VU Research Portal

Subgrouping patients with low back pain

Apeldoorn, A.T.

2012

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Apeldoorn, A. T. (2012). *Subgrouping patients with low back pain*.

8

General discussion

This thesis describes a series of studies about sub-grouping patients with low back pain (LBP), and can be divided into two parts. The first part (Chapters 2 to 4) focuses on the development of a screening instrument to identify a subgroup of patients in secondary care with chronic LBP who might benefit from additional psychological assessment. This study resulted in a screening instrument that is based on four signs and symptoms (Chapter 2). Information about the measurement properties of one of these signs, the Waddell score, is presented in Chapters 3 and 4. The second part (Chapters 5 to 7) describes the effectiveness and cost-effectiveness of a classification-based system when applied to patients with LBP in primary care. This classification-based system aims to subgroup these patients and matches them to treatments they are most likely to benefit from. The main findings of the research, its background, some supplementary findings, and important issues that were not explored in the previous chapters will be discussed in this chapter. Finally, this chapter will discuss the implications of our research and recommendations for future research.

Main findings of the thesis, part 1

The development of a screening instrument and the measurement properties of the Waddell score in patients with chronic LBP in secondary care.

- A screening instrument based on four signs and symptoms was developed for patients in secondary care with chronic LBP who might benefit from additional psychological assessment. The instrument showed good overall discriminative power and was regarded clinically useful (Chapter 2).
- The inter-rater reliability and the intra-rater reliability of the Waddell score were moderate and good, respectively (Chapter 3).
- The cross-sectional construct validity of the Waddell score was satisfactory with 15 out of 20 a priori formulated hypotheses accepted. The associations between the Waddell score and psychological measurement instruments were weak. Therefore, the Waddell score cannot be regarded as a straightforward psychological 'screener' (Chapter 4).

Is our screening instrument ready to be implemented in secondary care?
We tried to adhere to the principles that are essential in the development phase of a screening instrument. We defined an outcome of interest that is important (relevant

psychological disturbances [yes or no] according to our psychologists), selected signs and symptoms that are widely used in LBP research and may be associated with the outcome, selected a relevant sample of patients, blinded the physical therapists and psychologists to each other's findings and used appropriate statistical analyses. Our screening instrument showed a good overall discriminative power (area under the curve = 0.81) indicating that it has potential to determine which patients could benefit from additional psychological assessment. Further research is needed to replicate our findings in a similar patient population in order to validate our screening instrument and subsequently, show to what extent this screening instrument increases consistency in decision-making, improves patient outcomes and is cost-effective.

Is our screening instrument expected to be relevant for primary care?

It can be argued that our screening instrument might be useful in primary care too. Clinicians in both primary and secondary care have difficulties in making consistent decisions about the management of individuals with LBP,[1] and generally under-diagnose the existence of distress, anxiety and depression.[2-4] However, we must realize that the prevalence of relevant psychological disturbances differs between patients in primary and secondary care.[5] The prevalence of relevant psychological disturbances in our secondary care study was 53%, but has been estimated between 6-32% for various psychological diagnoses in Dutch primary care patients.[6] Also, the prevalence of illness behaviour (≥ 2 Waddell's non-organic signs) in our secondary care population (Chapter 4) was much higher compared to our primary care population (Chapter 6), 48% and 13%, respectively.

The statistical efficiency of screening instruments is greatest when the prevalence of the outcome of interest (reference standard) is between 40% to 60%.[7] The prevalence in our secondary care study was thus optimal for our screening instrument to produce relevant changes from pre-test to post-test probabilities. It is expected that the information gain in primary care will be less, because of a lower prevalence of relevant psychological disturbances in primary care patients. However, for patients, the clinical impact may be important, regardless of the underlying prevalence of psychological disturbances.

Why studying Waddell's non-organic signs?

