a need to improve and harmonise the assessment process, as this will help in gathering comparable data and increase applicability across settings.

In the European COST Action TD-1005 "*Pain assessment in patients with impaired cognition, especially dementia*", experimental and clinical researchers together with health professionals aimed to develop a comprehensive and internationally agreed-upon pain assessment scale for older adults with impaired cognition. It was anticipated that the development of this new scale would require an iterative process, in which the loop of evaluation, adaptation and re-testing of items is followed several times [4]. The novel idea was to synthesise existing knowledge about observations of pain in older adults with dementia. For that purpose, all existing observational pain behaviour scales were identified and their items categorised in three groups: facial expressions, vocalizations, and body movements for the *research version of the Pain Assessment in Impaired Cognition* (PAIC, 36 items) [5]. In this way, we built further on the best available expertise. As such, the PAIC can be considered as a 'meta-tool'. For the final PAIC scale, further reduction of number of items was anticipated, using results from various psychometric studies to enhance usability [6].

The setting in which an observational scale will be used will vary between and even within countries [3]. The goal of the EU COST initiative was to develop a scale that can be used by a variety of health professionals in their clinical practice to rate a range of behaviours considered to be indicative of pain in people with dementia. It is important to examine items by using observations of health professionals working in a variety of real-life healthcare settings, in various European countries, as this will result in more robust findings. Specific aims of the present study were: a) to evaluate the interrater agreement and intrarater agreement of individual items and b) to study the factor structure of the PAIC item pool. Factor analysis is used to explore whether individual items can be grouped into meaningful components, for example, pain specific reactions and affective pain consequences.

Methods

Procedure

This was a multicentre, observational study in four countries covering various regions within Europe: Italy, Serbia, Spain, and The Netherlands. Each country was provided with the same study protocol, but implementation varied slightly due to different local conditions.

Health professionals performed observations among persons with dementia in everyday, real-life settings in two conditions: at rest and during movement. Observation was carried out under both conditions as it was expected that movement might induce pain. Also, some items can only be rated during movement of the whole body (e.g., pacing), while others (e.g., facial expressions) are more difficult to assess during gross movement. Examples of situations at rest include sitting in a chair or lying in bed, but excluded moments when drinking, eating, or sleeping. Situations during movement could include repositioning, thus observing a person when he/she moved or was being moved or transferred as part of his/her usual care. On day 1, all participants were seen by two observers who rated all items independently (preferably by observing the same situation together or one after the other within 10 minutes). All patients were rated a third time by one of the health care professionals on day 2. The observations at rest and during movement were on different subsequent days (the exact schedule depended on the situation and feasibility in each country; appendix 1).

Participants – Patients

For each country, participating patients were sought in the health care setting that has a high prevalence of patients with dementia, and in which future use of the PAIC was anticipated, e.g., nursing homes, geriatric hospital wards, or rehabilitation hospitals. It was a convenience sample of patients with a clinical diagnosis of dementia. Pain in any form was no inclusion or exclusion criterion. Given the high prevalence of pain in old individuals, we assumed that there would be a mix of patients with and without pain, in whom a range of items would be observed. We further assumed different levels of cognitive impairment (mild to severe dementia) in patients, and different levels of acquaintance (e.g., no previous, intermittent, or constant contact) of health care professionals with the patient. We excluded patients with Parkinson's disease, Huntington's disease, schizophrenia, Korsakov syndrome, patients in a vegetative state, coma patients and stroke patients with facial impairments that may hamper facial expressions. These groups were excluded either because observation of pain signs is more difficult (because of strong behavioural limitations), or because a substantial number of behaviours covered by the items would not occur in these groups.

Participants – observers

Depending on the care situation in each country, healthcare professionals who would likely use the new scale in the future were chosen as observers. They could be either physicians, nurses, nurse assistants or psychologists (Table 1). A brief training session of 15-30 minutes was held in each facility to inform the observers about the new assessment scale and about the type of items. The PAIC-scoring forms contained a brief written instruction on scoring. The instructions for using the PAIC were intentionally brief as we wanted to determine if the scale could be used reliably with minimal training.

Measures

The research version of the PAIC (*Pain Assessment in Impaired Cognition*) is an observational scale that includes facial expressions of pain (15 items), body movements (10 items), and vocalizations (11 items). The items were chosen following a process that included an extensive literature review of existing tools and several consultation rounds with experts- this process is described in detail elsewhere [5 6].

On the scoring form, for each item a short description of the meaning of the item was provided, for example, frowning 'lowering and drawing brows together', rubbing 'tugging or massaging affected area', shouting 'using a loud voice to express words'. Items were scored on a 4-point scale: 0 'not at all', 1 'slight degree', 2 'moderate degree', and 3 'great degree'. There was an additional column 'not scored', with the options: a 'item is not clear', b 'situation is unsuitable', c 'physical status of person not suitable for scoring', d 'other'. The text was translated and culturally adapted using a forward-backward procedure in seven European languages. For each country, the translation has been checked with a think aloud test [7 8].

Several characteristics of the rating situation, the observer and the patient were measured to describe the study sample: profession of the rater, experience in pain rating, duration of acquaintance with patient, facility (community care, institutional long term care (LTC), hospital care, hospice care), sex and age of the patient, and type of dementia (as stated in the medical chart). Severity of cognitive impairment was measured with the Reisberg Global Deterioration Scale (GDS). This scale describes seven stages of cognitive impairment, where stages 1-3 are pre-dementia stages and stages 4-7 are dementia stages [9].

Ethics and data collection

In each country, a supervising researcher coordinated the study. Ethics approval was obtained in each country, consistent with local procedures (for Italy by the Ethic Committee of Policlinico General Hospital, Bari in February 2015; for Serbia by the ethics committee of the Rehabilitation Clinic of the University of Belgrade School of Medicine 03-2212; for Spain by the Germanes Hospitalàries Hospital Sagrat Cor Martorell Medical Ethics Committee PR-2015-04; for The Netherlands: LUMC Medical Ethical Committee P14.245). Depending on local procedures, appropriate informed (proxy) consent was obtained. Each country collected and archived data on paper, and registered data in a local database. All

datasets were sent to one location in The Netherlands (to MWMdW at LUMC), to form one central research database from which data-cleaning and analyses were conducted. See also publication of Dutch results on reliability [10].

Sample size and statistical analyse

We aimed to recruit 50 patients per country, in total 200 patients from four countries, which is sufficient for factor analysis [4].

First, we examined the ratings of each individual PAIC item: the degree to which certain items were endorsed (or not) on the 4-point scale, missing items, and floor/ceiling effects of the items. In this context, a floor effect emerges when the behaviour described in an item is almost never present. The ceiling effect results from the opposite when a behaviour is almost always present. In both cases, the affected item is of limited value because it cannot indicate variance between persons. Second, reliability was analysed by percentage of agreement in scores on the 4-point scale between raters [11]. Missing scores were recoded to 0, thus assuming that items that were not scored meant that behaviour was not shown. More than 5% missing scores were discussed. For sensitivity analyses, first, percentage agreement was also calculated with dichotomized scores (0=absent; 1,2,3=present), and this was compared with percentage agreement of scores with the 4-point scale. Second, pairs of observations with missing scores were excluded, and this was compared with the percentage agreement of scores (on the 4-point scale) with missing scores recoded to 0. Percentages agreements below 70% were regarded as poor agreement.

