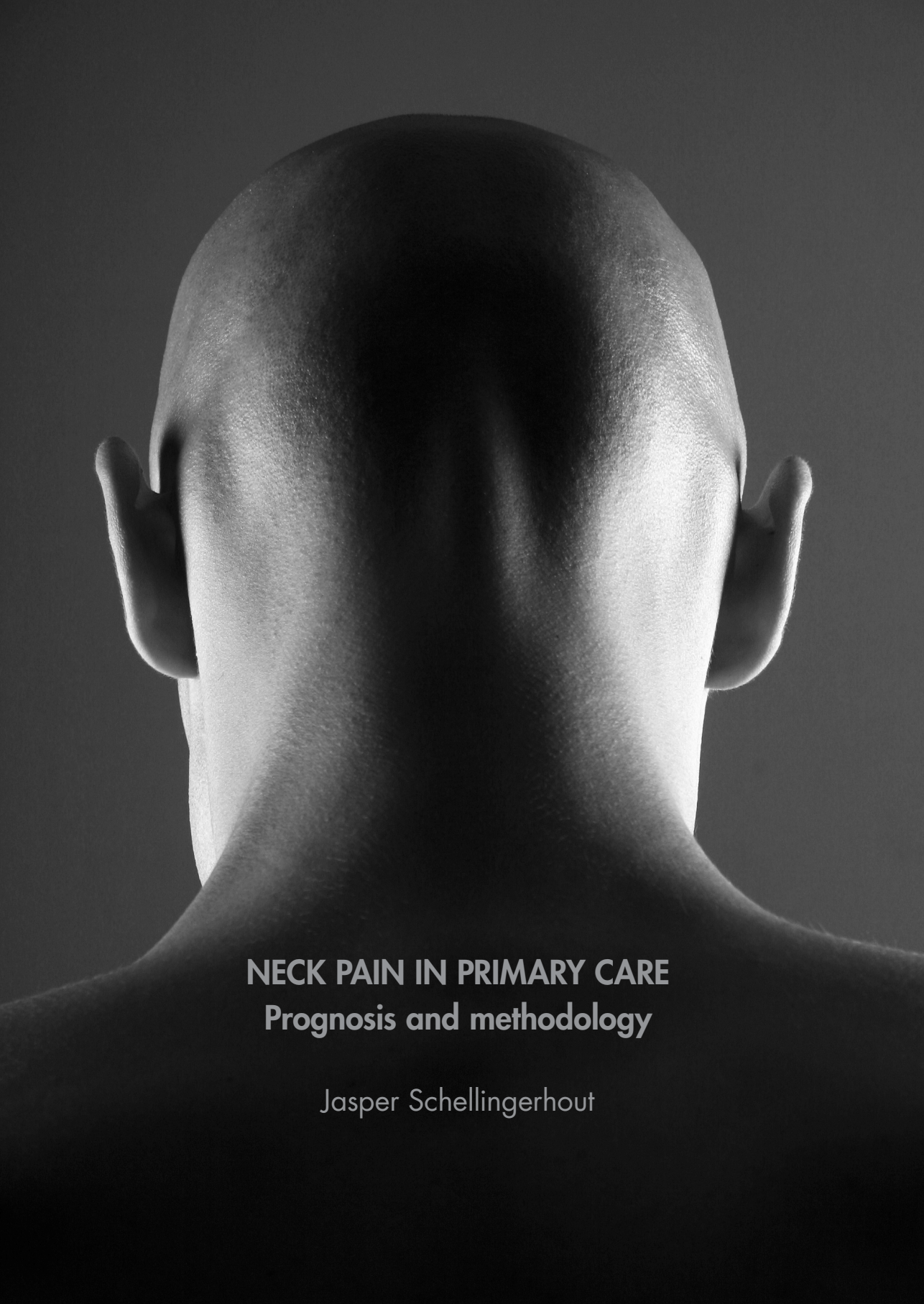

NECK PAIN IN PRIMARY CARE
Prognosis and methodology

Jasper Schellingerhout

Neck Pain in Primary Care: Prognosis and methodology

Jasper Schellingerhout

SBOH, werkgever van artsen in opleiding

Erasmus Universiteit Rotterdam

Erasmus MC, afdeling Huisartsgeneeskunde

ISBN: 978-94-6169-122-4
Layout and printing: Optima Grafische Communicatie, Rotterdam, The Netherlands

Neck Pain in Primary Care: Prognosis and methodology

Nekpijn in de eerste lijn: prognose en methodologie

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam

op gezag van de rector magnificus
Prof.Dr. H.G. Schmidt
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 12 oktober 2011 om 15.30 uur

door

Jasper Mattijs Schellingerhout
geboren te Dordrecht

PROMOTIECOMMISSIE

Promotoren:	Prof.dr. B.W. Koes
	Prof.dr.ir. H.C.W. de Vet
Overige leden:	Prof.dr. H.J. Stam
	Prof.dr. E.W. Steyerberg
	Prof.dr. D.A.W.M. van der Windt
Copromotoren:	Dr. A.P. Verhagen
	Dr. M.W. Heymans

CONTENTS

Chapter 1

General introduction

EPIDEMIOLOGY

Neck pain is one of the most common musculoskeletal complaints in adults in the general population, with an estimated point prevalence of 20.6-22.2% and a 1-year prevalence of 31.4-35.6%.[1-5] About half of those who experience neck pain consult their general practitioner (GP) for this complaint.[1] In daily practice this means that a GP is consulted about five times a week for an episode of neck pain.[6] The risk of developing neck pain is highest in women and in people of middle age (40-49 years).[7] The etiology of neck pain is largely unknown, but the involvement of genetic factors has been suggested.[7]

DIAGNOSIS

When patients report that they have neck pain it means that they experience pain with or without stiffness in the region between the back of the head and the shoulders. Most of the episodes of neck pain are of unknown origin, usually referred to as non-specific neck pain.[8] However, GPs have to be alert for specific causes of the neck pain like infection, fracture, rheumatic disease, or malignancy.[9]

Therefore, patient history and physical examination should be aimed at evaluating possible signs ("red flags") of a specific cause like a preceding trauma, unexplained weight loss, neurologic signs of spinal compromise, signs of infection, a previous history of cancer or neck surgery.[9] In case there is no sign of an underlying specific disorder, it is not useful to perform a physical examination of the neck, because the reproducibility and predictive value of physical tests are poor to moderate.[9-10] The same accounts for additional testing like blood tests or diagnostic imaging: if used in patients with non-specific neck pain they do not add useful information for further management of the patient.[9-10]

PROGNOSIS

Although usually harmless in origin, neck pain can be a real burden. The course of neck pain is characterised by exacerbations and remissions and only a small part of patients experience complete resolution of their symptoms within one year.[5] A proportion of neck pain patients will develop chronic neck pain. The definition of chronicity differs between studies, in terms of either 3 or 6 months duration of complaints. Nevertheless, estimates of their 1-year prevalence in the general population are similar: 8.7 - 17.8% for the 3 months definition,[1-2 4] and 8 - 13.8% for the 6 months definition.[2 11]

Several studies have been carried out to identify characteristics associated with persistence of neck complaints in the general population.[5 12-16] Noteworthy is that none of the predictors (e.g. age, gender, duration of complaints, pain intensity) identified in these studies consistently had a (large) impact on prognosis.[5 12-16] One of the possible explanations for this is that continuous candidate variables were frequently split into two or more categories in these studies. It has previously been described that categorization of candidate variables has some serious statistical drawbacks in (logistic) regression models.[17-20] The estimated magnitude of the association of a variable with the outcome in the regression model might be under- or overestimated.[18-20] These statistical problems might result in erroneously identifying a characteristic as a predictor for persistence of neck pain and vice versa. Therefore, we will evaluate the influence of various categorization strategies for candidate variables on the prognostic value of characteristics associated with persistence of non-specific neck pain.

Furthermore, the aforementioned studies only looked at the separate association of predictors with the outcome, which makes it hard to use the information derived from these studies in daily practice.[5 13-16] A prediction rule that quantifies the persistence of complaints by using a combination of predictors would contain more information and makes it possible to translate the findings for more direct use in daily practice. It would be helpful for a physician to gain insight into the prognosis of an individual patient with neck pain, and would aid in informing patients more accurately about their expected prognosis. Furthermore, it would aid researchers in selecting patients at high risk in studies on prevention of chronic neck pain. Such a prediction rule to quantify prognosis has been developed for shoulder pain and low back pain in the past, but is not yet available for non-specific neck pain.[21-22] Therefore, we decided to develop a prediction rule that estimates the probability of neck complaints persisting for at least 6 months.

Not only will we develop such a prediction rule, but we will also externally validate our prediction rule using the data of another longitudinal study on neck pain. External validation is of utmost importance, because a model that accurately predicts persistence of complaints in the development population may not do so in a different group of patients (e.g. other point in time, or region of origin).

TREATMENT

Management of non-specific neck pain by GPs, in the sense of treatment, consists of usual care in about 40% of the patients (i.e. advice on self-care combined with medication (NSAIDs, muscle relaxation medication)) and of referral for physiotherapy or spinal manipulation therapy in about 50% of the patients.[23] Systematic reviews

show that there is a positive effect of physiotherapy and spinal manipulation therapy compared to placebo or usual care in patients with non-specific neck pain, but these overall effects are relatively small.[24-25] There are decision-making algorithms for people with neck pain suggesting specific within-treatment variations in patients with neck pain referred for physical therapy.[26-27] However, these algorithms are consensus based and lack validation, which makes it uncertain if they are beneficial for patients with neck pain. This is probably the reason why available clinical practice guidelines for the management of neck pain do not give more specific treatment suggestions than recommendations derived from systematic reviews.[28-30] For clinical practice this means that GPs still do not know which treatment is optimal for a patient with non-specific neck pain.

Since the group of non-specific neck pain patients is heterogeneous, it seems reasonable that they will not all respond to treatment in the same way. So, instead of the usual approach of evaluating effectiveness of interventions in the overall population of patients with non-specific neck pain, it might be useful to look at the (differences in) effectiveness of physiotherapy, spinal manipulation therapy, and usual care in subgroups of patients. This has not been done before and could be of great benefit for clinical practice. Therefore, we decided to develop a decision model, based on patient characteristics, that points out which subgroups of patients with non-specific neck pain are more likely to benefit from either physiotherapy, spinal manipulation therapy, or usual care.

EVALUATION

Questionnaires are most frequently used for evaluative purposes in non-specific neck pain patients. Beside the generic instruments, like the Numerical Rating Scale (NRS) and the Visual Analogue Scale (VAS), several disease-specific questionnaires have been developed to measure pain and disability in patients with neck pain (e.g. Neck Disability Index (NDI), and Neck Pain and Disability Scale (NPDS)) . To determine the usefulness of these disease-specific questionnaires, it is essential to know the quality of their measurement properties (e.g. reliability, validity, and responsiveness).[31-33] A systematic review, published in 2002, showed that almost all disease-specific questionnaires were lacking psychometric information.[34] However, it is likely that new information is available, since neck pain research has increased in the last couple of years, amongst other things stimulated by the World Health Organisation's (WHO) initiative "the Bone and Joint Decade 2000-2010". Furthermore, recently the "COnsensus-based Standards for the selection of health status Measurement INstruments" (COSMIN) checklist, an instrument to evaluate the methodological quality of studies

on measurement properties of health status questionnaires, has become available.[35] Using the COSMIN checklist it is now possible to critically appraise and compare the quality of these studies.

A renewed systematic review could present an updated critical assessment of available disease-specific questionnaires for evaluation of non-specific neck pain. Therefore, we decided to critically appraise and compare the measurement properties of neck-specific questionnaires.

Previous systematic reviews on neck-specific questionnaires combined the results of studies on measurement properties, regardless of the language version of the questionnaire.[9 34 36-37] This may lead to inconsistent results for measurement properties, as was demonstrated in a recent review of the cross-cultural adaptations of the McGill Pain Questionnaire.[38] Therefore, we will evaluate the measurement properties of the original and translated versions of neck-specific questionnaires in separate systematic reviews.

AIMS

Based on the lacking knowledge and insights described above, the aims of this thesis are:

- To evaluate the influence of various categorization strategies of candidate variables on the final model content and performance when developing a multivariable logistic regression model. (Chapter 2)
- To develop and externally validate a prediction rule that estimates the probability of persistent complaints in non-specific neck pain patients. (Chapter 3)
- To develop a decision model that points out which subgroups of patients with non-specific neck pain are more likely to benefit from either physiotherapy, spinal manipulation therapy, or usual care. (Chapter 4)
- To systematically review the measurement properties of neck-specific questionnaires. (Chapter 5 and 6)

REFERENCES

1. Picavet HS, Schouten JS. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. Pain 2003;102:167-78.
2. Bovim G, Schrader H, Sand T. Neck pain in the general population. Spine 1994;19:1307-9.
3. Palmer KT, Walker-Bone K, Griffin MJ, Syddall H, Pannett B, Coggon D, et al. Prevalence and occupational associations of neck pain in the British population. Scand J Work Environ Health 2001;27:49-56.
4. Luime JJ, Koes BW, Miedem HS, Verhaar JA, Burdorf A. High incidence and recurrence of shoulder and neck pain in nursing home employees was demonstrated during a 2-year follow-up. J Clin Epidemiol 2005;58:407-13.
5. Cote P, Cassidy JD, Carroll LJ, Kristman V. The annual incidence and course of neck pain in the general population: a population-based cohort study. Pain 2004;112:267-73.
6. Bot SD, van der Waal JM, Terwee CB, van der Windt DA, Schellevis FG, Bouter LM, et al. Incidence and prevalence of complaints of the neck and upper extremity in general practice. Ann Rheum Dis 2005;64:118-23.
7. Hogg-Johnson S, van der Velde G, Carroll LJ, Holm LW, Cassidy JD, Guzman J, et al. The burden and determinants of neck pain in the general population: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine 2008;33:S39-51.
8. Bogduk N. Regional musculoskeletal pain. The neck. Baillieres Best Pract Res Clin Rheumatol 1999;13:261-85.
9. Nordin M, Carragee EJ, Hogg-Johnson S, Weiner SS, Hurwitz EL, Peloso PM, et al. Assessment of neck pain and its associated disorders: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine 2008;33:S101-22.
10. Rubinstein SM, van Tulder M. A best-evidence review of diagnostic procedures for neck and low-back pain. Best Pract Res Clin Rheumatol 2008;22:471-82.
11. Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. European Journal of Pain 2006;10:287-333.
12. Pool JJ, Ostelo RW, Knol D, Bouter LM, de Vet HC. Are psychological factors prognostic indicators of outcome in patients with sub-acute neck pain? Manual Therapy 2010;15:111-6.
13. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain 1998;77:1-13.
14. Hill J, Lewis M, Papageorgiou AC, Dziedzic K, Croft P. Predicting persistent neck pain: a 1-year follow-up of a population cohort. Spine 2004;29:1648-54.
15. Kjellman G, Skargren E, Oberg B. Prognostic factors for perceived pain and function at one-year follow-up in primary care patients with neck pain. Disabil Rehabil 2002;24:364-70.
16. Hoving JL, de Vet HC, Twisk JW, Deville WL, van der Windt D, Koes BW, et al. Prognostic factors for neck pain in general practice. Pain 2004;110:639-45.
17. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127-41.
18. Becher H. The concept of residual confounding in regression models and some applications. Stat Med 1992;11:1747-58.
19. Peter C. Austin LJB. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. Stat Med 2004;23:1159-78.
20. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. Epidemiology 1997;8:429-34.

21. Kuijpers T, van der Windt DA, Boeke AJ, Twisk JW, Vergouwe Y, Bouter LM, et al. Clinical prediction rules for the prognosis of shoulder pain in general practice. Pain 2006;120:276-85.
22. Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Can rate of recovery be predicted in patients with acute low back pain? Development of a clinical prediction rule. Eur J Pain 2009;13:51-5.
23. Vos C, Verhagen A, Passchier J, Koes B. Management of acute neck pain in general practice: a prospective study. Br J Gen Pract 2007;57:23-8.
24. Gross AR, Hoving JL, Haines TA, Goldsmith CH, Kay T, Aker P, et al. Manipulation and mobilisation for mechanical neck disorders. Cochrane Database Syst Rev 2004:CD004249.
25. Kay TM, Gross A, Goldsmith C, Santaguida PL, Hoving J, Bronfort G. Exercises for mechanical neck disorders. Cochrane Database Syst Rev 2005:CD004250.
26. Childs JD, Fritz JM, Piva SR, Whitman JM. Proposal of a classification system for patients with neck pain. J Orthop Sports Phys Ther 2004;34:686-96.
27. Wang WT, Olson SL, Campbell AH, Hanten WP, Gleeson PB. Effectiveness of physical therapy for patients with neck pain: an individualized approach using a clinical decision-making algorithm. Am J Phys Med Rehabil 2003;82:203-18.
28. Gross AR, Kay TM, Kennedy C, Gasner D, Hurley L, Yardley K, et al. Clinical practice guideline on the use of manipulation or mobilization in the treatment of adults with mechanical neck disorders. Man Ther 2002;7:193-205.
29. Childs JD, Cleland JA, Elliott JM, Teyhen DS, Wainner RS, Whitman JM, et al. Neck pain: Clinical practice guidelines linked to the International Classification of Functioning, Disability, and Health from the Orthopedic Section of the American Physical Therapy Association. J Orthop Sports Phys Ther 2008;38:A1-A34.
30. Philadelphia Panel. Philadelphia Panel evidence-based clinical practice guidelines on selected rehabilitation interventions for neck pain. Phys Ther 2001;81:1701-17.
31. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd edn. Oxford: Oxford University Press, 2003.
32. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
33. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine 1999;24:1290-4.
34. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. Spine 2002;27:515-22.
35. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-49.
36. Vernon H. The Neck Disability Index: State-of-the-Art, 1991-2008. J Manipulative Physiol Ther 2008;31:491-502.
37. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther 2009;39:400-17.
38. Menezes da Costa L, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol 2009;62:934-43.

Chapter 2

Categorizing continuous variables resulted in different predictors in a prognostic model for non-specific neck pain.

Published as:

Schellingerhout JM, Heymans MW, de Vet HC, Koes BW, Verhagen AP. Categorising continuous variables resulted in different predictors in a prognostic model for non-specific neck pain. J Clin Epid 2009;62:868-74.

ABSTRACT

Objective

To evaluate whether different categorization strategies for introducing continuous variables in multivariable logistic regression analysis results in prognostic models that differ in content and performance.

Methods

Backward multivariable logistic regression ($p<0.05$ and $p<0.157$) was performed with possible predictors for persistent complaints in patients with non-specific neck pain. The continuous variables were introduced in the analysis in three separate ways: 1. Continuous 2. Splitted into multiple categories 3. Dichotomized. The different models were compared with regard to model content, goodness of fit, explained variation, and discriminative ability. We also compared the effect on performance of categorization before and after the selection procedure.

Results

For $p<0.05$ the final model with continuous variables, containing five predictors, disagreed on three predictors with both categorization strategies. For $p<0.157$ the model with continuous variables, containing six predictors, disagreed on three predictors with the model containing stratified continuous variables and on six predictors compared to the model with dichotomized variables. The models in which the variables were kept continuous performed best. There was no clear difference in performance between categorization before and after the selection procedure.

Conclusion

Categorization of continuous variables resulted in a different content and poorer performance of the final model.

INTRODUCTION

In medical research continuous variables are frequently splitted into two or more categories when multivariable logistic regression models are developed. The supposed advantage of categorization is that it simplifies the interpretation of the model and the application in clinical practice.[1]

It has previously been described that categorization has some serious statistical drawbacks in multivariable (logistic) regression models.[1-4] The performance of multivariable regression models, in terms of explained variation and discriminative ability, becomes worse with dichotomization.[1] Furthermore, the estimated magnitude of the association of a covariate with the outcome might be under- or overestimated, due to residual confounding (e.g. by inflation of the type I error rate).[2-4] These statistical problems were all demonstrated in multivariable models with an identical model content (i.e. multivariable models with an identical set of predictors were compared).

Researchers usually perform a selection procedure to identify the predictors in the multivariable model and continuous variables are almost always categorized prior to the model selection procedure. So, the question rises if categorization of continuous variables before model selection results in a different content of the final multivariable model and if this raises additional problems. It is conceivable that an inflated type I error keeps a variable in the model erroneously. Likewise, non-linearity of a continuous variable might result in an unjustified removal or preservation of that variable or correlated variables, due to masking or exaggeration of the differences by categorization. If categorization indeed results in a different model content, it is very likely that it will influence the performance of the model in a negative way. In that case categorization after the selection procedure might even result in a better performance.

Therefore, we aim to evaluate the influence of various categorization strategies on the final model content and performance in multivariable logistic regression analysis. We compare continuous variables, retained as such, with categorization in multiple categories and dichotomization. We also assess the performance of the models when categorization takes place before and after variable selection. The evaluation will be carried out in the data of a study on prognosis of non-specific neck complaints.

METHODS

Setting

The study population consisted of 468 adults (18-70 years) with non-specific neck pain from the primary care population in the Netherlands. We combined data from 3 recently finished randomized controlled trials (RCTs) investigating treatments for pa-

tients with non-specific neck pain.[5-7] The RCTs were similar in design and setting. The assigned interventions were usual care, physiotherapy, spinal manipulation therapy, or a behavioral graded activity program.

Candidate variables

The following continuous variables, measured at baseline, were used in the analysis: age, pain intensity (11-point numerical rating scale [NRS-11], scale 0-10), functional disability (Neck Disability Index (NDI), scale 0-50), health related quality of life (Euro-QOL, 100mm VAS), and fear of movement (Tampa scale for kinesiophobia [TSK], scale 17-68). The following categorical variables were included: gender, duration of neck pain, previous episode of neck complaints, employment status, cause of neck pain,

Table 1 – Baseline characteristics of the study population

Variable	Overall		Hoving [5]	Pool [6]	Vonk [7]
	N=468	Missing	N=183	N=146	N=139
Age, in years (mean ± sd)	45.4 ± 11.8	0%	45.3 ± 11.6	45.1 ± 11.5	45.8 ± 12.4
Gender (male)	182	0%	72 (39%)	57 (39%)	53 (38%)
Level of education		3%			
high	140		50 (27%)	56 (38%)	34 (25%)
medium	202		91 (50%)	60 (41%)	51 (37%)
low	126		42 (23%)	30 (21%)	54 (38%)
Neck pain in the past (yes)	306	2%	119 (65%)	80 (55%)	107 (77%)
Previous treatment (yes)	268	4%	102 (56%)	72 (49%)	94 (68%)
Duration current episode		5%			
< 1 month	61		58 (32%)	0 (0%)	3 (2%)
1-3 months	234		78 (42%)	146 (100%)	10 (7%)
> 3 months	173		47 (26%)	0 (0%)	126 (91%)
Radiating pain (yes)	303	2%	102 (56%)	104 (71%)	97 (70%)
Cause neck pain (trauma)	65	2%	30 (16%)	8 (6%)	27 (19%)
Treatment preference		2%			
no	307		94 (51%)	95 (65%)	118 (85%)
yes, physiotherapy	70		36 (20%)	18 (12%)	16 (11%)
yes, manual therapy	91		53 (29%)	33 (23%)	5 (4%)
Employment status (employed)	342	3%	135 (74%)	113 (77%)	94 (68%)
Headache (yes)	323	2%	127 (69%)	97 (66%)	99 (71%)
Dizziness (yes)	161	2%	67 (37%)	47 (32%)	47 (34%)
Low back pain (yes)	98	2%	44 (24%)	9 (6%)	45 (32%)
Pain intensity, NRS 0-10 (mean ± sd)	5.7 ± 2.1	1%	6.0 ± 1.9	5.3 ± 2.2	5.8 ± 2.2
NDI (mean ± sd)	14.5 ± 6.7	3%	14.5 ± 7.0	14.0 ± 6.8	15.1 ± 6.3
EuroQOL 100mm VAS (mean ± sd)	69.7 ± 17.4	5%	71.2 ± 16.4	70.2 ± 17.1	67.4 ± 18.8
TSK (mean ± sd)	34.3 ± 7.2	43%	35.1 ± 7.3	32.2 ± 6.1	35.4 ± 7.6
Persistent complaints at 26 weeks (n, %)	201 (43%)	8%	80 (44%)	45 (31%)	76 (55%)

SD = standard deviation, NRS = numerical rating scale, NDI = neck disability index, VAS = visual analogue scale, TSK = Tampa scale for kinesiophobia

concomitant headache, concomitant low back pain, concomitant dizziness, level of education, treatment preference, previous treatment for neck complaints, and radiation of the pain to the elbow or shoulder. The included variables are listed in Table 1.

Two of the continuous variables, age and pain intensity, had no linear relationship with the outcome. Therefore, these variables were introduced in the analysis combined with their quadratic term.

All continuous variables were introduced in the analysis in 3 ways: continuous, splitted into >2 categories, further referred to as stratification, and dichotomized at the median (see Table 2).[8 9]

Duration of neck pain at baseline, although continuous in origin, was introduced in 3 categories in every analysis. This was inevitable, as it was not obtained as a continuous variable in one of the three RCTs from which the data originated.

Outcome measure

The outcome measure was (self reported) global perceived recovery,[10] measured on a 6- or 7-point ordinal Likert scale (0="completely recovered", 1="much improved", 2="slightly improved", 3= "no change", 4="slightly worsened", 5="much worsened", and on the 7-point scale: 6="worse than ever"). The outcome was dichotomized into "recovered" and "not recovered", with "not recovered" defined as a score >1 on the Likert scale.

We measured persistent complaints (i.e. "not recovered") at 26 weeks of follow-up.

Model development

To develop our prognostic model we performed multivariable backward logistic regression analysis and initially included all 17 candidate variables. The variables with the highest p-value were removed one-by-one, until all remaining variables had a p-value < 0.05 (Wald-test).[11] The backward selection was performed with the default values of the backward (Wald) logistic regression function of SPSS version 11.0 (SPSS Inc., Chicago IL).

In all of our initial multivariable models the sample size (n=468, with 201 patients with persistent complaints) is rather small relative to the number of parameters (K=17) (i.e. 201/17=11.8 events per variable).[12] Therefore, we also performed the analysis with a p-value of 0.157, which is suitable for small sample sizes.[13] The analyses were adjusted for all assigned interventions that were evaluated in the RCTs.

Reasoning for the backward regression analysis is that it starts out with the "full" model, which reflects the most accurate estimate of the coefficients by taking into account the correlation between variables. Furthermore, the backward selection pro-

Table 2 – Categories for stratification and dichotomisation

Variable	Stratification											Dichotomisation	
	Median	0		1		2		3		4		0	1
Age (in years)	45	18-29	(n=48)	30-39	(n=107)	40-49	(n=138)	50-59	(n=111)	60-70	(n=64)	18-44	45-70
Pain NRS-11 (scale 0-10)	6	0-3	(n=75)	4-6	(n=212)	7-10	(n=181)					0-5	6-10
NDI (scale 0-50)	14	0-4	(n=19)	5-14	(n=241)	15-24	(n=169)	25-34	(n=35)	35-50	(n=4)	0-13	14-50
TSK (scale 17-68)	32	17-33	(n=244)	34-50	(n=214)	51-68	(n=10)					17-31	32-68
EuroQOL VAS (0-100mm)	70	0-25	(n=9)	26-50	(n=69)	51-75	(n=204	76-100	(n=186)			0-69	70-100

NRS = numerical rating scale, NDI = neck disability index, TSK = Tampa scale for kinesiophobia, VAS = visual analogue scale

Table 3 – Multivariable association with persistent complaints at 26 weeks

Variable		Continuous†				Stratified‡					Dichotomised‡				
		P < 0.05		P < 0.157		P < 0.05		P < 0.157			P < 0.05		P < 0.157		
		OR	95%-CI	OR	95%-CI	Cat.	OR	95%-CI	OR	95%-CI	Cat.	OR	95%-CI	OR	95%-CI
Age (in years)															
Gender (1=female)															
Previous neck pain (1=yes)		1.68	1.09 - 2.51	1.67	1.08 - 2.56										
Previous treatment for neck pain (1=yes)							1.64	1.10 - 2.45	1.69	1.13 - 2.54		1.64	1.10 - 2.45	1.59	1.06 - 2.39
Duration current episode (< 1, 1-3, >3 months)															
Treatment preference (no/yes PT/yes MT)															
Pain intensity (NRS-11)	*linear*	0.62	0.42 - 0.92	0.61	0.41 - 0.91	*0-3*			ref.						
	square	1.06	1.02 - 1.09	1.06	1.02 - 1.10	*4-6*			1.05	0.59 - 1.88					
						7-10			1.67	0.92 - 3.05					

Table 3 – Multivariable association with persistent complaints at 26 weeks

Variable	Continuous†				Stratified‡					Dichotomised‡				
	P < 0.05		P < 0.157		P < 0.05		P < 0.157			P < 0.05		P < 0.157		
	OR	95%-CI	OR	95%-CI	Cat.	OR	95%-CI	OR	95%-CI	Cat.	OR	95%-CI	OR	95%-CI
Level of education (high/middle/low)														
Employment status (1=employed)	0.56	0.36 - 0.88	0.58	0.37 - 0.90		0.59	0.38 - 0.90	0.59	0.38 - 0.91		0.59	0.38 - 0.90	0.57	0.37 - 0.89
Concomitant low backpain (1=yes)	1.67	1.02 - 2.75	1.65	1.00 - 2.46		1.74	1.07 - 2.83	1.62	0.99 - 2.66		1.74	1.07 - 2.83	1.66	1.02 - 2.72
Concomitant headache (1=yes)	1.61	1.03 - 2.51	1.57	1.01 - 2.46		1.68	1.09 - 2.59	1.59	1.03 - 2.47		1.68	1.09 - 2.59	1.56	1.00 - 2.43
Concomitant dizziness (1=yes)														
Radiating pain (1=yes)	0.58	0.38 - 0.89	0.59	0.39 - 0.90		0.65	0.43 - 0.97	0.61	0.40 - 0.92		0.65	0.43 - 0.97	0.62	0.41 - 0.94
Cause of neck pain (1=trauma)													1.50	0.86 - 2.64
Functional disability (NDI)														
Quality of life (EuroQOL 100mm VAS)										*0-69*			ref.	
										70-100			0.70	0.46 - 1.06
Fear of movement (TSK)			1.02	0.99 - 1.05										

OR = Odds Ratio, 95%-CI = 95% confidence interval, cat. = category, ref. = reference category, **bold** = continuous variable

† Continuous variables are introduced as such in the logistic regression analysis

‡ Continuous variables were categorised prior to the selection procedure

cedure provides information on the extent of deviation from the performance of the full model, when a variable is removed from the model.

Imputation of missing values in the data was carried out by multiple imputation using all observed information and was performed using R software.[14-16]

Model performance

We checked the performance of the different models with regard to the goodness of fit, the explained variation, and the discriminative ability of the model.

Goodness of fit of the model is estimated by Akaike's Information Criterion (AIC). Alower value for the AIC indicates a better performance of the model. We used the AIC_c , which is the small-sample version of AIC.[17] The AIC_c should be used if the sample size (n) is small relative to the number of parameters (K) (i.e. $n / K \leq 40$).[17]

The explained variation of the model is estimated by Nagelkerke's R-square. Explained variation is the extent to which the outcome can be predicted by (the predictors in) the model.[18] The discriminative ability is reflected by the area under the receiver operating characteristics curve (AUC). The AUC represents the ability of the prognostic model to point out the patient that will have persistent symptoms in two patients with different outcomes, and ranges from 0.5 (chance) to 1.0 (perfect discrimination).[19]

To evaluate if categorization of continuous variables prior to the selection procedure results in a different performance compared to models in which the continuous variables are categorized after the selection procedure, we compare the performance of the models resulting from both procedures. Categorization "a priori" means that the continuous variables are categorized before the backwards selection procedure is carried out. Categorization "afterwards" means that the continuous variables are categorized after the backwards selection procedure is carried out (i.e. the predictors in each multivariable model are identical).

All performance measures were estimated using SPSS version 11.0 (SPSS Inc., Chicago IL).

RESULTS

The baseline characteristics of our study population and number of patients that had persistent complaints at 26 weeks of follow-up are presented in Table 1, for the total population and per trial. The eligibility criteria were identical in the three RCTs, except for one aspect: duration of complaints at baseline.[5-7] However, since duration of complaints was introduced as a covariate in our analysis, this did not affect our results.

The Tampa scale for kinesiophobia was not obtained in one of the RCTs from which our data originated,[5] resulting in 43% missing values for this variable.

The content of the final model for the different categorization strategies, as well as the multivariable association of the predictors with the outcome is presented in Table 3.

