



Universiteit  
Leiden  
The Netherlands

## **The cervical radiculopathy impact scale: development and evaluation of a new functional outcome measure for cervical radicular syndrome**

Gartner, F.R.; Marinus, J.; Hout, W.B. van den; Vleggeert-Lankamp, C.; Stiggelbout, A.M.

### **Citation**

Gartner, F. R., Marinus, J., Hout, W. B. van den, Vleggeert-Lankamp, C., & Stiggelbout, A. M. (2020). The cervical radiculopathy impact scale: development and evaluation of a new functional outcome measure for cervical radicular syndrome. *Disability And Rehabilitation*, 42(13), 1894-1905. doi:10.1080/09638288.2018.1534996

Version: Publisher's Version

License: [Creative Commons CC BY-NC-ND 4.0 license](#)






Downloaded from: <https://hdl.handle.net/1887/3181975>

**Note:** To cite this publication please use the final published version (if applicable).

ASSESSMENT PROCEDURE



## The Cervical Radiculopathy Impact Scale: development and evaluation of a new functional outcome measure for cervical radicular syndrome

Fania R. Gärtner<sup>a</sup> , Johan Marinus<sup>b</sup> , Wilbert B. van den Hout<sup>a</sup> , Carmen Vleggeert-Lankamp<sup>c</sup>  and Anne M. Stiggelbout<sup>a</sup> 

<sup>1</sup>Department of Medical Decision Making, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands; <sup>2</sup>Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands; <sup>3</sup>Department of Neurosurgery, Leiden University Medical Center, Leiden, The Netherlands

### ABSTRACT

**Objective:** To develop and validate an outcome scale for the cervical radicular syndrome and to build a mapping, predicting EQ-5D utility from the new scale.

**Study design and setting:** An item pool was developed based on literature and patient and clinician interviews. Item selection was based on symptomatology, factor analysis, and internal consistency. We assessed: (a) test–retest reliability by standard error of measurement and intraclass correlation coefficients; (b) construct validity by testing 22 hypotheses on relationships with existing measures and known-group differences. For the mapping, performance was assessed by mean absolute error and root mean squared error.

**Results:** A total of 254 patients with cervical radicular syndrome completed the first questionnaire, 61 stable patients a retest. Item selection led to a 21-item questionnaire consisting of three subscales: Symptoms, Energy and postures, and Actions and activities. Standard error of measurement values ranged from 6.7 to 11.2 on a 0 to 100 scale. All subscales showed good reliability (intraclass correlation coefficients: 0.84, 0.87, and 0.94). All hypotheses for construct validity were confirmed. A linear utility mapping was preferred, with reasonable statistical performance.

**Conclusion:** We developed a reliable and valid cervical radicular syndrome specific outcome scale, called the Cervical Radiculopathy Impact Scale (CRIS). This new questionnaire may facilitate (cost-)effectiveness studies in this field.

### ARTICLE HISTORY

Received 5 December 2017

Revised 5 October 2018

Accepted 8 October 2018

### KEYWORDS

Cervical radiculopathy; herniated disk; disability; pain; PROM; questionnaire



### ► IMPLICATIONS FOR REHABILITATION

- The cervical radicular syndrome is a frequently occurring and invalidating health problem, which causes severe radiating pain in the arm and/or hand, which can be accompanied by motor and/or sensory deficits.
- The Cervical Radiculopathy Impact Scale (CRIS) is a newly developed self-report questionnaire which covers measurement of symptoms and limitations in patients with cervical radiculopathy due to irradiating pain, tingling sensations and sensory loss in the arm in combination with neck disability.
- The CRIS consists of 21 items divided over three subscales: (i) symptoms, (ii) energy and postures, and (iii) actions and activities.
- The CRIS shows good content validity, test-retest reliability, construct validity and is able to discriminate between groups.
- The CRIS predicts EQ-5D utility and is therefore useful for (cost)effectiveness studies in this field.

## Introduction

The cervical radicular syndrome (CRS) is a frequently occurring and invalidating health problem. CRS causes severe radiating pain in the arm and/or hand, which can be accompanied by motor and/or sensory deficits. These symptoms often have an intensity that prohibits normal functioning [1,2]. The symptoms are usually caused by compression of the nerve root by a cervical herniated disc or by degenerative osteophytes. The annual age-adjusted incidence of CRS is 0.8 per 1,000 inhabitants, based on a study in Minnesota, USA; the prevalence estimated at 3.5 per 1,000 inhabitants, based on an Italian study [1,3].

Usually patients are treated by their family doctor with pain medication. In the majority of patients symptoms demonstrate spontaneous relief in the first weeks after onset of symptoms [4]. If the pain is persistent and disabling, patients can be referred to the neurologist, and subsequently the option for surgical decompression can be explored. A systematic review comparing surgery with conservative treatment strategies published in 2012 pointed out the lack of evidence on the effectiveness of both options. Only one low-quality randomized controlled trial was found, which provided no evidence for effectiveness in favor of one of these options [5]. Controversies persist regarding the best timing of diagnostic procedures, timing of referral of

**CONTACT** Fania R. Gärtner  [fr.gartner@lumc.nl](mailto:fr.gartner@lumc.nl)  Leiden University Medical Center, Department of Medical Decision Making, P.O. Box 9600, 2300RC Leiden, The Netherlands

 Supplemental data for this article can be accessed [here](#).

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Table 1. Characteristics of participants in expert-interviews and the expert check.

	Expert-interviews (N = 10)		Expert check (N = 18)	
	N	(%)	N	(%)
Gender (male)	8	(80)	16	(89)
Age (in years)				
(mean(SD))	48.4	(9.9)	49.1	(8.7)
min.–max.	34–62		34–64	
Occupation (multiple answers possible)				
Neurosurgeon	3	(30)	9	(50)
Neurologist	3	(30)	4	(22)
Physiotherapist	2	(20)	1	(6)
Chiropractor	–	–	1	(6)
Researchers	2	(20)	3	(17)
Movement scientist	1	(10)	1	(6)
Orthopedic surgeon	1	(10)	1	(6)
Physician for rehabilitation	2	(20)	2	(11)
Expertise in treating or studying CRS (scale: 0 = none to 4 = a lot) (mean(SD))	3	(1.12)	3.17	(0.857)

Note: Overlap between interview and expert-check sample N = 9.

patients to neurologists or neurosurgeons, and the timing of surgical interventions [4,5].

Because of the limited evidence and the growing interest in new intervention strategies, such as dynamic cervical disk prosthesis, CRS is high on today's research agenda. Example of trials are numerous, e.g., a previous randomized trial about three conservative treatments [6], an ongoing trial comparing three surgical strategies [7], and an ongoing trial on surgical versus nonsurgical treatment [8]. Such research is hindered by the absence of a validated patient-reported outcome scoring system for CRS [5]. A patient-reported outcome measure scale to measure burden of disease for CRS patients is lacking. Currently available measures do not provide a comprehensive view of the functional limitations due to CRS. The most commonly used outcome measure is the Visual Analog Scale (VAS) or Numerical Rating Scale (NRS) for arm pain. The VAS is limited in that it measures only pain and not disability, which moreover might be difficult to quantify for some patients and therefore result in missing values [9]. Yet, most important to patients might be functional limitations, for which no adequate scale is available. In absence of such a scale, the Neck Disability Index (NDI) is often used, which focuses on neck disability only, without including disability due to radiating pain [10]. The Disability of the Arm Shoulder and Hand Questionnaire (DASH) on the other hand, only focuses on disability due to arm, hand and shoulder pain and therefore misses other aspects of disability of CRS patients [11,12]. Therefore, both questionnaires are not optimal for use in research on CRS patients.