An exploratory factor analysis was performed on the sample containing the first observation of each patient in a rest situation, and with no missing scores. We chose not to recode missing scores to 0 as this would influence the correlation between items. The rest situation was chosen as it had the largest sample size, and because situations at rest are not as diverse as situations during movement, meaning that conditions of the measurements can be better standardized. Principal component analysis (PCA) was used with orthogonal (varimax) rotation. The Kaiser-Meyer-Olkin (KMO) statistics were checked to determine the adequacy of the sample size, and also to check KMO values of individual items to be above the limit of 0.5 [12]. The final decision about the number of factors was based on Eigenvalues and scree plot, combined with interpretability of the factors.

Results

Description of setting, observers and patients

In total, 50 healthcare professionals in four countries performed observations in 190 patients, 40 - 57 patients in each country (Table 1). In Italy, observations were done in different hospitals by three physicians, one nurse assistant, and eight psychologists with various degrees of experience of using pain measurement scales in daily practice. Observers in Italy had not known the patients before (56%) or had known them for less than a month (32%). In Serbia, observations were also done in a hospital setting by two nurses and two physicians that were well trained in the use of pain measurement scales. Serbian observers had known the patients for at least 1 week (18%) and up to 6 months (45%). In Spain, observations were done in a community day-care centre and in a day-care hospital facility by two nurses and four nurse assistants who all had experience with using pain measurement scales in daily practice. Spanish observers had known 96% of the patients for several months. In The Netherlands, 14 nursing assistants and 10 registered nurses observed residents in nursing homes. Forty-six percentage of them lacked experience with using pain measurement scales in daily practice, and 42% used these scales less than once a month. The observers had known 78% of the patients for 6 months or more.

Patients were on average 74 – 86 years old. In Italy and Serbia, half were women, and in Spain and The Netherlands, more than three quarters were women. The severity of dementia varied somewhat between countries with an average GDS-score of 4.6 (moderate) to 6.1 (severe). The majority of patients had Alzheimer's disease, except for Italy where the majority had vascular dementia.

Table 1. Characteristics of study population and observers.

	Italy		Serbia		Spain		The Netherlands	
Study population	(n=57)		(n=40)		(n=48)		(n=45)	
Period of data collection	2015		Sep'14-Aug'17		Oct'15-May'17		Nov'14-Oct'15	
Setting								
Community day care	0		0		34	71%	0	
Long-term residential care	0		0		14	29%	45	100%
Hospital care	57	100%	40	100%	0		0	
Length of stay in months, mean (SD)	-		-		-		29.5	(24.5)
Age in years, mean (SD) (range)	74.4	(11.5) (33-89)	81.5	(3.9) (75-89)	77.3	(7.8) (45-92)	85.7	(7.0) (69-103)
Gender, female	28	49%	22	55%	37	77%	36	80%
Dementia severity: Reisberg GDS								
Mean score (SD) (min-max score)	4.8	(2.0) (1-9)	5.7	(0.7) (5-7)	4.6	(0.9) (3-6.5)	6.1	(0.9) (4-7)

Table 1. Characteristics of study population and observers *(continued)*.

	Italy		Serbia		Spain		The Netherlands	
Study population	(n=57)		(n=40)		(n=48)		(n=45)	
Period of data collection	2015		Sep'14-Aug'17		Oct'15-May'17		Nov'14-Oct'15	
Type of dementia								
Alzheimer's disease	5	9%	19	48%	33	67%	25	57%
Vascular dementia	29	52%	13	33%	3	6%	3	7%
Mixed dementia	6	11%	6	15%	5	10%	3	7%
Other	9	13%	0		7	15%	1	2%
Not specified or unknown	7	16%	2	5%	0		12	27%
Acquaintance first observer with client								
Do not know this client	32	56%	0	0%	0	0%	7	16%
Less than 1 week	10	18%	7	18%	0	0%	0	0%
1 week to 1 month	8	14%	18	45%	2	4%	1	2%
Months	4	7%	15	38%	18	38%	2	4%
6 months or more	3	5%	0	0%	28	58%	35	78%
Observers	(n=12)		(n=4)		(n=6)		(n=28)	
Profession								
Physician	3	25%	2	50%	0		0	
Registered nurse	0		2	50%	2	33%	8	33%
Nursing assistant	0		0		4	67%	14	50%
Nurse in training	1	8%	0		0		2	8%
Psychologist	8	67%	0		0		0	
Confidence identifying pain mean (SD)	9.1	(1.4)			8.3	(1.0)	7.4	(2.0)
(min-max score)		(6-10)				(7-10)		
Pain measurement scales used in organization, yes	10	91%	4	100%	6	100%	13	54%
How often do you use pain measurement scales in daily practice?								
Never	2	18%			0		13	54%
Less than once a month	1	9%			0		10	42%
Once or twice a month	0				2	33%	0	
Around once a week	0				1	17%	1	4%
Most days	6	55%			3	50%	0	
Every day	2	18%			0		0	

Note: Missing values for Reisberg GDS n=6 (IT 4, NL 2), type of dementia n=2 (IT 1, NL 1), observer profession n=4 (NL 4), confidence identifying pain n=8 (SB 4, NL 4), pain measurement scales in organization n=5 (IT 1, NL 4), pain measurement scales in daily practice n=9 (IT 1, SB 4, NL 4).

Description of observation

In all countries, patients were rated at rest by one pair of observers. Rest situations could be lying in bed or sitting in a chair. Except for Italy, patients were also observed during movement. Movement situations comprised a short walk, e.g., down a corridor (Serbia, Spain, The Netherlands), transfer from bed to chair or wheelchair, or repositioning in bed (Serbia, The Netherlands).
In Serbia and Spain, patients were rated by one pair of observers. In The Netherlands, the same participants were seen by two pairs of observers, a different pair of observers at rest and during movement situations. In Italy, pairs of observers were not all the same for intrarater and interrater analyses (Appendix 1).

Item scores

Table 2 gives an overview of the distribution of scores on each PAIC item for the first observation of each patient at rest. More categories were used to grade the facial expressions compared to body movements and vocalizations. Facial expressions showed no floor effects: scores 0 'not at all present' for individual items ranged between 44.2% and 89.5% of observations. For body movements and vocalizations, floor effects were acceptable: 3 out of 10 body movements and 3 out of 11 vocalizations had scores of 0 for more than 90% of observations, with the item 'using offensive words' reaching 97.4% with a score of 0. For body movements, score 3 ('great degree') was not used very often: in 6 out of 10 items <1% of observations. There were four items in facial expressions and one item in vocalizations with 0.5% or 1.1% missing scores (that is missing scores in 1 or 2 out of 190 observations). In body movements, two items showed high numbers of missing items: 'guarding' (4.2% missing) and 'limping' (5.8% missing). This was also seen in movement situations, with respectively 5.3% and 8.3% (Appendix 2). The reason mostly given was that the physical status of the patient was not suitable for scoring this item.