Backward regression with a p-value of 0.05 resulted in disagreement on 3 predictors included in the model, if we compare the model with continuous variables to those with the stratified or dichotomized continuous variables: previous neck pain, previous treatment for neck pain, and pain intensity.

Repetition of the analysis with a p-value of 0.157 resulted in disagreement on 3 predictors included in the final model, if we compare the model with continuous variables to the one with stratified variables: previous neck pain, previous treatment for neck pain, and fear of movement. Dichotomization of the continuous variables even increased the number of dissimilarities, with respect to the model content, to six variables: previous neck pain, previous treatment for neck pain, pain intensity, cause of neck pain, quality of life, and fear of movement.

A striking observation, based on Table 3, is that none of the continuous predictors is selected in the final model for all categorization strategies with a corresponding p-value. Comparison of the results using a p-value of < 0.05 and a p-value of < 0.157 shows that the dissimilarities in model content increased with an increasing p-value.

The effect of the different strategies of categorization on model performance is presented in Table 4 for both p-values. Table 4 shows that analyses in which continu-

Table 4 – Performance model with different strategies of categorisation

	Continuous	Stratification		Dichotomisation	
		a priori†	afterwards‡	a priori†	afterwards‡
P < 0.05					
AIC_c (K¶)	600.9 (11)	607.3 (9)	608.0 (11)	607.3 (9)	609.4 (10)
Nagelkerke's R-square	0.164	0.137	0.146	0.137	0.137
AUC	0.696	0.678	0.688	0.678	0.683
P < 0.157					
AIC_c (K¶)	601.0 (12)	606.2 (11)	610.0 (13)	606.4 (11)	611.4 (11)
Nagelkerke's R-square	0.169	0.151	0.152	0.150	0.138
AUC	0.698	0.688	0.691	0.690	0.683

† The continuous variables are categorised before the backward selection procedure

‡ The continuous variables are categorised after the model content has been determined with continuous variables retained as such in the selection procedure.

¶ K = number of parameters (i.e. coefficients + intercept) in the model. The number of parameters = (number of OR's in Table 3) + 4, due to correction for treatment (3 coefficients) and addition of the intercept.

ous variables are retained as such perform better, with respect to goodness of fit, explained variation, and discriminative ability, than all strategies in which continuous variables are categorized. The models with stratified continuous variables perform slightly better than their dichotomized counterparts. This observation holds for both p-values.

The goodness of fit of the models in which the continuous variables are categorized after the selection procedure is worse than that of the models in which continuous variables are categorized a priori. However, the explained variation and discriminative ability are slightly better, except for dichotomization with $p < 0.157$.

DISCUSSION

Categorization of continuous variables into two or multiple categories prior to introducing them into backward logistic regression analysis resulted in a different model content and a decrease in model performance compared to the selection procedure with continuous variables.

It is likely that residual confounding and inflation of type I error lead to overestimation or underestimation of the association of variables with the outcome and in that way contribute to the observed dissimilarities, as shown by other studies.[2-4]

Decrease in explained variation and discriminative ability of models with dichotomized continuous variables was reported previously for Cox regression analysis.[1] This corresponds with the results in our study, although the differences in our study are somewhat smaller. This is probably due to the weak association of the actual predictors with the outcome in our study. We consider it likely that these differences are larger in studies with multiple (continuous) variables, that are highly correlated with each other and/or have a stronger univariable relation with the outcome.

In our models there were more dissimilarities in model content when we used a p-value of < 0.157 than with a p-value of < 0.05. In general, using a higher p-value will result in more variables that are retained in the model. These larger models will perform better, but their fit is usually worse.[13] This was confirmed in our study: models that were built with a p-value of < 0.157, instead of 0.05, showed a better discriminative ability and explained variation, but a worse fit (i.e. higher values for the AIC_c). Only for the models where stratification and dichotomization was done a priori the model fit slightly increased with the p-value of < 0.157.

In clinical research continuous variables are most frequently categorized before the analysis. However, if categorization is unavoidable it seems advisable to do it after the selection procedure; this will ensure the most accurate model content. An associated advantage of categorization after the selection procedure will be improvement of

consistency in variables identified as a predictor across different studies, as they are no longer influenced by chosen categories.

It was not possible to split all continuous variables into categories that are suggested or frequently used in medical literature. For the NRS-11 and NDI categories have been suggested in people with neck pain.[8 9] For age we used generally accepted categories. However, no such categories are established for the EuroQOL 100mm VAS and the TSK. We decided to split them into equal parts.

For dichotomization we chose the median as cut off-point. Another frequently used cut off-point for dichotomization is a so-called "optimal" cut off-point. Since these "optimal" cut off-points introduce additional biases in the analysis, we did not consider these in our study.[1 3 20]

Continuous candidate variables were checked for linearity before the analysis, instead of assuming linearity. This was done, because an incorrect linearity assumption might lead to a misspecified model. It is even possible in case of non-linearity that multiple categories give a better reflection of the association than the linear term.

To reflect non-linear relationships of continuous variables we chose quadratic polynomials. It would have been more sophisticated to use so-called fractional polynomials.[21 22] However, quadratic polynomials were chosen because they are used more often in clinical research.

The poorer statistical performance of models in which continuous variables are categorized, implies a poorer performance of the model in clinical practice. In our model, for example, a lower explained variation means that the estimated probability of persistent neck pain is less accurate and a lower discriminative ability makes it harder to distinguish patients that will have persisting neck pain from those who will recover. Therefore, we recommend for clinical research to introduce continuous variables as such in backward stepwise logistic regression analysis. This will result in a model with the best representation of the actual associations and with the highest performance. If categorization of continuous variables is desirable we suggest to do this after the selection procedure. In case of categorization we recommend to use multiple categories and disease-specific or generally accepted cut off-points based on the literature.

For future research we suggest to evaluate the effect of categorization of continuous variables on model content and performance more thoroughly. Also the effect of categorization of the outcome needs further attention.

CONCLUSION

Categorization of continuous variables prior to introducing them into backward multivariable logistic regression analysis resulted in different predictors remaining in the final model and loss of model performance. Every categorization strategy results in a loss of information, but stratification of continuous variables seems to result in a somewhat better performance than dichotomization.

It is advisable to retain continuous variables as such during the development of the model and, if unavoidable, categorize afterwards. This will at least provide the most accurate model content.

REFERENCES

1. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006; 25:127-41.
2. Becher H. The concept of residual confounding in regression models and some applications. Stat Med 1992; 11:1747-58.
3. Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. Stat Med 2004; 23:1159-78.
4. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. Epidemiology 1997;8:429-34.
5. Hoving JL, Koes BW, de Vet HC, van der Windt DA, Assendelft WJ, van Mameren H, et al. Manual therapy, physical therapy, or continued care by a general practitioner for patients with neck pain. A randomized, controlled trial. Ann Intern Med 2002;136:713-22.
6. Pool JJ, Ostelo RW, Koke AJ, Bouter LM, de Vet HC. Comparison of the effectiveness of a behavioral graded activity program and manual therapy in patients with sub-acute neck pain: design of a randomized clinical trial. Man Ther 2006;11:297-305.
7. Vonk F, Verhagen AP, Geilen M, Vos CJ, Koes BW. Effectiveness of behavioral graded activity compared with physiotherapy treatment in chronic neck pain: design of a randomized clinical trial [ISRCTN88733332]. BMC Musculoskelet Disord 2004;5:34.
8. Fejer R, Jordan A, Hartvigsen J. Categorizing the severity of neck pain: establishment of cut-points for use in clinical and epidemiological research. Pain 2005;119:176-182.
9. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
10. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.
11. Altman D. Practical statistics for medical research. London: Chapman and Hall; 1991.
12. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49:1373-9.
13. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19(8):1059-79.
14. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59:1087-91.
15. Moons KG, Donders RA, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59:1092-101.
16. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2007.
17. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods Research 2004; 33:261-304.
18. Hu B, Palta M, Shao J. Properties of R(2) statistics for logistic regression. Stat Med 2006; 25:1383-95.
19. Harrell Jr. FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-87.
20. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. J Natl Cancer Inst 1994; 86:829-835.

21. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Applied statistics 1994; 43:429-67.
22. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J Royal Stat Society Series A 1999; 162:71-94.

Chapter 3

Prognosis of patients with non-specific neck pain: development and external validation of a prediction rule for persistence of complaints.

Published as:

Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HC, Koes BW. Prognosis of patients with non-specific neck pain: development and external validation of a prediction rule for persistence of complaints. Spine 2010;35:E827-35.

ABSTRACT

Objective

Development and validation of a prediction rule that estimates the probability of complaints persisting for at least 6 months in patients presenting with non-specific neck pain in primary care.

Methods

The study population consisted of a sample (n=468) from the adult primary care population (18-70 years) in The Netherlands presenting with non-specific neck pain. The primary outcome measure was global perceived recovery measured at 6 months of follow-up. Seventeen baseline characteristics of the patients were included in the analysis. Significant predictors were identified by multivariable backward stepwise logistic regression analysis. A score chart was constructed by using the regression coefficient estimates. The score chart was externally validated in a cohort of patients with non-specific neck pain (n=315), who participated in a randomized controlled trial in the United Kingdom (PANTHER-trial).

Results

The multivariable analysis resulted in a set of 9 predictors. The score chart has a discriminative ability of 0.66. External validation of the score chart showed a discriminative ability of 0.65, an adequate calibration, a good fit, and a low explained variation.

Conclusion

We developed a score chart, estimating the probability of persistent complaints at 6 months follow-up for patients with non-specific neck pain. This chart performed well in the study population and external validation population. The prediction which patients are more likely to develop persistent complaints is significantly improved by the score chart.

INTRODUCTION

Neck pain is one of the most common musculoskeletal disorders, with an estimated 1-year prevalence of 31.4 - 35.6% in adults in the general population.[1-4] The course of neck pain is characterised by exacerbations and remissions and only a small part of the patients experience complete resolution of their symptoms within one year.[5]

A substantial proportion of the neck pain patients will thus develop chronic neck pain. The definition of chronicity differs between studies, in terms of either 3 or 6 months duration of complaints. Nevertheless, estimates of their 1-year prevalence in the general population are similar: 8.7 - 17.8% for the 3 months definition,[1 2 4] and 8 - 13.8% for the 6 months definition.[1 6]

An important question is: can we identify patients at risk of persistent complaints at the first consultation with the physician, based on their personal characteristics? This information could be helpful for a physician to gain insight into the prognosis of an individual patient with neck pain, and would aid in informing patients more accurately about their expected prognosis. Furthermore, this would aid researchers in selecting patients at high risk in studies on prevention of chronic neck pain.

Some studies on characteristics associated with persistence of neck complaints in the general population have been conducted.[5 7-10] Characteristics identified as a predictor in more than one study were: age, duration of complaints, previous episode of neck pain, pain intensity, physical functioning, and accompanying low back pain.[5 7-10] However, none of the studies constructed a prediction model that quantifies prognosis.

Therefore, the purpose of this study was to develop and externally validate a prediction rule that estimates the probability of complaints persisting for at least 6 months in patients consulting their physician for non-specific neck pain.

METHODS

Data collection

Data from three recently finished randomized controlled trials (RCTs) evaluating the effectiveness of interventions for non-specific neck pain were combined.[11-13] These RCTs were all carried out in a primary care setting in The Netherlands and were similar in design. The patients (n=468) in the different RCTs were assigned to one of the following treatments: usual care by a general practitioner (n=64), physiotherapy (n=130), manual therapy (n=135), or a behavioral graded activity program (n=139).[11-13]

Study population

The three RCTs had similar eligibility criteria and consisted of a primary care population, aged 18-70 years, with non-specific neck pain. Non-specific neck pain was defined as neck pain without an identified pathological basis. So people with a specific disorder (e.g. herniated disc, neurological disorder, rheumatological disorder, malignancy, infection, or fracture) were excluded from the study populations.[11-13]

Outcome measure

The outcome measure is (self reported) global perceived recovery and is measured on a 6- or 7-point ordinal Likert scale (0="completely recovered", 1="much improved", 2="slightly improved", 3= "no change", 4="slightly worse", 5="much worse", and on the 7-point scale: 6="worse than ever").[14] The outcome was dichotomized into "recovered or much improved" and "persistent complaints"; the latter defined as a score of 2-6 on the Likert scale.[11-13] The outcome was measured at 6 months of follow-up.

Identification of potential predictors

As potential predictors we selected characteristics of the patient and the complaint.

To comply with the rule of at least ten events per variable in the analysis, we restricted the number of candidate variables to seventeen.[15] The most recently published systematic review and prospective studies on predictors for non-specific neck pain in the general population identified fourteen predictors, of which eleven predictors were available in our data.[5 7-10] The eleven predictors were: age, gender, employment status (employed/not employed), pain intensity at baseline (11-point Numerical Rating Scale (NRS-11), scale 0-10), duration of neck pain at baseline (< 1 month/1-3 months/> 3 months) , previous episode of neck complaints (yes/no), cause of neck pain (trauma/no trauma), accompanying headache (yes/no), accompanying low back pain (yes/no), function (Neck Disability Index (NDI), scale 0-50),[16] and quality of life (EuroQOL 100 mm VAS, scale 0-100).[17]

Another six characteristics were added to complete the set of seventeen: level of education (high/medium/low), treatment preference (no/yes, physiotherapy/yes, manual therapy), previous treatment for neck complaints (yes/no), accompanying dizziness (yes/no), radiation of pain to the shoulder or elbow (yes/no), and fear of movement (TAMPA-scale for Kinesiophobia (TSK), scale 17-68).[18] These six characteristics have not been evaluated in previous studies, but could, in our opinion, be related to persistence of complaints.[5 7-10]

On the NRS-11, NDI, and TSK a higher score indicates a higher burden of disease. A higher score on the EuroQOL VAS indicates better general health status.

Selection of model content

Candidate variables were checked for their univariable association by using logistic regression analysis. Continuous variables were checked for linearity by adding quadratic terms to the model and testing these for statistical significance ($p<0.05$).[19] This revealed a non-linear relationship for age and pain intensity. Therefore, they were introduced in the analysis together with their squared coefficient.

To develop our model we performed multivariable backward stepwise logistic regression analysis. Initially all candidate predictors were included. The variables with the highest p-value were removed one-by-one (Wald-test), until all remaining variables had a p-value < 0.157 (Akaike Information Criterion).[20 21] The additivity assumption was checked by adding interaction terms to the model and evaluating if they contributed significantly to the model ($p<0.157$).[19]

The analyses were adjusted for all assigned interventions that were evaluated in the RCTs, by introducing "treatment" as a variable in the model. For practical reasons, "treatment" and the squared coefficients were removed after the significant interaction terms were added to the logistic regression equation (LE).

Following removal of "treatment" and the squared coefficients, the coefficients of the regression model were adjusted using the slope (β) of the calibration line (formula: log odds (persistent complaints) = intercept + β * LE).[22] Subsequently the intercept was adjusted, to make the average predicted probability correspond with the observed overall event rate.[22]

Imputation of missing values

Imputation of missing values in the data was carried out by multiple imputation using all observed information.[23 24] A total of 5 imputed datasets were created.[25] Selection of variables was performed in each dataset separately. A variable was included in the model if it was selected as a predictor in at least two out of five imputed datasets.

To develop the final model the regression coefficients and standard errors of the predictors and interaction-terms were averaged over the 5 imputed datasets, according to Rubin.[25]

Construction of score chart

To make the model suitable for use in clinical practice, we transformed the logistic regression equation into a score chart. The coefficients in the logistic regression equation were multiplied by 25 and rounded to the nearest integer to obtain the score per predictor. Multiplication by 25 was chosen to get the majority of the coefficients close to an integer, thereby minimising the effects of rounding. The sum of all scores reflects the probability of persistent complaints at 6 months of follow-up.

Model performance

The discriminative ability of the logistic regression model and score chart was determined with the area under the receiver operating characteristics curve (AUC). The AUC represents the ability of the prediction rule to distinguish between patients with or without persistent complaints at 6 months follow-up and ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination).[26] An AUC $\geq$ 0.7 is considered as good discrimination and an AUC < 0.7 as moderate discrimination.[26]

To express the discriminative ability of the score chart in clinical terms we will calculate sensitivity, specificity, and positive predictive value (PPV) for different cut off-points. Sensitivity is the proportion of patients with persistent complaints after 6 months of follow-up that really have a score above the cut off-point.[20] Specificity is the proportion of patients with self-reported recovery after 6 months of follow-up that really have a score below the cut off-point.[20] PPV is the proportion of patients with a score above the cut off-point that really will have persistent complaints after 6 months of follow-up.[20]

An indication of the goodness of fit of the score chart was obtained by creating a calibration plot and by performing the Hosmer-Lemeshow test. In a calibration plot the predicted probability of persistent complaints is set out against the probabilities associated with observed frequency of persistent complaints.[27] An adequate calibration means that the predicted probabilities agree with the observed probabilities over the whole range of values. The Hosmer-Lemeshow test assesses the overall goodness of fit by comparing the predicted and observed frequencies throughout the deciles of risk.[26] The Hosmer-Lemeshow test results in a p-value that gives an indication of the model fit; a higher p-value represents a better fit and $p < 0.05$ indicates a significant lack of fit.[26]

The explained variation of the score chart was determined with Nagelkerke's R^2.[19] Nagelkerke's R^2 reflects the proportion of variation in the outcome explained by the predictors in the model.

The analysis was performed using SPSS version 11.0 (SPSS Inc., Chicago IL) and R software.[28]

External validation

External validation was performed using the data of an RCT (PANTHER-trial, n=346) carried out in the United Kingdom.[29] This RCT was carried out in a primary care setting and evaluated the effectiveness of electrotherapy and manual therapy in people with non-specific neck pain.[29] Thirty-one patients were excluded from the analysis, because they were outside the scope of the prediction rule due to their age (>70 years of age).

We determined the discriminative ability and goodness of fit of the score chart in the external population.

RESULTS

The baseline characteristics of the development population (n=468) and the validation population (n=315) are shown in Table 1. Persistent complaints were reported by 43% of the patients in the development population after 6 months of follow-up.

The uni- and multivariable association of the candidate predictors with the outcome is presented in Table 2. Introduction of interaction-terms in the model revealed interaction of "accompanying headache" with pain intensity, employment status, radiation of pain, and previous neck complaints. Combination of the predictors and interaction terms results in the logistic regression equation, with adjusted coefficients, as presented in Table 3. The discriminative ability of the equation is 0.66 (95% confidence interval: 0.61-0.71).

Table 1 – Baseline characteristics

Variable	Development set (n=468)		PANTHER-trial (n=315)	
	Value	Missing (n,%)	Value	Missing (n,%)
Age, in years (mean ± sd)	45.4 ± 11.8	0 (0%)	48.8 ± 12.1	0 (0%)
Gender (male, %)	182 (39%)	0 (0%)	114 (36%)	0 (0%)
Level of education		16 (3%)	na	
High	135 (29%)			
Medium	195 (42%)			
Low	122 (26%)			
Neck pain in the past (yes, %)	301 (64%)	10 (2%)	195 (62%)	15 (5%)
Previous treatment (yes, %)	258 (55%)	17 (4%)	110 (35%)	5 (2%)
Duration current episode		25 (5%)		0 (0%)
< 1 month	58 (13%)		17 (5%)	
1-3 months	225 (48%)		58 (18%)	
> 3 months	160 (34%)		240 (76%)	
Radiating pain (yes, %)	296 (63%)	10 (2%)	238 (76%)	2 (1%)
Cause neck pain (trauma, %)	63 (14%)	10 (2%)	48 (15%)	2 (1%)
Treatment preference		10 (2%)	na	
No	301 (64%)			
yes, physiotherapy	69 (15%)			
yes, manual therapy	88 (19%)			
Employment status (employed, %)	334 (71%)	12 (3%)	199 (63%)	0 (0%)
Headache (yes, %)	317 (68%)	10 (2%)	83 (26%)	2 (1%)
Dizziness (yes, %)	156 (33%)	10 (2%)	na	
Low backpain (yes, %)	96 (21%)	10 (2%)	94 (30%)	2 (1%)
Pain NRS-11, scale 0-10 (mean ± sd)	5.7 ± 2.1	6 (1%)	5.0 ± 2.3	0 (0%)
NDI, scale 0-50 (mean ± sd)	14.5 ± 6.7	12 (3%)	na	
EuroQOL VAS, 0-100mm VAS (mean± sd)	69.9 ± 17.3	23 (5%)	69.4 ± 17.2	1 (0%)
TAMPA-scale, scale 17-68 (mean± sd)	33.8 ± 7.1	199 (43%)	na	

na = not available in data of PANTHER-trial

Table 2 – Uni- and multivariable association of candidate variables with persistent neck complaints at 6 months follow-up

Variable	Univariable			Multivariable			Selection
	OR	95% - CI	p-value	OR	95% - CI	p-value	Frequency[1]
Received treatment							
usual care				ref.			
physiotherapy				0.89	0.46 - 1.72	0.728	
manual therapy				0.44	0.22 - 0.86	0.016	
behavioural graded activity				0.53	0.28 - 1.03	0.060	
Age (in years)							2
age	0.88	0.79 - 0.99	0.028	0.93	0.81 - 1.06	0.262	
age-square	1.00	1.00 - 1.06	0.020	1.00	1.00 - 1.00[2]	0.266	
Gender (female=1)	0.98	0.65 - 1.46	0.906				0
Level of education							0
high	ref.						
middle	1.26	0.79 - 2.02	0.330				
low	1.23	0.75 - 2.02	0.409				
Neck pain in the past (yes=1)	1.79	1.18 - 2.72	0.007	1.71	1.08 - 2.72	0.024	4
Previous treatment for neck pain (yes=1)	1.77	1.20 - 2.61	0.004				1
Duration current episode							1
< 1 month	ref.						
1-3 months	0.68	0.38 - 1.22	0.201				
> 3 months	1.25	0.68 - 2.31	0.472				
Radiating pain (yes=1)	0.66	0.45 - 0.97	0.033	0.58	0.37 - 0.89	0.013	5
Cause neck pain (trauma=1)	1.94	1.12 - 3.34	0.018	1.59	0.87 - 2.90	0.132	3
Treatment preference							0
no	ref.						
yes, physiotherapy	0.77	0.45 - 1.32	0.344				
yes, manual therapy	1.19	0.73 - 1.93	0.482				

Table 2 – Uni- and multivariable association of candidate variables with persistent neck complaints at 6 months follow-up (continued)

Variable	Univariable			Multivariable			Selection
	OR	95% - CI	p-value	OR	95% - CI	p-value	Frequency[1]
Employment status (employed=1)	0.60	0.39 - 0.92	0.019	0.65	0.37 - 1.17	0.156	3
Accompanying headache (yes=1)	1.92	1.23 - 3.00	0.005	1.55	0.95 - 2.54	0.080	5
Accompanying dizziness (yes=1)	1.47	0.99 - 2.17	0.058				0
Accompanying low back pain (yes=1)	2.07	1.31 - 3.27	0.002	1.49	0.89 - 2.48	0.128	3
Pain intensity at baseline (NRS 0-10)							5
pain	0.70	0.46 - 1.07	0.101	0.63	0.40 - 1.01	0.059	
pain-square	1.05	1.01 - 1.09	0.022	1.05	1.01 - 1.10	0.023	
Neck Disability Index (0-50)	1.05	1.02 - 1.08	0.001				1
TAMPA-scale (17-68)	1.03	1.01 - 1.06	0.019				1
EuroQOL 100mm VAS (0-100)	0.99	0.98 - 1.00	0.034	0.99	0.98 - 1.00	0.147	2

ref. = reference category, 95%-CI = 95% confidence interval

OR = odds ratio, an OR > 1 reflects a higher probability of persistent complaints and an OR < 1 a lower probability of persistent complaints, compared to the reference category

[1] number of imputed databases in which the variable is selected as predictor

[2] 95% confidence interval = 0.999 – 1.002

Table 3 – Final model of variables associated with persistent complaints at 6 months

Variable	Beta	s.e.	OR	84.3%-CI[1]
Constant	-1.704	0.838		
Age	0.029	0.015	1.03	1.01 - 1.05
Pain intensity (NRS-11)	-0.042	0.058	0.96	0.88 - 1.04
Accomp. headache (1=yes)	0.198	0.498	1.22	0.60 - 2.46
Radiation of pain to elbow/shoulder (1=yes)	-0.564	0.212	0.57	0.42 - 0.77
Previous neck complaints (1=yes)	0.515	0.207	1.67	1.25 - 2.24
Cause of complaints (1=trauma)	0.234	0.149	1.26	1.02 - 1.56
Accomp. low back pain (1=yes)	0.829	0.415	2.29	1.27 - 4.12
Employment status (1=employed)	0.372	0.244	1.45	1.03 - 2.05
EuroQOL 100mm VAS	-0.005	0.003	1.00	0.99 - 1.00
Accomp. headache * Pain intensity	0.116	0.073	1.12	1.01 - 1.24
Accomp. headache * Previous neck complaints	-0.376	0.252	0.69	0.48 - 0.98
Accomp. headache * Radiation of pain	0.392	0.253	1.48	1.03 - 2.12
Accomp. headache * Employment status	-0.815	0.276	0.44	0.30 - 0.65

s.e. = standard error, 84.3%-CI = 84.3% confidence interval, all coefficients are adjusted with the slope of 0.4818

OR = odds ratio, an OR > 1 reflects a higher probability of persistent complaints and an OR < 1 a lower probability of persistent complaints, compared to the reference category

[1] corresponds with the p-value of 0.157

Construction of score chart

The score chart that we derived from the logistic regression model is presented in Table 4. The weight of an item in the score chart is based on its related coefficient in the logistic regression equation. Table 4 also provides the score chart legend to convert the total score into the predicted probability of persistent complaints. An example of how to calculate the score for an individual patient is presented in Appendix A.

The discriminative ability of the score chart is 0.66 (95% confidence interval: 0.62-0.71). The sensitivity, specificity, and PPV for different cut off-points are presented in Table 5. Table 5 shows that for a score $\geq$ 35 on the chart, the probability of persistent complaints is significantly higher than the overall probability of persistent complaints of 43%. The calibration plot is shown in Figure 1. Calibration of the score chart seems adequate. The corresponding Hosmer-Lemeshow test resulted in a p-value of 0.61, which indicates that the model fits quite well.[26] The explained variation of the score chart is 0.12.

External validation

The baseline characteristics of the patients in the PANTHER-trial were largely similar to our baseline characteristics (see Table 1).[29] The proportion of patients with persistent complaints in the PANTHER-trial was 39%. Application of the logistic regression equation (from Table 3) to the PANTHER-data showed a discriminative ability of 0.65 (95% confidence interval: 0.59-0.71). The score chart had a discriminative ability (AUC) of

Table 4 – Score chart

		Score
Age	+ 7	/ 10 yr[1]
Accompanying low back pain	+ 21	
Traumatic cause neck complaints	+ 6	
Health status (scale 0-100)[2]	- 3	/ 25 points[3]
Accompanying headache	+ 5	
No accompanying headache		
Radiation of pain to elbow/shoulder	- 14	
Previous neck complaints	+ 13	
Paid employment	+ 9	
Pain intensity (scale 0-10)[4]	- 1	/ point
Accompanying headache		
Radiation of pain to elbow/shoulder	- 4	
Previous neck complaints	+ 4	
Paid employment	- 11	
Pain intensity (scale 0-10)[4]	+ 2	/ point
		Total score

Total score	Probability[5]
< 10	0 - 20%
10 - 34	20 - 40%
35 - 54	40 - 60%
55 - 79	60 - 80%
> 79	80 - 100%

An example of how to get the score of an individual patient is demonstrated in Appendix A

[1] The score increases with 7 points per 10 year (e.g. a 40-year old person receives a score of 4x7=28 points)

[2] Question: "Can you rate your own health status today?" (0=worst imaginable, 100=best imaginable)

[3] The score decreases with 3 points per 25 points on the health status scale (e.g. a person with a score of 75 receives a score of 3x-3=-9 points)

[4] Question: "Can you rate your current pain intensity?" (0=no pain, 10=worst imaginable pain)

[5] Probability that neck complaints will still be present at 6 months after the first consultation

Table 5 – Discriminative ability score chart in development set

	Persistent[1]	Recovered		Sensitivity	Specificity	Positive predictive
Score	(n=199)	(n=269)	Score	(95% - CI)	(95% - CI)	value (95% - CI)
> 79	9	1	> 79	0.05 (0.02-0.07)	1.00 (0.99-1.00)	90% (71-100%)
55 - 79	35	22	≥ 55	0.22 (0.16-0.28)	0.91 (0.88-0.95)	66% (54-77%)
35 - 54	78	82	≥ 35	0.61 (0.55-0.68)	0.61 (0.55-0.67)	54% (47-60%)
10 - 34	76	153	≥ 10	0.99 (0.99-1.00)	0.04 (0.02-0.06)	43% (39-48%)
< 10	1	11				

Overall population (n=468): probability persistent complaints = 43% (95%-CI: 38-47%), 95%-CI = 95% confidence interval

[1] Number of patients with persistent complaints after 6 months of follow-up

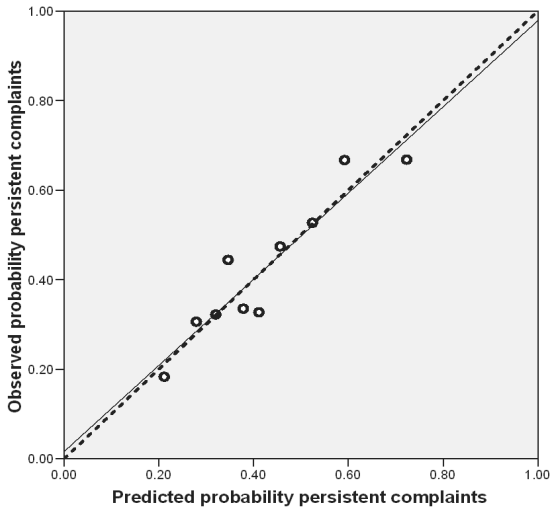

Figure 1 – Calibration score chart
o = deciles of risk. ---- = perfect calibration, __ = calibration line

0.66 (95% confidence interval: 0.59-0.72) in the population of the PANTHER-trial, which is the same as in the development population.

The more detailed characteristics for discriminative ability (see Table 6) are similar to those in the development population: for a score ≥ 35 the probability of persistent complaints is significantly higher than the overall probability of persistent complaints of 39%. The calibration plot is shown in Figure 2. This plot shows that the score chart tends to overestimate the probability of persistent complaints, with a maximum overestimation of 7% for the highest decile of risk. The Hosmer-Lemeshow test resulted in a p-value of 0.61, which indicates that the model fits quite well.[26] The explained variation of the score chart is 0.10.

Table 6 – Discriminative ability score chart in validation set

Score	Persistent[1] (n=124)	Recovered (n=191)	Score	Sensitivity (95% - CI)	Specificity (95% - CI)	Positive predictive value (95% - CI)
> 79	0	1	> 79	0 (na[2])	1.00 (0.99-1.00)	0% (na[2])
55 - 79	34	21	≥ 55	0.27 (0.20-0.35)	0.89 (0.84-0.93)	61% (48-74%)
35 - 54	44	54	≥ 35	0.63 (0.54-0.71)	0.60 (0.53-0.67)	51% (43-59%)
10 - 34	43	103	≥ 10	0.98 (0.95-1.00)	0.06 (0.03-0.10)	40% (35-46%)
< 10	3	12				

Overall population (n=315): probability persistent complaints = 39% (95%-CI: 34-43%), 95%-CI = 95% confidence interval , na = not available

[1] Number of patients with persistent complaints after 6 months of follow-up

[2] Confidence interval not available, because there is only 1 person in this group

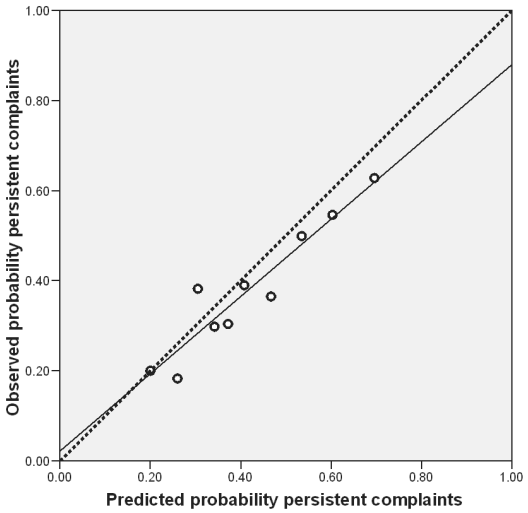

Figure 2 – Calibration in PANTHER-data
o = deciles of risk. ---- = perfect calibration, __ = calibration line

DISCUSSION

We developed and externally validated a score chart to estimate the probability of persistent complaints at 6 months follow-up for patients with non-specific neck pain in primary care. The chart has a moderate discriminative ability, an adequate calibration, a good fit, and a low explained variation in both the development and external population. The score chart predicts significantly better for every patient if he/she will have persistent complaints, than estimates for the overall population.