To improve the quality of studies on CRS we developed and validated the Cervical Radiculopathy Impact Scale (CRIS), which focuses on symptoms and functionality of the arm and neck. We also developed a mapping to predict EQ-5D utility from the CRIS, because of the high interest in cost-effectiveness studies on CRS treatment strategies [7,8]. This can be used to calculate quality-adjusted life years (QALYs) in economic evaluations.

## Materials and methods

### Development of the item pool

Relevant domains for the questionnaire were inventoried by a literature search, ten interviews with CRS patients, and ten interviews with experts in treating or researching CRS. The literature search was performed in the databases Pubmed and Embase. Search terms covering CRS were combined with terms on health

related outcomes such as "Questionnaire", "Quality of life", "Activities of Daily Living" or "Pain measurement", see the [Supplementary Material](#) for the search terms. Peer-reviews articles, written in English, German, Dutch or French were eligible for inclusion. The search resulted in 526 unique hits of which 208 were included. To inventory all possible domains of symptoms and limitations relevant for research on CRS patients we extracted the outcomes measured in the included articles and the (sub)-scales of the instruments used.

The topic list for the one-hour interviews with patients and experts included symptoms, functioning in daily life, emotions and cognitions, the burden and frequency of the domains discussed, and desired treatment aims. With the patients a "walk through the day" was held in which the patient described daily activities and how these were limited by their symptoms. For a description of the interview samples (see [Tables 1 and 2](#)). For the analyses, we transcribed the interviews verbatim and followed a purpose-driven approach, in which we applied open-coding to distinguish as many different symptoms and (aspects of) limitations as possible. In a process of re-reading and constant comparison, we refined the codes. Subsequently, we categorized all codes into domains covering related aspects [13]. Coding and categorization was done by one researcher and checked by a second (AS). Finally, we synthesized the domains deriving from the literature search and the interviews, and formulated items for all symptoms and limitations covered by the domains. In this step again we adhered to the principle of being as inclusive as possible [14]. Results of this approach led to a first pool with 47 items grouped under four domains: (1) physical symptoms other than pain (2) pain, (3) general limitations; (4) limitations in postures, actions and activities.

### Content validity and pilot test

In an online survey among 18 experts in treating or researching CRS ([Table 1](#)) the items and domains were checked for clarity, relevance and completeness in an expert check. Relevance was rated on a 3-point scale (insufficient, sufficient, good). The items that remained in the item pool were rated sufficient or good on their relevance by more than two-thirds of the experts, except for eight items, which were kept in the item pool because of their importance stressed in the interviews with patients. Three of these eight items remained in the final version of the CRIS: item 3 "Stiffness in your neck or shoulder" (relevance sufficient or good: 44%), item 18 "Opening a jar with a screw-top lid" (relevance sufficient or good: 65%), and item 21 "Finding a comfortable position

Table 2. The CRIS item pool.

Items in the original item pool		Item number in the definitive CRIS questionnaire
<b>Did you experience...</b>		
a)	A tingling or numb sensation or 'pins and needles' in your arm, hand or fingers	1.
b)	Loss of strength in your arm, hand or fingers	2.
c)	Stiffness in your neck or shoulder	3.
d)	How often did you experience pain in your neck?	4.
e)	What was the degree of the pain in your neck, as a whole?	5.
f)	What was the degree of the pain in your neck, when it was at its worst?	
g)	What was the degree of the pain in your neck when it was at its least?	
h)	How often did you experience pain in your shoulder?	6.
i)	What was the degree of the pain in your shoulder, as a whole?	7.
j)	What was the degree of the pain in your shoulder, when it was at its worst?	
k)	What was the degree of the pain in your shoulder, when it was at its least?	
l)	How often did you experience pain in your arm, hand or fingers?	8.
m)	What was the degree of the pain in your arm, hand or fingers, as a whole?	9.
n)	What was the degree of the pain in your arm, hand or fingers, when it was at its worst?	
o)	What was the degree of the pain in your arm, hand or fingers, when it was at its least?	
p)	How often did you experience headache?	
q)	What was the degree of pain due to movements (Please choose an option even if you avoided movement due to pain)	
<b>Because of my complaints...</b>		
r)	...all activities took longer, for example domestic tasks, personal care or work.	10.
s)	... everything I did used more energy.	11.
t)	... I had to lie down in bed or on the sofa more often.	12.
u)	... I had no interest in doing things.	
v)	... I avoided postures or activities out of fear that my symptoms would worsen.	13.
<b>Where you limited in...</b>		
w)	Looking over your shoulder	14.
x)	Keep your head upright or gaze upward	
y)	Flex your neck and gaze downward	
z)	Lifting up things, like children, a laundry-basket or heavy boxes	
aa)	Carrying things, like groceries or a shoulder-bag	
bb)	Walking longer distances	
cc)	Writing using a pen	15.
dd)	Working at a computer	16.
ee)	Cutting using a knife, e.g., vegetables or meat	17.
ff)	Opening a jar with a screw-top lid	18.
gg)	Holding a book or a newspaper	19.
hh)	Holding things in your hands without dropping them	20.
ii)	Finding a comfortable position when lying in bed	21.
jj)	Using means of transportations, like the car, scooter, bike or public transport.	
kk)	Self-care like taking a shower, washing hair, putting on cloth, or brushing teeth	
ll)	Domestic chores, like making dinner, doing laundry, cleaning, small chores or taking care of children or family	
mm)	Social activities with friends or family	
nn)	Your usual recreational activities, like sports or hobbies.	
oo)	Focus on the tasks you are performing	
pp)	Work (paid work, caregiving or voluntary work)	

Table 3. Characteristics of participants in patient-interviews and pilot test with patients.

	Patient-interviews (N=10)		Pilot test (N=9)	
	N	(%)	N	(%)
Gender (male)	5	(50)	5	(56)
Age (in years)				
(mean(SD))	53	(9.5)	50	(7.5)
min.–max.	34–65		41–63	
Time since diagnosis (in months) (mean(SD))	21	(9.8)	9	(8)
Currently experiencing symptoms	7	(70)	8	(89)
Underwent surgery	6	(60)	5	(56)

when lying in bed" (relevance sufficient or good: 65%) (see outer right column in Table 2).

In a pilot test, we held think-aloud interviews with nine patients with CRS (Table 3), to assess the clarity of the introduction, the items and response categories, as well as the applicability of the items [15]. Also, patients were asked if items were missing. Based on the results of the expert check and this pilot test, we reformulated items, merged two items, and deleted four. No extra items were added. This resulted in an item pool for the CRIS of 42 items for four domains: Physical symptoms other than

pain ( $n=3$ ), Pain ( $n=14$ ), General Limitations ( $n=5$ ), and Limitations in postures, actions and activities ( $n=20$ ) (see Table 3). The items all have a 5-point response scale, with labels varying by domain, e.g., for frequencies (1 = never to 5 = always) or level of limitations (1 = not limited to 5 = extremely limited).