Table 2. Scores per item (in percentages) in first observations in rest (n=190)

	Score:	0	1	2	3
	Not rated (missing)	Not at all	Slight degree	Moderate degree	Great degree
Facial expressions					
Pained expression		72.6	14.2	12.6	0.5
Frowning	0.5	70.5	19.5	7.9	1.6
Narrowing eyes		76.8	16.8	5.8	0.5
Closing eyes		76.3	11.6	3.7	8.4
Raising upper lip		89.5	8.4	1.1	1.1
Opened mouth	0.5	77.9	15.3	4.7	1.6
Tightened lips		62.1	23.2	11.1	3.7
Clenched teeth		88.9	7.9	1.6	1.6
Empty gaze	1.1	44.2	35.8	12.1	6.8
Seeming disinterested	1.1	44.7	24.2	20.0	10.0
Pale face		57.9	21.6	18.4	2.1
Teary eyed		87.9	10.0	1.6	0.5
Looking tense		63.7	22.6	12.6	1.1
Looking sad		45.8	37.4	14.2	2.6
Looking frightened		84.2	10.5	4.7	0.5
Body movements					
Freezing		80.0	14.7	4.2	1.1
Curling up		83.7	14.2	1.6	0.5
Clenching hands		78.4	16.8	3.7	1.1
Resisting care		85.8	11.6	2.1	0.5
Pushing		94.7	3.7	1.6	0.0
Guarding	4.2	82.6	10.0	2.6	0.5
Rubbing		89.5	7.9	2.6	0.0
Limping	5.8	90.0	3.2	0.5	0.5
Restlessness		76.8	15.8	4.7	2.6
Pacing		96.8	2.1	1.1	0.0
Vocalizations					
Using offensive words		97.4	1.1	1.6	0.0
Using pain related words		85.8	10.0	3.2	1.1
Repeating words		85.8	11.1	2.6	0.5
Complaining		80.0	15.3	2.1	2.6
Shouting		94.7	3.7	.5	1.1
Mumbling		84.2	12.1	2.6	1.1
Screaming	0.5	95.3	2.1	1.6	0.5
Groaning		81.1	14.7	2.6	1.6
Crying		87.4	8.4	4.2	0.0
Gasping		84.7	13.2	2.1	0.0
Sighing		74.2	20.0	4.7	1.1

Observer agreement of individual items

In both rest and movement situations, there were items of facial expressions with low agreement between observers with percentages below 70 (Table 3), especially in The Netherlands. Five items showed low interrater agreement in three or four countries: 'looking sad' (four countries), 'tightened lips', 'empty gaze', 'seeming disinterested', and 'looking tense'. In The Netherlands, facial items also showed low intrarater agreement for the same observers in two consecutive days (Table 4).

Body movement items generally showed good reliability for both interrater agreement and intrarater agreement, with 7 out of 10 items showing percentages of 70 or higher for all countries. The items 'freezing' and 'clenching hands' showed low interrater agreement in movement in The Netherlands and low intrarater agreement at rest in Spain. 'Restlessness' showed low intrarater and interrater agreement in The Netherlands. Note that for the items 'guarding' and 'limping', missing pairs of observations were above 5%. Sensitivity analyses on observations without pairs of observations that included missing scores showed that percentages agreement were 0-2% lower.

Vocalization items showed good reliability with a few exceptions, for example, for interrater agreement in Serbia at rest for the items 'groaning', 'gasping' and 'sighing'.

In a sensitivity analysis, percentage agreement was analysed after dichotomization of scores, indicating that pain-related behaviours were either present (scores 1 or higher) or absent (scores 0 or missing). As expected, compared to percentages agreement using scores on the 4-point scale, this resulted in higher intrarater and interrater agreement. For Italy and Serbia, all interrater agreement improved over 70% (Appendix 3 and 4).

Table 3. Interrater agreement in percentages.

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement	(n=39)	(n=40)	(n=40)	(n=48)	(n=48)	(n=45)	(n=45)	(n=172)	(n=133)
Facial expressions									
Pained expression	84	90	95	96	79	82	60	88	77
Frowning	85	93	93	81	77	53	29	77	65
Narrowing eyes	87	93	90	81	90	69	51	82	77
Closing eyes	85	95	93	56	90	69	56	75	79
Raising upper lip	90	98	98	98	90	91	84	94	90
Opened mouth	74	93	100	94	85	69	51	83	78
Tightened lips	77	63	73	50	52	69	60	64	61
Clenched teeth	87	83	95	83	83	82	69	84	82
Empty gaze	85	48	68	67	77	51	40	62	62
Seeming disinterested	80	48	68	46	65	56	56	56	62
Pale face	72	93	100	83	90	60	69	77	86
Teary eyed	77	85	88	98	100	89	84	88	91
Looking tense	77	85	93	52	54	67	47	69	63
Looking sad	67	68	78	52	71	53	49	59	65
Looking frightened	87	75	88	83	92	87	56	83	78
Body movements									
Freezing	100	80	68	73	81	84	44	84	65
Curling up	100	78	88	98	100	84	69	90	86
Clenching hands	92	83	90	79	81	76	60	82	77

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement	(n=39)	(n=40)	(n=40)	(n=48)	(n=48)	(n=45)	(n=45)	(n=172)	(n=133)
Body movements									
Resisting care	95	70	73	96	98	98	71	90	81
Pushing	95	95	90	100	100	100	89	98	93
Guarding #	95	98	98	96	98	78	82	91	93
Rubbing	90	100	100	100	98	78	89	92	96
Limping #	100	98	100	98	81	96	71	98	84
Restlessness	100	98	100	81	94	62	73	84	89
Pacing	92	98	95	98	90	98	96	97	93
Vocalizations									
Using offensive words	100	98	98	98	98	96	93	98	96
Using pain related words	92	73	70	100	96	89	73	89	81
Repeating words	95	85	83	94	98	96	82	92	88
Complaining	95	88	93	85	90	84	71	88	84
Shouting	97	98	98	96	94	98	78	97	90
Mumbling	95	83	93	98	92	69	58	86	81
Screaming	95	93	98	98	98	96	84	95	93
Groaning	90	65	85	98	92	89	73	86	84
Crying	95	70	75	98	100	89	93	88	90
Gasping	95	65	83	90	88	89	84	85	85
Sighing	90	68	85	79	81	73	60	77	75

Note: % Agreement with for missing score =0, percentage ≥ 70% in blue.
#Missing pairs of observations for guarding in rest 4.1% and in movement 5.3%; for limping in rest 6.4% and in movement 8.3%.

Table 4. Interrater agreement in percentages.