As mentioned in the introduction, there have been studies evaluating predictors for persistence of non-specific neck complaints in the general population.[5 7-10] Eight of the eleven literature-based candidate predictors were also identified as a predictor in our analysis. The direction of the associations of these predictors in our model (i.e. worse or better prognosis) is consistent with those found in the previous studies.[5 7-10] Gender, physical functioning, and duration of complaints were not identified as a predictor in our study, which could be caused by their borderline prognostic value: these variables showed a significant association with the outcome in only 10-45% of the analyses performed in the previous studies.[5 7-10] This "borderline behaviour" of variables seems to be a common phenomenon for predictors in neck pain: none of the predictors, identified in this study or previous studies, consistently has a (major) impact on prognosis.[5 7-10] Of the six variables that were added to the literature-based variables, only "radiation of pain" was identified as a predictor in our study. Four of

these six variables had a significant univariable association with the outcome, but (strong) correlation with other variables is probably the reason that they did not end up in the multivariable model.

We are the first to develop a multivariable model that quantifies persistence of non-specific neck complaints. This hampers the comparison with other studies. The perfect model has a sensitivity and specificity of 1.0. Lower values for sensitivity and specificity result in overestimation or underestimation of the risk for persistent complaints. Our model is not perfect, but does make it possible to identify groups of patients with a significantly higher risk of persistent complaints. This facilitates future research on the prevention of chronicity in populations at high risk of developing persistent complaints. The possible influence of treatment on prognosis in the overall population is evaluated by us in another study.[30]

The three RCTs from which we derived our data had similar eligibility criteria, except for one aspect: duration of complaints at baseline. One trial included only patients with chronic complaints (> 3 months),[12] whereas another included only patients with subacute complaints (4-12 weeks).[13] However, this did not affect the model, because duration of complaints was introduced as a variable in the analysis.

Inclusion of literature-based candidate variables was limited by the variables obtained in the three RCTs. For this reason the following variables with a possible predictive value were left out of the analysis: change in neck pain in previous 2 weeks, treatment expectations and daily cycling.[9 10]

The TAMPA-scale for kinesiophobia was not obtained in one of the RCTs from which our data originated,[11] resulting in 43% missing values for this variable after combination of the data from the three RCTs (see Table 1). The possible loss of information was taken care of by the multiple imputation procedure, which requires the Missing At Random assumption, i.e. missings can be explained by the available data in the dataset.[23] This latter statement is acceptable in our study, because there was no specific reason in the trial to exclude the TAMPA-scale.[11] Furthermore, most variables in our study have shown to be related to neck pain prognosis, which means that they can be used in the imputation model to estimate the missing TAMPA values. For other variables the proportion of missing items was $\leq$ 5% and there was no reason to suspect this data to be Missing Not At Random.

Physical characteristics were not included in our analysis. None of the available studies on predictors considered the predictive value of physical characteristics.[5 7-10] This is probably one of the reasons that two of the RCTs from which we derived our data did not obtain any physical characteristics, which made it impossible to include such characteristics in our analysis.

To assure optimal statistical strength we refrained from categorizing continuous variables.[31 32] However, duration of neck pain at baseline was not obtained as a continuous variable in one RCT and was therefore divided into three strata for the other trials as well: < 1 month, 1-3 months, and > 3 months.

Global perceived recovery is a frequently used outcome measure. Although the reliability and validity of perceived recovery are questioned,[33] we prefer this outcome over the measures for pain and function (e.g. NRS-11, Neck Disability Index).[16 34] The reason for this is that it has the huge advantage for clinical practice of reflecting the patient's personal assessment of overall improvement, instead of only one aspect of the complaint (e.g. pain or function).

The squared coefficients of non-linear continuous predictors were retained until the model content was selected, because an incorrect assumption of linearity during the selection procedure might lead to a misspecified model in which a relevant variable is not included.[32 35] However, the squared coefficients were removed prior to constructing the score chart, because utilisation of quadratic terms in a score chart is impractical; they lead to a calculation with squares and larger numbers. Removal of the squared coefficients decreased the discriminative ability (AUC) with only 0.01 (from 0.67 to 0.66).

The use of multiple RCTs with a similar setting makes us confident that our study population gives an accurate reflection of the Dutch primary care population. External validation shows that our model performs equally in a sample from the primary care population in the United Kingdom, despite the different distribution of characteristics. This makes it likely that our model is generalizable to adults in primary care populations in Western society. The sample size of the external study population (n=315) was adequate for external validation.[36]

The clinical value of our study is that it provides a tool for physicians to estimate the risk of persistent complaints in patients with non-specific neck pain in primary care, based on easily obtainable data at baseline. The discriminative ability is moderate, which is reflected by the small percentage of patients in the two extreme risk categories. However, the calibration and fit of the model are good, and the score chart shows significantly better for every patient if he/she will have persistent complaints, than estimates for the overall population. This makes it informative for patients and physicians. For researchers, identification of high risk groups facilitates research on effectiveness of interventions to prevent persisting complaints.

The low explained variation does not change the clinical applicability of the model in terms of discrimination in our study, but indicates that there are still factors missing that can improve the model. Therefore, future research should examine whether the model should be extended by variables that were not included in our study (e.g. physical characteristics).

CONCLUSION

We developed and externally validated a score chart to estimate the probability of persistent complaints at 6 months follow-up for patients with non-specific neck pain in primary care. The chart has a moderate discriminative ability, an adequate calibration, a good fit, and a low explained variation in both the development and external population. The score chart predicts significantly better for every patient whether he/she will develop persistent complaints, than estimates for the overall population. This makes the score chart informative for patients and physicians and facilitates future research in patients with high risk of developing chronic neck pain.

REFERENCES

1. Bovim G, Schrader H, Sand T. Neck pain in the general population. Spine 1994;19:1307-9.
2. Picavet HS, Schouten JS. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. Pain 2003;102:167-78.
3. Palmer KT, Walker-Bone K, Griffin MJ, Syddall H, Pannett B, Coggon D, et al. Prevalence and occupational associations of neck pain in the British population. Scand J Work Environ Health 2001;27:49-56.
4. Luime JJ, Koes BW, Miedem HS, Verhaar JA, Burdorf A. High incidence and recurrence of shoulder and neck pain in nursing home employees was demonstrated during a 2-year follow-up. J Clin Epidemiol 2005;58:407-13.
5. Cote P, Cassidy JD, Carroll LJ, Kristman V. The annual incidence and course of neck pain in the general population: a population-based cohort study. Pain 2004;112:267-73.
6. Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. Eur J Pain 2006;10:287-333.
7. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain 1998;77:1-13.
8. Hoving JL, de Vet HC, Twisk JW, Deville WL, van der Windt D, Koes BW, et al. Prognostic factors for neck pain in general practice. Pain 2004;110:639-45.
9. Kjellman G, Skargren E, Oberg B. Prognostic factors for perceived pain and function at one-year follow-up in primary care patients with neck pain. Disabil Rehabil 2002;24:364-70.
10. Hill J, Lewis M, Papageorgiou AC, Dziedzic K, Croft P. Predicting persistent neck pain: a 1-year follow-up of a population cohort. Spine 2004;29:1648-54.
11. Hoving JL, Koes BW, de Vet HC, van der Windt DA, Assendelft WJ, van Mameren H, et al. Manual therapy, physical therapy, or continued care by a general practitioner for patients with neck pain. A randomized, controlled trial. Ann Intern Med 2002;136:713-22.
12. Vonk F, Verhagen AP, Twisk JW, Koke AJ, Luiten MW, Koes BW. Effectiveness of a behaviour graded activity program versus conventional exercise for chronic neck pain patients. Eur J Pain 2009;13:533-41..
13. Pool JJ, Ostelo RW, Koke AJ, Bouter LM, de Vet HC. Comparison of the effectiveness of a behavioral graded activity program and manual therapy in patients with sub-acute neck pain: design of a randomized clinical trial. Man Ther 2006;11:297-305.
14. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.
15. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373-9.
16. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
17. The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.
18. Swinkels-Meewisse EJ, Swinkels RA, Verbeek AL, Vlaeyen JW, Oostendorp RA. Psychometric properties of the Tampa Scale for kinesiophobia and the fear-avoidance beliefs questionnaire in acute low back pain. Man Ther 2003;8:29-36.
19. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-87.

20. Altman D. Practical statistics for medical research. London: Chapman and Hall, 1991.
21. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19:1059-79.
22. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23:2567-86.
23. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087-91.
24. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092-101.
25. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.
26. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley, 2000.
27. Poses RM, Cebul RD, Centor RM. Evaluating physicians' probabilistic judgments. Med Decis Making 1988;8:233-40.
28. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2007.
29. Dziedzic K, Hill J, Lewis M, Sim J, Daniels J, Hay EM. Effectiveness of manual therapy or pulsed shortwave diathermy in addition to advice and exercise for neck disorders: a pragmatic randomized controlled trial in physical therapy clinics. Arthritis Rheum 2005;53:214-22.
30. Schellingerhout JM, Verhagen AP, Heymans MW, Pool JJ, Vonk F, Koes BW, et al. Which subgroups of patients with non-specific neck pain are more likely to benefit from spinal manipulation therapy, physiotherapy, or usual care? Pain 2008;139:670-80.
31. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127-41.
32. Schellingerhout JM, Heymans MW, de Vet HC, Koes BW, Verhagen AP. Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain. J Clin Epidemiol 2009;62:868-74.
33. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869-79.
34. Bolton JE. Sensitivity and specificity of outcome measures in patients with neck pain: detecting clinically significant improvement. Spine 2004;29:2410-7.
35. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society. Series A (Statistics in Society) 1999;162(1):71-94.
36. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58:475-83.

APPENDIX A – PRACTICAL EXAMPLE SCORE CHART

John Doe, a 40-year old blue-collar worker with neck pain radiating to the shoulder, wants to know his expected probability of persistent complaints. He never had complaints of his neck before, the complaints were not preceded by a trauma, he has no low back pain, but the neck pain is accompanied by a headache. He rates the current severity of pain as 4 on the 11-point scale and his health status as 85mm on the 100mm VAS.

This results in the following score:

			Score
Age	+ 7	/ 10 yr	28
Accompanying low back pain	+ 21		0
Traumatic cause neck complaints	+ 6		0
Health status (scale 0-100)	- 3	/ 25 points	-10
Accompanying headache	+ 5		5
No accompanying headache			
Radiation of pain to elbow/shoulder	- 14		
Previous neck complaints	+ 13		
Paid employment	+ 9		
Pain intensity (scale 0-10)	- 1	/ point	
Accompanying headache			
Radiation of pain to elbow/shoulder	- 4		-4
Previous neck complaints	+ 4		0
Paid employment	- 11		-11
Pain intensity (scale 0-10)	+ 2	/ point	8
		Total score	16

The total score of 16 points implies that the expected probability is 20-40% that he still has neck complaints at 6 months from the initial consultation session and is significantly more likely to recover (score < 35) than the "average" patient.

Chapter 4

Which subgroups of patients with non-specific neck pain are more likely to benefit from spinal manipulation therapy, physiotherapy, or usual care?

Published as:

Schellingerhout JM, Verhagen AP, Heymans MW, Pool JJ, Vonk F, de Vet HC, Koes BW. Which subgroups of patients with non-specific neck pain are more likely to benefit from spinal manipulation therapy, physiotherapy, or usual care? Pain 2008;139:670-80

ABSTRACT

Objective

To identify subgroups of patients with non-specific neck pain who are more likely to benefit from either physiotherapy, spinal manipulation therapy, or usual care, on the short- and long-term.

Methods

Data of 3 recently finished randomized controlled trials, with similar design and setting, were combined. The combined study population consisted of 329 patients with non-specific neck pain in an adult (18 –70 years) primary care population in the Netherlands. The primary outcome measure was global perceived recovery and was measured at the end of the treatment period and after 52 weeks of follow-up. Fourteen candidate variables were selected for the analysis. Predictors were identified by multivariable logistic regression analysis and were tested for interaction with treatment. Based on the multivariable models with interaction-terms a decision model for treatment choice was developed.

Results

The analysis revealed three predictors for recovery of which the effect is modified by treatment: pain intensity (0-10 scale) in the short-term model, age and (no) accompanying low back pain in the long-term model. With these predictors a clinically relevant improvement in recovery rate (up to 25% improvement) can be established in patients receiving a tailored instead of a non-advised treatment.

Conclusion

We identified three characteristics that facilitate a deliberate treatment-choice, to optimize benefit of treatment in patients with non-specific neck pain: age, pain intensity, and (no) accompanying low back pain.

INTRODUCTION

Neck pain is one of the most common musculoskeletal disorders, with an estimated point prevalence of 5.9 to 22.2% and a 1-year cumulative incidence of 14.6-17.9% in adults in the general population.[1-3]

There are many known specific causes for neck pain (e.g. herniated disc, rheumatic disease, malignancy), but most of the episodes of neck pain are of unknown origin, usually referred to as non-specific neck pain.[4]

Most cases of non-specific neck pain are treated in primary care by general practitioners, with as most frequently used interventions: a "wait and see" policy, referral for physiotherapy, and referral for spinal manipulation therapy.[5] Systematic reviews show that there is a positive effect of physiotherapy and spinal manipulation therapy in comparison to placebo or watchful waiting, in patients with non-specific neck pain, but these effects are relatively small.[6-8] Heterogeneity of the included study populations with non-specific neck pain might be a reason for the small effect. It could well be that certain subgroups within this population have a larger benefit of one of these treatments due to their prognostic status.

In the past, some characteristics have shown to be related to prognosis in patients with non-specific neck pain, but none of these studies evaluated whether the effect of these prognostic factors was modified by treatment.[1 9-12] If there are prognostic factors of which the effect varies depending on treatment, that would be of great clinical value. It would facilitate a deliberate treatment choice based on the optimal probability of recovery of an individual patient.

Therefore, we decided to identify patient characteristics predictive for recovery from non-specific neck pain and to test if their effect on prognosis is modified by treatment. Characteristics that interact with treatment will be used to develop a decision-making algorithm for treatment choice.

So the purpose of this study is to develop a decision model, based on patient characteristics, that points out which subgroups of patients with non-specific neck pain are more likely to benefit from either physiotherapy, spinal manipulation therapy, or usual care.

METHODS

Data collection

Data from three recently finished randomized controlled trials (RCTs) were combined.[13-15] These RCTs were all carried out in a primary care setting in The Netherlands and were similar in design. The assigned intervention was usual care (n=64), physiotherapy (n=130), spinal manipulation therapy (n=135), or a behavioral graded activity program (n=139). The patients treated with a behavioral graded activity (BGA) program were left out, because BGA is an infrequently used treatment and requires an extensive additional training.

Study population

The three RCTs had similar selection criteria and consisted of an adult (18-70 years) primary care population with non-specific neck pain. Non-specific neck pain was defined as neck pain without a known pathological basis. People with a specific disorder (e.g. herniated disc, neurological disorder, rheumatological disorder, malignancy, infection, or fracture) were excluded from the study populations, except for patients with whiplash-associated disorders.[13-15]

Treatment protocols

Physiotherapy: Consisted of active exercises, with the aim to improve strength or range of motion. The exercises could be preceded by, or combined with, manual traction or stretching, or massage. Techniques like spinal manipulation and mobilisation were excluded from treatment.[13-15] In one study physiotherapeutic applications (e.g. interferential current or heat applications) could also precede treatment.[13] The program consisted of 30-minute sessions, with a maximum of 18 sessions.[13 15]

Spinal manipulation therapy: Consisted of several mobilisation techniques applied at the cervical spine, with the aim to restore function and relieve pain. The mobilisation techniques consisted of low-velocity passive movements within or at the limit of joint range of motion. High-velocity thrust techniques in the spinal region were not used. The program consisted of 30- to 45-minute sessions, with a maximum of 6 sessions.[13-14]

Usual care: Consisted of information about prognosis and advice on self-care from the general practitioner. Patients also received an educational booklet containing ergonomic advice and exercises to improve strength and function. Medication, including paracetamol and NSAIDs, were prescribed if necessary. Follow-up visits were optional.[13]

Outcome measures

The outcome measure (self reported) global perceived recovery was measured on a 6- or 7-point ordinal Likert scale (0="completely recovered", 1="much improved", 2="slightly improved", 3= "no change", 4="slightly worsened", 5="much worsened" , and for the 7-point scale: 6="worse than ever").[16] The outcome was dichotomized into "recovered" and "not recovered", with "recovered" defined as "completely recovered" or "much improved".[13-15]

Since differences in effect between treatments tend to change over time, we decided to develop a separate model for the short- and long-term. For the short-term we measured the outcome at the end of the treatment period (6-9 weeks) and for the long-term at 52 weeks of follow-up.

Candidate variables

As potential predictive variables for recovery we chose sociodemographic variables and clinical characteristics that can easily be obtained by a physician at the first consultation session with a patient.

To comply with the rule of at least ten events per variable in the analysis, we had to restrict the number of candidate variables to fourteen.[17] Nine variables were based on the most recently published systematic review[9] and prospective studies[1 10-12] on prognostic factors for non-specific neck pain. Another five available variables were added to complete the set of fourteen variables. The sociodemographic variables are: age, gender, level of education, treatment preference of the patient, and employment status. The clinical characteristics are: duration of neck pain at baseline, previous episode of neck complaints, pain intensity at baseline (on a 11-point Numerical Rating Scale (NRS-11)), cause of neck pain (trauma/non-trauma), concomitant headache, concomitant low back pain, concomitant dizziness, treatment, and radiation of the pain to the elbow or shoulder.

The possible values/categories of the candidate variables are displayed in Table 1.

To assure optimal statistical strength we refrained from categorizing the continuous variables into two or multiple categories.[18] Duration of neck pain at baseline was not obtained as a continuous variable in one RCT[14] and was therefore divided into three strata: < 1 month, 1-3 months, > 3 months.

Statistical analysis

Candidate variables were checked for their univariable association with the outcome through univariable logistic regression analysis. Continuous predictors were checked

for linearity by adding quadratic terms to the model and testing them for statistical significance ($p<0.05$).[19] This revealed no non-linear relations.

To develop our model we performed multivariable backward stepwise logistic regression analysis. Initially all candidate variables were included. The variables with the highest p-value were removed one-by-one (Wald test), until all remaining variables had a p-value < 0.157 (Akaike Information Criterion).[20] This p-value is regarded suitable, because of our relatively small sample size.[21] Subsequently the predictors included in the model were checked for interaction with treatment by introducing interaction-terms into the model and evaluating if they contributed significantly to the model.[19]

Imputation of missing values in the data was carried out by multiple imputation using all observed information.[22-23] A total of 5 imputed databases were created.[24] Selection of variables was performed in each dataset. If a variable was selected in at least two out of five imputed databases as a predictor, it was included in the model.

To develop the final model the regression coefficients and standard errors of the predictors and interaction-terms were averaged over the 5 imputed datasets, according to Rubin.[24]

The internal validity of the final model was determined by a bootstrapping procedure with 200 replications. In each replication a random sample from the original dataset is drawn with replacement. The stepwise selection process is repeated in each replication dataset. Subsequently, the coefficients estimated in the replication dataset are compared to those in the final model. This results in a so-called shrinkage factor that reflects the amount of overoptimism of the model.[19 25] Overoptimism means that the coefficients in the final model will on average be smaller if this study is repeated in a similar population. The intercept and coefficients in the final model were adjusted for overoptimism with the shrinkage factor.

The general performance of the final model was checked with Nagelkerke's R^2, which estimates the explained variation of the model.[19] The discriminative ability was determined with the concordance (c) statistic, which in logistic regression is identical to the area under the receiver operating characteristics curve (AUC). The AUC represents the ability of the prediction rule to distinguish between patients that will and will not recover from neck pain and ranges from 0.5 (chance) to 1.0 (perfect discrimination).[19]

The analysis was performed using SPSS version 11.0 (SPSS Inc., Chicago IL) and R software.[26]

Construction of the decision model

Both models (i.e. for the short-term and long-term outcome) will be presented as a decision model. To develop the decision-model we will calculate the probability of recovery, including 84.3% confidence intervals (corresponds with the p-value of 0.157), for different categories of the (combination of) predictor(s) interacting with treatment.[27] This will be performed for the prognostically worst and most favorable combination of patient characteristics.

In case a continuous variable (age, pain intensity) interacts with treatment we will create age and pain score categories based on cut off-points extracted from the literature and calculate the probability of recovery for the extreme values of each category. For pain intensity (NRS-11) we will use the following categories[28] : 0-4, 5-7 and 8-10 points. For age we will use: 18-30, 31-40, 41-50, 51-60 and 61-70 years of age.[18]

A treatment (physiotherapy, manipulation therapy, or usual care) will be included in the decision model if one of the following criteria is met for the prognostically best or worst combination of patient characteristics:

The treatment results in a significantly higher probability of recovery than (one of) the other treatments (i.e. the estimates for both treatments lie outside each other's 84.3% confidence interval) for a specific combination of predictors. For continuous variables this has to be the case for both extreme values of a category (e.g. a significant higher probability of recovery at both 31 and 40 years of age, for the category 31-40 years).

The treatment doesn't result in a significantly lower probability of recovery compared with both other treatments. For continuous variables this has to be the case for one of the extreme values of a category.

Since clinicians don't tend to make decisions based on two models, we will also combine both decision models into one model. Treatment preference in this model is based on corresponding treatment suggestions for specific subgroups in the separate models.

In case the short- and long-term model have no overlapping preferred treatment(s) for a specific subgroup, we will include all suggested treatments. For example, if for a certain subgroup solely physiotherapy is suggested for the short-term and solely usual care for the long-term, than both will be included in the combined model.

Performance of the decision model

To estimate the clinical value of our models we will compare the recovery rates in our study population of those treated as suggested in the model, with those receiving a non-advised treatment. The differences in recovery rate will be expressed in absolute differences.

Table 1 – Baseline characteristics and number of missing values

Variable	MT		PT		UC		Overall	
	n=135	missing (%)	n=130	missing (%)	n=64	missing (%)	n=329	missing (%)
Age, in years (mean ± sd)	45.2 ± 11.7	0 (0%)	45.6 ± 12.4	0 (0%)	45.9 ± 10.5	0 (0%)	45.5 ± 11.7	0 (0%)
Gender (male, %)	54 (40%)	0 (0%)	46 (35%)	0 (0%)	28 (44%)	0 (0%)	128 (39%)	0 (0%)
Level of education		2 (1%)		5 (4%)		0 (0%)		7 (2%)
High	48 (36%)		30 (23%)		17 (27%)		95 (29%)	
medium	54 (40%)		59 (45%)		33 (51%)		146 (44%)	
Low	31 (23%)		36 (28%)		14 (22%)		81 (25%)	
Neck pain in the past (yes, %)	81 (60%)	0 (0%)	86 (66%)	3 (2%)	46 (72%)	0 (0%)	213 (65%)	3 (1%)
Duration current episode		0 (0%)		13 (10%)		0 (0%)		13 (4%)
< 1 month	18 (13%)		17 (13%)		23 (36%)		58 (18%)	
1-3 months	99 (73%)		26 (20%)		29 (45%)		154 (47%)	
> 3 months	18 (13%)		74 (57%)		12 (19%)		103 (31%)	
Radiating pain[1] (yes, %)	95 (70%)	0 (0%)	75 (59%)	3 (2%)	33 (52%)	0 (0%)	203 (62%)	3 (1%)
Cause neck pain (trauma, %)	13 (10%)	0 (0%)	23 (18%)	3 (2%)	9 (14%)	0 (0%)	45 (14%)	3 (1%)
Treatment preference		0 (0%)		3 (2%)		0 (0%)		3 (1%)
no	80 (59%)		86 (66%)		38 (60%)		204 (62%)	
yes, physiotherapy	21 (16%)		16 (12%)		11 (17%)		48 (15%)	
yes, spinal manipulation therapy	34 (25%)		25 (19%)		15 (23%)		74 (22%)	
Employment status (employed, %)	100 (74%)	2 (1%)	86 (66%)	3 (2%)	46 (72%)	0 (0%)	232 (71%)	5 (2%)
Headache[1] (yes, %)	88 (65%)	0 (0%)	90 (71%)	3 (2%)	49 (77%)	0 (0%)	227 (69%)	3 (1%)
Dizziness[1] (yes, %)	37 (27%)	0 (0%)	47 (37%)	3 (2%)	26 (41%)	0 (0%)	100 (30%)	3 (1%)
Low backpain[1] (yes, %)	17 (13%)	0 (0%)	46 (36%)	3 (2%)	12 (19%)	0 (0%)	75 (23%)	3 (1%)
Pain intensity, NRS 0-10 (mean ± sd)	5.5 ± 2.0	1 (1%)	5.8 ± 1.9	2 (2%)	6.3 ± 2.1	0 (0%)	5.8 ± 2.0	3 (1%)
Perceived recovery								
short-term (yes, %)	86 (65%)	2 (1%)	56 (46%)	9 (7%)	23 (36%)	0 (0%)	165 (52%)	11 (3%)
long-term (yes, %)	97 (74%)	4 (3%)	62 (51%)	9 (7%)	36 (56%)	0 (0%)	195 (62%)	13 (4%)

MT = manipulation therapy, PT = physiotherapy, UC = usual care

[1] reported to be present at baseline

Table 2 – Uni- and multivariable association of candidate variables with short-term and long-term recovery

Variable	Short-term					Long-term				
	Univariable		Multivariable			Univariable		Multivariable		
	OR	95% CI	OR	95% CI	no.[1]	OR	95% CI	OR	95% CI	no.[1]
Treatment					5					5
usual care	ref.		ref.			ref.		ref.		
physiotherapy	1.58	0.84 – 2.96	1.78	0.88 – 3.61		0.84	0.45 – 1.54	1.15	0.58 – 2.29	
spinal manipulation therapy	3.27	1.76 – 6.10	3.20	1.60 – 6.39		2.21	1.18 – 4.14	2.01	1.01 – 3.99	
Age	0.99	0.97 – 1.01			1	1.00	0.98 – 1.02	1.02	0.99 – 1.04	2
Gender (female=1)	1.08	0.68 – 1.70			0	1.23	0.78 – 1.95	1.56	0.93 – 2.62	4
Level of education					0					2
high	ref.					ref.		ref.		
middle	0,88	0.52 – 1.49				0.59	0.34 – 1.03	0.61	0.34 – 1.10	
low	0,71	0.39 – 1.29				0.70	0.36 – 1.33	0.76	0.38 – 1.54	
Neck pain in the past (yes=1)	0,93	0.58 – 1.51			0	0.76	0.47 – 1.25			1
Duration current episode					5					5
< 1 month	ref.		ref.			ref.		ref.		
1-3 months	0.65	0.34 – 1.22	0.42	0.21 – 0.87		0.82	0.41 – 1.64	0.65	0.31 – 1.36	
> 3 months	0.46	0.23 – 0.93	0.39	0.18 – 0.86		0.35	0.18 – 0.66	0.37	0.17 – 0.82	
Radiating pain (yes=1)	2.08	1.31 – 3.30	2.20	1.33 – 3.65	5	1.11	0.69 – 1.79			0
Cause neck pain (trauma=1)	0.90	0.44 – 1.82			0	0.60	0.31 – 1.16			0
Treatment preference					0					0
no	ref.					ref.				
yes, physiotherapy	0.94	0.49 – 1.78				1.10	0.57 – 2.13			
yes, spinal manipulation therapy	1.01	0.59 – 1.73				0.83	0.48 – 1.43			
Employment status (employed=1)	1.40	0.87 – 2.28			0	1.46	0.89 – 2.39	1.71	0.91 – 3.23	2
Headache (yes=1)	0.61	0.37 – 1.00	0.67	0.39 – 1.16	3	0.71	0.43 – 1.16			0
Dizziness (yes=1)	0.97	0.61 – 1.53			0	0.77	0.48 – 1.23			0
Low backpain (yes=1)	0.92	0.54 – 1.56			0	0.49	0.29 – 0.84	0.61	0.35 – 1.09	4
Pain intensity at baseline (NRS 0-10)	0.87	0.77 – 0.97	0.87	0.77 – 0.99	5	0.87	0.77 – 0.97	0.88	0.77 – 0.99	5

ref. = reference category, 95%-CI = 95% confidence interval

OR = odds ratio, an OR > 1 reflects a higher probability of recovery and an OR < 1 a lower probability of recovery, compared to the reference category

[1] number of imputed databases in which the variable is selected as predictor

RESULTS

The baseline characteristics of the candidate variables and number of patients that recovered for each treatment group and the overall group are shown in Table 1.

The univariable and multivariable association of the different variables with the outcome are presented in Table 2. For the short-term there is a significant interaction of treatment with pain intensity at baseline and for the long-term a significant interaction with accompanying low back pain and age (see Table 3).

Table 3 – Interaction of predictors with treatment (p-values)

Interaction term	UC vs. PT	UC vs. MT	PT vs. MT
Short-term			
Duration of complaints			
1-3 months	0.719	0.513	0.797
> 3 months	0.589	0.924	0.643
Radiating pain	0.515	0.715	0.729
Pain intensity	**0.043**	**0.108**	0.573
Accomp. headache	0.590	0.211	0.375
Long-term			
Duration of complaints			
1-3 months	0.579	0.919	0.532
> 3 months	0.861	0.486	0.571
Pain intensity	0.328	0.552	0.690
Education level			
medium	0.944	0.982	0.960
low	0.216	0.439	0.594
Employment status	0.268	0.232	0.847
Age	**0.029**	**0.047**	0.876
Sexe	0.827	0.920	0.719
Accomp. low back pain	0.881	0.298	**0.140**

UC = usual care, PT = physiotherapy, MT = manipulation therapy
Bold = selected as interaction term in the final model ($p < 0.157$)

The bootstrapping procedure resulted in a shrinkage factor of 0.826 for the short-term and 0.725 for the long-term regression model. Combination of predictors and interaction terms results in the logistic regression equations, with shrinkage adjusted coefficients, as presented in Appendix A.

A clinical example of how to calculate the probability of recovery by using one of the equations is also presented in Appendix A.