### Procedure

We aimed to collect 250 patients with CRS, using the rule of thumb of 10 respondents per item (e.g., [16]), with the limitations

scale of 20 items being the largest scale. Recruitment took place from February 2013 until October 2015. Data was collected in three samples, two clinical samples for which data was collected in two ongoing clinical trials, and a population-based sample. In one trial (the CASINO trial: CervicAI radiculopathy: Surgical or Nonsurgical Treatment) [8], CRS patients suffering from symptoms for at least two months were either randomly assigned to surgery or prolonged conservative care with surgery if needed, or were assigned to a cohort group if the patient resisted to be randomized to a specific treatment arm. In the second trial (NECK trial: Netherlands Cervical Kinetics) [7], on the effectiveness and safety of cervical disk prosthesis, patients were randomly assigned to three types of surgical intervention. In both clinical samples, patients were enrolled via participating hospitals.

The third sample was population-based. Patients were recruited via ads in newspapers and on the hospital's internet and intranet site, as well as via general practitioners and physiotherapists. One part of this sample (the CROSS pencil-and-paper sample) included telephone screening for eligibility by a research nurse and paper questionnaires. The other part (the CROSS online sample) included eligibility screening and data collection via online questionnaires.

To be included in the study, patients had to have at least minimal symptoms. Therefore, to select patients, we measured neck pain, arm pain, and tingling in the arm, hand, or fingers using three VAS items ranging from 0 (no pain/tingling) to 100 (worst possible pain/tingling). For eligibility, patients had to score at least 1 on the sum of these three items. Also, patients had to be between 18 and 75 years old, had to have at least two months of symptoms, and had had to be diagnosed by a neurologists based on MRI-scan (this criterion was based on self-report for the CROSS online sample). Patients who were pregnant, or were not able to speak and read the Dutch language, were excluded from the study.

Data were collected by questionnaires handed to patients during hospital visits or sent via mail or e-mail. Not all measurement instruments required for our study were included in the patient questionnaires of all three samples. Therefore, sample sizes vary in our analyses and are noted separately in the tables.

Test-retest data were collected in two of the three samples. In the CASINO sample, test and retest data were collected at 6 and 8 weeks after baseline, respectively. In the population-based sample the retest was filled out 10-17 days after the first measurement.

## Data analysis

### Item selection

Different item selection strategies were applied for the pain and symptoms items ( $N=17$ ) versus the limitations items ( $N=25$ ). For the former, the underlying model of the relationship between construct and items, is formative, i.e., the items together form the construct, and thus the construct is a result of the items. For the latter, the limitations items, the model is reflective, i.e., the construct is reflected by the items, since the construct causes them [17]. For this set of items dimensionality and internal consistence are relevant. Therefore, for item selection a classical test theory approach was followed. For formative scales it is characteristic that items are independent from each other [17]. Therefore, the internal consistency and factor structure are not leading for the development and less relevant for the reliability and validity of this subscale. For item selection symptomatology was leading. First, based on any comments in the open response fields, items were nominated for deletion. Second, items were screened on response distribution, and any items with 95% or more or with

0.5% or less of the responses in one response category were deleted [18]. Third, for pain in three localizations (neck, shoulder, and arm, hand or fingers) three items about the degree of pain had been included in the item pool: i) as a whole, ii) when pain was at its worst, iii) when pain was at its least. Of these three items, the item was selected that had the highest item-total correlation within the set of three items for that localization.

For the limitation items, steps taken were, first, screening on response distribution: any items with 95% or more in one response category were deleted [18]. Second, an explorative factor analysis with an orthogonal rotation approach, using principal component analysis and Varimax Rotation was performed for the reduction of items and the determination of the underlying factors [19,20]. We determined the optimum number of factors by Cattell's screetest [21] and the Kaiser's criterion (retain factors with Eigenvalue  $>1$ ) [22]. Based on the rotated factor loadings, items were assigned to the factor they had the highest loading on. Items with high loadings on both factors (double factor loadings) were dropped if the difference in factor loadings was less than 0.10 or if the item had factor loadings above 0.50 on more than one factor (implying that they could not clearly be assigned to one of the two factors). Also, items were dropped that hindered a clear interpretation of the factor meaning (as judged by two raters).

Subsequently, the number of items per factor was further reduced to a minimum number of items that still guaranteed Cronbach's alpha above 0.90, to fulfill the requirement to be used at the individual level [17]. Items with inter-item correlation coefficients above 0.7 were deleted first, then items that had the lowest item-total correlation but for which deletion would guarantee a sufficiently high Cronbach's alpha were deleted. Final choices about item deletion were based on consensus discussion within the research team.

We transformed total sum scores on the subscales to range from 0 (no pain or symptoms) to 100 (highest possible pain and symptoms). For calculating sum scores, we required that more than 50% of the subscale items had to be filled out.

### Test-retest reliability

For the test-retest reliability, we analyzed two properties for each of the final three CRIS subscales: level of agreement and test-retest reliability. For level of agreement we assessed the absolute measurement error, by calculating the standard error of measurement (SEM). The SEM equals the square root of the error variance of an ANOVA analysis, including systematic differences:  $SEM = \sqrt{(\sigma^2_{time} + \sigma^2_{error})}$  [23].

The test-retest reliability gives an indication of how well subjects can be distinguished from each other despite measurement errors. We computed intraclass correlation coefficients (ICCs) using data of two measurement points. To determine the ICC a two-way random effects model was used, the ICC(A0.1) according to McGraw and Wong [24]. The ICC calculation method in which systematic differences are considered to be part of the measurement error was used, called the ICC<sub>absolute agreement</sub>. The formula used was:  $ICC = \sigma^2_p / (\sigma^2_p + \sigma^2_{time} + \sigma^2_{error})$  [23]. For the single measure ICC values, we expected a minimum of 0.70 as sufficient and 0.80 as good [14].

An assumption in reliability analysis is that the sample used is stable regarding the studied concept [14]. We expected our sample to be stable during the 10-17 days interval of the retest. Additionally, to control for stability, we asked the subjects in the follow-up questionnaire three questions about any change in pain, symptoms, and limitations since they filled out the previous questionnaire. Subjects who answered "no" on all three items were regarded as stable subjects. We based our conclusions about reliability of the CRIS on the results of the stable sample.



Box 1. Hypotheses for construct validity of the three CRIS subscales.

Symptoms subscale		Confirmed
1.	Positive correlation with the VAS arm pain $\geq 0.50$	Yes
2.	Positive correlation with the VAS tingling in arm hand or fingers $\geq 0.50$	Yes
3.	Positive correlation with the VAS neck pain is $\geq 0.50$	Yes
4.	Negative correlation with the SF36 subscale bodily pain $\geq 0.50$	Yes
5.	Lower subscale score for working subjects that are not sick-leave than working subjects that are on sick-leave.	Yes
6.	Higher subscale score for subjects in the clinical samples before operation compared to subjects in the population-based sample.	Yes
<b>Sum</b>	<b>Minimally 75% of the six hypotheses need to be confirmed.</b>	<b>100%</b>
Energy and postures subscale		
1.	Positive correlation with the NDI $\geq 0.50$	Yes
2.	Positive correlation with the QuickDASH $\geq 0.50$	Yes
3.	Negative correlation with the SF36 subscale vitality $\geq 0.30$	Yes
4.	Negative correlation with the SF36 subscale physical role functioning $\geq 0.30$	Yes
5.	Absolute correlation with the SF36 subscale vitality is smaller than with the NDI and QuickDASH	Yes
6.	Absolute correlation with the SF36 subscale physical role functioning is smaller than with the NDI and QuickDASH	Yes
7.	Lower subscale score for working subjects that are not sick-leave than working subjects that are on sick-leave.	Yes
8.	Lower subscale score for subjects with a low illness perception score (experience illness as less threatening) compared to subjects with a high illness perception score (experience illness as less threatening).	Yes
9.	Higher subscale score for subjects in the clinical sample before operation compared to subjects in the population-based sample.	Yes
<b>Sum</b>	<b>Minimally 75% of the nine hypotheses need to be confirmed.</b>	<b>100%</b>
Actions and activities subscale		
1.	Positive correlation with the NDI $\geq 0.50$	Yes
2.	Positive correlation with the QuickDASH $\geq 0.50$	Yes
3.	Negative correlation with the SF36 subscale physical functioning $\geq 0.30$	Yes
4.	Absolute correlation with the SF36 subscale physical functioning is smaller than with the NDI and QuickDASH	Yes
5.	Lower subscale score for working subjects that are not sick-leave than working subjects that are on sick-leave.	Yes
6.	Lower subscale score for subjects with a low illness perception score (experience illness as less threatening) compared to subjects with a high illness perception score (experience illness as less threatening).	Yes
7.	Higher subscale score for subjects in the clinical sample before operation compared to subjects in the population-based sample.	Yes
<b>Sum</b>	<b>Minimally 75% of the seven hypotheses need to be confirmed.</b>	<b>100%</b>