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement	(n=46)	(n=40)	(n=39)	(n=48)	(n=48)	(n=40)	(n=40)	(n=174)	(n=127)
Facial expressions									
Pained expression	87	98	90	98	90	78	50	90	77
Frowning	91	100	95	75	79	60	35	82	70
Narrowing eyes	85	100	95	88	92	70	55	86	81
Closing eyes	87	98	95	71	96	55	73	78	88
Raising upper lip	94	100	100	98	85	88	80	95	88
Opened mouth	89	90	95	96	79	70	50	87	75
Tightened lips	91	83	87	73	73	70	60	79	73
Clenched teeth	94	93	100	96	88	83	70	91	86
Empty gaze	85	70	77	58	83	65	68	70	76
Seeming disinterested	80	70	74	75	96	65	70	73	81
Pale face	85	93	97	100	100	65	63	86	87
Teary eyed	87	83	87	96	98	95	85	90	91
Looking tense	89	90	97	65	75	68	53	78	75
Looking sad	87	73	77	71	75	68	58	75	70
Looking frightened	78	83	92	85	90	78	68	81	84
Body movements									
Freezing	96	95	74	69	75	80	65	85	72
Curling up	91	85	85	100	100	83	80	90	89
Clenching hands	87	85	97	69	92	85	70	81	87
Resisting care	85	90	77	100	98	98	73	93	84

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement	(n=46)	(n=40)	(n=39)	(n=48)	(n=48)	(n=40)	(n=40)	(n=174)	(n=127)
Body movements									
Pushing	87	93	100	100	100	98	80	94	94
Guarding #	89	98	97	96	98	80	75	91	91
Rubbing	96	98	97	100	100	80	88	94	95
Limping #	96	98	100	100	88	98	70	98	86
Restlessness	89	98	100	79	90	53	75	80	88
Pacing	96	95	97	98	96	95	93	96	95
Vocalizations									
Using offensive words	98	95	97	100	100	95	95	97	98
Using pain related words	94	88	80	100	92	85	63	92	79
Repeating words	96	88	87	98	98	85	80	92	89
Complaining	85	88	97	98	96	80	63	88	86
Shouting	94	100	100	98	100	88	78	95	93
Mumbling	85	88	90	96	90	78	73	87	84
Screaming	98	93	97	100	98	95	88	97	95
Groaning	83	78	82	98	90	78	55	85	76
Crying	96	90	85	100	98	93	80	95	88
Gasping	85	83	82	92	92	85	85	86	87
Sighing	89	83	87	81	75	73	75	82	79

Note: % agreement with for missing score =0, Percentage ≥ 70% in blue.
Missing pairs of observations for guarding in rest 4.0% and in movement 5.5%; for limping in rest 6.3% and in movement 8.7%

Factor analyses

Exploratory factor analyses were performed to explore whether individual items could be grouped into underlying components. This was done in 172 observations, the first observation at rest for each patient. For 18 of the 190 patients, observations were left out due to missing scores.

First, checks were performed to look whether all items could be included in the analysis. A visual check of the correlation matrix showed highest correlation between face (facial expression) item 1 'pained expression' and face item 3 'narrowing eyes' (0.72), and low correlations (majority <0.3 with all other items) for face item 4 'closing eyes', face item 6 'opening mouth', face item 8 'clenched teeth', bm (body movement) item 1 'freezing', bm item 9 'restlessness', bm item 10 'pacing', and voc (vocalization) item 1 'using offensive words'. KMO values of individual items were mostly above 0.7 ('good' for 25 items) or between 0.5-0.7 ('mediocre' for 10 items, with face item 4 'closing eyes' 0.58, bm item 10 'pacing' 0.54, and voc item 1 'using offensive words' 0.58), and below 0.5 for one item (0.48 for face item 8 'clenched teeth'). The four items with KMO values below 0.6 were removed [12] and we also excluded the two items with floor effects of <95% with scores 0 (bm item 10 'pacing' and voc item 1 'using offensive words').

Factor analyses was performed on the remaining 32 items. A KMO statistic of 0.830 confirmed that the sample size was adequate. Correlations between items were sufficiently large, according to Bartlett's test of sphericity (Chi square=3,372 (*df* 496), $p<0.001$). Eigenvalues were >1 for eight components. Visual inspection of the scree plot showed that six components should be retained. Analyses were rerun with this solution enforced on the data. Table 5 shows the factor loadings of the components after rotation. The six components explained 62.6% of the variance.

After inspection of factor loadings, we named the components as follows: 'Vocal pain expression' with seven vocalization items such as sighing, using pain related words, and gasping; 'Face anatomical descriptors' with highest factor loadings on narrowing eyes, teary eyed, and pained expression; 'Protective body movements' with pushing, resisting care, and guarding; 'Vocal defence' with items shouting and screaming; 'Tension' with items tightening lips, looking sad, looking tense, and freezing; and 'Lack of affect' with empty gaze and seeming disinterested. Note that although the item 'curling up' is grouped under component 1, it also has a high loading on component 3 'Protective body movements' (Table 5).

Table 5. Rotated Component Matrix from factor analysis on 32 PAIC items# in 172 observations in rest. Factor loading above 0.5 appear in bold and coloured cell.

PAIC items		Component					
		1	2	3	4	5	6
		Vocal pain expression	Face anatomical descriptors	Protective body movements	Vocal defence	Tension	Lack of affect
F1- pained expression	V11- sighing	**0.71**	0.18	0.08	0.05	0.30	0.00
F2- frowning	V2- using pain related words	**0.69**	0.19	0.14	0.44	0.15	0.04
F3- narrowing eyes	V10- gasping	**0.64**	0.41	0.04	0.11	0.10	0.05
F5 – raising upper lip	V8- groaning	**0.63**	0.23	-0.06	0.33	-0.02	0.24
F6- opening mouth	V6- mumbling'	**0.62**	0.08	0.34	0.47	0.04	0.04
F7- tightening lips	V3- repeating words	**0.61**	0.16	0.33	0.07	-0.01	0.13
F9- empty gaze	V4- complaining	**0.60**	0.30	0.08	0.32	-0.15	0.24
F10- seeming disinterested	BM7- rubbing	**0.58**	-0.14	0.29	0.09	0.20	0.10
F11- pale face	BM8- limping	**0.53**	0.09	0.06	-0.17	0.20	-0.01
F12- teary eyed	F3- narrowing eyes	0.20	**0.76**	0.17	0.12	0.19	0,11
F13- looking tense	F12- teary eyed	0.14	**0.66**	0.04	0.08	0.02	-0.03
F14- looking sad	F1 – pained expression	0.38	**0.64**	0.13	0.10	0.20	0.20
F15- looking frightened	F5 – raising upper lip	0.08	**0.57**	0.29	0.42	0.02	-0.02
BM1- freezing	V9- crying	0.43	**0.55**	0.37	0.18	0.14	0.05
BM2- curling up	F2- frowning	0.25	0.48	0.37	0.18	0.35	0.16
BM3- clenching hands	BM5- pushing	0.01	0.18	**0.75**	0.32	0.11	0.11
BM4- resisting care	BM4- resisting care	0.36	0.14	**0.74**	0.06	0.01	0.12