The short-term model has an adjusted explained variation of 11.7% and a discriminative ability (AUC) of 0.71. The long-term model has an adjusted explained variation of

Table 4 – Probability of recovery for different categories of pain intensity in the short-term model

	Best profile[1]							Worst profile[2]					
Score NRS-11		UC		PT		MT		UC		PT		MT	
Interval	Score[3]	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI
0-4 points	0	48%	22-75%	**90%**	79-95%	**91%**	81-96%	14%	5-35%	**61%**	43-77%	**64%**	44-80%
(n=87)	4	57%	41-72%	**80%**	70-87%	**85%**	77-91%	19%	11-31%	**42%**	32-52%	**51%**	39-64%
5-7 points	5	59%	45-72%	**76%**	66-84%	**83%**	75-89%	21%	13-31%	**37%**	29-46%	**47%**	36-59%
(n=171)	7	63%	51-74%	69%	56-79%	**80%**	70-87%	24%	16-34%	28%	21-37%	**42%**	31-54%
8-10 points	8	65%	52-77%	64%	50-76%	**77%**	65-86%	25%	16-37%	25%	17-34%	**38%**	26-51%
(n=71)	10	69%	51-83%	55%	37-71%	**73%**	57-85%	29%	16-47%	18%	10-30%	**33%**	20-49%

UC = usual care, PT = physiotherapy, MT = manipulation therapy, **bold** = significantly higher probability of recovery than (one of) the other treatments

[1] Complaints for <1 month, radiation of pain to shoulder/elbow, no accompanying headache

[2] Complaints for >3 months, no radiation of pain to shoulder/elbow, accompanying headache

[3] The probability of recovery for that specific score on the NRS-11

Table 5 - Probability of recovery for different combinations of interactions in the long-term model

Combinations		Best profile[1]						Worst profile[2]					
		UC		PT		MT		UC		PT		MT	
	Age[3]	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI	prob.	84.3%-CI
Accomp. low back pain													
18-30 years	18	**91%**	75-97%	74%	55-87%	72%	51-86%	**38%**	16-67%	15%	7-30%	14%	6-31%
(n=7)	30	87%	72-95%	78%	63-88%	76%	58-87%	31%	14-54%	19%	10-32%	17%	8-33%
31-40 years	31	87%	72-94%	78%	63-88%	76%	59-87%	30%	14-53%	19%	10-32%	17%	8-33%
(n=18)	40	84%	68-93%	81%	68-90%	79%	63-89%	25%	12-44%	22%	13-35%	19%	9-35%
41-50 years	41	83%	68-92%	82%	69-90%	79%	64-89%	24%	12-43%	22%	13-35%	20%	10-35%
(n=18)	50	79%	61-90%	84%	72-92%	81%	67-90%	20%	10-36%	25%	16-38%	22%	11-39%
51-60 years	51	79%	60-90%	84%	72-92%	82%	68-91%	20%	10-36%	26%	16-39%	22%	11-39%
(n=17)	60	74%	52-89%	86%	75-93%	84%	70-92%	16%	7-32%	29%	19-43%	25%	13-43%
61-70 years	61	74%	51-89%	87%	75-93%	84%	70-92%	15%	7-31%	30%	19-44%	26%	13-44%
(n=15)	70	69%	41-88%	**89%**	77-95%	86%	71-94%	12%	5-29%	**34%**	21-50%	29%	14-49%
No low back pain													
18-30 years	18	**92%**	81-97%	77%	61-88%	86%	73-93%	**43%**	21-70%	18%	9-33%	28%	14-48%
(n=26)	30	89%	80-95%	81%	69-89%	88%	79-94%	35%	19-56%	22%	12-35%	32%	18-50%
31-40 years	31	89%	79-95%	81%	69-89%	88%	79-94%	35%	19-54%	22%	12-36%	33%	19-50%
(n=68)	40	86%	77-93%	84%	73-91%	90%	82-94%	29%	17-45%	25%	15-38%	36%	23-52%
41-50 years	41	86%	76-92%	84%	74-91%	90%	82-94%	29%	17-44%	25%	16-38%	36%	23-52%
(n=78)	50	83%	71-90%	86%	77-92%	**91%**	84-95%	24%	14-37%	29%	19-42%	**40%**	26-56%
51-60 years	51	82%	70-90%	86%	77-92%	**91%**	84-95%	23%	14-36%	29%	19-42%	**41%**	27-56%
(n=54)	60	78%	62-89%	88%	79-94%	**92%**	86-96%	19%	10-32%	**33%**	21-47%	**44%**	29-61%
61-70 years	61	78%	63-88%	**89%**	79-94%	**93%**	86-96%	19%	10-32%	**33%**	22-48%	**45%**	30-61%
(n=28)	70	73%	50-88%	**90%**	80-95%	**94%**	87-97%	15%	7-30%	**37%**	23-54%	**49%**	31-66%

UC = usual care, PT = physiotherapy, MT = manipulation therapy, **bold** = significantly higher probability of recovery than (one of) the other treatments

[1] Female, complaints for <1 month, employed, 0 points on NRS-11, high level of education

[2] Male, complaints for >3 months, unemployed, 10 points on NRS-11, medium level of education

[3] The probability of recovery at that specific age

8.6% and a discriminative ability (AUC) of 0.72. The explained variation is rather low, but both c-statistics indicate a good discriminative ability of the model.

Decision-model

The probability of recovery on the short-term for each treatment, and for the different categories of pain intensity, is presented in Table 4. Table 4 shows that the probability of recovery for patients receiving physiotherapy or spinal manipulation therapy decreases with an increasing score on the NRS-11, whereas the benefit of usual care increases with an increasing score on the NRS-11.

The probability of recovery on the long-term for each treatment, and for the different combinations of age and accompanying low back pain, is presented in Table 5. Table 5 shows that with increasing age the probability of recovery decreases for patients treated with usual care, whereas it increases for physiotherapy and manipulation therapy. However, the distinguishing effect of age is almost neutralised by the co-existence of low back pain: none of the treatments is superior at both extremes of an age category. If low back pain is absent the modifying effect of age becomes visible: people of higher age benefit more from physiotherapy and manipulation therapy than usual care.

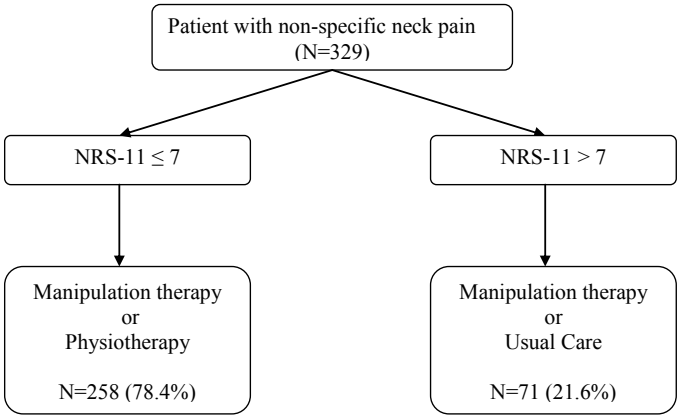

Figure 1 – Decision model short-term

The decision-models for the short- and long-term are presented in Figure 1 and 2, respectively. Both also provide the proportion of patients in our study population with a profile leading to the suggested treatment(s). The proportions show that it is possible to optimize the probability of recovery, by allocating patients to a specific treatment, in all patients on the short-term and in 24.9% of the patients on the long-term.

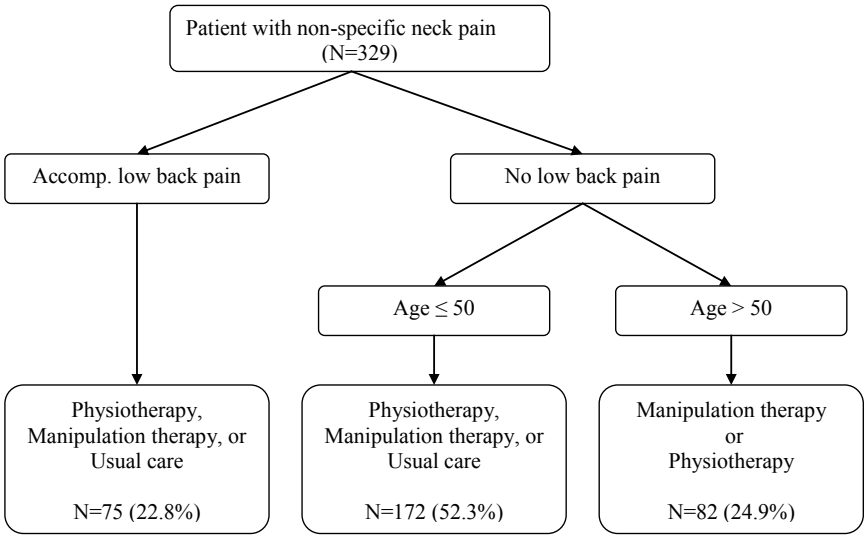

Figure 2 – Decision model long-term

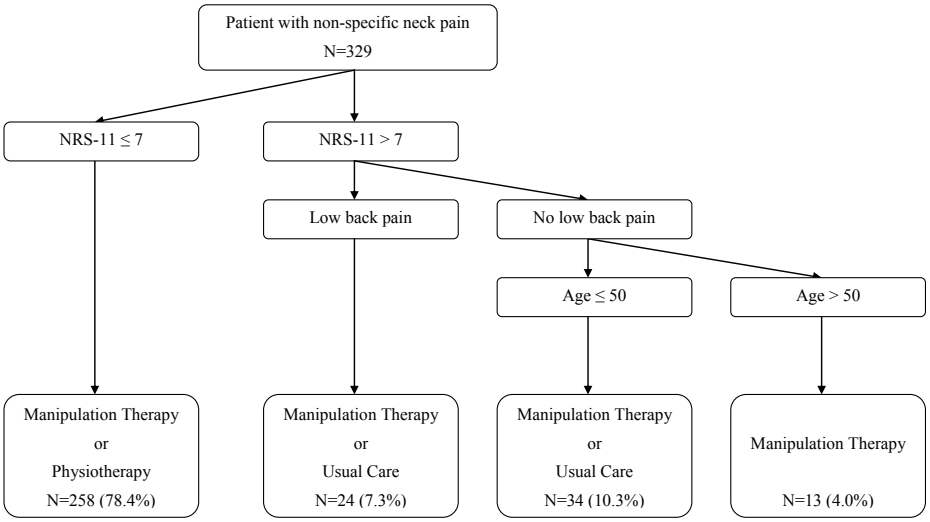

Figure 3 – Short- and long-term model combined

Combination of the short- and long-term model leads to the decision-model as presented in Figure 3. This model shows that manipulation therapy is a good treatment-choice for everyone with non-specific neck pain. Next to this, it shows that

physiotherapy is beneficial for patients with a medium or low pain intensity. Patients with a high pain intensity and of younger age benefit from usual care.

Model performance

The short-term recovery rate of patients in our study population receiving a treatment not advised by the model is 32.4%. For patients receiving a treatment suggested by the model the recovery rate is 57.6%. So, application of the short-term model will improve the probability of recovery with 25.2% (95% confidence interval: 12.9 – 37.6%) in patients receiving a non-advised treatment.

The long-term recovery rate of patients receiving a non-advised treatment is 50.0%, whereas it is 62.4% in patients receiving a tailored treatment. So, the long-term model improves the recovery rate of patients receiving a non-advised treatment with 12.4% (95% confidence interval: -12.7 – 37.4%).

The combined model shows similar differences in recovery rate. Short-term recovery increases with 26.1% (95% confidence interval: 14.0 – 38.2%) in patients receiving a non-advised treatment. Long-term recovery increases with 16.5% (95% confidence interval: 3.8 – 29.2%) in patients receiving a non-advised treatment.

DISCUSSION

The results of our study show that there are several predictors for recovery in patients with non-specific neck pain and that three of them are useful to guide treatment-choice: pain intensity at baseline for short-term recovery, and (absence of) low back pain and age for recovery on the long-term. With these predictors a clinically relevant improvement in recovery rate (up to 25% improvement) can be established in patients receiving a tailored instead of a non-advised treatment. This will result in an increase of overall recovery as well, but the exact gain is uncertain: treatment allocation in our study does not resemble daily practice, because of the randomization process.

As mentioned in the introduction, systematic reviews show only a small positive effect of physiotherapy and manipulation therapy in patients with non-specific neck pain.[6-8] We hypothesised that the contrast between treatments might be larger in subgroups of patients. Both statements are supported by our data: physiotherapy and manipulation therapy have a positive effect on prognosis, but patient characteristics have a larger impact on prognosis than treatment choice.

In the literature we identified two other decision-making algorithms for people with neck pain.[29-30] Both algorithms suggest specific within-treatment variations in patients with neck pain referred for physical therapy and are consensus based, without any validation.[29-30] Our study is situated in general practice and is based on statistical

reasoning. The differences in the suggested treatments and the different reasoning for introducing characteristics into the model make it hard to compare the models.

Due to the setting and selection criteria used in the three RCTs from which our data originated, we are confident that our study population is a good reflection of the Dutch primary care population with non-specific neck pain. However, external validation of our model is desirable to assess the usefulness in other populations.

The selection criteria were identical in the three RCTs, except for one aspect: duration of complaints at baseline. One trial included only patients with chronic complaints (> 3 months),[15] whereas another included only patients with subacute complaints (4-12 weeks).[14] However, since duration of complaints was introduced as a covariate in our analysis, this did not affect our results.

In our opinion, the advantage of using RCTs for building the model, instead of a prospective cohort study, is that it offers the possibility to introduce treatment as a covariate in the model, without the risk of biased results due to confounding by indication.

A remark could be made with regard to the selection of candidate variables. For reasons of practical applicability in daily clinical practice, we deliberately refrained from evaluating possible predictors that consist of multiple questions (e.g. questionnaires on disability, kinesiophobia, or quality of life). For this reason the Oswestry score, although suggested to be of predictive value in neck pain,[12] was left out of the analysis. Next to this, the choice of candidate variables was limited by the variables obtained in each of the three RCTs. For this reason the following variables, with a possible predictive value,[11-12] were not included in the analysis: well-being, treatment expectations, and cycling daily. Addition of the aforementioned variables might have improved the model.

Psychological factors and physical characteristics were also not included in our analysis. None of the studies on predictors found in the literature considered the predictive value of physical characteristics.[1 9-12] This is probably one of the reasons that two of the RCTs from which we derived our data did not obtain any physical characteristics, which made it impossible to include such characteristics in our analysis. Previous studies on predictors that evaluated psychological factors did not find any with a significant predictive value.[11-12]

Global perceived recovery is a frequently used outcome measure. In contrast to measures for pain and function (e.g. NRS-11, Neck Disability Index),[31-32] the reliability and validity of perceived recovery are not yet established in people with neck pain.[33] However, because of its huge advantage for clinical practice of reflecting the patient's personal assessment of overall improvement, instead of only one aspect of the complaint (e.g. pain or function), we decided to use it as our main outcome measure.

It is theoretically possible that a certain characteristic has no predictive value in itself, but that it is predictive in combination with treatment, thereby enlarging the contrast in recovery rate between the three treatments. These characteristics were not considered in our analyses, as we limited our study, for sample size reasons, to "overall" predictors of the course of neck pain.

The combined decision-model has the advantage of an easier application in clinical practice than two separate models. Based on the performance of this model in our data, it results in an equal increase in recovery rate on the short- and long-term compared to the separate models.

The value of our models for clinical practice is that they point out which subgroups of patients with non-specific neck pain benefit more from physiotherapy, spinal manipulation therapy, or usual care. By this, the models optimize the probability of recovery and provide a more deliberate treatment choice. From the different models can be concluded that all patients benefit from spinal manipulation therapy, that physiotherapy should be applied to patients with a medium to low pain intensity, and that usual care should be applied to patients with a high pain intensity and of younger age.

In case there is no preference in our decision-model for one of the included treatments we propose to start with usual care, because of the additional costs of active treatment.[34]

CONCLUSION

This study shows that it is beneficial to assign treatment based on predictors for recovery in patients with non-specific neck pain. The probability of recovery, after treatment with physiotherapy, spinal manipulation therapy, or usual care can be optimized using pain intensity at baseline for the short-term, and (absence of) low back pain and age for the long-term.

REFERENCES

1. Cote P, Cassidy JD, Carroll LJ, Kristman V. The annual incidence and course of neck pain in the general population: a population-based cohort study. Pain 2004;112:267-73.
2. Croft PR, Lewis M, Papageorgiou AC, Thomas E, Jayson MI, Macfarlane GJ, et al. Risk factors for neck pain: a longitudinal study in the general population. Pain 2001;93:317-25.
3. Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: a systematic critical review of the literature. Eur Spine J 2006;15:834-48.
4. Bogduk N. Regional musculoskeletal pain. The neck. Baillieres Best Pract Res Clin Rheumatol 1999;13:261-85.
5. Vos C, Verhagen A, Passchier J, Koes B. Management of acute neck pain in general practice: a prospective study. Br J Gen Pract 2007;57:23-8.
6. Philadelpia Panel. Philadelphia Panel evidence-based clinical practice guidelines on selected rehabilitation interventions for neck pain. Phys Ther 2001;81:1701-17.
7. Gross AR, Hoving JL, Haines TA, Goldsmith CH, Kay T, Aker P, et al. Manipulation and mobilisation for mechanical neck disorders. Cochrane Database Syst Rev 2004:CD004249.
8. Kay TM, Gross A, Goldsmith C, Santaguida PL, Hoving J, Bronfort G. Exercises for mechanical neck disorders. Cochrane Database Syst Rev 2005:CD004250.
9. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain 1998;77:1-13.
10. Hoving JL, de Vet HC, Twisk JW, Deville WL, van der Windt D, Koes BW, et al. Prognostic factors for neck pain in general practice. Pain 2004;110:639-45.
11. Hill J, Lewis M, Papageorgiou AC, Dziedzic K, Croft P. Predicting persistent neck pain: a 1-year follow-up of a population cohort. Spine 2004;29:1648-54.
12. Kjellman G, Skargren E, Oberg B. Prognostic factors for perceived pain and function at one-year follow-up in primary care patients with neck pain. Disabil Rehabil 2002;24:364-70.
13. Hoving JL, Koes BW, de Vet HC, van der Windt DA, Assendelft WJ, van Mameren H, et al. Manual therapy, physical therapy, or continued care by a general practitioner for patients with neck pain. A randomized, controlled trial. Ann Intern Med 2002;136:713-22.
14. Pool JJ, Ostelo RW, Koke AJ, Bouter LM, de Vet HC. Comparison of the effectiveness of a behavioral graded activity program and manual therapy in patients with sub-acute neck pain: design of a randomized clinical trial. Man Ther 2006;11:297-305.
15. Vonk F, Verhagen AP, Geilen M, Vos CJ, Koes BW. Effectiveness of behavioral graded activity compared with physiotherapy treatment in chronic neck pain: design of a randomized clinical trial [ISRCTN88733332]. BMC Musculoskelet Disord 2004;5:34.
16. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.
17. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373-9.
18. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127-41.
19. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-87.
20. Altman D. Practical statistics for medical research. London: Chapman and Hall, 1991.

21. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19:1059-79.
22. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092-101.
23. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087-91.
24. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.
25. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Statistics in Medicine 1990;9:1303-25.
26. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2007.
27. Sofroniou N, Hutcheson GD. Confidence intervals for the predictions of logistic regression in the presence and absence of a variance-covariance matrix. Understanding Statistics 2002;1:3-18.
28. Fejer R, Jordan A, Hartvigsen J. Categorizing the severity of neck pain: Establishment of cut-points for use in clinical and epidemiological research. Pain 2005;119:176-82.
29. Wang WT, Olson SL, Campbell AH, Hanten WP, Gleeson PB. Effectiveness of physical therapy for patients with neck pain: an individualized approach using a clinical decision-making algorithm. Am J Phys Med Rehabil 2003;82:203-18.
30. Childs JD, Fritz JM, Piva SR, Whitman JM. Proposal of a classification system for patients with neck pain. J Orthop Sports Phys Ther 2004;34:686-96.
31. Bolton JE. Sensitivity and specificity of outcome measures in patients with neck pain: detecting clinically significant improvement. Spine 2004;29:2410-7.
32. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
33. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869-79.
34. Borghouts JA, Koes BW, Vondeling H, Bouter LM. Cost-of-illness of neck pain in The Netherlands in 1996. Pain 1999;80:629-36.

APPENDIX A – LOGISTIC REGRESSION EQUATIONS

The logistic regression equation for the short-term model, with shrinkage adjusted coefficients, is:
log odds (recovery) = -0.697 + 2.253*physiotherapy + 2.348*spinal manipulation therapy – 0.653*duration (1-3 months) – 0.733*duration (>3 months) + 0.089*pain intensity + 0.611*radiating pain – 0.356*headache – 0.288*physiotherapy*pain intensity – 0.222*spinal manipulation therapy*pain intensity

and the equation for the long-term model is:
log odds (recovery) = 2.330 – 2.128*physiotherapy – 1.516*spinal manipulation therapy – 0.028*age + 0.295*gender – 0.216*low back pain – 0.256*duration (1-3 months) – 0.720*duration (>3 months) + 0.359*employment status – 0.102*pain intensity – 0.340*education level (medium) – 0.183*education level (low) + 0.043*physiotherapy*low back pain – 0.645*spinal manipulation therapy*low back pain + 0.048*physiotherapy*age + 0.046*spinal manipulation therapy*age

Calculation example
Patient X: 54 years of age, complaints for >3 months, no radiation of the pain to shoulder/elbow, accompanying headache, and a pain intensity of 7 (scale 0-10)

Treatment with physiotherapy would, on the short-term, result in a probability of recovery of:

log odds (recovery) = -0.697 + 2.253*physiotherapy + 2.348*spinal manipulation therapy - 0.653*duration (1-3 months) - 0.733*duration (>3 months) + 0.089*pain intensity + 0.611*radiating pain - 0.356*headache - 0.288*physiotherapy*pain intensity – 0.217*spinal manipulation therapy*pain intensity
= -0.697 + 2.253 * 1 + 2.348 * 0 - 0.653 * 0 - 0.733 * 1 + 0.089 * 7 + 0.611 * 0 - 0.356 * 1 - 0.288 * (1*7) - 0.222 * (0*7)
= -0.697 + 2.253 - 0.733 + 0.623 - 0.356 - 2.016 = -0.926

odds (recovery) = $e^{-0.926}$ = 0.396

probability (recovery) = 0.396 / (1 + 0.396) = 0.284 = 28.4%

Chapter 5

Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review.

Published as:

Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2011, doi:10.1007/s11136-011-9965-9

ABSTRACT

Objective

To critically appraise and compare the measurement properties of the original versions of neck-specific questionnaires.

Methods

Bibliographic databases were searched for articles concerning the development or evaluation of the measurement properties of an original version of a self-reported questionnaire, evaluating pain and/or disability, which was specifically developed or adapted for patients with neck pain. The methodological quality of the selected studies and the results of the measurement properties were critically appraised and rated using a checklist, specifically designed for evaluating studies on measurement properties.

Results

The search strategy resulted in a total of 3641 unique hits, of which 25 articles, evaluating 8 different questionnaires, were included in our study. The Neck Disability Index is the most frequently evaluated questionnaire and shows positive results for internal consistency, content validity, structural validity, hypothesis testing, and responsiveness, but a negative result for reliability. The other questionnaires show positive results, but the evidence for each measurement property is mostly limited and at least 50% of the information on measurement properties per questionnaire is lacking.

Conclusion

Our findings imply that studies of high methodological quality are needed to properly assess the measurement properties of the currently available questionnaires. Until high quality studies are available, we recommend using these questionnaires with caution. There is no need for the development of new neck-specific questionnaires until the current questionnaires have been adequately assessed.

INTRODUCTION

Several disease-specific questionnaires have been developed to measure pain and/or disability in patients with neck pain (e.g. Neck Disability Index (NDI), Neck Pain and Disability Scale (NPDS)).[1 2] In order to make a rational choice for the use of these questionnaires in clinical research and practice it is important to assess and compare their measurement properties (e.g. reliability, validity, and responsiveness).[3]

A systematic review, published in 2002, evaluated the measurement properties of several neck-specific questionnaires and showed that, except for the NDI, all questionnaires were lacking psychometric information and that comparison was therefore not possible.[4] Recent reviews show that the amount of studies evaluating measurement properties of neck-specific questionnaires has extended considerably in the past years.[5-7] However, all these reviews lack an adequate instrument to critically appraise the methodological quality of the included studies. Studies of high methodological quality are needed to guarantee appropriate conclusions about the measurement properties. Recently the "COnsensus-based Standards for the selection of health status Measurement INstruments" (COSMIN) checklist, an instrument to evaluate the methodological quality of studies on measurement properties of health status questionnaires, has become available.[8] Using the COSMIN checklist it is now possible to critically appraise and compare the quality of these studies.

A recent review of the cross-cultural adaptations of the McGill Pain Questionnaire showed that pooling of the measurement properties of different language versions results in inconsistent findings regarding the results for measurement properties, caused by differences in cultural context.[9] Since it is likely that the same accounts for the translated questionnaires in our review, we decided to evaluate them in a separate systematic review.[10]

The purpose of this study is to critically appraise and compare the measurement properties of the original version of neck-specific questionnaires.

METHODS

Search strategy

We searched the following computerized bibliographic databases: Medline (1966 to July 2010), EMbase (1974 to July 2010), CINAHL (1981 to July 2010), and PsycINFO (1806 to July 2010). We used the index terms "neck", "neck pain", and "neck injuries/injury" in combination with "research measurement", "questionnaire", "outcome assessment", "psychometry", "reliability", "validity", and derivatives of these terms.

The full search strategy used in each database is available upon request from the authors. Reference lists were screened to identify additional relevant studies.

Selection criteria

A study was included if it was a full text original article (e.g. not an abstract, review or editorial), published in English, concerning the development or evaluation of the measurement properties of an original version of a neck-specific questionnaire. The questionnaire had to be self-reported, evaluating pain and/or disability, and specifically developed or adapted for patients with neck pain.

For inclusion, neck pain had to be the main complaint of the study population. Accompanying complaints (e.g. low back pain or shoulder pain) were no reason for exclusion, as long as the main focus was neck pain. Studies considering study populations with a specific neck disorder (e.g. neurological disorder, rheumatological disorder, malignancy, infection, or fracture) were excluded, except for patients with cervical radiculopathy or whiplash associated disorder (WAD).

Two reviewers (JMS, APV) independently assessed the titles, abstracts, and reference lists of studies retrieved by the literature search. In case of disagreement between the two reviewers, there was discussion to reach consensus. If necessary, a third reviewer (HCV) made the decision regarding inclusion of the article.

Measurement properties

The measurement properties are divided over three domains: reliability, validity, and responsiveness.[11] In addition, the interpretability is described.

Reliability

Reliability is defined as the extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same questionnaire (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons on different occasions (intra-rater).[11]

Reliability contains the following measurement properties:

Internal consistency: The interrelatedness among the items in a questionnaire, expressed by Cronbach's α or Kuder-Richardson Formula 20 (KR-20).[8 11]

Measurement error: The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured, expressed by the standard error of measurement (SEM).[11 12] The SEM can be converted into the smallest detectable change (SDC).[12] Changes exceeding the SDC can be labeled as change beyond measurement error.[12] Another approach is to calculate the limits of agreement (LoA).[13] To determine the adequacy of measurement error the smallest detectable

change and/or limits of agreement is related to the minimal important change (MIC).[14] As measurement error is expressed in the units of measurements it is impossible to give one value for adequacy. However it is important that the measurement error (i.e. noise, expressed as SDC or limits of agreement) is not larger than the signal (i.e. MIC) that one wants to assess.

Reliability: The proportion of the total variance in the measurements which is due to 'true' differences between patients.[11] This aspect is reflected by the Intraclass Correlation Coefficient (ICC) or Cohen's Kappa.[3 11]

Validity

Validity is the extent to which a questionnaire measures the construct it is supposed to measure and contains the following measurement properties:[11]

Content validity: The degree to which the content of a questionnaire is an adequate reflection of the construct to be measured.[11] Important aspects are whether all items are relevant for the construct, aim, and target population and if no important items are missing (comprehensiveness).[15]

Criterion validity: The extent to which scores on an instrument are an adequate reflection of a gold standard.[11] Since a real gold standard for health status questionnaires is not available,[15] we will not evaluate criterion validity.

Construct validity is divided into three aspects:

Structural validity: The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured.[11] Factor analysis should be performed to confirm the number of subscales present in a questionnaire.[15]

Hypothesis testing: The degree to which a particular measure relates to other measures in a way one would expect if it is validly measuring the supposed construct, i.e. in accordance with predefined hypotheses about the correlation or differences between the measures.[11]

Cross-cultural validity: The degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument.[11] The cross cultural validity of neck specificity questionnaire is addressed in a separate systematic review.[10]

Responsiveness

Responsiveness is the ability of an instrument to detect change over time in the construct to be measured.[11] Responsiveness is considered an aspect of validity, in a longitudinal context.[15] Therefore, the same standards apply as for validity: the correlation between change scores of two measures should be in accordance with predefined hypotheses.[15] Another approach is to determine the area under the receiver operator characteristic curve (AUC).[15]

Interpretability

Interpretability is the degree to which one can assign qualitative meaning to quantitative scores.[11] This means that investigators should provide information about clinically meaningful differences in scores between subgroups, floor and ceiling effects, and the MIC.[15] Interpretability is not a measurement property, but an important characteristic of a measurement instrument.[11]

Quality assessment

To determine whether the results of the included studies can be trusted, the methodological quality of the studies was assessed. This step was carried out using the COSMIN checklist.[8] The COSMIN checklist consists of nine boxes with 5-18 items concerning methodological standards for how each measurement property should be assessed. Each item was scored on a 4-point rating scale (i.e. "poor", "fair", "good", or "excellent"), which is an additional feature of the COSMIN checklist (see www.cosmin.nl). An overall score for the methodological quality of a study was determined for each measurement property separately, by taking the lowest rating of any of the items in a box. The methodological quality of a study was evaluated per measurement property.

Data extraction and assessment of (methodological) quality were performed by two reviewers (JMS, CBT) independently. In case of disagreement between the two reviewers, there was discussion in order to reach consensus. If necessary, a third reviewer (HCV) made the decision.