### Construct validity

For the construct validity we tested six to nine hypotheses per subscale regarding the relatedness of the CRIS subscales with measures for similar constructs and regarding known-group differences (see Box 1).

For the subscale Symptoms, the VAS pain scores for arm pain, for numbness and tingling in the arm, hand of fingers and for neck pain were used. For the two limitations subscales, two disease-specific disability measures were used, the NDI and the shortened version of the Disability of Arm shoulder and Hand questionnaire (QuickDASH). Also, the subscales physical functioning, physical role functioning, and vitality of the generic Short Form 36 Health Survey (SF36), were used. To test these hypotheses correlation coefficients (Pearson's correlation calculation for normally distributed scales, Spearman's correlation calculation for non-normally distributed scales) were calculated between the CRIS subscale and the comparison measures.

For known-groups differences testing, we hypothesized higher scores on all three CRIS subscales for working subjects that were on sick-leave compared to working subjects that were not [25]. Also, for all three subscales we hypothesized higher scores for subjects from the two clinical samples that had not undergone surgery yet, compared to subjects from the population-based sample. Furthermore, for the two limitation subscales we hypothesized that CRIS subscale scores would be higher for subjects with a high, more threatening, illness perception compared to subjects with a low, less threatening, illness perception, measured by the Brief Illness perception Questionnaire [25]. For confirmation of these hypotheses, differences not only had to be statistically significant, but also exceed the SEM value for the relevant subscale, to confirm that the difference was not based on measurement error. We tested for differences between groups

using *t*-test symptoms (normally distributed subscales) or Mann-Whitney *U* tests (not normally distributed scales).

For good construct validity of the different subscales, we formulated the requirement that at least 75% of the hypotheses had to be confirmed [17].

### Prediction of EQ-5D utilities

A mapping formula was estimated to predict EQ-5D valuations (Dutch and UK tariffs) from the three CRIS subscale scores. In addition to the data used to develop the CRIS, longitudinal measurements were used from the patients in the two clinical trial samples. The data contained no missing items on the CRIS scales or the EQ-5D items. Several model specifications were analyzed: linear, quadratic with interactions, piecewise linear, and power functions. Models were estimated using Generalized Estimation Equations (with exchangeable covariance matrix), to account for the repeated measurements. Statistical performance of the models was compared by the prediction error, as measured by the mean absolute error (MAE) and root mean squared error (RMSE). External validity was assessed using cross-validation, by consecutively excluding one of the three subsamples (CASINO, NECK and CROSS), re-estimating the mapping, and assessing the average prediction error on the excluded external subsamples.

### Instruments

#### NDI

NDI, measures self-rated disability in patients with neck pain, where disability is understood as perceived effect of pain and impairment on the patient's performance and enjoyment of activities of daily living [10,26]. The NDI has been validated in Dutch

[27]. The ten items have a 6-point response scale ranging from 0 (no disability) to 5 (total disability). For all respondents who completed all 10 items a sum score was calculated, ranging from 0 to 50 points, with higher scores indicating a higher degree of disability.

#### QuickDASH

The DASH is a 30 item self-report questionnaire measuring patients' perceptions of disabilities and symptoms associated with any condition affecting the upper limb [11]. The full-length version of the DASH has been validated in Dutch [12]. We used a shortened 11 item version, with a five point response scale with varying category labels [28]. A sum score was calculated for all respondents who filled out at least 10 of the 11 items. The sum score was transformed into a score range from 0 (no disability) to 100 (most severe disability).

#### VAS for arm pain, tingling sensations in the arm and neck pain

We measured arm pain, tingling sensations in the arm, and neck pain in the past week, each with one item on a VAS, 100 mm line.

Scores ranged from 0 (no pain/tingling) to 100 (worst possible pain/tingling).

#### Brief illness perception questionnaire (brief IPQ)

Illness perception was measured by the Dutch version of the Brief Illness perception Questionnaire [29,30]. The eight items have an 11-point response scale ranging from 0 to 10 with varying category labels. For all respondents who filled out all eight items a sum score was calculated with a transformed score range from 0 (least threatening view of the illness) to 100 (most threatening view of the illness). For the analyses of known-group differences we dichotomized the IPQ scores based on a median split, scores of 53.75 or lower were recoded into the low score and scores higher than 53.75 into the high scores.

#### SF-36 subscales

Generic Health related Quality of life was measured by the Dutch version of the 36-item Short Form Health Survey (SF-36) [31–33]. Of its eight dimensions we used three subscales: the ten-item physical functioning with 3-point response scale, the

**Table 4.** Participant characteristics for the total sample and the stable sample for the 2-week test–retest.

	First measure sample			Test–retest sample		
	Total N	N	(%)	Total N	N	(%)
<b>Gender: female</b>	224	127	(57)	50	30	(60)
<b>Age in years (mean (SD))</b>	238	52	(10.4)	54	56	(9.3)
<b>Marital status: living together</b>	249	203	(82)	60	51	(85)
<b>Having children: yes</b>	249	211	(85)	60	53	(88)
<b>Highest level of education</b>	198			60		
No education/ primary school/ secondary school with entrance to vocational education/ lower vocational education		30	(15)		13	(22)
Vocational education or entrance level for vocational education		89	(45)		22	(37)
Secondary school with entrance level for higher education		16	(8)		6	(10)
Higher level education		58	(29)		17	(28)
Other		5	(3)		2	(3)
<b>Employment status</b>	248			60		
Housekeeping		26	(11)		6	(10)
Full-time (self)employed		99	(40)		18	(30)
Part-time (self)employed		76	(31)		16	(27)
Unpaid volunteer work		3	(1)		1	(2)
Not working (unemployed, retired, long-term sick-leave)		44	(18)		19	(32)
<b>Sick-leave</b>	234			59		
Yes		54	(23)		9	(15)
No		123	(53)		30	(51)
Not applicable		57	(24)		20	(34)
<b>Earlier treatment</b>						
Pain medication	247	138	(44)	60	29	(48)
Pain treatment at outpatient clinic	195	29	(15)	59	14	(24)
Physiotherapy	246	173	(70)	59	42	(71)
Manual therapy	246	58	(24)	59	11	(19)
Chiro practice	246	23	(9)	59	5	(9)
Neck brace	243	40	(16)	59	16	(27)
Acupuncture	195	12	(6)	59	4	(7)
Other	246	37	(15)	59	12	(20)
None	214	10	(4)	59	4	(7)
Surgery	245	52	(21)	60	7	(12)
<b>Number of months since first GP visit for CRS symptoms (mean(SD))</b>	192	38	(67.2)	60	57	(77.3)
<b>Clinical diagnoses (more than one categories are possible)</b>	137			16		
C4		0	(0)		0	(0)
C5		8	(6)		1	(6)
C6		72	(53)		8	(50)
C7		63	(46)		9	(56)
C8		14	(10)		1	(6)
Pyramid		1	(1)		0	(0)
<b>Lateralization HNP</b>	104			10		
Left		51	(49)		5	(40)
Right		44	(42)		4	(50)
Both		9	(9)		1	(10)

four-item physical role functioning subscale with dichotomous response scale, and the four-item vitality subscale with 6-point response scale. Subscale scores range from 0 to 100 with higher scores indicating better health related quality of life. They were calculated if at least 50% of items were completed and a missing item was replaced by the mean subscale score per case.