		Component					
		1	2	3	4	5	6
PAIC items		**Vocal pain expression**	**Face anatomical descriptors**	**Protective body movements**	**Vocal defence**	**Tension**	**Lack of affect**
BM5- pushing	BM6- guarding	0.35	0.08	**0.73**	0.00	0.02	0.08
BM6- guarding	F15- looking frightened	0.01	0.32	**0.56**	0.22	0.35	0.16
BM7- rubbing	BM2- curling up	**0.63**	0.24	**0.54**	-0.12	0.12	0.11
BM8- limping	V5- shouting	0.21	0.18	0.02	**0.81**	0.08	-0.04
BM9- restlessness	V7- screaming	0.08	0.26	0.07	**0.76**	0.00	-0.12
V2- using pain related words	BM9- restlessness	0.14	-0.16	0.24	**0.56**	0.24	0.22
V3- repeating words	F6- opening mouth	-0.01	0.20	0.16	**0.51**	-0.25	0.41
V4- complaining	F7- tightening lips	0.10	-0.02	0.12	0.10	**0.78**	0.04
V5- shouting	F14- looking sad	0.06	0.39	-0.06	-0.01	**0.65**	0.24
V6- mumbling'	F13- looking tense	0.17	0.23	0.23	0.30	**0.65**	0.12
V7- screaming	BM1- freezing	0.15	0.04	-0.01	-0.16	**0.63**	0.04
V8- groaning	BM3- clenching hands	0.42	0.17	0.22	0.21	0.42	0.07
V9- crying	F9- empty gaze	0.06	-0.03	0.14	0.04	0.11	**0.84**
V10- gasping	F10- seeming disinterested	0.12	-0.02	0.11	0.05	0.18	**0.83**
V11- sighing	F11- pale face	0.19	0.33	0.08	-0.08	0.09	**0.65**

Abbreviations: F = Facial expressions, in blue; BM = body movements, in green; V = vocalizations, in pink. # Items F4 closing eyes, F8 clenched teeth, BM10 pacing, VOC1 using offensive words, are excluded from the analysis.

Discussion and conclusions

Recognition of pain in persons with dementia might improve when observational scales are used in daily practice. This is the first study in a European setting to investigate the observer agreement of a large pool of behavioural pain items assembled in the PAIC scale (research version), derived from widely recognized observation scales. For items based on body movements and vocalizations, reliability was generally good. For a number of facial expression items though, agreement between observers was below 70%. This was the case for the items 'looking sad', 'tightened lips', 'empty gaze', 'seeming disinterested' and 'looking tense'. This was seen both in observations at rest and in movement. Poor agreement was especially found in The Netherlands, where the group of observers was large, and experience and education in use of observation scales was low. Facial responses are often quite subtle and fleeting and thus, observers might have had more difficulty noticing them during observation without extensive training. At the same time, it has to be considered that the face items proved to be especially valuable in grading the pain because they were almost free of floor effects, and a high variance of different categories were used to describe the behaviour. This favourable use of more categories for behavioural description by the observer, however, leads to a reduction of observer agreement.

There is strong evidence in the research literature that facial responses are valid for measuring pain and therefore these items are important in observational scales [13]. This suggests that training is probably necessary for the rating of items, especially in grading pain with use of several categories of severity. The need for training was also mentioned by healthcare professionals in a survey across Europe[14] and is planned for the short version of the PAIC scale. [6] For the details of PAIC15 and the associated e-training see: https://paic15.com/nl/start-nl/.

Factor analyses found that individual items could be grouped into six underlying components (Table 5). In the first component, 'vocal pain expression', the majority of vocal items were grouped together. The third group, 'protective body movements', contained many (four out of nine) of the body movement items. Then, we found a factor 'vocal defence', with two vocal items, one body movement, and one face item. The face items were grouped under three components, which we named 'face anatomical descriptors', 'tension', and 'lack of affect'. Lautenbacher et al [15] performed a factor analyses on face items only and found two quite similar components, that is, 'anatomical descriptors' and 'lack of affect', and we adopted the same names. The most important difference between that study and the present study was that the three face items grouping together in the component 'tension' fell in three different components: tightened lips fell in their component 'anatomical descriptors', looking sad into 'lack of affect' and looking tense into 'arousal'. Thus, these factors, which could not be replicated, may be unstable.

Zwakhalen et al. [16] looked at the factor structure of the 24-item PACSLAC-D and found three components. They suggest that some items are more universal pain cues for various target groups, such as facial expressions, while other items are more social-emotional cues, such as mood, aggression and agitation, which may be more specific for patients with dementia. From that perspective, our factors 1 ('vocal pain expression')

and 2 ('face anatomical descriptors') might reflect pain in general, and are the most specific expressions of pain. The body movements that we found in component 1 might also be more universal pain cues compared to body movement items in component 3 ('protective body movements'). These items might be directly or indirectly related to dementia, when the care situation or how people are approached induces protective behaviour. Furthermore, the component 'lack of affect' might also be more specific to dementia itself. This is in line with findings from interviews with health professionals in The Netherlands when studying construct validity [8]. Further validity studies are needed to resolve which items reflect pain in general, pain in dementia or other forms of distress in dementia.

A strength of this study is that it took place in four countries using four different European languages. In this way, it would reflect use of the scale in future daily care situations and patient populations across different cultures. Thus, the development of the PAIC has been a truly international project.

A limitation is that some countries had deviated slightly from the European protocol, with regards to the scheme and number of observations. For example, in The Netherlands two different pairs of observers were involved for each patient, and in Italy observations were only performed at rest and not all patients were observed simultaneously for interrater agreement. This makes comparison somewhat challenging. On the other hand, we planned in advance that the study should be performed in prevalent real-life healthcare conditions in participating countries. This is important, because assessment in daily practice is generally performed whilst providing nursing care [14].

Furthermore, we were most interested in aggregated data, not comparison of data between countries.

To maximize the number of observations to be analysed, we chose to recode missing scores to 0 for the analyses of interrater and intrarater agreement, as if behaviours were not shown. This might not be the case, and percentages present might thus be estimated too low. Another point is that for items that occur rarely, the level of agreement might give a false impression of good reliability. This is especially the case in the sensitivity analysis, where we dichotomized scores. We chose to perform the factor analyses on observations at rest, because we had less observations in movement and the rest condition was more standardized among countries. However, it is possible that different findings would emerge for the test items if we had done the analysis of the items during movement.

This study focussed on scoring and observer agreement of individual items. For intrarater agreement, observations on consecutive days were chosen rather than video recordings. As the observed construct (i.e., observed pain behaviour) is not stable, this might have negatively influenced observer agreement. The high agreement rates, which was achieved under these unfavourable conditions, show that it does not matter whether the patient is observed on one day or the next.

It should be noted that some observational scales score individual items (e.g., PACSLAC-II), some combine items in the response options (e.g., PAINAD), and some score overarching domains (e.g., Abbey Pain Scale and MOBID-2) with or without extensive listing of possible items. (Examples of the tools/forms can be found on internet, for PACLSAC-II, PAINAD, and MOBID-2 on URL: https://geriatricpain.org/assessment/pain-behavior-tool-critique/

list-nonverbal-pain-behavior-tools-2019 and for Abbey pain scale on URL: https://www.apsoc.org.au/PDF/Publications/Abbey_Pain_Scale.pdf (accessed August 6th 2019)). In the latter, pre-existing assumptions (without education) might play a large role in scoring and as such affect the reliability of the scale. Thus, for the PAIC we decided to score individual items. These differences make comparison of former results with the present study difficult. Lichtner et al. [3] reviewed the psychometric properties of observational pain scales, including their reliability. Scale sum-scores and not scores on individual items have been studied: overall, the majority of the assessed tools had moderate to good inter-rater reliability (but limitations in sample sizes) and moderate to good temporal stability.