Best evidence synthesis – levels of evidence

To summarize all the evidence on the measurement properties of the different questionnaires we synthesized the different studies by combining their results, taking the number and methodological quality of the studies and the consistency of their results into account. The possible overall rating for a measurement property is "positive",

Table 1 – Levels of evidence for the overall quality of the measurement property [17]

Level	Rating	Criteria
strong	+++ or ---	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
moderate	++ or --	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
limited	+ or -	One study of fair methodological quality
conflicting	+/-	Conflicting findings
unknown	?	Only studies of poor methodological quality

[..] = reference number, + = positive result, - = negative result

Table 2 – Quality criteria for measurement properties [18]

Property	Rating	Quality Criteria
Reliability		
Internal consistency	+	(Sub)scale unidimensional AND Cronbach's alpha(s) ≥ 0.70
	?	Dimensionality not known OR Cronbach's alpha not determined
	-	(Sub)scale not unidimensional OR Cronbach's alpha(s) < 0.70
Measurement error	+	MIC > SDC OR MIC outside the LOA
	?	MIC not defined
	-	MIC ≤ SDC OR MIC equals or inside LOA
Reliability	+	ICC / weighted Kappa ≥ 0.70 OR Pearson's r ≥ 0.80
	?	Neither ICC / weighted Kappa, nor Pearson's r determined
	-	ICC / weighted Kappa < 0.70 OR Pearson's r < 0.80
Validity		
Content validity	+	The target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be complete
	?	No target population involvement
	-	The target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete
Construct validity		
- Structural validity	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	-	Factors explain < 50% of the variance
- Hypothesis testing	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
Responsiveness		
Responsiveness	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlation with related constructs is lower than with unrelated constructs

[..] = reference number , MIC = minimal important change, SDC = smallest detectable change, LOA = limits of agreement, ICC = intraclass correlation coefficient, AUC = area under the curve
+ = positive rating, ? = indeterminate rating, - = negative rating

"indeterminate", or "negative", accompanied by levels of evidence, similarly as was proposed by the Cochrane Back Review Group (see Table 1).[16 17]

To assess whether the results of the measurement properties were positive, negative, or indeterminate, we used criteria based on Terwee et al. (see Table 2).[18]

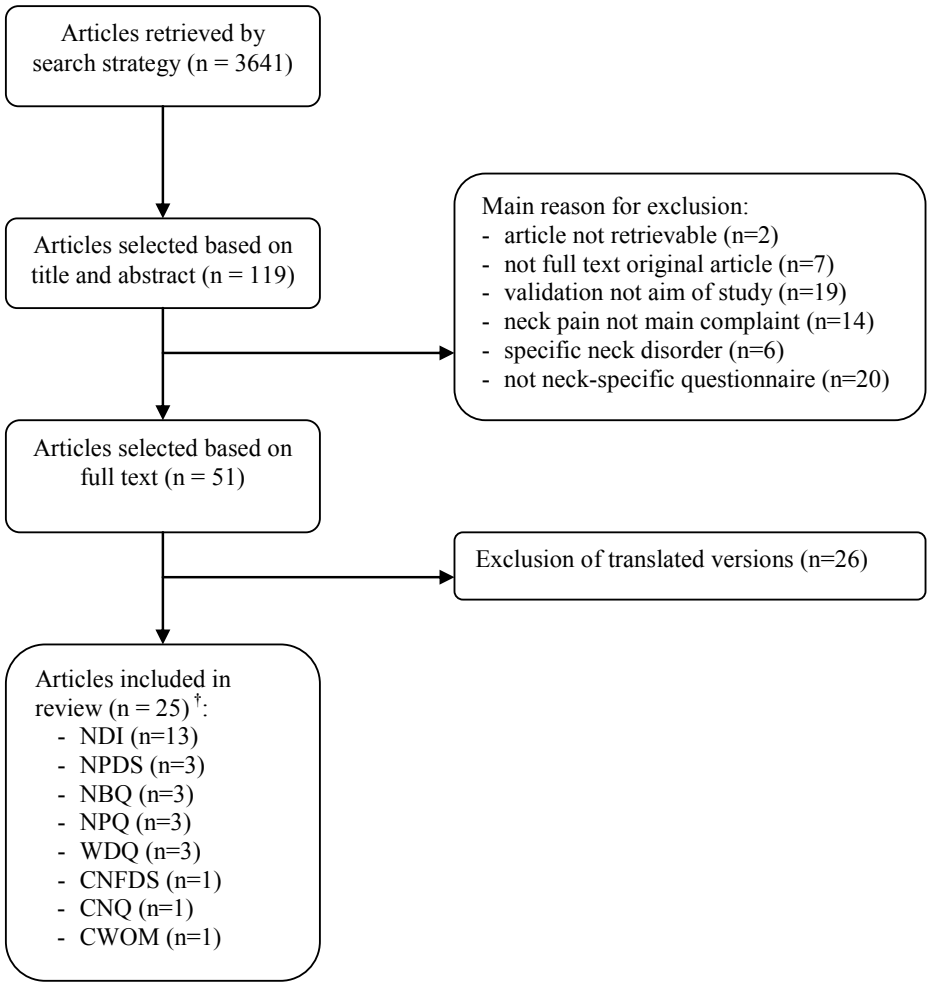

Figure 1 – Flowchart search and selection

† The sum of the different questionnaires is higher than 25, because some studies evaluate more than one questionnaire

RESULTS

The search strategy resulted in a total of 3641 unique hits, of which 119 articles were selected based on their title and abstract. The full text of these 119 articles was evaluated, which resulted in exclusion of another 68 articles. Reference checking did not result in additional included articles. Twenty-six articles concerned translated versions of neck-specific questionnaires, which were evaluated in a separate systematic review and therefore excluded.[10] Finally, 25 articles, evaluating 8 different questionnaires, were included in our study (see Figure 1). All original versions were developed in

Table 3 – Characteristics of the included studies

Study	Population	Country	Setting
Bolton et al. [36]	non-specific neck pain	England	chiropractor
Bolton et al. [37]	non-specific neck pain	England	chiropractor
Chan Ci En et al. [21]	> 3 month non-traumatic neck pain	Australia	physiotherapist
Chok et al. [22]	neck pain	Singapore	physiotherapist
Cleland et al. [29]	non-specific neck pain	USA	physiotherapist
Cleland et al. [31]	cervical radiculopathy	USA	physiotherapist
Ferrari et al. [41]	motor vehicle collision victims	Canada	primary care
Gay et al. [23]	chronic, uncomplicated neck pain	USA	physiotherapist
Goolkasian et al.-1 [19] †	mechanical neck pain	USA	orthopedist
Goolkasian et al.-2 [19] †	chronic mechanical neck pain	USA	orthopedist
Hains et al. [26]	neck pain	Canada	chiropractor
Hoving et al. [32]	WAD	Australia	physiotherapist/GP/ rheumatology
Jordan et al.-1 [20] †	chronic mechanical neck pain	Denmark	primary care
Jordan et al.-2 [20] †	chronic mechanical neck pain	Denmark	physiotherapist
Leak et al. [38]	mechanical neck pain	England	rheumatologist
Pinfold et al. [40]	WAD	Australia	physiotherapist
Rebbeck et al. [45]	WAD	Australia	primary care/insurance cohort
Riddle et al. [33]	non-specific neck pain	USA	physiotherapist
Sim et al. [39]	non-specific neck pain	England	physiotherapist
Stewart et al. [34]	> 3 month whiplash	Australia	physiotherapist
Stratford et al. [28]	neck pain of suspected musculoskeletal origin	Canada/ USA	physiotherapist
van der Velde et al. [27]	mechanical neck pain	USA	general population/ chiropractor
Vernon et al. [1]	WAD or chronic non traumatic neck complaints	England	chiropractor
Wheeler et al. [2]	mechanical neck pain	USA	orthopedist
White et al. [43]	chronic mechanical neck pain	England	physiotherapist/ rheumatologist
Willis et al. [42]	WAD	Australia	physiotherapist
Young et al. [30]	mechanical neck pain	USA	physiotherapist

[..] = reference number , GP = general practitioner, WAD = whiplash associated disorder

† study is mentioned twice, because they evaluated a questionnaire in two different populations

Table 4 – Methodological quality of each study per measurement property and questionnaire

Study	Internal Consistency	Measurement Error	Reliability	Content Validity	Structural Validity	Hypotheses Testing	Responsiveness
NDI							
Chan Ci En et al. [21]				poor		poor	
Chok et al. [22]			poor				poor
Cleland et al. [29]		fair	fair				fair
Cleland et al. [31]		poor	poor				fair
Gay et al. [23]	poor	poor				poor	poor
Hains et al. [26]	excellent				good	good	
Hoving et al. [32]				poor		fair	
Riddle et al. [33]						good	poor
Stewart et al. [34]							fair
Stratford et al. [26]	fair	poor	poor				poor
van der Velde et al. [27]	fair				fair	poor	
Vernon et al. [1]	poor		poor	fair		fair	poor
Young et al. [30]			poor				good
NPDS							
Chan Ci En et al. [21]				poor		poor	
Goolkasian et al.-1 [19]			poor				
Goolkasian et al.-2 [19]			poor				fair
Wheeler et al. [2]	poor			poor	fair	fair	
NBQ							
Bolton et al. [36]							poor
Bolton et al. [37]	poor	poor	poor			fair	fair
Gay et al. [23]	poor	poor				poor	poor

Table 4 – Methodological quality of each study per measurement property and questionnaire (continued)

Study	Internal Consistency	Measurement Error	Reliability	Content Validity	Structural Validity	Hypotheses Testing	Responsiveness
NPQ							
Hoving et al. [32]				poor		fair	
Leak et al. [38]	poor	poor	poor	poor			fair
Sim et al. [39]	poor						fair
WDQ							
Ferrari et al. [41]	fair					poor	
Pinfold et al. [40]	good			poor	fair		
Willis et al. [42]		poor	poor				fair
CNFDS							
Jordan et al.-1 [20]	poor		poor	poor			
Jordan et al.-2 [20]				poor		fair	poor
CNQ							
White et al. [43]		fair		poor		fair	
CWQ							
Rebbeck et al. [45]	poor				fair		fair

[..] = reference number

English, except for the Copenhagen Neck Functional Disability Scale (CNFDS), which was originally developed in Danish. The general characteristics of these studies are presented in Table 3. Two studies evaluated measurement properties for different populations and are therefore mentioned twice in Table 3.[19 20]

The methodological quality of the studies is presented in Table 4 for each questionnaire and measurement property. The synthesis of results per questionnaire and their accompanying level of evidence is presented in Table 5.

Below we will discuss the results per questionnaire. The results from studies of poor methodological quality are not mentioned.[21-24]

Table 5 – Quality of measurement properties per questionnaire

Questionnaire	Internal Consistency	Measurement Error	Reliability	Content Validity	Structural Validity	Hypothesis Testing	Responsiveness
NDI	+++	?	-	+	++	+++	++
NPDS	?	na	?	?	+	+	+
NBQ	?	?	?	na	na	+	+
NPQ	?	?	?	?	na	+	++
WDQ	++	?	?	?	+	?	+
CNFDS	?	na	?	?	na	+	?
CNQ	na	na	+	?	na	+	na
CWOM	?	na	na	na	na	+	+

+++ or --- = strong evidence positive/negative result, ++ or -- = moderate evidence positive/negative result, + or - = limited evidence positive/negative result, +/- = conflicting evidence, ? = unknown, due to poor methodological quality, na = no information available

Neck Disability Index (NDI)

The NDI was designed to measure activities of daily living (ADL) in patients with neck pain and was derived from the Oswestry low back pain Disability Index (ODI).[1 25] The 10 items have 6 response categories (range 0-5, total score range 0-50).[1] We did not find studies evaluating the average time needed to fill out the English version of the NDI.

Exploratory factor analysis shows that there is moderate evidence that the NDI has a 1-factor structure,[26] but there is also limited evidence that it is not unidimensional.[27] Both studies evaluating internal consistency assume a 1-factor structure, which resulted in a Cronbach α of 0.87-0.92.[26 28] The result of the only methodologically sound study evaluating measurement error is indeterminate,[29] because information is needed on the MIC for judging the measurement error. A value for the MIC cannot be provided yet, as the estimates for the MIC are too diverse (i.e. 3.5, 7.5 and 9.5 on a 0-50 scale).[29-31] There is limited evidence that reliability of the NDI is inadequate (ICC = 0.50).[29] There is limited positive evidence for the content validity of the NDI.[1] Hypothesis testing shows that NDI has a positive correlation with instruments measuring pain

and/or physical functioning (r = 0.53-0.70).[1 26 32 33] There is moderate positive evidence for responsiveness of the NDI (AUC = 0.79).[30] Two studies of lower methodological quality confirm this positive finding,[29 34] and one study of lower quality reports a negative result (AUC = 0.57).[31] Regarding interpretability: no floor or ceiling effects have been detected,[1 21 28 33] and differences in score between subgroups (e.g. same work status vs. altered work status) have been reported.[30 33]

Neck Pain and Disability Scale (NPDS)

The NPDS was designed to measure pain and disability in patients with neck pain and was developed using the Million Visual Analogue Scale as a template.[2 35] It consists of 20 items and each item is scored on a 10 cm visual analogue scale. Each item is converted to a score from 0 to 5 (total score range 0-100). We did not find studies evaluating the average time needed to fill out the English version of the NPDS.

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, or content validity of the NPDS. Exploratory factor analysis shows a 4-factor structure for the NPDS.[2] There is limited positive evidence for hypothesis testing (r = 0.52-0.78) and responsiveness (r = 0.59).[2 19] No floor or ceiling effects have been detected,[2 21] and differences in scores between subgroups (neck pain vs. no pain vs. lower back pain) have been reported.[2] There is no information regarding the MIC.

Neck Bournemouth Questionnaire (NBQ)

The NBQ was designed to measure pain, physical functioning, social functioning, and psychological functioning in patients with nonspecific neck pain and was developed using the Bournemouth Questionnaire for back pain as a template.[36 37] It consists of 7 items, each scored on a 0-10 numerical scale (total score range: 0-70).[36] We did not find studies evaluating the average time needed to fill out the English version of the NBQ.

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, or structural validity of the NBQ. There is limited positive evidence for hypothesis testing (r = 0.63) and responsiveness (r_{items} = 0.42-0.82).[36] Floor or ceiling effects, differences in scores between subgroups, and the MIC have not been studied.

Northwick Park Neck Pain Questionnaire (NPQ)

The NPQ was designed to measure the influence of non-specific neck pain on daily activities and was developed using the ODI as a template.[25 38] Each of the nine items consists of five ordinal responses (score 0-4) and the total (percentage) score is calculated by the following formula: (total score/maximum possible score) x 100%.[38] We

did not find studies evaluating the average time needed to fill out the English version of the NPQ.

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, or structural validity of the NPQ. There is a positive correlation (r = 0.56) between the NPQ and Problem Elicitation Technique (PET).[32] There is moderate positive evidence for responsiveness (r = 0.60).[38 39] No floor or ceiling effects have been detected.[38 39] Differences in scores between subgroups have not been evaluated. The MIC is unclear, because the single study on this property does not quantify it.[39]

Whiplash Disability Questionnaire (WDQ)

The WDQ was designed to measure disability in patients with WAD and was derived from the NDI.[1 40] It consists of 13 items, each scored on a 0-10 numerical scale (total score range: 0-130).[40] We did not find studies evaluating the average time needed to fill out the English version of the WDQ.

There were no methodologically sound studies evaluating the measurement error, reliability, content validity, or hypothesis testing of the WDQ. Exploratory factor analysis shows that the WDQ probably has a 1-factor structure.[40] There is moderate positive evidence for internal consistency (Cronbach's α = 0.95-0.96).[40 41] These high values indicate that the WDQ might contain redundant items. There is limited positive evidence for responsiveness (r = 0.67).[42] No floor or ceiling effects have been detected,[40-42] information on other aspects of interpretability is lacking.

Copenhagen Neck Functional Disability Scale (CNFDS)

The CNFDS was designed by a group of experts in the field of neck pain to measure disability in patients with neck pain.[20] It consists of 15 items with three possible ordinal responses per item (score 0-2). The total score ranges from 0 to 30.[20] The average time needed to fill out the Danish version of the CNFDS is 10 minutes.[20]

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, structural validity, or responsiveness of the CNFDS. The CNFDS correlates with a Numerical Rating Scale (NRS) for pain (r = 0.64).[20] There are no comparisons available between the CNFDS and other instruments measuring pain or disability. No ceiling effect has been detected,[20] there is no information available on floor effects, differences in scores between subgroups, or the MIC.

Core Neck Questionnaire (CNQ)

The CNQ was designed to measure outcomes of care in patients with non-specific neck pain and was developed using the Core Outcome Measure for back pain (COM)

as a template.[43 44] The CNQ consists of seven items, scored from 1 to 5, which are added up to a total score.[43] We did not find studies evaluating the average time needed to fill out the English version of the CNQ.

There were no methodologically sound studies evaluating the internal consistency, measurement error, content validity, structural validity or responsiveness of the CNFDS. The reliability of the total score of the CNQ has not been studied, but four of the six items have an ICC > 0.70.[43] There was a positive correlation of the CNQ with the NDI ($r > 0.60$).[43] No floor or ceiling effects have been detected,[43] there is no information on other aspects of interpretability.

Core Whiplash Outcome Measure (CWOM)

The CWOM was designed to measure relevant health outcomes in patients with whiplash associated disorder (WAD).[45] The COM was used as a template to develop the CWOM.[44 45] The CWOM consists of 5 items, each scored on a 1-5 scale (total score range: 5-25).[45] We did not find studies evaluating the average time needed to fill out the English version of the CWOM.

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, or structural validity of the CWOM. There is limited positive evidence for correlation with instruments measuring pain and/or physical functioning ($r = 0.65$-0.82) and for responsiveness (AUC = 0.73-0.81).[45] The scores for different stages of whiplash have been reported,[45] but other aspects of interpretability are not mentioned.

DISCUSSION

Eight different questionnaires have been developed to measure pain and/or disability in patients with neck pain. All original versions are in English, except for the CNFDS, which was developed in Danish. The NDI is the most frequently evaluated questionnaire and its measurement properties seem adequate, except for reliability. The other questionnaires show positive results, but the evidence is mostly limited and at least half of the information on measurement properties per questionnaire is lacking. Therefore, the results should be treated with caution.

The COSMIN checklist has recently been developed and is based on consensus between experts in the field of health status questionnaires.[8] The COSMIN checklist facilitates a separate judgment of the methodological quality of the included studies and their results. This is in line with the methodology of systematic reviews of clinical trials.[16] The inter-rater agreement of the COSMIN-checklist is adequate.[46] The inter-

rater reliability for many COSMIN items is poor, which is suggested to be due to interpretation of checklist items.[46] To minimise differences between reviewers (JMS, CBT, and HCV) in interpretation of checklist items, decisions were made in advance on how to score the different items.

The criteria in Table 1 are based on the levels of evidence as previously proposed by the Cochrane Back Review Group.[17] The criteria are originally meant for systematic reviews of clinical trials, but we believe that they are also applicable for reviews on measurement properties of health status questionnaires.

Exclusion of non-English papers may introduce selection bias. However, the leading journals, and as a consequence the most important studies, are published in English. So, research performed in populations with a different native language is generally still published in English. This is illustrated by the large number of articles we retrieved regarding translations of neck-specific questionnaires (see Figure 1). In these papers we did not find a reference to an original version of a neck-specific questionnaire that was not included in our systematic review. This makes us confident that chances are small that we have missed any original versions of neck-specific questionnaires.

The different studies showed similar methodological shortcomings. A small sample size, for example, frequently led to indeterminate results. We do not discuss these flaws in detail here, but elaborate on this subject in a separate publication.[47]

A problem we encountered during the rating of "hypothesis testing" and "responsiveness" was that most studies do not formulate hypotheses regarding expected correlations in advance. Moreover, none of the development studies specified the supposed underlying constructs of the questionnaire. Therefore, it is difficult to judge content validity, which is one of the most important measurement properties. We dealt with this problem by reaching agreement about what we thought were the supposed underlying constructs, based on the items in the questionnaire, before we rated the studies.

The assumption that pooling of results from original and translated versions could result in inconsistent findings regarding the results for measurement properties is confirmed in our systematic review of translated versions of neck-specific questionnaires.[10] A poor translation process and/or lack of cross-cultural validation seem to affect the measurement properties of the questionnaire, particularly the validity (i.e. structural validity and hypothesis testing).[10] This is not surprising, as the importance and/or meaning of questionnaire items (e.g. driving, depressed mood) may depend on setting and context. So, a simple translation of the original questionnaire is not sufficient and might affect the measurement of the underlying constructs.[10]

Since the review in 2002 17 of the 25 included studies in our review were published and four new neck-specific questionnaires have been developed.[4 36 40 43 45] These stud-

ies added new information, but due to their poor to fair methodological quality a substantial amount of uncertainty about the quality of the measurement properties remains.

The quality of the measurement properties of several neck-specific questionnaires was recently evaluated in a best evidence synthesis, which showed positive results for the NDI, NPDS, NBQ, NPQ, CNFDS, and WDQ.[5] However, these results were partially based on methodologically flawed studies and this study contained only a small part of the manuscripts included in our study.

A state-of-the-art review evaluating the NDI reported that its reliability, internal consistency, factor structure (i.e. unidimensional scale), construct validity, and responsiveness are well described and of very high quality,[7] which is not completely in agreement with our findings. Possible explanations for the discrepancies are that the study reporting the negative result for reliability was published after the search of the state-of-the-art review ended and that they did not critically appraise the methodological quality or results of the included studies.[7 29] A more recent systematic review evaluating the NDI reports a good internal consistency, acceptable reliability, good construct validity and responsiveness, and inconsistent results regarding the structural validity of the NDI.[6] The differences with our findings are probably attributable to the fact that they did not take the methodological quality of the included studies into account.[6]

It is difficult to determine the content validity of the different neck-specific questionnaires, because almost all retrieved studies on this subject were of poor methodological quality. Furthermore, the underlying constructs were not clear. However, a recent content analysis showed that correspondence between the symptoms expressed by neck pain patients and the content of the questionnaires was low, mainly due to lack of patient involvement in development of the questionnaire.[48] The importance of content validity for a questionnaire makes it desirable that this measurement property is evaluated in a high quality study for each questionnaire. The results from these studies will show which questionnaires are suitable for neck pain patients and if development of a new neck-specific questionnaire is necessary.

The most frequently studied measurement property is responsiveness. This is not surprising, since these questionnaires are often used as an outcome measure. However, except for the NDI and NPQ, there is only limited positive evidence for responsiveness.

For clinical practice and research we advise to use the original version of neck-specific questionnaires with caution: the majority of the results are positive, but the evidence is mostly limited and for each questionnaire, except for the NDI, at least half of the information regarding measurement properties is lacking. Provisionally we recommend using the NDI, because it is the questionnaire for which the most information is

available and the results are mostly positive. However, research is needed to clarify its underlying constructs, measurement error, reliability, and to improve the interpretation of its scores.

No clinician should make decisions regarding management of neck pain patients solely on unvalidated instruments. However, neck-specific questionnaires can provide a broader and deeper understanding of the impact of neck pain on the individual patients.

For future research we recommend performing high quality studies to evaluate the unknown measurement properties, especially content validity, and provide strong evidence for the other measurement properties. It seems advisable to refrain from developing new neck-specific questionnaires until high quality studies evaluating the measurement properties of current questionnaires show shortcomings that make it necessary to develop a new questionnaire.

CONCLUSION

A lot of information regarding the measurement properties of the original version of the different neck-specific questionnaires is still lacking or of poor methodological quality. The available evidence on the measurement properties is mostly limited. The NDI is the most frequently evaluated questionnaire and its measurement properties seem adequate, except for reliability and the fact that there is information lacking regarding its underlying constructs and measurement error.

Our findings do not mean that the current questionnaires are poor, but imply that studies of high methodological quality are needed to properly assess their measurement properties. It is recommendable to use the COSMIN checklist when designing these studies. There is no need for the development of new neck-specific questionnaires until the measurement properties of the current questionnaires have been adequately assessed.

REFERENCES

1. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
2. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine 1999;24:1290-4.
3. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press, 2003.
4. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. Spine 2002;27:515-22.
5. Nordin M, Carragee EJ, Hogg-Johnson S, Weiner SS, Hurwitz EL, Peloso PM, et al. Assessment of neck pain and its associated disorders: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine 2008;33:S101-22.
6. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther 2009;39:400-17.
7. Vernon H. The Neck Disability Index: State-of-the-Art, 1991-2008. J Manipulative Physiol Ther 2008;31:491-502.
8. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-49.
9. Menezes da Costa L, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol 2009;62:934-43.
10. Schellingerhout JM, Heymans MW, Verhagen AP, de Vet HC, Koes BW, Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. BMC Med Res Methodol 2011;11:87.
11. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.
12. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.
13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1307-10.
14. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol 2009;62:1062-7.
15. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
16. Furlan AD, Pennick V, Bombardier C, van Tulder M, Editorial Board CBRG. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. Spine 2009;34:1929-41.
17. van Tulder M, Furlan A, Bombardier C, Bouter L, Editorial Board CBRG. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine 2003;28:1290-9.

18. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
19. Goolkasian P, Wheeler AH, Gretz SS. The neck pain and disability scale: test-retest reliability and construct validity. Clin J Pain 2002;18:245-50.
20. Jordan A, Manniche C, Mosdal C, Hindsberger C. The Copenhagen neck functional disability scale: A study of reliability and validity. J Manipulative Physiol Ther 1998;21:520-7.
21. Chan Ci En M, Clair DA, Edmondston SJ. Validity of the Neck Disability Index and Neck Pain and Disability Scale for measuring disability associated with chronic, non-traumatic neck pain. Man Ther 2009;14:433-8.
22. Chok B, Gomez E. The reliability and application of the Neck Disability Index in physiotherapy. Physiotherapy Singapore 2000;3:16-19.
23. Gay RE, Madson TJ, Cieslak KR. Comparison of the Neck Disability Index and the Neck Bournemouth Questionnaire in a sample of patients with chronic uncomplicated neck pain. J Manipulative Physiol Ther 2007;30:259-62.
24. Bolton JE. Sensitivity and specificity of outcome measures in patients with neck pain: detecting clinically significant improvement. Spine 2004;29:2410-7.
25. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980;66:271-3.
26. Hains F, Waalen J, Mior S. Psychometric properties of the neck disability index. J Manipulative Physiol Ther 1998;21:75-80.
27. van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A. Rasch analysis provides new insights into the measurement properties of the neck disability index. Arthritis Rheum 2009;61(4):544-51.
28. Stratford P, Riddle D, Binkley J, Spadoni G, Westaway M, Padfield B. Using the Neck Disability Index to make decisions concerning individual patients. Physiother Canada 1999;51:107.
29. Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. Arch Phys Med Rehabil 2008;89:69-74.
30. Young BA, Walker MJ, Strunce JB, Boyles RE, Whitman JM, Childs JD. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. Spine J 2009;9:802-8.
31. Cleland JA, Fritz JM, Whitman JM, Palmer JA. The reliability and construct validity of the Neck Disability Index and patient specific functional scale in patients with cervical radiculopathy. Spine 2006;31:598-602.
32. Hoving JL, O'Leary EF, Niere KR, Green S, Buchbinder R. 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;102:273-81.
33. Riddle DL, Stratford PW. Use of generic versus region-specific functional status measures on patients with cervical spine disorders. Phys Ther 1998;78:951-63.
34. Stewart M, Maher CG, Refshauge KM, Bogduk N, Nicholas M. Responsiveness of pain and disability measures for chronic whiplash. Spine 2007;32:580-5.
35. Million R, Nilsen KH, Jayson MI, Baker RD. Evaluation of low back pain and assessment of lumbar corsets with and without back supports. Ann Rheum Dis 1981;40:449-54.
36. Bolton JE, Humphreys BK. The Bournemouth Questionnaire: A short-form comprehensive outcome measure. II. Psychometric properties in neck pain patients. J Manipulative Physiol Ther 2002;25:141-8.

37. Bolton JE, Breen AC. The Bournemouth Questionnaire: a short-form comprehensive outcome measure. I. Psychometric properties in back pain patients. J Manipulative Physiol Ther 1999;22:503-10.
38. Leak AM, Cooper J, Dyer S, Williams KA, Turner-Stokes L, Frank AO. The Northwick Park Neck Pain Questionnaire, devised to measure neck pain and disability. Br J Rheumatol 1994;33:469-74.
39. Sim J, Jordan K, Lewis M, Hill J, Hay EM, Dziedzic K. Sensitivity to change and internal consistency of the Northwick Park Neck Pain Questionnaire and derivation of a minimal clinically important difference. Clin J Pain 2006;22:820-6.
40. Pinfold M, Niere KR, O'Leary EF, Hoving JL, Green S, Buchbinder R. Validity and Internal Consistency of a Whiplash-Specific Disability Measure. Spine 2004;29:263-8.
41. Ferrari R, Russell A, Kelly AJ. Assessing whiplash recovery--the Whiplash Disability Questionnaire. Aust Fam Physician 2006;35:653-4.
42. Willis C, Niere KR, Hoving JL, Green S, O'Leary EF, Buchbinder R. Reproducibility and responsiveness of the Whiplash Disability Questionnaire. Pain 2004;110:681-8.
43. White P, Lewith G, Prescott P. The core outcomes for neck pain: Validation of a new outcome measure. Spine 2004;29:1923-30.
44. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research. A proposal for standardized use. Spine 1998;23:2003-13.
45. Rebbeck TJ, Refshauge KM, Maher CG, Stewart M. Evaluation of the core outcome measure in whiplash. Spine 2007;32:696-702.
46. Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med Res Methodol 2010;10:82.
47. Terwee CB, Schellingerhout JM, Verhagen AP, de Vet HC, Koes BW. Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: a systematic review. J Manipulative Physiol Ther 2011;43:261-72.
48. Wiitavaara B, Bjorklund M, Brulin C, Djupsjobacka M. How well do questionnaires on symptoms in neck-shoulder disorders capture the experiences of those who suffer from neck-shoulder disorders? A content analysis of questionnaires and interviews. BMC Musculoskelet Disord 2009;10:30.

Chapter 6

Measurement properties of translated versions of neck-specific questionnaires: a systematic review.

Published as:

Schellingerhout JM, Heymans MW, Verhagen AP, de Vet HC, Koes BW, Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. BMC Med Res Methodol 2011;11:87.

ABSTRACT

Objective

To critically appraise the quality of the translation process, cross-cultural validation and the measurement properties of translated versions of neck-specific questionnaires.

Methods

Bibliographic databases were searched for articles concerning the translation or evaluation of the measurement properties of a translated version of a neck-specific questionnaire. The methodological quality of the selected studies and the results of the measurement properties were critically appraised and rated using the COSMIN checklist and criteria for measurement properties.

Results

The search strategy resulted in a total of 3641 unique hits, of which 27 articles, evaluating 6 different questionnaires in 15 different languages, were included in this study. Generally the methodological quality of the translation process is poor and none of the included studies performed a cross-cultural adaptation. A substantial amount of information regarding the measurement properties of translated versions of the different neck-specific questionnaires is lacking. Moreover, the evidence for the quality of measurement properties of the translated versions is mostly limited or assessed in studies of poor methodological quality.

Conclusion

Until results from high quality studies are available, we advise to use the Catalan, Dutch, English, Iranian, Korean, Spanish and Turkish version of the NDI, the Chinese version of the NPQ, and the Finnish, German and Italian version of the NPDS. The Greek NDI needs cross-cultural validation and there is no methodologically sound information for the Swedish NDI. For all other languages we advise to translate the original version of the NDI.

INTRODUCTION

Several disease-specific questionnaires have been developed to measure pain and disability in patients with neck pain (e.g. Neck Disability Index (NDI), Neck Pain and Disability Scale (NPDS)).[1-2] To make them suitable for use in other languages, several of these neck-specific questionnaires have been translated. However, a simple translation of the original version doesn't guarantee similar measurement properties, because differences in cultural context have to be taken into account as well.[3-4]

Previous reviews of neck-specific questionnaires have not paid sufficient attention to possible differences in performance, caused by differences in cultural context, and combine the results of studies that evaluate measurement properties of different language versions of the same questionnaire.[5-6] This may lead to inconsistent results for measurement properties, as was demonstrated in a recent review of the cross-cultural adaptations of the McGill Pain Questionnaire.[7]

Since it is possible that the measurement properties of neck-specific questionnaires vary between different nationalities, we decided to evaluate them per language. This reduces inconsistency in results due to cultural differences and also facilitates a choice for the best questionnaire per language. The measurement properties of original versions of the different neck-specific questionnaires were evaluated in a separate systematic review.[8]

The purpose of this study is to critically appraise the quality of the translation process, cross-cultural validation and the measurement properties of translated versions of neck-specific questionnaires.

METHODS

Search strategy

We searched the following computerized bibliographic databases: Medline (1966 to July 2010), EMbase (1974 to July 2010), CINAHL (1981 to July 2010), and PsycINFO (1806 to July 2010). We used the index terms "neck", "neck pain", and "neck injuries/injury" in combination with "research measurement", "questionnaire", "outcome assessment", "psychometry", "reliability", "validity", and derivatives of these terms. The full search strategy used in each database is available upon request from the corresponding author. Reference lists were screened to identify additional relevant studies.

Selection criteria

A study was included if it was a full text original article (e.g. not an abstract, review or editorial), published in English, concerning the translation or evaluation of the measurement properties of a translated version of a neck-specific questionnaire. The questionnaire had to be self-reported, evaluating pain and/or disability, and specifically developed or adapted for patients with neck pain.

For inclusion, neck pain had to be the main complaint of the study population. Accompanying complaints (e.g. low back pain or shoulder pain) were no reason for exclusion, as long as the main focus was neck pain. Studies considering study populations with a specific neck disorder (e.g. neurological disorder, rheumatological disorder, malignancy, infection, or fracture) were excluded, except for patients with cervical radiculopathy or whiplash associated disorder (WAD).

Two reviewers (JMS, APV) independently assessed the titles, abstracts, and reference lists of studies retrieved by the literature search. In case of disagreement between the two reviewers, there was discussion to reach consensus. If necessary, a third reviewer (HCV) made the decision regarding inclusion of the article.

Measurement properties

The measurement properties are divided over three domains: reliability, validity, and responsiveness.[9] In addition, the interpretability is described.

Reliability

Reliability is defined as the extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same questionnaire (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons on different occasions (intra-rater).[9]

Reliability contains the following measurement properties:

Internal consistency: The interrelatedness among the items in a questionnaire, expressed by Cronbach's α or Kuder-Richardson Formula 20 (KR-20).[9-10]

Measurement error: The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured, expressed by the standard error of measurement (SEM).[9 11] The SEM can be converted into the smallest detectable change (SDC).[11] Changes exceeding the SDC can be labeled as change beyond measurement error.[11] Another approach is to calculate the limits of agreement (LoA).[12] For determining the adequacy of measurement error the SDC and/or LoA is related to the minimal important change (MIC).[13]

Reliability: The proportion of the total variance in the measurements which is due to 'true' differences between patients.[9] This aspect is reflected by the Intraclass Correlation Coefficient (ICC) or Cohen's Kappa.[9 14]

Validity

Validity is the extent to which a questionnaire measures the construct it is supposed to measure and contains the following measurement properties:[9]

Content validity: The degree to which the content of a questionnaire is an adequate reflection of the construct to be measured.[9] Important aspects are whether all items are relevant for the construct, aim, and target population and if no important items are missing (comprehensiveness).[15]

Criterion validity: The extent to which scores on an instrument are an adequate reflection of a gold standard.[9] Since a real gold standard for health status questionnaires is not available,[15] we will not evaluate criterion validity.

Construct validity is divided into three aspects:

Cross-cultural validity: The degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument.[9] This is assessed by means of multi-group factor analysis or differential item functioning using data from a population that completed the questionnaire in the original language, as well as data from a population that completed the questionnaire in the new language.

Structural validity: The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured.[9] Factor analysis should be performed to confirm the number of subscales present in a questionnaire.[15]

Hypothesis testing: The degree to which a particular measure relates to other measures in a way one would expect if it is validly measuring the supposed construct, i.e. in accordance with predefined hypotheses about the correlation or differences between the measures.[9]

Responsiveness

Responsiveness is the ability of an instrument to detect change over time in the construct to be measured.[9] Responsiveness is considered an aspect of validity, in a longitudinal context.[15] Therefore, the same standards apply as for validity: the correlation between change scores of two measures should be in accordance with predefined hypotheses.[15] Another approach is to consider the measurement instrument as a diagnostic test to distinguish improved and non-improved patients. The responsiveness of the instrument is then expressed as the area under the receiver operator characteristic curve (AUC).[15]

Interpretability

Interpretability is the degree to which one can assign qualitative meaning to quantitative scores.[9] This means that investigators should provide information about clinically meaningful differences in scores between subgroups, floor and ceiling effects, and the MIC.[15] Interpretability is not a measurement property, but an important characteristic of a measurement instrument.[9]

Quality assessment

Assessment of the methodological quality of the selected studies was carried out using the COSMIN checklist.[10] The COSMIN checklist consists of nine boxes with methodological standards for how each measurement property should be assessed. Each item was scored on a 4-point rating scale (i.e. "poor", "fair", "good", or "excellent", see www.cosmin.nl). An overall score for the methodological quality of a study was determined by taking the lowest rating of any of the items in a box. The methodological quality of a study was evaluated per measurement property. Special attention was paid to the methodological quality of the translation process and cross-cultural validation. The COSMIN box concerning this measurement property is presented in Table 1.