#### EQ-5d

The three-level EQ-5D is a measure that provides health state descriptions. It consists of five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension is described by three levels of severity: no problems (1), some problems (2), and extreme problems (3). For each EQ-5D filled out by the patients, we calculated Dutch and UK utility values [34,35]. Utility represents the valuation by the general public (of the Netherlands and UK) of the quality of life described by the patients, on a scale anchored at 0 (as bad as death) and 1 (perfect health).

## Results

### Sample

In total, 254 patients completed the questionnaire and were eligible for item selection and construct validity analyses, 142 in the two clinical samples (CASINO:  $N=91$ ; NECK:  $N=51$ ), 112 in the population-based sample (CROSS online:  $N=33$ ; Cross paper  $N=79$ ). For the test-retest analyses we had follow-up data of 125 respondents who completed the second questionnaire within ten to 17 days after the first questionnaire. See Table 4 for details about these two samples. The mapping was estimated in the same sample of patients, but with 241 and 81 additional follow-up measurements for the patients in the CASINO and the NECK trial, respectively.

### Item selection

#### Selection of symptoms items

Based on comments in the open response field, item q) was deleted, as multiple respondents found this item confusing or did

Table 5. Item-total correlation of three items on degree of pain for three different localizations.

	Neck pain ( $N=249$ )	Shoulder pain ( $N=254$ )	Pain in arm hand or fingers ( $N=253$ )
Overall pain	0.688	0.921	0.939
When pain was at its worst	0.666	0.846	0.891
When pain was at its least	0.612	0.795	0.858

Table 6. Descriptive results for the three subscales.

Subscales	$N$	No. of items	Cronbach's alpha	Mean (SD)	Median (range)
CRIS Symptoms	246	9	0.927	48.3 (23.5)	50 (0–94.4)
CRIS Energy and postures	253	6	0.909	47.2 (28.5)	50 (0–100)
CRIS Actions and activities	253	6	0.902	29.2 (23.9)	25 (0–100)

Values are based on transformed scores ranging from 0 to 100.

Table 7. Factor loadings of the 24 CRIS limitation items on two factors after varimax rotation, ( $N=248$ ).

Items about functional limitations		Factor loading		Reason for item deletion
		Factor 1	Factor 2	
u)	I had no desire to do things.	<b>0.827</b>	0.323	
w)	Looking over your shoulder.*	<b>0.765</b>	0.246	
v)	I avoided postures or activities out of fear that my symptoms would worsen.*	<b>0.750</b>	0.289	
s)	Everything I did used more energy.*	<b>0.749</b>	0.430	
t)	I had to lie down in bed or on the sofa more often.*	<b>0.744</b>	0.289	
r)	All activities took longer, for example domestic tasks, personal care or work.*	<b>0.742</b>	0.417	
x)	Keep your head upright or gaze upward.	<b>0.740</b>	0.216	
z)	Lifting up things, like children, a laundry-basket or heavy boxes	0.683	0.541	Double loading
y)	Flex your neck and gaze downward	<b>0.668</b>	0.381	
aa)	Carrying things, like groceries or a shoulder-bag	0.661	0.576	Double loading
ii)	Finding a comfortable position when lying in bed.*	<b>0.631</b>	0.415	
nn)	Your usual recreational activities, like sports or hobbies.	0.622	0.523	Double loading
oo)	Keeping your attention at the tasks you are doing.	0.578	0.475	Interpretability
jj)	Using means of transportations, like the car, scooter, bike or public transport.	0.556	0.537	Double loading
bb)	Walking longer distances	0.555	0.539	Double loading
ee)	Cutting using a knife, e.g., vegetables or meat.*	0.268	<b>0.842</b>	
gg)	Holding a book or a newspaper.*	0.288	<b>0.779</b>	
ff)	Opening a jar with a screw-top lid.*	0.378	<b>0.775</b>	
cc)	Writing using a pen.*	0.186	<b>0.730</b>	
dd)	Working at a computer.*	0.403	<b>0.672</b>	
ll)	Domestic chores, like making dinner, doing laundry, cleaning, small chores or taking care of children or family	0.566	0.654	Double loading
hh)	Holding things in your hands without dropping them.*	0.344	<b>0.637</b>	
kk)	Self-care like taking a shower, washing hair, putting on cloth, or brushing teeth.	0.440	0.627	Interpretability
mm)	Social activities with friends or family.	0.559	0.624	Double loading
Eigenvalue		14.46	1.39	
Total variance explained		60.24%	5.77%	

\*Item included in final version of the CRIS questionnaire. Bold printed factor loadings indicate on the factor the item is assigned to. Eigenvalues refer to the total variance explained by each factor. Percentage of variance explained by the factor presents the values before varimax rotation.



**Table 8.** Test–retest descriptive data and ICC values for the three subscales.

Subscale	First measure (T1)			Second measure (T2)			Change scores (T1–T2)			ICC 95% CI		
	Mean (SD)	Median (range)	N	Mean (SD)	Median (range)	N	Mean (SD)	N	SEM	ICC	Lower bound	Upper bound
<b>Stable sample</b>												
CRIS Symptoms	47.59 (23.83)	50.00 (0–97.22)	61	46.73 (23.83)	47.22 (0–94.44)	61	0.86 (12.40)	61	8.7	0.867	0.786	0.917
CRIS energy and postures	47.32 (28.01)	45.83 (0–100)	61	44.88 (28.26)	41.67 (0–100)	61	2.45 (15.78)	61	11.2	0.842	0.750	0.902
CRIS Actions and activities	30.67 (25.76)	20.83 (0–95.83)	61	31.49 (28.84)	20.83 (0–95.83)	61	–0.82 (9.47)	61	6.7	0.941	0.903	0.964
<b>Total sample</b>												
CRIS Symptoms	45.15 (22.15)	44.44 (0–97.22)	125	43.26 (23.27)	41.67 (0–94.44)	125	1.89 (12.42)	125	8.8	0.849	0.791	0.891
CRIS energy and postures	45.86 (26.08)	45.00 (0–100)	125	42.63 (26.91)	41.67 (0–100)	125	3.23 (15.50)	125	11.2	0.824	0.756	0.874
CRIS Actions and activities	27.50 (24.26)	20.83 (0–100)	125	26.87 (26.41)	16.67 (0–95.83)	125	0.63 (11.95)	125	8.4	0.889	0.846	0.921