What are the implications of this study? The EU-COST Action working group set out to study individual items for an observational scale, PAIC. This scale was designed as a meta-tool, systematically looking for and extracting the best items in existing observational scales for pain assessment in dementia [5]. This idea was recently echoed by a US-American research group following a similar line of methodological reasoning [17]. Together with results from other psychometric studies, results of the present study will be used in the item reduction process by means of a Delphi procedure, to form the final PAIC-scale [6]. This is also necessary for feasibility of the measurement scale in daily practice. Training, which has already been planned for the short version of the PAIC scale (PAIC15 [6]) should not only focus on the use of assessment tools but also on the interpretation of the results [14]. For this, further research on total scores will be necessary, for example, how can item scores best be summed and what are the implications of certain (changes in) scores. As individuals and professionals are challenged to understand their role in the dynamic interplay among biological, psychological, and social determinants of pain, training even might embrace this broader context [18]. Ultimately, training should focus on how to incorporate assessments into daily practice when use of observational scale is intended to improve pain management [2 19 20].

Acknowledgments

We want to thank al participating patients, their families, and healthcare professionals. In particular we want to thank our collaborators Martina Amanzio, Sara Invitto, Jorge Navarro, and Felice Sdanganelli (from Italy); Wieke Rijkmans (from The Netherlands); and Antoni Monllau and Manel Sánchez (from Spain). We also would like to thank all other members of working-group 3 of the EU-COST Action TD1005 on their preparatory work on the protocol.

Author contributions

Substantial contributions to conception and design: MdW, AHvD, RdV, TF, AL, BM, WPA. Acquisition of data: MdW, AHvD, LG-L, LK, MdT.
Analysis and interpretation of the data: MdW, AHvD, RdV, LG-L, LK, MdT, MK, SL, FL, BM, JS, WPA. MdW made the first draft, and all authors critially revised the manuscript. All authors approved the final version of the manuscript.

Statement funding sources

Members of the EU-COST action “Pain in impaired cognition, especially dementia” received EU-funding for collaborative activities such as meetings. AHvD was supported by the employer of elderly care medicine/general practitioner trainees (SBOH).

Conflict of interest disclosures

No conflict of interest was declared.

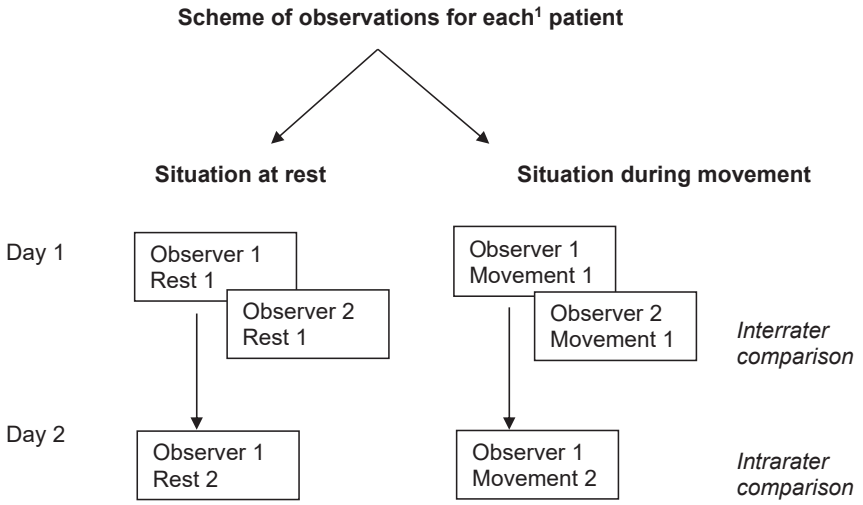
References

1. Corbett A, Husebo B, Malcangio M, et al. Assessment and treatment of pain in people with dementia. *Nature reviews Neurology* 2012;8(5):264-74. doi: 10.1038/nrneurol.2012.53 [published Online First: 2012/04/11]
2. Achterberg WP, Pieper MJ, van Dalen-Kok AH, et al. Pain management in patients with dementia. *Clin Interv Aging* 2013;8:1471-82. doi: 10.2147/CIA.S36739
3. Lichtner V, Dowding D, Esterhuizen P, et al. Pain assessment for people with dementia: a systematic review of systematic reviews of pain assessment tools. *BMC Geriatr* 2014;14:138. doi: 10.1186/1471-2318-14-138
4. de Vet, Terwee CB, Mokking LB, et al. Measurement in Medicine: a practical guide. . Cambridge University Press2011.
5. Corbett A, Achterberg W, Husebo B, et al. An international road map to improve pain assessment in people with impaired cognition: the development of the Pain Assessment in Impaired Cognition (PAIC) meta-tool. *BMC neurology* 2014;14:229. doi: 10.1186/s12883-014-0229-5 [published Online First: 2015/03/03]
6. Kunz M, De Waal MWM, Achterberg WP, et al. The Pain Assessment in Impaired Cognition scale (PAIC-15): a multidisciplinary and international approach to develop and test a meta-tool for pain assessment in impaired cognition, especially dementia. Submitted along with present article. *Eur J of Pain* 2019
7. Ohrbach R, Bjorner J, Jezewski M, et al. Guidelines for Establishing Cultural Equivalence of Instruments. Buffalo, University of Buffalo.2009.
8. van Dalen-Kok AH, Achterberg WP, Rijkmans WE, et al. Pain Assessment in Impaired Cognition (PAIC): content validity of the Dutch version of a new and universal tool to measure pain in dementia. *Clinical interventions in aging* 2018;13:25-34. doi: 10.2147/cia.s144651 [published Online First: 2018/01/11]
9. Reisberg B, Ferris SH, de Leon MJ, et al. The Global Deterioration Scale for assessment of primary degenerative dementia. *Am J Psychiatry* 1982;139(9):1136-9.
10. Van Dalen-Kok AH, Achterberg WP, Rijkmans WE, et al. Pain assessment in impaired cognition: observer agreement in a long-term care setting in patients with dementia. *Pain Management* 2019 doi: 10.2217/pmt-2019-0025 [published Online First: 12-08-2019]
11. de Vet HC, Mokkink LB, Terwee CB, et al. Clinicians are right not to like Cohen's kappa. *BMJ* 2013;346:f2125. doi: 10.1136/bmj.f2125
12. Field A. Discovering statistics using SPSS (thrid edition).2009.
13. Lautenbacher S, Kunz M. Facial Pain Expression in Dementia: A Review of the Experimental and Clinical Evidence. *Curr Alzheimer Res* 2017;14(5):501-05. doi: 10.2174/1567205013666160603010455 [published Online First: 2016/06/24]
14. Zwakhalen S, Docking RE, Gnass I, et al. Pain in older adults with dementia : A survey across Europe on current practices, use of assessment tools, guidelines and policies. *Schmerz* 2018 doi: 10.1007/s00482-018-0290-x [published Online First: 2018/06/23]
15. Lautenbacher S, Sampson EL, Pahl S, et al. Which Facial Descriptors Do Care Home Nurses Use to Infer Whether a Person with Dementia Is in Pain? *Pain medicine (Malden, Mass)* 2017;18(11):2105-15. doi: 10.1093/pm/pnw281 [published Online First: 2016/12/31]