Data extraction and assessment of (methodological) quality were performed by two reviewers (JMS, CBT) independently. In case of disagreement between the two reviewers, there was discussion in order to reach consensus. If necessary, a third reviewer (HCV) made the decision.

Best evidence synthesis – levels of evidence

To determine the overall quality of the measurement properties of the different questionnaires we synthesized the different studies per language by combining their results, adjusted for methodological quality of the studies and the consistency of their results. The possible overall rating for a measurement property is "positive", "indeterminate", or "negative", accompanied by levels of evidence, similarly as was proposed by the Cochrane Back Review Group (see Table 2).[16-17]

To assess whether the results of the measurement properties were positive, negative, or indeterminate, we used criteria based on Terwee et al. (see Table 3).[18]

RESULTS

The search strategy resulted in a total of 3641 unique hits, of which 119 articles were selected based on their title and abstract. The full text assessment resulted in exclusion of another 68 articles. Reference checking did not result in additional articles. Twenty-four articles concerned original versions of neck-specific questionnaires, which were

Table 1 – Methodological criteria for the translation process and cross-cultural validation [10]

Item	Methodological Criteria
1	Was the percentage of missing items given?
2	Was there a description of how missing items were handled?
3	Was the sample size included in the analysis adequate?
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, in the construct to be measured, or in both languages
6	Did the translators work independently from each other?
7	Were items translated forward and backward?
8	Was there an adequate description of how differences between the original and translated versions were resolved?
9	Was the translation reviewed by a committee (e.g. original developers)?
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?
11	Was the sample used in the pre-test adequately described?
12	Were the samples similar for all characteristics except language and/or cultural background?
13	Were there any important flaws in the design or methods of the study?
14	for CTT: Was confirmatory factor analysis performed?
15	for IRT: Was differential item function (DIF) between language groups assessed?

CTT = Classical Test Theory, IRT = Item Response Theory

Table 2 – Levels of evidence for the overall quality of the measurement property [17]

Level	Rating	Criteria
strong	+++ or ---	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
moderate	++ or --	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
limited	+ or -	One study of fair methodological quality
conflicting	+/-	Conflicting findings
unknown	?	Only studies of poor methodological quality

[..] = reference number
+ = positive result, - = negative result

evaluated in a separate systematic review.[8] Finally, 27 articles on translated questionnaires, evaluating 6 different questionnaires in 15 different languages, were included in this study (see Figure 1).

The general characteristics of these studies are presented in Table 4. None of the included studies performed a cross-cultural validation (Table 1, items 14 and 15), i.e. no studies performed multi-group factor analysis or differential item functioning. Therefore, we were only able to rate the methodological quality of the translation process (Table 1, items 4-11). The methodological quality of the studies is presented

Table 3 – Quality criteria for measurement properties [Based on Terwee et al., 18]

Property	Rating	Quality Criteria
Reliability		
Internal consistency	+	(Sub)scale unidimensional AND Cronbach's alpha(s) ≥ 0.70
	?	Dimensionality not known OR Cronbach's alpha not determined
	-	(Sub)scale not unidimensional OR Cronbach's alpha(s) < 0.70
Measurement error	+	MIC > SDC OR MIC outside the LOA
	?	MIC not defined
	-	MIC ≤ SDC OR MIC equals or inside LOA
Reliability	+	ICC / weighted Kappa ≥ 0.70 OR Pearson's r ≥ 0.80
	?	Neither ICC / weighted Kappa, nor Pearson's r determined
	-	ICC / weighted Kappa < 0.70 OR Pearson's r < 0.80
Validity		
Content validity	+	The target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be complete
	?	No target population involvement
	-	The target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete
Construct validity		
- Cross-cultural validity	+	Original factor structure confirmed OR no important DIF
	?	Confirmation original factor structure AND DIF not mentioned
	-	Original factor structure not confirmed OR important DIF
- Structural validity	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	-	Factors explain < 50% of the variance
- Hypothesis testing	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
Responsiveness		
Responsiveness	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlation with related constructs is lower than with unrelated constructs

[..] = reference number , MIC = minimal important change, SDC = smallest detectable change, LOA = limits of agreement, ICC = intraclass correlation coefficient, DIF = differential item functioning, AUC = area under the curve

† + = positive rating, ? = indeterminate rating, - = negative rating

Table 4 – General information per study

Study	Language	Country	Population	Setting
Nieto et al. [26]	Catalan	Spain	< 3 months whiplash	rehabilitation unit
Chiu et al. [27]	Chinese	Hong Kong	neck pain	physiotherapist
Lee et al. [28]	Chinese	Hong Kong	neck pain	physiotherapist
Jorritsma et al. [20]	Dutch	Netherlands	> 3 months non-specific neck pain	rehabilitation unit
Pool et al. [30]	Dutch	Netherlands	non-specific neck pain	general practitioner
Schmitt et al. [31]	Dutch	Netherlands	> 3 weeks whiplash	general population
Vos et al. [32]	Dutch	Netherlands	< 6 weeks non-specific neck pain	general practitioner
Stewart et al. [34]	English	Australia	> 3 months whiplash	physiotherapist
Salo et al. [36]	Finnish	Finland	neck pain	physiotherapist/rehabilitation unit
Forestier et al. [19]	French	France	> 3 months mechanical neck pain	general population
Martel et al. [38]	French	Canada	> 12 weeks mechanical neck pain	general population
Wlodyka-Demaille et al. [37]	French	France	> 15 days non-specific neck pain	rehabilitation unit/ rheumatologist
Wlodyka-Demaille et al. [21]	French	France	> 15 days non-specific neck pain	rehabilitation unit/ rheumatologist
Bremerich et al. [25]	German	Switzerland	> 3 months non-specific neck pain	rheumatologist
Scherer et al. [39]	German	Germany	neck pain	general practitioner
Trouli et al. [40]	Greek	Greece	non-specific neck pain	primary care
Agarwal et al. [41]	Hindi	India	cervical radiculopathy	physiotherapist
Mousavi et al. [42]	Iranian	Iran	non-specific neck pain	primary care/physiotherapist
Monticone et al. [43]	Italian	Italy	> 4 weeks non-specific neck pain	rehabilitation unit
Lee et al. [44]	Korean	South Korea	non-specific neck pain	physiotherapist
Andrade et al. [47]	Spanish	Spain	non-specific neck pain	rehabilitation unit
Gonzalez et al. [45]	Spanish	Spain	> 4 months non-specific neck pain	physiotherapist
Kovacs et al. [24]	Spanish	Spain	non-specific neck pain	primary care/hospital outpatient clinic
Ackelman et al. [23]	Swedish	Sweden	acute/chronic neck pain	emergency room/ physiotherapist
Aslan et al. [48]	Turkish	Turkey	> 3 months non-specific neck pain	physiotherapist/rehabilitation unit
Bicer et al. [22]	Turkish	Turkey	> 6 months non-specific neck pain	rehabilitation unit
Kose et al. [49]	Turkish	Turkey	> 6 weeks non-specific neck pain	primary care

[..] = reference number

Table 5 – Methodological quality of each study per measurement property

Language Study	Instrument	Translation process	Internal Consistency	Measurement Error	Reliability	Content Validity	Structural Validity	Hypotheses Testing	Responsiveness
Catalan									
Nieto et al. [26]	NDI	poor	good				fair	good	
Chinese									
Chiu et al. [27]	NPQ	poor	poor		excellent	poor		fair	poor
Lee et al. [28]	NPQ							fair	poor
Dutch									
Jorritsma et al. [20]	NDI			poor	poor				
	NPDS	fair		poor	poor				
Pool et al. [30]	NDI			fair					fair
Schmitt et al. [31]	NBQ	excellent	poor	fair	fair			poor	
Vos et al. [32]	NDI			fair	fair				poor
English									
Stewart et al. [34]	CNFDS								fair
Finnish									
Salo et al. [36]	NDI	poor	excellent		poor		good	poor	
	NPDS	poor	excellent		poor		good	poor	
French									
Forestier et al. [19]	CNFDS	poor	poor						poor
Martel et al. [38]	NBQ	poor			poor			fair	moderate
Wlodyka et al. [37]	NDI	poor	poor	poor	poor		fair	fair	
	NPDS	poor	poor	poor	poor		fair	fair	
	NPQ	poor	poor	poor	poor		fair	fair	
Wlodyka et al. [21]	NDI								poor
	NPDS								poor
	NPQ								poor
German									
Bremerich et al. [25]	NPDS	fair		poor	poor				
Scherer et al. [39]	NPDS	poor	excellent				good	good	

Table 5 – Methodological quality of each study per measurement property (continued)

Language Study	Instrument	Translation process	Internal Consistency	Measurement Error	Reliability	Content Validity	Structural Validity	Hypotheses Testing	Responsiveness
Greek									
Trouli et al. [40]	NDI	good	good	poor	poor		good		fair
Hindi									
Agarwal et al. [41]	NPDS	fair	poor	poor	poor	poor		fair	
Iranian									
Mousavi et al. [42]	NDI	excellent	fair		fair	poor			fair
	NPDS	excellent	fair		fair	poor	fair		fair
Italian									
Monticone et al. [43]	NPDS	poor	fair		fair		fair	poor	
Korean									
Lee et al. [44]	NDI	poor	fair	poor	poor			fair	poor
	NPDS	poor	poor	poor	poor			fair	poor
Spanish									
Andrade et al. [47]	NDI		fair	poor	poor		fair	fair	fair
Gonzalez et al. [45]	NPQ	poor	poor		fair			poor	poor
Kovacs et al. [24]	NDI	excellent	poor		poor			poor	poor
	NPQ		poor					poor	poor
	CNQ	excellent	poor		poor			poor	poor
Swedish									
Ackelman et al. [23]	NDI				poor	poor		poor	
Turkish									
Aslan et al. [48]	NDI	excellent			fair			fair	
Bicer et al. [22]	NPDS	poor	poor					poor	
Kose et al. [49]	NDI	fair	poor		fair			poor	fair
	NPDS	fair	poor		fair			poor	fair
	NPQ	fair	poor		fair			poor	fair
	CNFDS	fair	poor		fair			poor	fair

[..] = reference number

Table 6 – Quality of the measurement properties per language and questionnaire

Language	Instrument	Internal Consistency	Measurement Error	Reliability	Content Validity	Structural Validity†				Hypotheses Testing	Responsiveness
						1	2	3	4		
Catalan	NDI	++	na	na	na	-	+			++	na
Chinese	NPQ	?	na	+++	?	na				++	?
Dutch	NDI	na	-	+	na	na				na	+
	NPDS	na	?	?	na	na				na	na
	NBQ	?	?	+	na	na				?	na
English	CNFDS	na	na	na	na	na				na	+
Finnish	NDI	?	na	?	na	--				?	na
	NPDS	+++	na	?	na			++		?	na
French	NDI	na	?	?	na		+			-	?
	NPDS	na	?	?	na			+		+/-	?
	NBQ	na	na	?	na	na				+/-	-
	NPQ	na	?	?	na		+			+/-	?
	CNFDS	?	na	na	na	na				na	?
German	NPDS	?	?	?	na	--		++		++	na
Greek	NDI	?	?	?	na	--				na	-
Hindi	NPDS	?	?	?	?	na				+/-	na
Iranian	NDI	+	na	+	?	na				na	+
	NPDS	+	na	+	?				+	na	-
Italian	NPDS	+	na	+	na			+		?	na
Korean	NDI	+	?	?	na	na				?	?
	NPDS	?	?	?	na	na				?	?
Spanish	NDI	+	na	?	na	+				+	+
	NPQ	?	na	-	na	na				?	?
	CNQ	?	na	?	na	na				?	?
Swedish	NDI	na	na	?	?	na				?	na
Turkish	NDI	?	na	++	na	na				+	+
	NPDS	?	na	+	na	na				?	+
	NPQ	?	na	+	na	na				?	+
	CNFDS	?	na	+	na	na				?	+

+++ or --- = strong evidence positive/negative result, ++ or -- = moderate evidence positive/negative result, + or - = limited evidence positive/negative result, +/- = conflicting evidence, ? = unknown, due to poor methodological quality, na = no information available

† the numbers reflect the number of factors that are mentioned in the underlying studies

in Table 5 for each measurement property, arranged per language. Generally the methodological quality of the studies was poor to fair. The synthesis of the results per questionnaire and their accompanying level of evidence is presented in Table 6 for each language. For each questionnaire, except for the Iranian NPDS and Spanish NDI, at least half of the information regarding measurement properties is lacking. Moreover, the evidence for the quality of measurement properties is mostly limited, due to methodological shortcomings of the included studies.

Below we will discuss the results for the different questionnaires per language. The results regarding measurement properties from studies of poor methodological quality are not mentioned.[19-25]

Catalan

The NDI is the only neck-specific questionnaire that has been translated in Catalan.[26] The NDI was originally designed to measure activities of daily living (ADL) in patients with neck pain.[1] The methodological quality of the translation process is poor.[26] Confirmatory factor analysis showed that the NDI is not unidimensional and there is limited evidence that the NDI has a 2-factor structure.[26] Assuming a 2-factor structure, there is moderate positive evidence for internal consistency: Cronbach's α is 0.70 for "pain and interference with cognitive functioning" and 0.83 for "functional disability".[26] There is a positive correlation (r = 0.51) between the NDI and the Pain Intensity Index.[26]

The available evidence on measurement properties of the Catalan NDI is positive, despite the poor methodological quality of the translation process.

Chinese

The Northwick Park Neck Pain Questionnaire (NPQ) is the only neck-specific questionnaire that has been translated in Chinese.[27-28] The NPQ was originally designed to measure the influence of non-specific neck pain on daily activities.[29] The methodological quality of the translation process is poor.[27]

There is strong positive evidence for the reliability of the NPQ (ICC = 0.95).[27] Hypothesis testing resulted in moderate positive evidence for correlation between the NPQ and instruments measuring pain and physical functioning (r = 0.59-0.75).[27-28] Differences in score between subgroups have been reported (e.g. healthy persons vs. neck pain patients, and patients who sought medical consultation vs. those who did not).[27] The average time needed to fill out the NPQ is 5.5 minutes.[27]

The available information on measurement properties of the Chinese NPQ looks promising, despite the poor methodological quality of the translation process.

Dutch

The NDI, NPDS, and Neck Bournemouth Questionnaire (NBQ) have been translated in Dutch.[20 30-32] The NPDS was originally designed to measure pain and disability in patients with neck pain.[2] The NBQ was originally designed to measure pain, physical functioning, social functioning, and psychological functioning in patients with non-specific neck pain.[33] The translation process of the NDI is not described, so the quality of this process is unknown. The methodological quality of the translation process of the NDPS is fair,[20] and of the NBQ is excellent.[31]

There is limited positive evidence for the reliability of the NDI (ICC = 0.90),[32] and for responsiveness (sensitivity = 0.9 and specificity = 0.7 for a clinically important change of 3.5).[30] There is limited negative evidence for its measurement error (MIC = 3.5 and SDC = 10.5 on a 0-50 scale).[30] There is limited positive evidence for the reliability of the NBQ (ICC = 0.92).[31] The result for measurement error of the NBQ is indeterminate, because the MIC is not defined.[31] No floor or ceiling effects have been detected for the NDI or NBQ, and for both questionnaires differences in score between subgroups have been reported (men vs. women).[31-32]

The lack of information derived from these studies makes it difficult to point out the best available neck-specific questionnaire in Dutch. Based on the information available on the measurement properties of the original version of the NDI and NBQ, we advise to use the Dutch NDI.[8]

English

The, originally Danish, Copenhagen Neck Functional Disability Scale (CNFDS) is the only neck-specific questionnaire that has been translated in English.[34] The CNFDS was originally designed to measure disability in patients with neck pain.[35] The translation process is not described, so the quality of this process is unknown. There is limited positive evidence for the responsiveness of the CNFDS (AUC = 0.73).[34] Many neck-specific questionnaires have originally been developed in English.[8] We advise to use one of these questionnaires, preferably the NDI.[8]

Finnish

The NDI and NPDS have been translated in Finnish.[36] The methodological quality of the translation process of these questionnaires is poor.[36]

There is moderate evidence that the NDI is not one-dimensional and that the NPDS has a 3-factor structure.[36] The result for internal consistency of the NDI is indeterminate, because the authors unjustly assume a 1-factor model.[36] There is strong positive evidence for the internal consistency of the NPDS (Cronbach α = 0.82-0.84).[36] No floor or ceiling effects have been detected for the NDI or NPDS and for both questionnaires differences in score between subgroups have been reported (stable vs. improved patients).[36]

The available information suggests that the Finnish NPDS has better measurement properties than the Finnish NDI.

French

The following neck-specific questionnaires have been translated in French: NDI,[21 37] NPDS,[21 37] NBQ,[38] NPQ,[21 37] and CNFDS.[19] The methodological quality of all these translation processes is poor.[19 37-38]

There is limited evidence that the NDI has a 2-factor structure.[21] Hypothesis testing showed that the correlation of the NDI with an instrument measuring psychological functioning is somewhat higher (r = 0.55), than with instruments measuring pain (r = 0.48), and physical functioning (r = 0.50).[21] There is limited evidence that the NPDS has a 3-factor structure.[21] Hypothesis testing showed a positive result for correlation of the NPDS with instruments measuring pain (r = 0.52), and physical functioning (r = 0.63), and a negative result (results slightly below the pre-set criterion of r = 0.5) for correlation with instruments measuring psychological functioning (r = 0.40-0.49).[21] Hypothesis testing showed a positive result for correlation of the NBQ with an instrument measuring pain and physical functioning (r = 0.61-0.67), and a negative result for correlation with an instrument measuring psychological functioning (r = 0.17-0.25).[38] There is limited negative evidence for the responsiveness of the NBQ (r = 0.42).[38] There is limited evidence that the NPQ has a 2-factor structure.[21] Hypothesis testing showed a positive result for correlation of the NPQ with an instrument measuring physical functioning (r = 0.53), and a negative result for correlation with an instrument measuring pain (r = 0.43).[21]

No floor or ceiling effects have been detected for the NDI, NPDS, and NPQ.[21 37] The average time needed to fill out the NDI, NPDS, and NPQ is 7.4, 6.4, and 7.2 minutes, respectively.[37]

The lack of information derived from these studies makes it difficult to point out the best available neck-specific questionnaire in French. Based on the information available on the measurement properties of the original version of the NDI, NPDS, NBQ, NPQ, and CNFDS,[8] we advise to develop a high quality translation of the NDI.

German

The NPDS is the only neck-specific questionnaire that has been translated in German.[25 39] There are two translations of the NPDS in German: one translation process of poor and one of fair methodological quality.[25 39]

Factor analysis provided moderate evidence that the NPDS has a 3-factor structure.[39] The result for internal consistency is indeterminate,[39] because the authors unjustly assume a 1-factor model. There is moderate positive evidence for hypothesis testing (>75% of results in accordance with predefined hypotheses).[39] No floor or ceiling effects have been detected for the NPDS.[39]

The available information on measurement properties of the German NPDS looks promising, despite the poor methodological quality of the translation process.

Greek

The NDI is the only neck-specific questionnaire that has been translated in Greek.[40] The methodological quality of the translation process is good.[40]

Exploratory factor analysis provided moderate evidence that the NDI does not have a 1-factor structure.[40] The result for internal consistency is indeterminate,[40] because the authors unjustly assume a 1-factor model. There is limited negative evidence for responsiveness (r = 0.30 with Global Rating of Change).[40]

Based on the good quality of the translation process and the negative results for unidimensionality and responsiveness, we advise to perform a cross-cultural validation of the Greek NDI.

Hindi

The NPDS is the only neck-specific questionnaire that has been translated in Hindi.[41] The methodological quality of the translation process is fair.[41]

Hypothesis testing showed a positive result for correlation of the NPDS with an instrument measuring psychological functioning (r = 0.80), and a negative result for correlation with an instrument measuring pain (r = 0.30), and an instrument measuring physical functioning (r = 0.15). The average time needed to fill out the NPDS was 8 minutes.[41]

Based on the information derived from this study, we advise to develop a high quality translation of the NDI.

Iranian

The NDI and NPDS have been translated in Iranian.[42] The methodological quality of the translations processes is excellent.[42]

There is limited positive evidence for the internal consistency (Cronbach alpha = 0.88, assuming a 1-factor structure), reliability (ICC = 0.97), and responsiveness (r = 0.65 for physical functioning and r = 0.70 for pain) of the NDI.[42] Exploratory factor analysis resulted in limited positive evidence for a 4-factor structure of the NPDS.[42] There is limited positive evidence for internal consistency (Cronbach alpha = 0.75-0.94 for the four subscales), and reliability (ICC = 0.97).[42] There is limited negative evidence for responsiveness of the NPDS, because correlation with change scores on instruments measuring the same constructs was lower than correlation with instruments measuring other constructs.[42]

No floor or ceiling effects have been detected for the NDI or NPDS.[39]

The Iranian NDI and NPDS both seem to have adequate measurement properties, but we advise using the NDI, based on the negative result for responsiveness of the NPDS and the good measurement properties of the original version of the NDI.[8]

Italian

The NPDS is the only neck-specific questionnaire that has been translated in Italian.[43] The methodological quality of the translation process is poor.[43]

There is limited evidence that the NPDS has a 3-factor structure (variance = 63%).[43] A confirmatory analysis with 4 factors showed a small improvement in variance (67%).[43] Assuming a 3-factor structure, there is limited positive evidence for internal consistency: Cronbach α was 0.92 for "neck dysfunction related to general activities", 0.86 for "cognitive-behavioral aspects", and 0.89 for "neck dysfunction related to activities of the cervical spine".[43] There is limited positive evidence for the reliability of the NPDS (r = 0.89-0.93).[43] The average time needed to fill out the NPDS is 7.5 minutes.[43]

The available information on measurement properties of the Italian NPDS looks promising, despite the poor methodological quality of the translation.

Korean

The NDI and NPDS have been translated in Korean.[44] The methodological quality of the translation processes is poor.[44]

There is limited positive evidence regarding the internal consistency of the NDI (Cronbach α = 0.92, assuming a 1-factor structure).[44] No floor or ceiling effects have been detected for the NDI or NPDS and differences in score between subgroups have been reported (neck pain patients vs. healthy persons).[44]

Lack of information makes it difficult to point out whether the Korean NDI or NPDS has the best measurement properties. Based on the information available on the measurement properties of the original version of the NDI and NPDS,[8] we advise to use the Korean NDI.

Spanish

The NDI, NPQ, and Core Neck Questionnaire (CNQ) have been translated in Spanish.[24,45] The CNQ was originally designed to measure outcomes of care in patients with non-specific neck pain.[46] The methodological quality of the translation process of the NPQ is poor,[45] and of the NDI and CNQ is excellent.[24]

There is limited positive evidence for a 1-factor structure of the NDI and its internal consistency (Cronbach α = 0.89).[47] Hypothesis testing showed a positive result for correlation of the NDI with an instrument measuring pain (r = 0.65), and an instrument measuring physical functioning (r = 0.89).[47] There is limited positive evidence for the responsiveness of the NDI.[47] There is limited negative evidence regarding the reliability of the NPQ (ICC = 0.63).[45] No floor or ceiling effects have been detected for the NDI, NPQ, or CNQ, and scores across different categories of pain intensity have been reported.[24] The average time needed to fill out the NDI and CNQ is 4.0 and 2.1 minutes, respectively.[24]

Based on the available information, we advise to use the Spanish NDI.

Swedish

The NDI is the only neck-specific questionnaire that has been translated in Swedish.[23] The methodological quality of the translation process is unknown. No floor or ceiling effects have been detected for the NDI.[23]

Based on the lack of information, we advise to perform high quality studies to fill in the missing information on the measurement properties of the Swedish NDI.

Turkish

The following neck-specific questionnaires have been translated and evaluated in Turkish: NDI,[48-49] NPDS,[22 49] NPQ,[49] and CNFDS[49]. There are two translations of the NDI in Turkish: one translation process was of excellent methodological quality,[48] and one of fair methodological quality.[49] There are two translations of the NPDS as well: one translation process was of poor methodological quality,[22] and one of fair methodological quality.[49] The translation processes of the NPQ and CNFDS are both of fair methodological quality.[49]

There is moderate positive evidence for the reliability of the NDI (ICC = 0.86-0.98),[48-49] and limited positive evidence for hypothesis testing (r = 0.66-0.73 with instruments measuring pain and/or disability) and responsiveness (r = 0.79, with a physician's assessment of health).[48-49] There is limited positive evidence for the reliability (ICC_{NPDS} = 0.81, ICC_{NPQ} = 0.85, ICC_{CNFDS} = 0.84) and responsiveness (r_{NPDS} = 0.79, r_{NPQ} = 0.81, and r_{CNFDS} = 0.65, with a physician's assessment of health on a scale of 0 to 100) of the NPDS, NPQ, and CNFDS.[49]

The average time needed to fill out the NDI, NPDS, NPQ, and CNFDS is 8.8, 10.2, 8.4, and 6.8 minutes, respectively.[49] All 4 translated questionnaires show promising results, but we advise using the NDI, because of the excellent methodological quality of the translation process and the good measurement properties of the original version.[8]

DISCUSSION

Translated versions of neck-specific questionnaires have been evaluated in 15 different languages. Generally the methodological quality of the translation process is poor, which was mainly due to the fact that the translated version was not pre-tested in the target population. Furthermore, none of the included studies performed a cross-cultural validation. This is necessary to evaluate whether the constructs underlying the original questionnaire are represented adequately by the questionnaire items in the new language. For each questionnaire, except for the Iranian NPDS and Spanish NDI, at least half of the information regarding measurement properties was lacking. More-

over, the evidence for the quality of measurement properties of the translated versions is mostly limited, due to methodological shortcomings of the included studies.

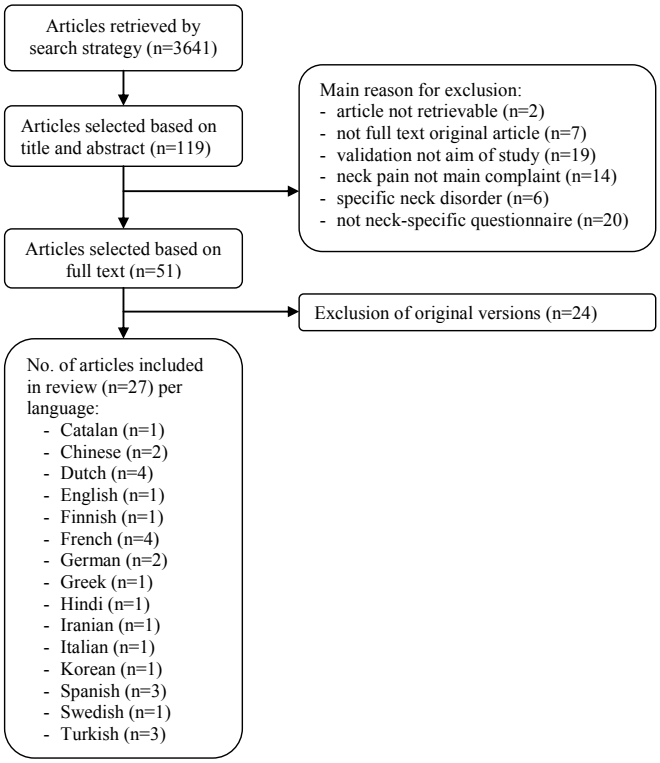

Figure 1 – Flowchart search and selection

The COSMIN checklist has recently been developed and is based on consensus between experts in the field of health status questionnaires.[10] The COSMIN checklist facilitates a separate judgment of the methodological quality of the included studies and their results. This is in line with the methodology of systematic reviews of clinical trials.[16] The criteria in Table 2 are based on the levels of evidence as previously proposed by the Cochrane Back Review Group.[17] The criteria are originally meant for systematic reviews of clinical trials, but we believe that they are also applicable for reviews on measurement properties of health status questionnaires.

Exclusion of non-English papers may introduce selection bias. However, the leading journals, and as a consequence the most important studies, are published in English. So, research performed in populations with a different native language is generally still published in English. This is illustrated by the large number of articles we retrieved

regarding translations of neck-specific questionnaires (see Figure 1). Thus, we argue that the most important translations have been included in our study.

Many studies showed similar methodological shortcomings. Some methodological aspects that need to be improved are: assessment of unidimensionality in internal consistency analysis, the use of stable patients and similar test conditions in studies on reliability and measurement error, and studies on construct validity and responsiveness should be based on predefined hypotheses. We do not discuss these flaws here, because we have elaborated on this subject in a separate paper.[50]

We pooled the results per language, which neglects the fact that populations might share the same language, but differ in cultural context.[3] However, we think that this did not affect our results, because the only inconsistency in results for the same language version was found for the Chinese NPQ and the populations in the two studies evaluating the Chinese NPQ came from the same region in China and were similar in context.[27-28]

A systematic review of the measurement properties of the original version of neck-specific questionnaires showed that for each questionnaire, except for the NDI, at least half of the information regarding measurement properties was lacking.[8] The available results were mainly positive, but the evidence was mostly limited.[8] This systematic review of translated questionnaires shows similar findings, except that the results for construct validity and responsiveness are more frequently inconsistent or negative. These inconsistencies are in correspondence with those found for translations of the McGill Pain Questionnaire.[7] A possible explanation for this difference in results between original questionnaires and their translated counterparts is the poor methodological quality of the translation process and/or lack of cross-cultural validation.[3-4]

A poor translation process and/or lack of cross-cultural validation seem to primarily affect the validity of the questionnaire. This is illustrated by the differences found between the results for structural validity of the translated versions and their original counterparts, and the negative/inconsistent results for hypothesis testing of the translated questionnaires. This is not surprising, as the importance and/or meaning of questionnaire items (e.g. driving, depressed mood) may depend on setting and context. So, a simple translation of the original questionnaire is not sufficient and might affect the underlying constructs. The translation process does not seem to affect the reliability of the questionnaire. This is illustrated by the fact that 95% of the results for internal consistency and reliability are positive, regardless of the methodological quality of the translation process.

A recent review concluded that the translated versions of the NDI into Brazilian-Portuguese, Dutch, French, Korean, and Spanish are of high quality.[6] A possible

explanation for discrepancies with our findings is that the methodological quality of the translation process was not taken into account in that review. The same accounts for a state-of-the-art review of the NDI, in which a list of available translations is recommended, without critical appraisal of the quality of the translation process and cross-cultural validation, nor the quality of the measurement properties.[5]

This study evaluates the measurement properties of translated versions of neck-specific questionnaires, thereby providing an overview of their availability and making it possible to choose the best questionnaire for a specific study population. However, it is advisable to use them cautiously, since the evidence is mostly limited and for each of these translations, except for the Spanish NDI, at least half of the information regarding measurement properties is lacking. For clinical research and practice we advise to use the following questionnaires: the Catalan, Dutch, English, Iranian, Korean, Spanish and Turkish version of the NDI, the Chinese version of the NPQ, and the Finnish, German and Italian version of the NPDS. This is based on the available results for the measurement properties of these translations, and in the case of the Dutch, English, and Korean NDI on the measurement properties of the original version.[8] The Greek NDI needs cross-cultural validation and due to poor methodological quality of the available study there is no information on the Swedish NDI. For all other languages it is advisable to first choose the best available original version of the neck-specific questionnaires and perform a high quality translation of this questionnaire. Our previous systematic review on the original versions of all neck-specific questionnaires showed that the NDI was the best questionnaire.[8]

For future research we recommend performing high quality studies to fill in the information on the unknown measurement properties.

CONCLUSION

Translated versions of neck-specific questionnaires have been evaluated in 15 different languages. Generally the methodological quality of the translation process is poor and none of the included studies performed a cross-cultural validation. A substantial amount of information regarding the measurement properties of translated versions of the different neck-specific questionnaires is still lacking or assessed in studies of poor methodological quality. As a result the available evidence on the measurement properties is mostly limited. So, it is advisable to use the available translated questionnaires cautiously. For the time being we advise to use the following questionnaires in clinical research and practice: the Catalan, Dutch, English, Iranian, Korean, Spanish and Turkish version of the NDI, the Chinese version of the NPQ, and the Finnish,

German and Italian version of the NPDS. The Greek NDI needs cross-cultural validation and there is no methodologically sound information for the Swedish NDI. Studies of high methodological quality are needed to fill in the unknown measurement properties.

For all other languages we advise to translate the original version of the NDI.