Non-organic sign-testing was first described around 100 years ago, and was originally used to identify malingering.[8,9] With improvement in our understanding of medical and psychosocial aspects of low back pain, this view has been modified. In 1980, Waddell

et al.[10] reintroduced and standardized a group of non-organic signs, and discussed them in the light of new psychological knowledge. Waddell et al.[10] found that the signs correlated with psychological data, and suggested that the signs could be used as a screening tool to help identify patients who require more detailed psychological assessment. Since then, around 100 studies investigating or discussing Waddell's non-organic signs have been published. Although most of the researchers were positive about the clinical relevance of non-organic sign-testing, a few authors addressed the limitations.[11-15] Of those that did, most notable is the review of Fishbain et al.[16] in 2003, which challenged the entire scientific basis of the non-organic signs. Since then, only a handful of publications have mentioned the use of Waddell's non-organic signs. Although the review by Fishbain et al.[16] suffered from some methodological shortcomings itself, it appears to have had a major impact on the general opinion regarding non-organic signs. In our research, we found that the reliability of Waddell's non-organic sign-testing was sufficient and the cross-sectional construct validity was satisfactory. Our research also showed that it is hard to define what exactly the Waddell signs are measuring at a conceptual and clinical level. However, the Waddell score provides additional contribution to our screening tool above other signs and symptoms, and therefore we think there is a place for non-organic sign-testing in clinical practice.

Recommendations for further research
- Further research is needed to validate our screening instrument in a similar population. If the screening instrument appears to perform well, a next step is to assess if the instrument improves initial decision-making, reduces costs and improves patient outcomes. If so, one might be interested to assess its performance in a primary care setting.
- More research is needed to determine the reliability of the Waddell score. Ideally, this would be performed by several clinicians with a different degree of experience in non-organic sign-testing.

The main finding of this thesis, part 2

The effectiveness and cost-effectiveness of a classification-based system in patients with LBP in primary care

- The classification-based treatment approach as used in this study was not more effective (Chapter 6) and not cost-effective (Chapter 7) in comparison with usual physical therapy care in a population of patients with sub-acute and chronic LBP in primary care.

Why did we conduct this study?

In 1995, Delitto and colleagues described a classification-based treatment approach for patients with LBP that was based on signs and symptoms. Since this proposal, several studies have been conducted that provide evidence in support of this subgrouping approach (Chapter 5, Table 1). In 2003, Fritz et al.[17] compared the classification-based treatment approach to treatment according to US clinical guidelines using a randomized controlled trial (RCT). They randomized 78 patients with acute (< 3 weeks) work-related LBP of sufficient severity to necessitate modification of work duties. They found that the classification-based treatment approach was more effective than treatment based on the guidelines for disability measured with the Oswestry Disability Index (ODI) and on the physical component of the short form 36 health survey questionnaire (SF-36) at the 4-week follow-up, but not after 1 year. This raised the question whether this approach would be effective in another health care system and a related but different population. We compared the effectiveness of the classification-based treatment approach to usual physical therapy care in patients with sub-acute and chronic LBP in Dutch primary physical therapy care. We hypothesized that patients would benefit most from classification-based treatment, however we found no support for this hypothesis (Chapter 6).

Why did our results deviate from results of previous studies?

The somewhat different results between our study and that of Fritz et al.[17] should be considered in the light of several differences in the designs of the studies. Probably most important is the difference in patient selection criteria. The classification-based treatment approach was originally tested on patients with acute symptoms and disability scores of > 30 on the ODI (range 0-100).[18] Delitto et al.[18] based acuteness on a criterion related to the severity of disability rather than days since onset.

They reported that the majority of patients so categorized have an ODI score between 40 and 60. The study by Fritz et al.[17] included patients with a short duration of LBP and moderate to severe disability (mean ODI at baseline of 43). In contrast, 79% of our participating patients had chronic LBP (> 12 weeks) and the mean ODI at baseline was 20. Another difference between the studies was the use of slightly different classification algorithms. We adapted the classification algorithm based on the context in which physical therapy is provided within the Dutch health care system. For example, we did not include the subgroup traction, because Dutch private practices generally do not have mechanical traction equipment and the Dutch physical therapy guidelines discourage traction in patients with LBP.[19] Further, we modified the version used by Fritz et al.[17] based on updated evidence.[20] For example, we used the broader operational definition directional preference instead of centralization[21] to classify patients in the subgroup direction-specific exercises. Because directional preference identifies a broader category of patients, this likely shifted more patients into a classification of direction-specific exercises compared to the study of Fritz et al.[17] A further difference was the difference in comparison treatment. Although both studies compared classification-based treatment with clinical guidelines-based treatment, including the advice to remain active, Fritz et al. treated their patients with low-stress aerobic exercises and general muscle reconditioning whereas in our study patients received personalised, individually tailored sessions.