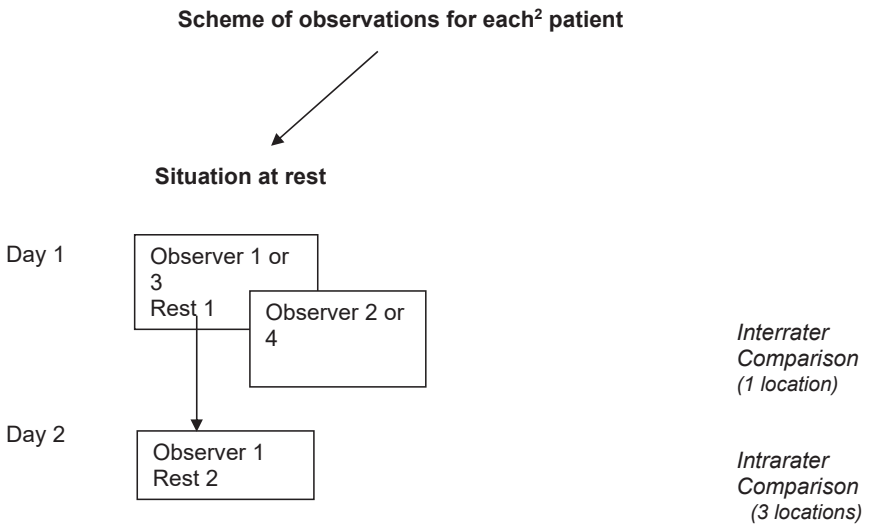
16. Zwakhalen SM, Hamers JP, Berger MP. Improving the clinical usefulness of a behavioural pain scale for older people with dementia. *Journal of advanced nursing* 2007;58(5):493-502. doi: 10.1111/j.1365-2648.2007.04255.x [published Online First: 2007/04/20]
17. Ersek M, Herr K, Hilgeman MM, et al. Developing a Pain Intensity Measure for Persons with Dementia: Initial Construction and Testing. *Pain medicine (Malden, Mass)* 2018 doi: 10.1093/pm/pny180 [published Online First: 2018/10/05]
18. Craig KD. Social communication model of pain. *Pain* 2015;156(7):1198-9. doi: 10.1097/j.pain.0000000000000185 [published Online First: 2015/06/19]
19. Hadjistavropoulos T, Herr K, Prkachin KM, et al. Pain assessment in elderly adults with dementia. *Lancet Neurol* 2014;13(12):1216-27. doi: 10.1016/S1474-4422(14)70103-6 [published Online First: 2014/12/03]
20. Pieper MJ, van der Steen JT, Francke AL, et al. Effects on pain of a stepwise multidisciplinary intervention (STA OP!) that targets pain and behavior in advanced dementia: A cluster randomized controlled trial. *Palliative medicine* 2018;32(3):682-92. doi: 10.1177/0269216316689237 [published Online First: 2017/02/02]

Appendix 1: scheme of observations (version A, B, and C).

Version A (Serbia, Spain)



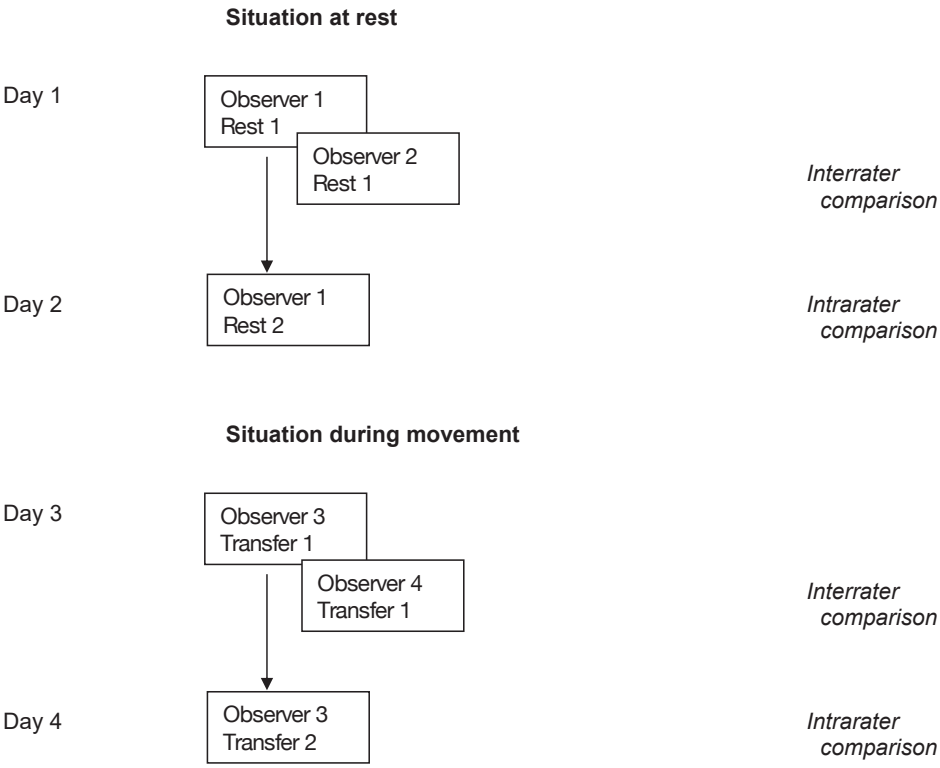
Version B (Italy)



[1] For 1 out of 40 patients in Serbia, observation on day 2 during movement was missing.
[2] Of 57 patients in Italy, 46 were observed twice by the same rater (intrarater comparison), and 39 patients were observed by two observers (interrater comparison).

Version C (The Netherlands)

Scheme of observations for each[3] patient



[3] For 5 out of 45 patients in The Netherlands observations were missing on day 2 and day 4.

Appendix 2. Scores per item (in %) in first observations in movement (n=133).