REFERENCES

1. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
2. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine 1999;24:1290-4.
3. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25:3186-91.
4. Wang WL, Lee HL, Fetzer SJ. Challenges and strategies of instrument translation. West J Nurs Res 2006;28:310-21.
5. Vernon H. The Neck Disability Index: State-of-the-Art, 1991-2008. J Manipulative Physiol Ther 2008;31:491-502.
6. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther 2009;39:400-17.
7. Menezes da Costa L, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol 2009;62:934-43.
8. Schellingerhout JM, Heymans MW, Verhagen AP, De Vet HC, Koes BW, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2011, doi: 10.1007/s11136-011-9965-9.
9. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.
10. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-49.
11. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.
12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
13. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol 2009;62:1062-7.
14. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press, 2003.
15. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
16. Furlan AD, Pennick V, Bombardier C, van Tulder M, Editorial Board CBRG. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. Spine 2009;34:1929-41.
17. van Tulder M, Furlan A, Bombardier C, Bouter L, Editorial Board CBRG. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine 2003;28:1290-9.

18. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
19. Forestier R, Francon A, Arroman FS, Bertolino C. French version of the Copenhagen neck functional disability scale. Joint Bone Spine 2007;74:155-59.
20. Jorritsma W, de Vries GE, Geertzen JH, Dijkstra PU, Reneman MF. Neck Pain and Disability Scale and the Neck Disability Index: reproducibility of the Dutch Language Versions. Eur Spine J 2010;19:1695-701.
21. Wlodyka-Demaille S, Poiraudeau S, Catanzariti JF, Rannou F, Fermanian J, Revel M. The ability to change of three questionnaires for neck pain. Joint Bone Spine 2004;71:317-26.
22. Bicer A, Yazici A, Camdeviren H, Erdogan C. Assessment of pain and disability in patients with chronic neck pain: reliability and construct validity of the Turkish version of the neck pain and disability scale. Disabil Rehabil 2004;26:959-62.
23. Ackelman BH, Lindgren U. Validity and reliability of a modified version of the neck disability index. J Rehabil Med 2002;34:284-7.
24. Kovacs FM, Bago J, Royuela A, Seco J, Gimenez S, Muriel A, et al. Psychometric characteristics of the Spanish version of instruments to measure neck pain disability. BMC Musculoskelet Disord 2008;9:42.
25. Bremerich FH, Grob D, Dvorak J, Mannion AF. The neck pain and disability scale: Cross-cultural adaptation into german and evaluation of its psychometric properties in chronic neck pain and C1-2 fusion patients. Spine 2008;33:1018-27.
26. Nieto R, Miro J, Huguet A. Disability in subacute whiplash patients: usefulness of the neck disability index. Spine 2008;33:E630-5.
27. Chiu TT, Lam TH, Hedley AJ. Subjective health measure used on Chinese patients with neck pain in Hong Kong. Spine 2001;26:1884-9.
28. Lee KC, Chiu TT, Lam TH. Correlation between generic health status and region-specific functional measures on patients with neck pain. Int J Rehabil Res 2006;29:217-20.
29. Leak AM, Cooper J, Dyer S, Williams KA, Turner-Stokes L, Frank AO. The Northwick Park Neck Pain Questionnaire, devised to measure neck pain and disability. Br J Rheumatol 1994;33:469-74.
30. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. Spine 2007;32:3047-51.
31. Schmitt MA, de Wijer A, van Genderen FR, van der Graaf Y, Helders PJ, van Meeteren NL. The Neck Bournemouth Questionnaire cross-cultural adaptation into Dutch and evaluation of its psychometric properties in a population with subacute and chronic whiplash associated disorders. Spine 2009;34:2551-61.
32. Vos CJ, Verhagen AP, Koes BW. Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. Eur Spine J 2006;15:1729-36.
33. Bolton JE, Humphreys BK. The Bournemouth Questionnaire: A short-form comprehensive outcome measure. II. Psychometric properties in neck pain patients. J Manipulative Physiol Ther 2002;25:141-48.
34. Stewart M, Maher CG, Refshauge KM, Bogduk N, Nicholas M. Responsiveness of pain and disability measures for chronic whiplash. Spine 2007;32:580-5.

35. Jordan A, Manniche C, Mosdal C, Hindsberger C. The Copenhagen neck functional disability scale: A study of reliability and validity. J Manipulative Physiol Ther 1998;21:520-27.
36. Salo P, Ylinen J, Kautiainen H, Arkela-Kautiainen M, Hakkinen A. Reliability and validity of the finnish version of the neck disability index and the modified neck pain and disability scale. Spine 2010;35:552-6.
37. Wlodyka-Demaille S, Poiraudeau S, Catanzariti JF, Rannou F, Fermanian J, Revel M. French translation and validation of 3 functional disability scales for neck pain. Arch Phys Med Rehabil 2002;83:376-82.
38. Martel J, Dugas C, Lafond D, Descarreaux M. Validation of the French version of the Bournemouth Questionnaire. Journal of the Canadian Chiropractic Association 2009;53:102-10.
39. Scherer M, Blozik E, Himmel W, Laptinskaya D, Kochen MM, Herrmann-Lingen C. Psychometric properties of a German version of the neck pain and disability scale. Eur Spine J 2008;17(7):922-9.
40. Trouli MN, Vernon HT, Kakavelakis KN, Antonopoulou MD, Paganas AN, Lionis CD. Translation of the Neck Disability Index and validation of the Greek version in a sample of neck pain patients. BMC Musculoskelet Disord 2008;9:106.
41. Agarwal S, Allison GT, Agarwal A, Singer KP. Reliability and validity of the Hindi version of the Neck Pain and Disability Scale in cervical radiculopathy patients. Disabil Rehabil 2006;28:1405-11.
42. Mousavi SJ, Parnianpour M, Montazeri A, Mehdian H, Karimi A, Abedi M, et al. Translation and validation study of the Iranian versions of the neck disability index and the neck pain and disability scale. Spine 2007;32:E825-E31.
43. Monticone M, Baiardi P, Nido N, Righini C, Tomba A, Giovanazzi E. Development of the Italian version of the Neck Pain and Disability Scale, NPDS-I: cross-cultural adaptation, reliability, and validity. Spine 2008;33:E429-34.
44. Lee H, Nicholson LL, Adams RD, Maher CG, Halaki M, Bae SS. Development and psychometric testing of Korean language versions of 4 neck pain and disability questionnaires. Spine 2006;31:1841-5.
45. Gonzalez T, Balsa A, Sainz de Murieta J, Zamorano E, Gonzalez I, Martin-Mola E. Spanish version of the Northwick Park Neck Pain Questionnaire: reliability and validity. Clin Exp Rheumatol 2001;19:41-6.
46. White P, Lewith G, Prescott P. The core outcomes for neck pain: Validation of a new outcome measure. Spine 2004;29:1923-30.
47. Andrade Ortega JA, Delgado Martinez AD, Almecija Ruiz R. Validation of the Spanish version of the Neck Disability Index. Spine 2010;35:E114-8.
48. Aslan E, Karaduman A, Yakut Y, Aras B, Simsek IE, Yagly N. The cultural adaptation, reliability and validity of neck disability index in patients with neck pain: a Turkish version study. Spine 2008;33:E362-5.
49. Kose G, Hepguler S, Atamaz F, Oder G. A comparison of four disability scales for Turkish patients with neck pain. J Rehabil Med 2007;39:358-62.
50. Terwee CB, Schellingerhout JM, Verhagen AP, de Vet HC, Koes BW. Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: room for improvement. J Manipulative Physiol Ther 2011;43:261-72.

Chapter 7

General Discussion

A recent initiative of the Neck Pain Task Force (NPTF) to collect and combine the available evidence on neck pain, resulted in the overall conclusion that "neck pain is common, and its determinants and prognosis are multifactorial".[1] This conclusion is rather vague and suggests that our knowledge regarding the optimal management of patients with non-specific neck pain is scarce. Therefore, the aim of this thesis was to get insight in and to improve the prognosis of patients with non-specific neck pain. This was established by evaluating the following questions:

Is it possible to predict the probability of persistent complaints in patients with non-specific neck pain? (Chapter 3)

Are there subgroups of patients with non-specific neck pain that are more likely to benefit from either physiotherapy, spinal manipulation therapy, or usual care? (Chapter 4)

Furthermore, a reasonable amount of information regarding the etiology and clinical management of non-specific neck pain is of questionable quality, due to methodologic shortcomings of the underlying studies. This makes it necessary to be cautious with clinical application of information derived from these studies. Therefore, we also paid attention in this thesis to the methodology of studies on prognosis of neck pain and to the quality of measurement instruments used in studies on neck pain. The following questions were formulated:

What is the influence of various categorization strategies for continuous variables on the final model content and model performance in multivariable logistic regression analysis? (Chapter 2)

What is the quality of the measurement properties of the different neck-specific questionnaires? (Chapter 5)

Is there a difference in quality of measurement properties between original and translated versions of neck-specific questionnaires? (Chapter 6)

PREDICTION OF PERSISTENCE OF NON-SPECIFIC NECK PAIN

Is it possible to predict the probability of persistent complaints in patients with non-specific neck pain? Based on our findings presented in Chapter 3, we can say that it is possible. The score chart presented in Table 1 predicts significantly better for every patient whether he/she will develop persistent complaints, than estimates for the overall population. The score chart has a moderate discriminative ability, an adequate calibration, and a good fit. However, the explained variation of the score chart is low. This means that only a small part of the differences between patients regarding persistence of complaints is explained by the predictors in the score chart.

Table 1 – Score chart

			Score
Age	+ 7	/ 10 yr[1]	
Accompanying low back pain	+ 21		
Traumatic cause neck complaints	+ 6		
Health status (scale 0-100)[2]	- 3	/ 25 points[3]	
Accompanying headache	+ 5		
No accompanying headache			
Radiation of pain to elbow/shoulder	- 14		
Previous neck complaints	+ 13		
Paid employment	+ 9		
Pain intensity (scale 0-10)[4]	- 1	/ point	
Accompanying headache			
Radiation of pain to elbow/shoulder	- 4		
Previous neck complaints	+ 4		
Paid employment	- 11		
Pain intensity (scale 0-10)[4]	+ 2	/ point	
		Total score	

Total score	Probability[5]
< 10	0 - 20%
10 - 34	20 - 40%
35 - 54	40 - 60%
55 - 79	60 - 80%
> 79	80 - 100%

An example of how to get the score of an individual patient is demonstrated in Appendix A

[1] The score increases with 7 points per 10 year (e.g. a 40-year old person receives a score of 4x7=28 points)

[2] Question: "Can you rate your own health status today?" (0=worst imaginable, 100=best imaginable)

[3] The score decreases with 3 points per 25 points on the health status scale (e.g. a person with a score of 75 receives a score of 3x-3=-9 points)

[4] Question: "Can you rate your current pain intensity?" (0=no pain, 10=worst imaginable pain)

[5] Probability that neck complaints will still be present at 6 months after the first consultation

A remarkable observation is the prognostic behavior of accompanying headache, especially the counterintuitive interaction between accompanying headache and previous neck pain, and employment status. Accompanying headache improves the probability of recovery in those who are employed and patients with previous neck pain. This counterintuitive behavior of accompanying headache was also observed by van der Velde et al., who performed a Rasch analysis of the Neck Disability Index (NDI).[2] The prevalence of accompanying headache was higher than expected in neck pain patients with lower levels of disability, and lower than expected in those with higher levels of disability.[2]

Finally, the lack of prognostic value of duration of complaints is interesting, because patients with chronic neck pain are generally considered to differ prognostically from those with (sub)acute complaints. This observation could be attributable to the categorization of duration of complaints in our study. However, considering the inconsistent prognostic value of this characteristic in other studies as well,[3-7] it is also plausible that the clinical value of duration of complaints is overrated.

The direction of the associations of the individual predictors in the score chart (i.e. worse or better prognosis) is consistent with those found in previous studies.[3-7] However, it is noteworthy that none of the predictors in our study or previous studies consistently has a (major) impact on prognosis,[3-7] which explains the low explained variation of the score chart. Addition of predictors with a major impact on prognosis would improve the performance of the score chart. The search for predictors with a large impact on prognosis is complicated by the fact that the most plausible predictors have already been evaluated.[3-7] New insights in the origin of non-specific neck pain could provide new variables to be evaluated. These new variables might improve the performance of the score chart, because causally related characteristics usually have a large impact on prognosis. However, the etiology of non-specific neck pain is unknown and the fruitless search for causal factors in the last decades makes it unlikely that insights will improve in the short term. Psychological and genetic factors are largely unexplored and could have prognostic value. However, their usefulness must be established, because the clinical value of predictors depends on several characteristics of the predictor. First, predictors should be variables that are easily obtainable and reproducible. Genetic factors, for example, are not easily obtainable and variables derived from physical examination do not seem to be useful, due to their poor to moderate reproducibility.[8] Second, in an ideal situation predictors should be modifiable variables, to facilitate the possibility for intervention. Future research should be directed at finding additional predictors for persistence of complaints in patients with non-specific neck pain. However, it is unlikely that this search will result in easily obtainable, modifiable, and reproducible predictors with a large impact on prognosis, until we know more about the etiology of non-specific neck complaints.

In order to validate a clinical prediction rule it is of utmost importance to evaluate it in a group of patients, other than the development population (e.g. other point in time, or region of origin). This so-called external validation has been performed for only one of the other currently available clinical prediction rules for neck pain and resulted in negative findings regarding the validity of the prediction rule.[9-10] In contrast to these negative findings, external validation showed that our score chart performs well in a sample from the primary care population in the United Kingdom, despite the different

distribution of characteristics. These findings strengthen the clinical value of our score chart and makes it likely that the score chart is generalizable to adults in primary care populations in Western society.

Our score chart is the first clinically applicable prediction rule that quantifies the risk of persistent complaints in patients presenting with non-specific neck pain in primary care. For clinical practice it means that a physician now has insight in the prognosis of an individual patient with neck pain, and is able to inform patients more accurately about their expected prognosis. Furthermore, it helps researchers in selecting patients at high risk in studies on prevention of chronic neck pain. However, the "gut feeling" of general practitioners (GPs) regarding the risk of persistent complaints in patients with non-specific neck pain seems to be as adequate as our score chart at the moment (discriminative ability: 0.67 vs. 0.66).[11] GPs generally differ in clinical experience and expertise, resulting in a difference in adequacy of their "gut feeling". Our score chart will improve the discriminative ability of GPs, and other care providers, with a below average "gut feeling". The problem is that we do not know which GPs have a below average "gut feeling" and GPs probably don't know it themselves as well. So, it is impossible to indicate which GP should use our score chart to improve his/her discriminative ability. However, an advantage of the score chart is that it clearly states which factors influence prognosis, in contrast to the unknown origin of "gut feeling".

The application of our model might be hampered by the number of characteristics in the score chart and the resulting more difficult calculation of expected prognosis. However, the calculation is simplified by the availability of an online application (http://www.necksolutions.com/neck-pain-prognosis.html). This is not only helpful for physicians, but also makes it possible for patients with non-specific neck pain to get an estimate of their own prognosis.

A DECISION MODEL FOR TREATMENT OF NON-SPECIFIC NECK PAIN

Based on our findings in Chapter 4 there appear to be subgroups of patients with non-specific neck pain that are more likely to benefit from physiotherapy, spinal manipulation therapy, or usual care by a GP (i.e. advice on self-care combined with analgesics and/or muscle relaxation medication). We identified three characteristics that are useful to guide treatment selection in patients with non-specific neck pain: pain intensity at baseline, (absence of) low back pain and age. These characteristics do not only predict recovery like the predictors identified in Chapter 3, but also modify the effect of treatment. These three modifying characteristics were used to develop

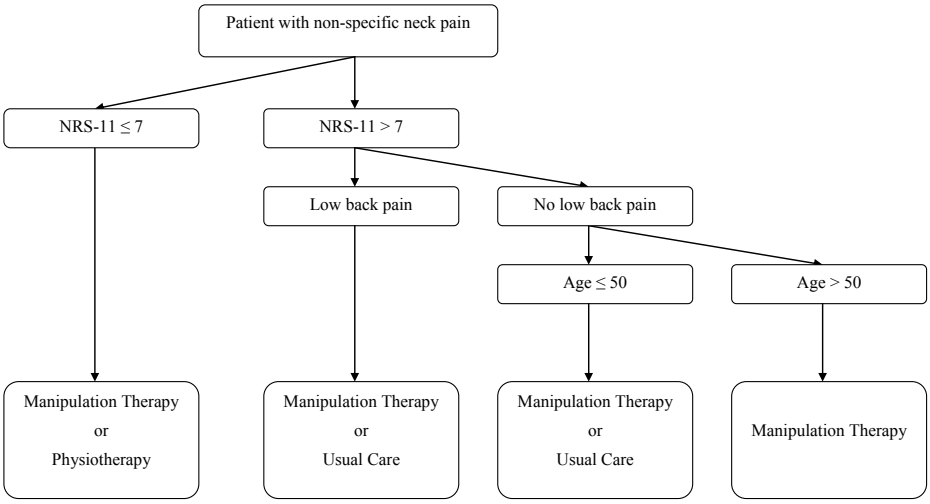

Figure 1 – Decision model for treatment of non-specific neck pain
NRS-11 = pain intensity at baseline on a 11-point Numerical Rating Scale

a decision model to improve the recovery rate in patients consulting their GP for non-specific neck pain (see Figure 1).

The most prominent distinction made by the decision model is between patients with mild to moderate neck pain (NRS ≤ 7) and those with severe neck pain (NRS-11 > 7). The model suggests to focus on restoring physical functioning or prevention of physical disability in those with mild to moderate neck pain and to focus on pain relief in those with severe pain (see Figure 1). To our knowledge, only one primary care guideline for management of acute neck pain is available. This guideline recommends to focus on pain relief and restoration of normal activities, but does not specify the corresponding treatment(s).[12] However, our findings are in correspondence with international primary care guidelines for the management of low back pain, which recommend to improve/restore physical functioning by applying exercise therapy or spinal manipulation therapy and to decrease pain by applying analgesics or spinal manipulation therapy.[13]

In the design of a study with the aim to develop a decision model some methodological aspects are of vital importance (e.g. sample size, rationale for the model, and internal and external validation).[9] The development of our decision model meets all methodological criteria, except for external validation.[9] Evaluation of the decision model in a group of patients other than the development population is important to determine its clinical value. Furthermore, the exact gain in recovery rate of application of our decision model is unknown, because the randomized treatment allocation in

our study did not resemble the choice for a specific therapy in daily practice. It would be useful to carry out a (cluster) randomized trial, comparing current practice of GPs with GPs applying the decision model. This will not only facilitate external validation of the model, but will also quantify the overall improvement in recovery rate.

Successful implementation of our decision model in clinical practice not only depends on a possible improvement in recovery rate, but also on other aspects like feasibility of the decision model. To evaluate feasibility we conducted a pilot study. We presented the decision model to 23 general practitioners (GPs) and physiotherapists, all practicing in the city of Krimpen aan de IJssel. We asked them whether they judged the decision model to be useful and if it was feasible for use in daily practice.[14] The poll showed that 74% of the GPs and physiotherapists considered the model to be (very) useful and that 78% said that it would be (very) easy to use.[14] These results are promising, but do not guarantee that they will actually use the decision model.

Our decision model points out two treatments as most effective in different subgroups (see Figure 1). This makes it possible to take patient preference, adverse events, and costs into account. Among the patients included in the study population 41% of the patients reported a preference for either physiotherapy or spinal manipulation therapy (see Chapter 4, Table 1), but we do not know if our sample is representative regarding patient preference.

Prescription of analgesics is frequently part of usual care. Adverse events associated with paracetamol consist mainly of allergic reactions.[15] Non steroidal anti-inflammatory drugs (NSAIDs) frequently cause gastro-intestinal complaints like nausea, epigastric pain and diarrhea and, in rare cases, may cause serious adverse events like gastric bleeding, myocardial infarction, and stroke.[15] The adverse events caused by physiotherapy and spinal manipulation therapy are usually benign and self-limiting (e.g. headache, transient neurological symptoms, nausea, and dizziness).[16-17] However, spinal manipulation therapy rarely results in serious cerebrovascular events like artery dissection, and stroke.[16 18]

The initial costs for usual care are lower than for physiotherapy or spinal manipulation therapy,[19] but in the long run spinal manipulation therapy is the most cost-effective.[20] The costs of neck pain for society are rapidly increasing,[21] so studies to improve effectiveness of interventions are not only very welcome from a clinical viewpoint, but also from an economic perspective.

A recent systematic review of all conservative interventions for non-specific neck pain showed that spinal manipulation therapy, exercise therapy, acupuncture, and analgesics provide a clinically relevant improvement in the short term (i.e. at the end of treatment), compared to placebo or a wait-and-see policy.[22] The same accounts

for laser therapy on the medium term (3-9 months).[22] However, long term benefits of these interventions have not been demonstrated.[22] The conclusions regarding the effectiveness of conservative interventions have not changed much over the years, despite the increasing amount of evidence. It seems unlikely that future RCTs evaluating interventions in the overall group of patients with non-specific neck pain will provide new insights. Therefore, directing future research towards the development/validation of decision models using patient characteristics, instead of repeating RCTs for interventions in the overall population, seems more useful. We recommend continuing with validating and/or extending our decision model, before considering the development of a new model.

CATEGORIZATION OF CONTINUOUS VARIABLES

In medical research continuous variables are frequently splitted into two or more categories when multivariable logistic regression models are developed. The advantage of categorization is that it simplifies the interpretation of the model and the application in clinical practice.[23] However, our findings in Chapter 2 show that this type of modification of continuous variables comes at a cost.

Continuous variables were introduced in our analysis in three ways: continuous, split into more than two categories, further referred to as stratification, and dichotomized at the median. Dichotomization or stratification of continuous variables prior to the model selection procedure not only resulted in different predictors in the final multivariable model, but resulted in a poorer performance of the model as well. Dichotomization or stratification of continuous variables after the selection procedure also resulted in a poorer performance of the model, but at least ensured the most adequate model content. So, we recommend introducing continuous variables as such in (logistic) regression analysis. If categorization of continuous variables is desirable we suggest doing this after the selection procedure.

We are the first to evaluate the effect on model content of categorization of continuous variables in logistic regression analysis. However, the negative impact on performance of dichotomization in regression analysis has been demonstrated before.[23] Dichotomization has even more impact if so-called "optimal" cutoff points are used, because they introduce additional biases in the analysis (e.g. overestimation of the discriminative ability).[23-25] In an attempt to decrease the negative influence of "optimal" cutoff points a polychotomization method has been developed for regression models.[26] This polychotomization method results in an essentially unbiased estimate of discriminative ability.[26]

We demonstrated the effect of categorization of continuous variables on model content and performance in our cohort of patients with non-specific neck pain. Comparison of the content of our model with predictors identified in previous studies, shows that a large part of the predictors for persistence of non-specific neck pain demonstrates a "borderline behavior": none of the predictors, identified in this study or previous studies, consistently has a (major) impact on prognosis.[5 7-10] It is possible that this "borderline behavior" contributed to the differences in model content.

We think that it is unlikely that our choice for a backward selection procedure, instead of a forward selection procedure, influenced the results. In a forward selection procedure categorization of continuous variables could similarly result in selection of a different set of predictors.

The observed differences in performance between the categorization strategies (continuous, stratification, dichotomization) were small. Especially the difference between dichotomization and stratification was marginal, as was the difference in performance between categorization before and after the selection procedure. This is probably due to the fact that there was only one continuous variable with a strong association with the outcome. Furthermore, the observed discriminative ability and explained variation were rather low, which might have decreased the contrast in performance between the different categorization strategies. It would be useful to repeat this study in a population with a complaint that has several continuous variables with a strong (non-linear) relationship with the outcome. This will facilitate better insight in possible differences between the different categorization strategies. For future research we suggest to evaluate the effect of categorization of continuous variables in models for other complaints and in models with much higher discriminative ability and explained variation.

NECK-SPECIFIC QUESTIONNAIRES

Several disease-specific questionnaires have been developed to measure pain and/or disability in patients with neck pain (e.g. Neck Disability Index (NDI), Neck Pain and Disability Scale (NPDS)).[27-28] In order to make a rational choice for the use of these questionnaires in clinical research and practice it is important to assess and compare their measurement properties (e.g. reliability, validity, and responsiveness).[29]

A recent review of the cross-cultural adaptations of the McGill Pain Questionnaire showed inconsistent findings for the quality of the measurement properties of different language versions.[30] These differences are probably caused by differences in cultural context.[30] We assume that the same accounts for neck-specific question-

naires. Therefore we decided to evaluate the quality of the measurement properties of original and translated versions of neck-specific questionnaires separately.

Original versions of neck-specific questionnaires

The results from the systematic review in Chapter 5 show that the quality of the measurement properties of the original version of the eight different neck-specific questionnaires is largely uncertain: except for the NDI, the evidence regarding measurement properties is mostly limited and for at least half of the measurement properties per questionnaire information is lacking.

To critically appraise the methodological quality of the included studies, we applied the recently developed "COnsensus-based Standards for the selection of health status Measurement INstruments" (COSMIN) checklist.[31] Appraisal of the methodological quality of the included studies showed that most of the studies have the same methodological shortcomings (e.g. inadequate sample size, no assessment of unidimensionality in internal consistency analysis, lack of predefined hypotheses in studies on construct validity and responsiveness).[32] The critical appraisal of the methodological quality of the included studies frequently resulted in poor and fair ratings, which indicates that most studies provide only limited evidence. However, the methodologically adequate studies showed predominantly positive results. So, the results, if available, are generally positive, but the evidence is mostly limited. The quality of measurement properties in previous reviews on neck-specific questionnaires was generally rated higher.[8 33-35] This is probably caused by the fact that previous reviews did not perform an (adequate) critical appraisal of the methodological quality of the included studies.[8 33-35]

So, one may ask what the value of neck-specific questionnaires is for research and clinical practice. For research purposes and in daily clinical practice neck-specific questionnaires are applied to measure response to treatment over time. However, we recommend being cautious with interpreting the results from these measurements, because the NDI and Northwick Park Neck Pain Questionnaire (NPQ) are the only questionnaires with at least moderate positive evidence for responsiveness. Another important aspect for clinical practice of neck-specific questionnaires is the interpretability of a questionnaire (i.e. the degree to which one can assign qualitative meaning to quantitative scores).[36] Currently we lack knowledge on the interpretability of most neck-specific questionnaires, because differences in scores between several subgroups have only been reported for the NDI, NPDS, and Core Whiplash Outcome Measure (CWOM) and the minimal important change (MIC) is unknown for all questionnaires. In other words, we do not know for any of the neck-specific questionnaires which change in score reflects an actual change in health status.

The increase in studies evaluating measurement properties of neck-specific questionnaires, since the first systematic review in 2002, has not led to more clarity regarding the quality of the measurement properties, due to methodological shortcomings. So, it is of vital importance to perform high quality studies, to gain insight in the quality of the measurement properties of neck-specific questionnaires. The methodological quality can be improved in future studies by applying the criteria mentioned in the COSMIN checklist.[31] Until these studies are available we recommend using the NDI, since this is the most frequently evaluated questionnaire and its measurement properties seem adequate.

For future research we recommend to refrain from developing new neck-specific questionnaires until high quality studies show strong evidence that the measurement properties of current questionnaires are poor. Only in that case development of a new questionnaire is indicated.

Translated versions of neck-specific questionnaires

Is there a difference in quality of measurement properties between original and translated versions of neck-specific questionnaires? The systematic review in Chapter 6 shows that there are similarities between the translated and original versions of neck-specific questionnaires: the evidence was mostly limited and generally at least half of the information regarding measurement properties was lacking. However, the results for measurement properties of translated versions of neck-specific questionnaires are more frequently negative or inconsistent, compared to the results for the original versions. The latter observation is in correspondence with our assumption and the review of the cross-cultural adaptations of the McGill Pain Questionnaire for low back pain.[30] The more frequently negative or inconsistent results for measurement properties of translated versions is probably caused by the generally poor methodological quality of the translation process and the fact that none of the studies included in our systematic review performed a cross-cultural validation.

The poor methodological quality of the translation process and lack of cross-cultural validation seem to primarily affect the validity of the questionnaire. This is demonstrated by the differences in results, between the translated versions and their original counterparts, regarding structural validity and hypothesis testing. This is not surprising, as the importance and/or meaning of questionnaire items (e.g. driving, depressed mood) may depend on culture, setting and context. The translation process does not seem to affect the reliability of the questionnaire; almost all results for internal consistency and reliability are positive, regardless of the methodological quality of the translation process (see Chapter 6). This could be explained by the fact that patients will interpret a question in the same way on different occasions, even if the interpretation is incorrect.

Previous systematic reviews on neck-specific questionnaires combined the results of studies on measurement properties, regardless of the language version of the questionnaire.[8 33 35 37] However, based on our findings we recommend to evaluate the questionnaires per language, as long as a proper translation process and cross-cultural validation have not been carried out. This will result in more consistency regarding the quality of measurement properties of a specific version, by decreasing the amount of negative results caused by a poor translation process and/or lack of cross-cultural validation.

IMPLICATIONS AND RECOMMENDATIONS FOR CLINICAL PRACTICE AND RESEARCH

Clinical guidelines in the field of neck pain are still scarce.[38] This in contrast to, for example, the number of clinical guidelines for the management of low back pain.[13] Development and implementation of clinical guidelines is imperative to achieve a worldwide improvement in the management of patients with non-specific neck pain in clinical practice. The new insights in prognosis and treatment of patients with non-specific neck pain provided by this thesis could contribute to the development of clinical guidelines.

The findings in this thesis have the following implications for clinical practice:

Insight in the probability of non-specific neck complaints lasting at least 6 months can be gained by applying the externally validated score chart presented in Table 1.

The characteristics "pain intensity at baseline", "(absence of) low back pain", and "age" modify the effect of physiotherapy, spinal manipulation therapy, and usual care. An improvement in recovery rate can be established in patients with non-specific neck pain, by applying the decision model for treatment presented in Figure 1.

The findings in this thesis have the following implications for research:

When developing prognostic models continuous variables should be introduced as such in (logistic) regression analysis and should not be categorized. This will result in a model with the best representation of the actual associations and with the highest performance. If categorization of continuous variables is desirable we suggest doing this after the selection procedure of the variables in the model.

The English version of the NDI is the most frequently evaluated neck-specific questionnaire and its measurement properties seem adequate. The other (translated versions of) neck-specific questionnaires should be used with caution.

In systematic reviews on measurement properties, the results for measurement properties of translated versions of neck-specific questionnaires and their original counterparts should not be combined, unless a methodologically adequate translation process and cross-cultural validation have been carried out.

For future research we recommend:

To perform studies on the etiology of non-specific neck complaints, because more insight in the origin of non-specific neck pain could provide new (modifiable) clinical characteristics that predict the course of disease or the effectiveness of treatment.

To focus on development/validation of decision models for treatment based on patient characteristics, instead of repeating RCTs for interventions in the overall population.

To refrain from developing new neck-specific questionnaires until high quality studies show strong evidence that the measurement properties of current questionnaires are poor. Only in that case development of a new questionnaire is indicated.