**Table 9.** Correlation coefficients of the three subscales and the comparison measures.

	CRIS pain and symptoms		CRIS energy and postures		CRIS actions and activities	
	Correlation coefficient	N	Correlation coefficient	N	Correlation coefficient	N
CRIS symptoms						
CRIS energy and postures	0.80	254				
CRIS actions and activities	0.73	254	0.76	254		
NDI	0.72	199	<b>0.82</b>	199	<b>0.73</b>	199
QuickDASH	0.76	207	<b>0.84</b>	207	<b>0.81</b>	207
SF36 subscale bodily pain	<b>–0.81</b>	254	–0.85	254	–0.72	254
SF36 subscale vitality	–0.53	254	<b>–0.63</b>	254	–0.54	254
SF36 subscale physical role functioning	–0.62	254	<b>–0.77</b>	254	–0.66	254
SF36 subscale physical Functioning	–0.63	254	–0.72	254	<b>–0.71</b>	254
VAS arm pain	<b>0.80</b>	254	0.69	254	0.64	254
VAS tingling arm hand or fingers	<b>0.65</b>	203	0.54	203	0.52	202
VAS neck pain	<b>0.75</b>	254	0.72	254	0.55	253

All correlation coefficients are based on Spearman's correlation calculation, except for the pain and symptoms scale with the NDI, QuickDASH and SF36 physical component scale, which are based on Pearson's correlations.

Bold printed correlation coefficients are used for hypotheses testing.

not understand it. Based on response distribution, all other items were kept. Concerning the degree of pain in three localizations (neck, shoulder, and arm/hand/fingers) the first item on overall degree of pain in the last week was kept for all three localizations since it had the highest item-total correlation (see Table 5). We decided to delete item p) about headache, because of its low item-total correlations (0.52), which confirmed the doubts about the relevance of this item expressed in the expert check.

The final subscale Symptoms comprises nine items (a, b, c, d, e, h, i, l, and m) out of the original 17 items. Transformed scale scores on this scale ranging from 0 (no pain or symptoms)–100 (highest possible pain and symptoms) followed a normal distribution. For descriptive results of the final 9-item subscale (see Table 6).

#### Selection of limitation items

Based on the response distribution, item oo was deleted as it was not applicable to 21.3% of respondents. Thus, the principal component analysis was performed with 24 items. Based on the Kaiser criterion and the interpretability of factors a two factor structure was chosen with a total of variance explained of 66.0%. Nine items were assigned to the first factor, called "Energy and postures", six items to factor two called "Actions and activities" and nine items were dropped based on double factor loadings and hinder of interpretability (see Table 7).

The nine items initially assigned to factor 1 had a Cronbach's alpha of 0.947 ( $N = 250$ ). Based on inter-item correlations, item-total correlations, and Cronbach's alpha three items were dropped, respectively, u, x, and y. The final subscale Energy and postures consisted of six items (item r, s, t, v, w, and ii) and had a Cronbach's alpha of 0.909 ( $N = 253$ ). For factor 2, the six items included (items: cc, dd, ee, ff, gg, hh) had a Cronbach's alpha of 0.902 ( $N = 253$ ). All six items were kept in the final subscale

Actions and activities. The CRIS thus consists of 21 items in three subscales. Very few items had missing data. The maximum number of missing items for the three subscales were two for symptoms (9 items), one for energy and postures (6 items), and one for actions and activities (6 items). Based on the transformed total scores ranging from 0 (no limitations)–100 (highest possible limitations), neither subscale showed a normal distribution.

The final questionnaire in Dutch and English is presented as [Supplementary Material](#).

#### Reliability testing

Of the 125 respondents who filled out two questionnaires within a period of 10–17 days 61 reported that no changes had occurred in the past 2 weeks in their pain, symptoms, or limitation. These respondents together form the "stable sample" which was used for the test–retest analyses. SEM scores in the stable sample range from 6.7 to 11.7. ICC scores for all three subscales exceeded 0.8, which was the requirement for good test–retest reliability (see Table 8).

#### Construct validity

Based on the correlation coefficients (Table 9) and testing of known-group differences, in combination with SEM value comparison (Table 10), all hypotheses were confirmed, indicating good construct validity of all three subscales.

#### Prediction of EQ-5D utilities

The linear prediction mapping was selected (Table 11) as it parsimonious and is user-friendly, with statistical performance very similar to the more complex models: the average prediction error

**Table 10.** Testing for differences in the subscale scores for sick-leave and illness perception, based on the unpaired *t*-test for the symptoms subscale and the Mann–Whitney *U* test for the energy and postures subscale and the actions and activities subscale.

	<i>N</i>	Mean (SD)	Median (range)	mean rank score	Mean difference	<i>p</i> -value	Confidence interval
<b>CRIS symptoms</b>							
<b>Sick-leave</b>	177	49.1 (23.1)			–10.7	0.004	(–18.0) – (–3.4)
No	123	45.8 (23.5)					
Yes	54	56.6 (20.2)					
<b>Clinical versus population-based sample</b>	206	53.3 (21.3)			–15.4	<0.001	(–20.91) – (–9.9)
Population-based	112	46.2 (20.9)					
Clinical	94	61.6 (18.7)					
<b>CRIS energy and postures</b>							
<b>sick-leave</b>	177		54.2 (0–100)			<0.001	
No	123		41.7 (0–100)	76.6			
yes	54		66.7 (4.2–95.8)	117.2			
<b>Illness perception</b>	195		58.3 (0–100)			<0.001	
Low (less threatening)	102		39.6 (0–100)	74.3			
High (more threatening)	93		66.7 (12.5–100)	124.0			
<b>Clinical versus population-based sample</b>	206		58.3 (0–100)			<0.001	
Population-based	112		45.4 (0–100)	85.59			
Clinical	94		66.7 (4.2–100)	124.84			
<b>CRIS actions and activities</b>							
<b>sick-leave</b>	177		25 (0–100)			<0.001	
No	123		16.7 (0–100)	78.2			
Yes	54		41.7 (0–83.3)	113.6			
<b>Illness perception</b>	195		29.2 (0–100)			<0.001	
Low (less threatening)	102		16.7 (0–100)	72.9			
High (more threatening)	93		45.0 (0–87.5)	125.5			
<b>Clinical versus population-based sample</b>	206		29.17 (0–100)			0.012	
Population-based	112		20.8 (0–100)	93.69			
Clinical	94		37.5 (0–87.5)	114.87			

**Table 11.** Linear prediction model for the Dutch and UK EQ-5D utility values (*N* = 576).

Linear model	Dutch EQ-5D			UK EQ-5D		
	$\beta^*$	SE	<i>p</i> -value	$\beta^*$	SE	<i>p</i> -value
Constant	0.991	0.018	0.000	0.977	0.020	0.000
CRIS symptoms	–0.385	0.072	0.000	–0.434	0.080	0.000
CRIS energy and postures	–0.274	0.067	0.000	–0.298	0.075	0.000
CRIS actions and activities	–0.209	0.083	0.012	–0.253	0.091	0.006

\*For the CRIS subscales transformed to range from 0 to 1 (instead of 0 to 100). For example, for CRIS subscale scores 42, 33 and 0, respectively, the mapped Dutch utility value is  $0.991 - 0.385 \times 0.42 - 0.274 \times 0.33 - 0.209 \times 0.00 = 0.739$ .