	Score:	0	1	2	3
	Missing	Not at all	Slight degree	Moderate degree	Great degree
Facial expressions					
Pained expression		68.4	16.5	12.8	2.3
Frowning		66.9	22.6	8.3	2.3
Narrowing eyes		79.7	14.3	4.5	1.5
Closing eyes		86.5	9.8	3.0	0.8
Raising upper lip		90.2	9.0	0.8	0.0
Opened mouth		78.9	13.5	6.0	1.5
Tightened lips		57.1	25.6	14.3	3.0
Clenched teeth		83.5	15.0	1.5	0.0
Empty gaze		54.9	24.1	15.8	5.3
Seeming disinterested		60.2	16.5	18.0	5.3
Pale face		57.1	24.1	15.0	3.8
Teary eyed		88.0	11.3	0.0	0.8
Looking tense		55.6	37.6	6.0	0.8
Looking sad		58.6	23.3	18.0	0.0
Looking frightened		66.2	25.6	6.8	0.5
Body movements					
Freezing		63.9	25.6	7.5	3.0
Curling up		82.7	11.3	5.3	0.8
Clenching hands		75.2	17.3	4.5	3.0
Resisting care		68.4	21.1	7.5	3.0
Pushing		92.5	4.5	1.5	1.5
Guarding	5.3	85.7	8.3	0.8	0.0
Rubbing		95.5	3.8	0.8	0.0
Limping	8.3	72.9	12.8	6.0	0.0
Restlessness		89.5	6.8	3.0	0.8
Pacing		94.7	4.5	0.8	0.0
Vocalizations					
Using offensive words		95.5	3.0	0.8	0.8
Using pain relates words		72.2	18.8	6.8	2.3
Repeating words		85.7	9.8	4.5	0.0
Complaining		79.7	13.5	5.3	1.5
Shouting		90.2	4.5	3.0	2.3
Mumbling		78.9	14.3	4.5	2.3
Screaming	0.8	89.5	6.0	1.5	2.3
Groaning		68.4	20.3	8.3	3.0
Crying		82.7	9.0	7.5	0.8
Gasping		83.5	14.3	2.3	0.0
Sighing		65.4	22.6	11.3	0.8

Appendix 3. Intrarater agreement, percentages for dichotomized scores.

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement after dichotomization	(n=46)	(n=40)	(n=39)	(n=48)	(n=48)	(n=40)	(n=40)	(n=174)	(n=127)
Facial expressions									
Pained expression	94	100	100	98	90	90	60	95	84
Frowning	96	100	100	75	79	70	55	85	78
Narrowing eyes	89	100	100	89	92	70	63	87	85
Closing eyes	91	98	97	81	96	63	75	83	90
Raising upper lip	94	100	100	98	85	88	85	95	90
Opened mouth	91	93	100	96	79	75	60	89	80
Tightened lips	94	90	97	81	85	78	65	86	83
Clenched teeth	94	93	100	93	90	85	73	92	87
Empty gaze	96	95	100	65	85	70	73	81	86
Seeming disinterested	96	100	100	79	100	78	73	88	91
Pale face	91	100	100	100	100	75	83	92	95
Teary eyed	89	85	90	96	98	98	85	92	91
Looking tense	96	93	100	75	83	75	70	85	84
Looking sad	91	90	92	88	79	83	70	88	80
Looking frightened	91	85	95	88	90	78	78	86	87
Body movements									
Freezing	100	95	97	79	75	83	75	89	82
Curling up	100	88	100	100	100	83	85	83	95
Clenching hands	94	95	100	69	92	88	75	86	89
Resisting care	91	90	100	100	98	98	83	95	94

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement after dichotomization	(n=46)	(n=40)	(n=39)	(n=48)	(n=48)	(n=40)	(n=40)	(n=174)	(n=127)
Pushing	91	98	100	100	200	98	85	97	95
Body movements									
Guarding #	89	98	100	96	98	80	80	91	93
Rubbing	96	100	100	100	100	83	90	95	97
Limping #	96	100	100	100	92	98	73	98	88
Restlessness	91	98	100	81	90	63	83	83	86
Pacing	96	95	97	98	96	95	93	96	95
Vocalizations									
Using offensive words	100	95	97	100	100	95	95	98	98
Using pain relates words	96	93	100	100	94	85	73	94	89
Repeating words	100	95	100	98	98	90	83	96	94
Complaining	89	90	100	98	96	85	70	91	89
Shouting	96	100	100	98	100	88	83	95	95
Mumbling	94	90	100	96	90	80	85	90	91
Screaming	100	93	97	100	98	95	93	97	96
Groaning	87	80	95	98	90	78	65	86	84
Crying	96	90	100	100	98	93	80	95	93
Gasping	89	83	97	92	92	85	85	87	91
Sighing	91	88	100	81	75	80	85	85	86

Note: % agreement for scores after dichotomization to not present (0) or present (1-3), for missings score=0, percentage ≥ 70% in blue.
missing pairs of observations for guarding in rest 4.0% and in transfer 5.5%; for limping in rest 6.3% and in transfer 8.7%

Appendix 4. Interrater agreement, percentages for dichotomized scores.

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement after dichotomization	(n=39)	(n=40)	(n=40)	(n=48)	(n=48)	(n=45)	(n=45)	(n=172)	(n=133)
Facial expressions									
Pained expression	90	98	100	96	83	84	69	92	84
Frowning	90	100	100	81	79	56	38	81	71
Narrowing eyes	97	98	100	81	90	69	69	86	86
Closing eyes	87	95	93	60	90	78	62	79	81
Raising upper lip	92	98	98	98	90	91	87	95	91
Opened mouth	74	93	100	94	88	73	60	84	82
Tightened lips	80	85	95	54	56	73	69	72	72
Clenched teeth	90	85	98	83	83	82	76	85	85
Empty gaze	87	95	100	71	81	67	53	79	77
Seeming disinterested	90	100	100	50	67	69	64	76	76
Pale face	80	100	100	83	100	67	73	82	87
Teary eyed	80	85	88	98	100	89	87	88	92
Looking tense	77	90	100	54	56	73	69	73	74
Looking sad	80	90	93	63	75	69	56	74	74
Looking frightened	92	80	95	83	92	89	76	86	87
Body movements									
Freezing	100	93	98	73	83	84	60	87	80
Curling up	100	80	98	98	100	89	71	92	90
Clenching hands	92	88	100	83	81	82	69	86	83
Resisting care	95	88	100	96	98	98	78	94	92

	Italy	Serbia		Spain		The Netherlands		Total	
	Rest	Rest	Movement	Rest	Movement	Rest	Movement	Rest	Movement
Interrater agreement after dichotomization	(n=39)	(n=40)	(n=40)	(n=48)	(n=48)	(n=45)	(n=45)	(n=172)	(n=133)
Pushing	95	95	95	100	100	100	89	98	95
Guarding #	95	98	100	96	98	80	84	92	94
Rubbing	90	100	100	100	98	82	91	93	96
Limping #	100	100	100	98	83	96	76	98	86
Restlessness	100	98	100	83	94	64	78	86	90
Pacing	92	98	98	98	92	98	96	97	95
Vocalizations									
Using offensive words	100	98	98	98	100	96	93	98	96
Using pain relates words	95	80	100	100	96	91	80	92	92
Repeating words	95	90	95	94	98	98	82	94	92
Complaining	95	90	98	85	90	87	73	89	87
Shouting	97	100	98	96	94	100	82	98	91
Mumbling	95	90	98	98	92	71	67	88	85
Screaming	97	93	100	98	98	96	87	96	95
Groaning	95	78	100	98	92	89	84	90	92
Crying	97	88	100	98	100	89	93	93	98
Gasping	100	70	95	90	88	89	84	87	89
Sighing	92	85	100	79	83	78	73	83	85

Note: % agreement for scores after dichotomization to not present (0) or present (1-3), percentage ≥ 70% in blue.
missing pairs of observations for guarding in rest 4.1% and in transfer 5.3%; for limping in rest 6.4% and in transfer 8.3%