To apply the criteria mentioned in the COSMIN-checklist when designing a study to evaluate the measurement properties of a neck-specific questionnaire.[31]

REFERENCES

1. Haldeman S, Carroll L, Cassidy JD. Findings from the bone and joint decade 2000 to 2010 task force on neck pain and its associated disorders. J Occup Environ Med 2010;52:424-7.
2. van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A. Rasch analysis provides new insights into the measurement properties of the neck disability index. Arthritis Rheum 2009;61:544-51.
3. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain 1998;77:1-13.
4. Hoving JL, de Vet HC, Twisk JW, Deville WL, van der Windt D, Koes BW, et al. Prognostic factors for neck pain in general practice. Pain 2004;110(3):639-45.
5. Kjellman G, Skargren E, Oberg B. Prognostic factors for perceived pain and function at one-year follow-up in primary care patients with neck pain. Disabil Rehabil 2002;24:364-70.
6. Hill J, Lewis M, Papageorgiou AC, Dziedzic K, Croft P. Predicting persistent neck pain: a 1-year follow-up of a population cohort. Spine 2004;29:1648-54.
7. Cote P, Cassidy JD, Carroll LJ, Kristman V. The annual incidence and course of neck pain in the general population: a population-based cohort study. Pain 2004;112:267-73.
8. Nordin M, Carragee EJ, Hogg-Johnson S, Weiner SS, Hurwitz EL, Peloso PM, et al. Assessment of neck pain and its associated disorders: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine 2008;33:S101-22.
9. Stanton TR, Hancock MJ, Maher CG, Koes BW. Critical appraisal of clinical prediction rules that aim to optimize treatment selection for musculoskeletal conditions. Phys Ther 2010;90:843-54.
10. Cleland JA, Mintken PE, Carpenter K, Fritz JM, Glynn P, Whitman J, et al. Examination of a clinical prediction rule to identify patients with neck pain likely to benefit from thoracic spine thrust manipulation and a general cervical range of motion exercise: multi-center randomized clinical trial. Phys Ther 2010;90:1239-50.
11. Vos CJ, Verhagen AP, Koes BW. General practitioner's gut feeling about the prognosis of acute neck pain patients: an accurate predictor. Submitted.
12. Australian Acute Musculoskeletal Pain Guidelines Group. Evidence-based management of acute musculoskeletal pain. Bowen Hills: Australian Academic Press, 2003.
13. Koes BW, van Tulder M, Lin CW, Macedo LG, McAuley J, Maher C. An updated overview of clinical guidelines for the management of non-specific low back pain in primary care. Eur Spine J 2010;19:2075-94.
14. Beentjes E, Schellingerhout JM, Verhagen AP. Feasibility of a decision model for treatment and referral of patients with non-specific neck pain in general and physiotherapy practice. Rotterdam, 2009.
15. American Society of Health-System Pharmacists. AHFS Consumer Medication Information. Bethesda, 2008.
16. Rubinstein SM. Adverse events following chiropractic care for subjects with neck or low-back pain: do the benefits outweigh the risks? J Manipulative Physiol Ther 2008;31:461-4.
17. Carlesso LC, Gross AR, Santaguida PL, Burnie S, Voth S, Sadi J. Adverse events associated with the use of cervical manipulation and mobilization for the treatment of neck pain in adults: a systematic review. Man Ther 2010;15:434-44.
18. Paciaroni M, Bogousslavsky J. Cerebrovascular complications of neck manipulation. Eur Neurol 2009;61:112-8.

19. Lewis M, James M, Stokes E, Hill J, Sim J, Hay E, et al. An economic evaluation of three physiotherapy treatments for non-specific neck disorders alongside a randomized trial. Rheumatology (Oxford) 2007;46:1701-8.
20. Korthals-de Bos IB, Hoving JL, van Tulder MW, Rutten-van Molken MP, Ader HJ, de Vet HC, et al. Cost effectiveness of physiotherapy, manual therapy, and general practitioner care for neck pain: economic evaluation alongside a randomized controlled trial. BMJ 2003;326:911.
21. Martin BI, Deyo RA, Mirza SK, Turner JA, Comstock BA, Hollingworth W, et al. Expenditures and health status among adults with back and neck problems. JAMA 2008;299:656-64.
22. Leaver AM, Refshauge KM, Maher CG, McAuley JH. Conservative interventions provide short-term relief for non-specific neck pain: a systematic review. J Physiother 2010;56:73-85.
23. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127-41.
24. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 1994;86:829-35.
25. Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. Stat Med 2004;23:1159-78.
26. Tsuruta H, Bax L. Polychotomization of continuous variables in regression models based on the overall C index. BMC Med Inform Decis Mak 2006;6:41.
27. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
28. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine 1999;24:1290-4.
29. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press, 2003.
30. Menezes da Costa L, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol 2009;62:934-43.
31. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-49.
32. Terwee CB, Schellingerhout JM, Verhagen AP, de Vet HC, Koes BW. Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: a systematic review. J Manipulative Physiol Ther 2011;43:261-72.
33. Vernon H. The Neck Disability Index: State-of-the-Art, 1991-2008. J Manipulative Physiol Ther 2008;31:491-502.
34. Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. Arch Phys Med Rehabil 2008;89:69-74.
35. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther 2009;39:400-17.
36. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.

37. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. Spine 2002;27:515-22.
38. Balagué F. The Bone and Joint Decade (2000–2010) Task Force on Neck Pain and Its Associated Disorders: A Clinician's Perspective. Spine 2008;33:S4.

Chapter 8.1

Summary

Chapter 1 is an introduction of the subject and aims of this thesis. In summary, the aims of this thesis were:

- To evaluate the influence of various categorization strategies of candidate variables on the final model content and performance when developing a multivariable logistic regression model. (Chapter 2)
- To develop and externally validate a prediction rule that estimates the probability of persistent complaints in non-specific neck pain patients. (Chapter 3)
- To develop a decision model that points out which subgroups of patients with non-specific neck pain are more likely to benefit from either physiotherapy, spinal manipulation therapy, or usual care. (Chapter 4)
- To systematically review the measurement properties of neck-specific questionnaires. (Chapter 5 and 6)

In medical research continuous variables are frequently splitted into two or more categories when multivariable logistic regression models are developed. The advantage of categorization is that it simplifies the interpretation of the model and the application in clinical practice.[1] In Chapter 2 the influence of various categorization strategies on the final model content and performance in multivariable logistic regression analysis is evaluated. Categorization of continuous variables prior to introducing them into backward multivariable logistic regression analysis resulted in different predictors remaining in the final model and loss of model performance. Every categorization strategy resulted in a loss of information, but stratification of continuous variables seems to result in a somewhat better performance than dichotomization. It is advisable to retain continuous variables as such during the development of the model and, if unavoidable, categorize after variable selection. This results in a slight reduction of the model performance, but at least will provide the most accurate model content.

Some studies on characteristics associated with persistence of neck complaints in the general population have been conducted.[2-6] However, none of the studies constructed a prediction model that quantifies prognosis. The objective in Chapter 3 was to develop and validate a prediction rule that estimates the probability of complaints persisting for at least 6 months in patients presenting with non-specific neck pain in primary care. The study population consisted of adults recruited from primary care practices (n=468) in The Netherlands presenting with non-specific neck pain. The outcome measure was global perceived recovery measured at 6 months of follow-up. Seventeen baseline characteristics of the patients were considered in the multivariable analysis. The multivariable analysis resulted in a set of 9 predictors. A score chart was constructed by using the regression coefficient estimates. The score chart has a discriminative ability (AUC) of 0.66. The score chart was then externally validated in a

cohort of patients with non-specific neck pain (n=315) recruited in primary care in the United Kingdom. External validation of the score chart showed a discriminative ability of 0.65, an adequate calibration, and a good fit. The prediction which neck pain patients are more likely to develop persistent complaints, or to recover, is significantly improved by the score chart, compared to the probability estimates based on the prevalence.

Systematic reviews show that there is a positive effect of physiotherapy and spinal manipulation therapy in comparison to placebo or usual care, in patients with non-specific neck pain, but these effects are relatively small.[7-9] Heterogeneity of the included study populations with non-specific neck pain might be a reason for the small effect. It could well be that certain subgroups within this population have a larger benefit of one of these treatments due to their prognostic status. The aim of Chapter 4 was to develop a decision model that points out which subgroups of patients with non-specific neck pain are more likely to benefit from physiotherapy, spinal manipulation therapy, or usual care. For that purpose several patient characteristics were examined to assess whether they modified the effect of treatment. It turned out that three predictors for recovery modified the treatment effect: pain intensity in the short-term model, age and (no) accompanying low back pain in the long-term model. From the different models it can be concluded that all patients benefit from spinal manipulation therapy, that physiotherapy should be applied to patients with a medium to low pain intensity, and that usual care should be applied to patients with high pain intensity and of younger age. Application of the decision model resulted in up to 25% improvement in recovery rate in patients receiving a tailored instead of a non-advised treatment.

Several disease-specific questionnaires have been developed to measure pain and/or disability in patients with neck pain (e.g. Neck Disability Index (NDI), Neck Pain and Disability Scale (NPDS)).[10-11] In order to make a rational choice for the use of these questionnaires in clinical research and practice it is important to assess and compare their measurement properties (e.g. reliability, validity, and responsiveness).[12] Pooling of the measurement properties of different language versions may result in inconsistent findings, caused by differences in cultural context.[13] Therefore, we decided to evaluate the translated versions of neck-specific questionnaires separately.

The objective in Chapter 5 was to critically appraise and compare the measurement properties of the original versions of neck-specific questionnaires. The literature search resulted in a total of 3641 unique hits, of which 25 articles, evaluating 8 different questionnaires, were included in our study. The NDI is the most frequently evaluated questionnaire and shows strong evidence for a positive result for internal consistency and structural validity, moderate evidence for a positive result for hypothesis testing

and responsiveness, and limited evidence for a positive result for content validity and a negative result for reliability. The other questionnaires show positive results, but the evidence for each measurement property is mostly limited and at least 50% of the information on measurement properties per questionnaire is lacking. Our findings imply that studies of high methodological quality are needed to properly assess the measurement properties of the currently available questionnaires. Until high quality studies are available, we recommend using these questionnaires with caution. There is no need for the development of new neck-specific questionnaires until the current questionnaires have been adequately assessed.

In Chapter 6 we critically appraised the quality of the translation process, cross-cultural validation and the measurement properties of translated versions of neck-specific questionnaires. The literature search resulted in a total of 3641 unique hits, of which 27 articles, evaluating 6 different questionnaires in 15 different languages, were included in this study. Generally the methodological quality of the translation process is poor and none of the included studies performed a cross-cultural validation. A substantial amount of information regarding the measurement properties of translated versions of the different neck-specific questionnaires is lacking. Moreover, the evidence for the quality of measurement properties of the translated versions is mostly limited or assessed in studies of poor methodological quality. Until results from high quality studies are available, we advise to use the following questionnaires in the different countries: the NDI in Catalan, Dutch, English, Iranian, Korean, Spanish and Turkish, the NPQ in Chinese, and the NPDS in Finnish, German and Italian. The Greek NDI needs cross-cultural validation and there is no methodologically sound information for the Swedish NDI. For all other languages we advise to translate the original version of the NDI.

Chapter 7 reflects on the findings in this thesis and makes recommendations for clinical practice and (future) research.

REFERENCES

1. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127-41.
2. Cote P, Cassidy JD, Carroll LJ, Kristman V. The annual incidence and course of neck pain in the general population: a population-based cohort study. Pain 2004;112:267-73.
3. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain 1998;77:1-13.
4. Hoving JL, de Vet HC, Twisk JW, Deville WL, van der Windt D, Koes BW, et al. Prognostic factors for neck pain in general practice. Pain 2004;110:639-45.
5. Kjellman G, Skargren E, Oberg B. Prognostic factors for perceived pain and function at one-year follow-up in primary care patients with neck pain. Disabil Rehabil 2002;24:364-70.
6. Hill J, Lewis M, Papageorgiou AC, Dziedzic K, Croft P. Predicting persistent neck pain: a 1-year follow-up of a population cohort. Spine 2004;29:1648-54.
7. Philadelphia Panel. Philadelphia Panel evidence-based clinical practice guidelines on selected rehabilitation interventions for neck pain. Phys Ther 2001;81:1701-17.
8. Gross AR, Hoving JL, Haines TA, Goldsmith CH, Kay T, Aker P, et al. Manipulation and mobilisation for mechanical neck disorders. Cochrane Database Syst Rev 2004:CD004249.
9. Kay TM, Gross A, Goldsmith C, Santaguida PL, Hoving J, Bronfort G. Exercises for mechanical neck disorders. Cochrane Database Syst Rev 2005:CD004250.
10. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
11. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine 1999;24:1290-4.
12. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press, 2003.
13. Menezes da Costa L, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol 2009;62:934-43.

Chapter 8.2

Samenvatting

Hoofdstuk 1 vormt een inleiding op het onderwerp en de doelstellingen van dit proefschrift. De doelstellingen van dit proefschrift waren:

- Het evalueren van de invloed van verschillende manieren waarop kandidaatsvariabelen worden gecategoriseerd op de samenstelling en statistische prestaties van het uiteindelijke voorspellende model. Hiervoor wordt een model gebruikt dat tot stand gekomen is door middel van multivariabele logistische regressieanalyse. (Hoofdstuk 2)
- Het ontwikkelen en extern valideren van een voorspellend model dat de kans op blijvende klachten schat bij patiënten met aspecifieke nekpijn. (Hoofdstuk 3)
- Het ontwikkelen van een beslismodel dat aangeeft welke subgroepen van patienten met aspecifieke nekpijn het meest gebaat zijn bij fysiotherapie, manuele therapie, of afwachtend beleid. (Hoofdstuk 4)
- Het geven van een systematisch literatuuroverzicht van de meetkarakteristieken van nekspecifieke vragenlijsten. (Hoofdstuk 5 en 6)

Bij het ontwikkelen van een voorspellend model met multivariabele logistische regressie, worden in geneeskundig onderzoek continue variabelen (bijv. leeftijd) vaak opgedeeld in twee of meer categorieën. Het voordeel van categoriseren is dat de interpretatie van het model en de toepassing in de praktijk eenvoudiger zijn dan bij een model met continue variabelen.[1] In Hoofdstuk 2 wordt de invloed onderzocht van het categoriseren van continue variabelen op de uiteindelijke samenstelling en de statistische prestaties van een multivariabel logistisch regressiemodel. Het categoriseren van continue variabelen vóór een achterwaartse (backward) selectie van variabelen in een multivariabel logistisch regressiemodel resulteerde in andere voorspellende factoren in het uiteindelijke model en slechtere statistische prestaties van het model. Elke vorm van categoriseren resulteerde in een slechtere statistische prestatie. Stratificeren (> 2 categorieën) van continue variabelen lijkt iets minder ongunstig te zijn dan dichotomiseren (2 categorieën). Het is aan te bevelen om continue variabelen niet te categoriseren tijdens de ontwikkeling van een voorspellend model en, indien onontkoombaar, pas te categoriseren na samenstelling van het model. Deze strategie resulteert in een iets verminderde statistische prestatie van het model, maar zorgt in ieder geval voor de meest accurate samenstelling van het model.

Er zijn enkele studies uitgevoerd naar de relatie tussen bepaalde karakteristieken en het al dan niet voortduren van nekklachten in de algemene bevolking.[2-6] Echter, geen van deze studies ontwikkelde een voorspellend model dat de prognose kwantificeert. Het doel in Hoofdstuk 3 was om een model te ontwikkelen en te valideren dat de kans schat dat aspecifieke nekklachten minimaal 6 maanden aanhouden bij patiënten in de eerste lijn. De onderzoekspopulatie bestond uit volwassenen in Nederland (n=468),

die hun huisarts consulteerden voor aspecifieke nekpijn. De uitkomstmaat was het ervaren herstel, gemeten 6 maanden na het eerste consult bij de huisarts. Zeventien variabelen werden meegenomen in de multivariabele logistische regressieanalyse. Uit de analyse kwamen 9 voorspellers naar voren. Op basis van de geschatte coëfficiënten uit het regressiemodel werd een scoretabel geconstrueerd. De scoretabel heeft een matig discriminerend vermogen (met een oppervlakte onder de ROC curve van 0.66). De scoretabel werd extern gevalideerd in een in eerstelijns praktijken in Engeland geworven groep patiënten met aspecifieke nekpijn (n=315). Externe validering toonde aan dat de kwaliteit van het voorspellende model hetzelfde bleef met een vergelijkbaar discriminerend vermogen (met een oppervlakte onder ROC curve van 0.65) en een adequate overeenstemming tussen voorspelde en geobserveerde kansen. De voorspelling of iemand klachten zal houden, of binnen 6 maanden zal herstellen is, in vergelijking met schattingen gebaseerd op de prevalentie in de algemene bevolking, significant verbeterd door de scoretabel.

Systematische literatuuroverzichten wijzen uit dat fysiotherapie en manuele therapie een positief effect hebben op de klachten bij mensen met aspecifieke nekpijn, in vergelijking met afwachtend beleid.[7-9] Het verschil in effect is echter relatief gering. Heterogeniteit van de onderzoekspopulatie van patiënten met aspecifieke nekpijn kan een verklaring vormen voor het geringe verschil in effect. Het zou goed kunnen dat bepaalde subgroepen binnen deze populatie op basis van hun prognostische kenmerken meer baat hebben bij één van deze behandelingen,. Het doel in Hoofdstuk 4 was om een beslismodel te ontwikkelen dat aangeeft welke subgroepen van patiënten met aspecifieke nekpijn het meest gebaat zijn bij fysiotherapie, manuele therapie of afwachtend beleid. Met dat doel werden verscheidene patiëntkarakteristieken geëvalueerd, om te zien of ze het effect van de behandeling beïnvloedden. Uit de analyse kwamen drie factoren naar voren, die een relatie toonden met herstel en die bovendien interactie vertoonden met behandeling: pijnintensiteit voor herstel op de korte termijn, leeftijd en aan-/afwezigheid van lage rugpijn voor herstel op de lange termijn. Op basis van de verschillende modellen kunnen we stellen dat alle patiënten met aspecifieke nekpijn baat hebben bij manuele therapie, dat fysiotherapie een goede keus is bij patiënten met een milde tot gemiddelde pijnintensiteit en dat afwachtend beleid vooral geschikt is bij patiënten met een hoge pijnintensiteit die jonger dan vijftig jaar zijn. Door toepassing van het beslismodel zouden patiënten een maximaal 25% grotere kans op herstel hebben dan wanneer ze een andere behandeling zouden hebben gekregen dan het beslismodel adviseert.

Er zijn in de loop der tijd verscheidene ziektespecifieke vragenlijsten ontwikkeld om pijn en/of functionele beperkingen te meten bij patiënten met nekpijn (bijv. Neck

Disability Index (NDI), Neck Pain and Disability Scale (NPDS)).[10-11] Om een weloverwogen keuze te maken voor één van deze instrumenten, is het belangrijk om hun meetkarakteristieken te beoordelen en vergelijken (bijv. betrouwbaarheid, validiteit en responsiviteit).[12] De meetkarakteristieken van versies in verschillende talen kunnen afwijken van die van de originele versies, door verschillen in culturele context.[13] Daarom hebben wij besloten om de vertaalde versies van nekspecifieke vragenlijsten apart te beoordelen.

Het doel in Hoofdstuk 5 was om de meetkarakteristieken van de originele versies van nekspecifieke vragenlijsten kritisch te beoordelen en te vergelijken. Het literatuuronderzoek resulteerde in een totaal aantal van 3641 unieke artikelen. Daarvan werden 25 artikelen, die 8 verschillende vragenlijsten evalueerden, geïncludeerd in dit onderzoek. De NDI is de meest onderzochte vragenlijst en toont sterk bewijs voor een positief resultaat voor structurele validiteit, matig bewijs voor een positief resultaat voor hypothesetesten en responsiviteit, en beperkt bewijs voor een positief resultaat voor inhoudsvaliditeit en een negatief resultaat voor betrouwbaarheid. De andere vragenlijsten tonen ook positieve resultaten voor de meetkarakteristieken, maar het bewijs is meestal beperkt en minimaal 50% van de informatie over de meetkarakteristieken ontbreekt per vragenlijst. Onze bevindingen geven aan dat onvoldoende studies van hoge methodologische kwaliteit uitgevoerd zijn om de meetkarakteristieken van de beschikbare vragenlijsten te beoordelen. Totdat resultaten van dergelijke studies beschikbaar zijn is het verstandig om voorzichtig om te gaan met de huidige vragenlijsten. Er is geen reden om nieuwe nekspecifieke vragenlijsten te ontwikkelen, voordat de huidige vragenlijsten adequaat beoordeeld zijn.

Het doel in Hoofdstuk 6 was om het vertaalproces, de transculturele validering en de meetkarakteristieken van de vertaalde versies van nekspecifieke vragenlijsten kritisch te beoordelen en te vergelijken. Het literatuuronderzoek resulteerde in een totaal aantal van 3641 unieke artikelen. Daarvan werden 27 artikelen, die 6 verschillende vragenlijsten in 15 verschillende talen evalueerden, geïncludeerd in dit onderzoek. In het algemeen is de kwaliteit van het vertaalproces slecht en geen van de studies voerde een transculturele validering uit. Een aanzienlijk deel van de informatie met betrekking tot de meetkarakteristieken van vertaalde versies van nekspecifieke vragenlijsten ontbreekt. Bovendien is het bewijs voor de kwaliteit van de meetkarakteristieken van die lijsten vaak beperkt of onderzocht in studies van slechte methodologische kwaliteit. Totdat er informatie beschikbaar komt uit studies van hoge kwaliteit adviseren wij één van de volgende versies van de vragenlijsten te gebruiken: de NDI in het Catalaans, Nederlands, Engels, Iraans, Koreaans, Spaans en Turks, de Northwick Park Neck Pain Questionnaire (NPQ) in het Chinees, en de NPDS in het Fins, Duits en Italiaans. De Griekse versie van de NDI moet transcultureel gevalideerd worden en er

is geen methodologisch adequate informatie over de Zweedse versie van de NDI. Voor alle andere talen adviseren we om de originele versie van de NDI te vertalen.

In Hoofdstuk 7 wordt ingegaan op de bevindingen in dit proefschrift en worden aanbevelingen gedaan voor de praktijk en toekomstig onderzoek.

REFERENCES

1. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127-41.
2. Cote P, Cassidy JD, Carroll LJ, Kristman V. The annual incidence and course of neck pain in the general population: a population-based cohort study. Pain 2004;112:267-73.
3. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain 1998;77:1-13.
4. Hoving JL, de Vet HC, Twisk JW, Deville WL, van der Windt D, Koes BW, et al. Prognostic factors for neck pain in general practice. Pain 2004;110:639-45.
5. Kjellman G, Skargren E, Oberg B. Prognostic factors for perceived pain and function at one-year follow-up in primary care patients with neck pain. Disabil Rehabil 2002;24:364-70.
6. Hill J, Lewis M, Papageorgiou AC, Dziedzic K, Croft P. Predicting persistent neck pain: a 1-year follow-up of a population cohort. Spine 2004;29:1648-54.
7. Philadelphia Panel. Philadelphia Panel evidence-based clinical practice guidelines on selected rehabilitation interventions for neck pain. Phys Ther 2001;81:1701-17.
8. Gross AR, Hoving JL, Haines TA, Goldsmith CH, Kay T, Aker P, et al. Manipulation and mobilisation for mechanical neck disorders. Cochrane Database Syst Rev 2004:CD004249.
9. Kay TM, Gross A, Goldsmith C, Santaguida PL, Hoving J, Bronfort G. Exercises for mechanical neck disorders. Cochrane Database Syst Rev 2005:CD004250.
10. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-15.
11. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine 1999;24:1290-4.
12. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press, 2003.
13. Menezes da Costa L, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol 2009;62:934-43.

Chapter 9

Dankwoord

Aan het einde van het proefschrift is het gebruikelijk om iedereen te bedanken die bijgedragen heeft aan het tot stand komen ervan. Ik breek uiteraard niet met deze traditie en wil daarom de volgende mensen bedanken:

Allereerst Siep, je hebt mij niet alleen enthousiast gemaakt om dit traject in te gaan, maar je hebt de eerste periode ook zeker als mentor en aanjager invloed gehad op de ontwikkeling van mijn wetenschappelijke activiteiten. Bedankt voor je steun en ik hoop je hiermee mild te stemmen als opponent.

Uiteraard wil ik ook mijn begeleiders Arianne, Bart, Martijn en Riekie bedanken.

Arianne, mijn eerste aanspreekpunt binnen het team. In het begin kon ik je behoorlijk irriteren met mijn eigenwijze gedrag, maar gelukkig bleek je geduldig en zelf ook niet eenvoudig van je standpunt af te krijgen. Dit resulteerde in uitgebreide discussies, waarvan dit proefschrift kwalitatief zeker geprofiteerd heeft. Het feit dat we op den duur aan elkaars (eigen)aardigheden gewend raakten had gelukkig geen negatief effect op het niveau van onze discussies. Bedankt voor je geduld en bereikbaarheid.

Martijn, mijn vraagbaak voor statistische kwesties. We hebben heel wat gesprekken gevoerd over de voors en tegens van bepaalde statistische technieken, waarvan ik veel geleerd heb en wat mijn vocabulaire zeker uitgebreid heeft. Daarnaast had je ook altijd aandacht voor ontwikkelingen op het persoonlijke vlak. Het was een genoegen om met je samen te werken.

Bart, die altijd elk commentaar op een manuscript begint met een positieve opmerking over de inhoud, gevolgd door de kritische kanttekeningen. Deze positieve grondhouding in combinatie met een nuchtere kijk op de dingen is waarschijnlijk ook wat je zo aanspreekt in Australië. Ik heb veel van je commentaar geprofiteerd. Hartelijk dank voor je bijdrage.

Riekie, als laatste van het officiële team, maar zeker niet de minste. Je was de enige die altijd (ruim) voor de deadline mijn manuscripten retourneerde. Daarnaast zag je altijd nog mogelijkheden om het manuscript te verbeteren. Je getoonde toewijding en perfectionisme hebben me zeker gemotiveerd om aan artikelen te blijven schaven tot het gewenste niveau bereikt was.

Caroline Terwee, hartelijk dank voor je hulp bij het uitvoeren van de systematische reviews. We waren de eerste review waarin de COSMIN checklist gebruikt werd en dan heb je natuurlijk last van de gebruikelijke kinderziektes. Uiteindelijk hebben we die overwonnen, met twee mooie reviews als resultaat.

Martyn Lewis and Krysia Dziedzic, thank you very much for your cooperation and hospitality. It was a pleasure to visit you at Keele University.

Mijn kamer- en afdelingsgenoten (in alfabetische volgorde): Bianca, Cindy, Diana, Dieuwke, Evelien, Heleen, Jos, Jurgen, Marieke, Rianne, Sita en Winnifred. Ik weet dat het dankwoord van dit proefschrift door jullie als eerste gelezen wordt. Daarom hou ik het bij: bedankt voor jullie steun en gezelligheid! Zo, nu kunnen jullie je aandacht gaan besteden aan de relevante stukken in dit proefschrift.

Rene en Marlies wil ik graag bedanken voor de administratieve ondersteuning en adviezen.

Ik wil Hans hartelijk bedanken voor het controleren en waar nodig corrigeren van de spelling en grammatica in dit manuscript.

Uiteraard zijn er ook altijd mensen die indirect hebben bijgedragen in de vorm van morele support. In dat kader wil ik allereerst mijn ouders bedanken. Jullie aanmoedigingen om mij ook op wetenschappelijk gebied te ontwikkelen en jullie interesse in de vorderingen van mijn onderzoek zijn mij zeer tot steun geweest.

Uiteraard wil ik ook mijn paranimfen, Janus en Pieter, bedanken voor het feit dat zij mij bij willen staan bij mijn verdediging. We hebben al vele memorabele momenten met elkaar meegemaakt en ik hoop dat deze dag daarop een waardige aanvulling wordt.

Als allerlaatste wil ik Liesbeth bedanken. Meestal volgt er nu een stuk dat de partner heeft gefaciliteerd dat de promovendus ongestoord aan zijn proefschrift kon werken. Niets is echter minder waar, want de dagen die ik thuis werkte zag jij vooral als goede mogelijkheid om samen gezellig uitstapjes te gaan maken. Ik heb je vaak daarin moeten teleurstellen. Het thuis werken is hier echter mee ten einde, dus als ik thuis ben kunnen we vanaf nu ook daadwerkelijk samen leuke dingen gaan doen.

CURRICULUM VITAE

Jasper Schellingerhout is geboren op 8 augustus 1978 in Dordrecht. Na het behalen van zijn diploma aan het Johan de Witt-gymnasium in Dordrecht ging hij in 1996 Geneeskunde studeren aan de Universiteit van Antwerpen (België). Deze studie vervolgde hij vanaf 1997 aan het Erasmus MC in Rotterdam en rondde hij af in 2003. Na zijn studie werkte hij een jaar als poortarts in het Amphia-ziekenhuis in Breda, waarna hij in 2005 startte met de opleiding tot huisarts aan het Erasmus MC. Tijdens zijn opleiding tot huisarts volgde hij een masterclass "wetenschappelijk schrijven", wat resulteerde in twee publicaties van zijn hand met als onderwerp (behandeling van) schouderklachten. Vanaf 2006 deed hij een onderzoek naar (de behandeling van) nekklachten op de afdeling huisartsgeneeskunde van het Erasmus MC, waaruit de artikelen in dit proefschrift voortkwamen. Aan dit onderzoek werkten ook het EMGO-instituut van het VU medisch centrum in Amsterdam en het "Institute for primary care & health sciences research" van Keele University in Keele (Engeland) mee. In 2007 behaalde hij de graad van "Master of Science" in klinische epidemiologie aan de NIHES in Rotterdam en in 2009 rondde hij zijn opleiding tot huisarts af. Sinds 2009 is hij als huisarts werkzaam in Etten-Leur.

PORTFOLIO

Courses

NIHES Clinical Epidemiology, 2006-2007	70 ECTS
Masterclass Scientific Writing, 2006-2007	80 hours

Conferences

Oral presentation

NHG conference, 2007	20 hours
Keele University, 2007	20 hours
KNGF conference, 2007	20 hours
IFOMT conference, 2008	20 hours
World Neck Pain Congress, 2008	20 hours
Neck Pain in Primary Care conference, 2008	20 hours
NHG conference, 2009	20 hours
KNGF conference, 2009	20 hours

Poster presentation

WEON, 2007	16 hours
World Neck Pain Congress, 2008	16 hours
WEON, 2009	16 hours

Workshop

NHG conference, 2008	20 hours
LOVAH conference, 2009	20 hours
Low Back Pain forum, 2009	20 hours
NHG conference, 2010	20 hours

Teaching activities

Workshop "Shoulder Complaints" for GP-trainees, 2008	20 hours
Instruction GPs/physiotherapists on use decision model, 2009	20 hours

Other

Peer review NHG-guideline "Shoulder complaints", 2008	8 hours

LIST OF PUBLICATIONS

Schellingerhout JM, Thomas S. Recensie: Niet zo nuttig informatorium. Huisarts & Wetenschap 2007;50:179

Schellingerhout JM, Thomas S, Verhagen AP. Aspecifieke schouderklachten: geen effectiviteit van de gangbare behandelingen; literatuurstudie. Ned Tijdschr Geneeskd 2007;151:2892-7

Schellingerhout JM, Thomas S. Recensie: Differential diagnosis for primary care: a handbook for health care practitioners. Ned Tijdschr Geneeskd 2008;152:718

Schellingerhout JM, Verhagen AP, Thomas S, Koes BW. Lack of uniformity in diagnostic labeling of shoulder pain: time for a different approach. Man Ther 2008;13:478-83

Schellingerhout JM, Verhagen AP, Heymans MW, Pool JJ, Vonk F, de Vet HC, Koes BW. Which subgroups of patients with non-specific neck pain are more likely to benefit from spinal manipulation therapy, physiotherapy, or usual care? Pain 2008;139:670-80

Schellingerhout JM, Heymans MW, de Vet HC, Koes BW, Verhagen AP. Categorising continuous variables resulted in different predictors in a prognostic model for non-specific neck pain. J Clin Epid 2009;62:868-74

Schellingerhout JM, Verhagen AP, Heymans MW, Pool JJ, Vonk F, de Vet HC, Koes BW. Een beslismodel voor aspecifieke nekpijn. Huisarts & Wetenschap 2009;52:384-90

Schellingerhout JM. Nekpijn. Huisarts & Wetenschap 2010;53:117.

Schellingerhout JM. Effectiviteit van pregabaline hangt af van de soort pijn en de dosering. Huisarts & Wetenschap 2010;53:120.

Schellingerhout JM, Verhagen AP. Letter to the Editor concerning "Development of a clinical prediction rule to identify patients with neck pain likely to benefit from cervical traction and exercise" by Raney N et al. (2009) Eur Spine J 2010;19:833.

Verhagen AP, Schellingerhout JM. RE: LLLT meta-analysis. Lancet 2010;375:721.

Verhagen AP, Karels CH, Schellingerhout JM, Willemsen SP, Koes BW, Bierma-Zeinstra SM. Pain severity and catastrophising modify treatment success in neck pain patients. Man Ther 2010;15:267-72.

Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HC, Koes BW. Prognosis of patients with non-specific neck pain: development and external validation of a prediction rule for persistence of complaints. Spine 2010;35:E827-35.

Schellingerhout JM. Beslisregels bij klachten bewegingsapparaat. Huisarts & Wetenschap 2010;53:582.

Verhagen AP, Lewis M, Schellingerhout JM, Heymans MW, Dziedzic KS, de Vet HC, Koes BW. Do whiplash patients differ from other patients with non-specific neck pain regarding pain, function or prognosis? Man Ther 2011;doi:10.1016/j.math.2011.02.009.

Terwee CB, Schellingerhout JM, Verhagen AP, de Vet HC, Koes BW. Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: a systematic review. J Manipulative Physiol Ther 2011;43:261-72.

Schellingerhout JM, Heymans MW, Verhagen AP, de Vet HC, Koes BW, Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. BMC Med Res Methodol 2011;11:87.

Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2011, doi: 10.1007/s11136-011-9965-9.