**Table 12.** Statistical performance of the EQ-5D prediction models.

	Dutch EQ-5D				UK EQ-5D			
	<i>N</i>	ME	MAE	RMSE	<i>N</i>	ME	MAE	RMSE
Prediction error	576		0.141	0.198	576		0.151	0.218
Among EQ-5D $\geq 0.75$	325	–0.063	0.096	0.115	291	–0.065	0.109	0.128
Among EQ-5D 0.5 to 0.75	131	–0.050	0.118	0.140	156	–0.071	0.137	0.163
Among EQ-5D < 0.5	120	0.284	0.286	0.195	129	0.313	0.313	0.204
External prediction error	576		0.145	0.205	576		0.164	0.226
Range for alternative nonlinear models								
Best alternative	576		0.136	0.194	576		0.149	0.214
Worst alternative	576		0.142	0.198	576		0.152	0.220

ME, mean signed error; MAE, mean absolute error; RMSE, root means squared error.

differed by at most 3% from the best alternative nonlinear model (Table 12). The estimated coefficients were all statistically significant, with reasonable model prediction error and explained variance (Dutch  $R^2=0.53$ ; UK  $R^2=0.55$ ). External validity was good: the prediction error in external samples increased by at most 4%. The range of the predicted utilities was quite wide (Dutch range 0.123 to 0.991; UK range –0.008 to 0.977), but with underestimation for higher utilities and overestimation for lower utilities.

## Discussion

To develop a CRS-specific questionnaire, an item pool consisting of 42 items was developed based on literature and interviews

with patients and experts. The item selection process led to the final questionnaire consisting of three subscales with in total 21 items. Subscale 1) Symptoms includes nine items covering pain in the neck, shoulder, and arm/hand/fingers, as well as items on, tingling, loss of strength, and stiffness in the neck. Subscales 2) Energy and postures (6 items) and 3) Actions and activities (6 items) cover items on functional limitations due to pain and symptoms. The results of the principal component analysis were of good quality, as the percentage of explained variance was over 50% and Cronbach's alpha scores exceeded 0.7 [14]. For all subscales, the requirement for good test–retest reliability was fulfilled, with ICC scores exceeding 0.80, as well as the requirements for good construct validity, as all 22 hypotheses (100%) were

confirmed. Also, all hypotheses about known-group differences were confirmed with differences between the predefined groups exceeding the SEM, we therefore can conclude that the CRIS is well able to discriminate between groups. A linear utility mapping with reasonable statistical performance was constructed that is easy to use in economic evaluations.

Some aspects that are typically included in quality of life measures did not survive the item selection phase of the limitations items. This concerns items covering emotional aspects (see Table 3, item u, aspects of self-care (see Table 3, items kk, ll) and social activities (see Table 3, items mm, nn) [32–34]. The main reason was the exclusion of items with double factor loadings, which applied to three of these five items. Another reason was interpretability.

Results of the two reliability measures might seem contradictory. ICC values indicated good reliability, but SEM values were rather large especially for the subscale Energy and postures (SEM = 11.2 on a scale range of 0–100). This phenomenon is explained by the great variation between patients (SDs vary between 23.5 and 28.5 for the three subscales), such that even with measurement error reliability can be high. This highlights that reliability is a characteristic of an instrument used in a population, not just of an instrument [17]. It might be valuable to assess the reliability and the SEM values in more detail for specific patient populations, for example newly diagnosed patients, patients that have recently been referred to secondary care, or patients that have persistent symptoms after initial treatment.

Concerning the EQ-5D mapping, the errors of the estimated mapping are somewhat larger compared to other studies that predicted utility values from disease-specific outcome measures. This may be related to the relatively low utility values in our population, resulting in larger ranges with larger errors for the lower utilities. Also, the CRIS is a disease-specific questionnaire that does not explicitly capture all domains of generic quality of life. Specifically, the subscales of the CRIS correlated better with the self-care, usual activities, and pain/discomfort domains of the EQ-5D (correlations >0.5) than with the mobility and the anxiety/depression domains (correlations <0.5). Therefore, the estimated utility mapping can be used in cost-effectiveness studies, but is only a second-best alternative to generic utility measurement [36].

Strengths of our study were first, the inclusion of various disciplines that are involved in CRS care during the development process of the CRIS. Second, the formulation of specific hypotheses for the three different CRIS subscales and the inclusion of multiple comparison measures. Third, we have ensured that our sample consists of patients with varying severity of symptoms and illness durations; the large variation in CRIS scores within our sample, makes the results of reliability and validity testing valid for the whole scale range. Fourth, to facilitate the use of the CRIS outside the Netherlands, we have translated the CRIS into English according to COSMIN standards including multiple forward backward translations and consensus discussion, see [Supplementary Material \[37,38\]](#). We expect the CRIS items to be to a large extent context and culture independent, but a careful cultural adaptation and validation is recommended before using the translated English version of the CRIS.

Our study also had some limitations. First, we have used the same dataset for the item selection and for the reliability and validity testing. Thus, respondents have filled out the items of the CRIS only when they were part of the item pool with 42 items, not as part of the final questionnaire with 21 items. Although this is not uncommon in the field of clinimetric studies, it is not ideal. Therefore, further validation of the CRIS is recommended. Second,

due to the inclusion of the population-based sample we could not confirm the diagnosis by clinical examination or MRI for all patients included in this study, but partly had to base diagnosis on self-report. Our approach was motivated by our wish to have a heterogeneous population-based sample, and not just the patients from our two trials on herniated disc. We may have included some patients who did not have a clinically or MRI-confirmed herniated disc, but did have CRS symptoms. The other side of the coin is that the instrument can be used for all patients with symptoms of CRS. Given that we selected as items the symptoms of CRS, performed extensive testing, and found good clinimetric properties, we believe the CRIS meets this goal. The final scores that we found should however not be used as reference values for patients with a clinically confirmed diagnosis.

With its good test-retest reliability, construct validity and discriminative ability, the CRIS is suitable for use in research on symptoms and limitations of CRS patients. With the estimated EQ5D utility mapping, the disease-specific CRIS can also be used to estimate QALYs in economic evaluations [36]. The CRIS might be suitable for the use in clinical practice as well. The role patient-reported outcomes play in improving patient care is increasingly accepted [39]. Individual patient's scores on the CRIS might provide a starting point for the clinician and patient to discuss the momentary symptoms and limitations and any improvements or deteriorations to subsequently adapt patient care to these changes. Also, the availability of three subscales makes it possible to distinguish symptoms and pain scores from limitations in daily life, which facilitates purposive interventions. With the Cronbach's alpha of the three subscales exceeding 0.90 (posthoc analysis showed a Cronbach's alpha of 0.93 ( $N = 246$ ) for subscale 1 Complaints), the CRIS fulfills the requirement to be used at the individual level [17].

To use the CRIS as an outcome measure in intervention studies, change scores need to be able to be interpreted well. For this purpose, we have calculated the SEMs for the separate subscales [40–42]. Additionally, the minimal important change score for improvement is crucial for interpreting change scores [43–45]. The minimal important change score is a score on the scale range of the instrument that indicates the lowest change score that is considered clinically relevant. For calculating the MIC for the CRIS it is necessary to have a patient sample that experiences improvement in their symptoms and limitations in functioning [44,46]. Since we did not have such a sample, we could not estimate minimal important change scores. However, in an ongoing trial, data on the experience of change as well as the CRIS items are collected to determine the minimal important change score in the future. Furthermore, future evaluation should include the assessment of responsiveness, which can be seen as validity of change scores.

## Acknowledgements

The authors thank the Spine Intervention Prognostic Study Group (SIPS) Leiden—the Hague and S. van Geest (Department of Neurosurgery, Leiden University Medical Center, Leiden, the Netherlands) for their contribution to this study during the data acquisition.

## Disclosure statement

The authors report no conflicts of interest.

## Ethical approval

The Medical Ethics Committee of the Leiden University Medical Centre gave approval for this study.

## Funding


This work was supported by the Netherlands Organization for Health Research and Development (ZonMw) [grant number 171202017].

## ORCID

Fania R. Gärtner  <http://orcid.org/0000-0002-0351-0204>

Johan Marinus  <http://orcid.org/0000-0002-3978-3183>

Wilbert B. van den Hout  <http://orcid.org/0000-0002-6425-0135>

Carmen Vleggeert-Lankamp  <http://orcid.org/0000-0001-9597-7225>

Anne M. Stiggelbout  <http://orcid.org/0000-0002-6293-4509>

## References

- [1] Radhakrishnan K, Litchy WJ, O'Fallon WM, et al. Epidemiology of cervical radiculopathy: a population-based study from Rochester, Minnesota, 1976 through 1990. *Brain*. 1994;117:325–335.
- [2] Yoss RE, Corbin KB, Maccarty CS, et al. Significance of symptoms and signs in localization of involved root in cervical disk protrusion. *Neurology*. 1957;7:673–683.
- [3] Salemi G, Savettieri G, Meneghini F, et al. Prevalence of cervical spondylotic radiculopathy: a door-to-door survey in a Sicilian municipality. *Acta Neurol Scand*. 1996;93:184–188.
- [4] Carrette S, Fehlings MG. Clinical practice: cervical radiculopathy. *N Engl J Med*. 2005;353:392–399.
- [5] Gebremariam L, Koes BW, Peul WC, et al. Evaluation of treatment effectiveness for the herniated cervical disc: a systematic review. *Spine (Phila Pa 1976)*. 2012;37:E109–E118.
- [6] Kuijper B, Tans JT, Beelen A, et al. Cervical collar or physiotherapy versus wait and see policy for recent onset cervical radiculopathy: randomised trial. *BMJ*. 2009;339:b3883.
- [7] Arts MP, Brand R, van den Akker E, et al. The NETHERLANDS Cervical Kinematics (NECK) trial. Cost-effectiveness of anterior cervical discectomy with or without interbody fusion and arthroplasty in the treatment of cervical disc herniation: a double-blind randomised multicenter study. *BMC Musculoskelet Disord*. 2010;11:122.
- [8] van Geest S, Kuijper B, Oterdoom M, et al. CASINO: surgical or nonsurgical treatment for cervical radiculopathy, a randomised controlled trial. *BMC Musculoskelet Disord*. 2014;15:129.
- [9] Loos MJ, Houterman S, Scheltinga MR, et al. Evaluating postherniorrhaphy groin pain: visual Analogue or Verbal Rating Scale? *Hernia*. 2008;12:147–151.
- [10] Vernon H. The Neck Disability Index: state-of-the-art, 1991–2008. *J Manipulative Physiol Ther*. 2008;31:491–502.
- [11] Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med*. 1996;29:602–608.
- [12] Veehof MM, Slegers EJ, van Veldhoven NH, et al. Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). *J Hand Ther*. 2002;15:347–354.
- [13] Pope C, Ziebland S, Mays N. Qualitative research in health care. Analysing qualitative data. *BMJ*. 2000;320:114–116.
- [14] Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60:34–42.
- [15] Willis GB. Cognitive interviewing: a tool for improving questionnaire design. *Cognitive interviewing in practice: think-aloud, verbal probing and other techniques*. Thousand Oaks, CA: Sage Publications; 2005. p. 42–63.
- [16] Boateng GO, Neilands TB, Frongillo EA, et al. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front Public Health*. 2018;6:149.
- [17] de Vet HC, Terwee CB, Mokkink LB, et al. *Measurement in Medicine*. 1st ed. Cambridge (UK): Cambridge University Press; 2011.
- [18] Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 4th ed. Oxford: Oxford University Press; 2008.
- [19] Tabachnick BG, Fidell LS. *Principal components and factor analysis*, 4th ed. Vol.13. Boston (MA): Allyn and Bacon; 2001.
- [20] Stevens JP. *Exploratory and confirmatory factor analysis: applied multivariate statistics for the social sciences*. 4th ed. Mahwah (NJ): Lawrence Erlbaum; 2002. p. 385–469.
- [21] Cattell RB. Citation classic—the screen test for the number of factors. *Cc/Soc Behav Sci*. 1983;5:16.
- [22] Kaiser HF. The application of electronic computers to factor analysis. *Educ Psychol Measure*. 1960;20:141–151.
- [23] de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59:1033–1039.
- [24] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1:30–46.
- [25] Wibault J, oberg B, Dederding A, et al. Individual factors associated with neck disability in patients with cervical radiculopathy scheduled for surgery: a study on physical impairments, psychosocial factors, and life style habits. *Eur Spine J*. 2014;23:599–605.
- [26] Ailliet L, Knol DL, Rubinstein SM, et al. Definition of the construct to be measured is a prerequisite for the assessment of validity. The Neck Disability Index as an example. *J Clin Epidemiol*. 2013;66:775–782.
- [27] Hoving JL, O'Leary EF, Niere KR, et al. Validity of the neck disability index, Northwick Park neck pain questionnaire, and problem elicitation technique for measuring disability associated with whiplash-associated disorders. *Pain*. 2003;102:273–281.
- [28] Gummesson C, Ward MM, Atroshi I. The shortened disabilities of the arm, shoulder and hand questionnaire (QuickDASH): validity and reliability based on responses within the full-length DASH. *BMC Musculoskelet Disord*. 2006;7:44.
- [29] Broadbent E, Petrie KJ, Main J, et al. The brief illness perception questionnaire. *J Psychosom Res*. 2006;60:631–637.
- [30] de Raaij EJ, Schroder C, Maissan FJ, et al. Cross-cultural adaptation and measurement properties of the Brief Illness

- Perception Questionnaire-Dutch Language Version. *Man Ther.* 2012;17:330–335.
- [31] Aaronson NK, Muller M, Cohen PD, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *J Clin Epidemiol.* 1998;51:1055–1068.
- [32] Brazier JE, Harper R, Jones NM, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ.* 1992;305:160–164.
- [33] Ware JE. Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care.* 1992;30:473–483.
- [34] Dolan P. Modeling valuations for EuroQol health states. *Med Care.* 1997;35:1095–1108.
- [35] Lamers LM, McDonnell J, Stalmeier PF, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. *Health Econ.* 2006;15:1121–1132.
- [36] Dakin H. Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. *Health Quality Life Outcomes.* 2013;11:151.
- [37] Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality Life Res.* 2012;21:651–657.
- [38] Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine.* 2000;25:3186–3191.
- [39] Snyder CF, Aaronson NK, Choucair AK, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Quality Life Res.* 2012;21:1305–1314.
- [40] Beckerman H, Roebroeck ME, Lankhorst GJ, et al. Smallest real difference, a link between reproducibility and responsiveness. *Quality Life Res.* 2001;10:571–578.
- [41] de Vet HC, Terwee CB. The minimal detectable change should not replace the minimal important difference. *J Clin Epidemiol.* 2010;63:804–805.
- [42] de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Quality Life Outcomes.* 2006;4:54.
- [43] Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol.* 2008;61:102–109.
- [44] Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol.* 2010;63:524–534.
- [45] Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumat.* 2002;14:109–114.
- [46] de Vet HC, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol.* 2010;63:37–45.