Another RCT that provides support for the classification-based approach was that by Brennan et al.[22] They included 123 patients with acute and sub-acute LBP and a mean ODI score of 43 at baseline. All patients were classified according to the classification algorithm and randomized to receive direction-specific exercises, manipulation or stabilization exercises, regardless of their subgroup. Outcomes were compared between patients who had received 'matched treatment' (according to the classification category) and 'unmatched' (non-classification) treatment. Brennan et al.[22] found a statistically significant, but small, reduction in disability, favouring the matched treatment group after 4 and 52 weeks. However, as the study was designed to validate the subgrouping used (i.e. to examine the effectiveness of the classification-based approach to target treatments compared to the use of the same treatments but not matched to patients) and not to validate the classification-based approach compared to another current approach, these results do not directly compare with the results of our RCT.[23]

What are the possible explanations as to why the classification-based approach was not more effective in comparison with usual physical therapy care?

An explanation for the lack of difference between classification-based treatment and usual physical therapy care in our RCT might be the high standard of usual care for LBP patients in the Netherlands. It might be that the contrast between the two treatment arms was smaller than anticipated.

Another explanation might be that our participating physical therapists were not trained well enough to treat the patients according to the different classification-based protocols. The physical therapists in our study received six hours of instruction. This instruction was provided by the author of this thesis and by experts who focussed on the protocols for direction-specific exercises and stabilization exercises. Concerning direction-specific exercises, at the start of this study only a few of the participating physical therapists were familiar with the mechanical diagnosis and therapy (MDT) method and its related direction-specific exercises. Therefore most physical therapists had to familiarize themselves with these exercises in a short period of time. It has been suggested however, that the level of training and experience in the MDT method may be essential to maximize treatment outcomes.[24] We tried to minimise the effect of this potential deficiency by providing our participating physical therapists with comprehensive direction-specific exercise protocols and support during the study from the author of this thesis (holder of MDT certificate) and a senior educator of the MDT Institute. Further, the initial classification procedure of 96% of the participating patients was performed by the author of this thesis with ample experience in the use of MDT methods. However, it cannot be ruled out that our education program and support may have been insufficient for optimal competence and may therefore have caused an under-estimation of the effectiveness of the classification-based treatment approach.

A further possible explanation is that the broadening of the criteria for the classification direction-specific exercises (directional preference instead of centralization) reduced the effectiveness of classification-based treatment. This modification was primarily based on a study by Long et al,[25] who found that direction-specific exercises matching patients' directional preference significantly and rapidly decreased pain. However, recently, Werneke et al.[26] found that while centralization was an important factor for predicting functional outcome when patients were treated with direction-specific exercises, directional preference on its own was not. This is an interesting finding, which underlines the continuous improvement that is necessary in existing classification systems. It is impossible to say which of abovementioned explanations is most probable.

Is the inter-rater reliability of the classification algorithm sufficient?

In our RCT, 34 patients (22%) were assessed by two research physical therapists. During the examination procedures, they discussed test results and additional testing and reached a very high agreement (kappa value 0.95, 95% confidence interval [CI] 0.85 to 1) about the final classification decision. These findings are better than preliminary findings which suggest that the inter-rater reliability is not more than acceptable for physical therapists who are familiar with the classification system.[27-29] In one of these studies, Stanton et al.[29] examined the inter-rater reliability of classification-based treatment between two examiners blinded to each other findings in 32 patients with LBP and found kappa values of 0.52 (95% CI 0.27 to 0.77). Interestingly, the reliability was better when they considered only patients with a 'clear classification' (i.e. patients that could be classified using only the first part of the classification algorithm) compared to the patients with an 'unclear classification' (kappa values 0.69 and 0.23, respectively).

Apart from our inter-rater reliability results presented in our RCT, we also conducted a separate inter-rater reliability study in a Dutch primary care back & neck centre (data not yet published). Three physical therapists familiar with the classification system participated and assessed 36 patients with LBP. Each patient was assessed twice on the same day, once by the author of this thesis and once by one of two other physical therapists. Assessments were conducted independently and examiners were blind to the other's classification decision. The inter-rater reliability was not more than fair with an agreement of 53% and a kappa value of 0.34 (95% CI 0.11 to 0.57). In contrast with the findings of Stanton et al,[29] we did not find a difference in reliability for patients with a 'clear classification' and patients with an 'unclear classification'.

The inter-rater reliability found within a subset of the study population in our RCT was much higher compared to the results from our separate inter-rater reliability study. This can be partially explained by the differences in examination procedures i.e. examiners assessed patients together or independently. It is not surprising that the inter-rater reliability is higher when patients are assessed together, because examiners can influence each other's opinion. Moreover, the rating is then based on the same patient examination, and therefore only concerns the reliability of the scoring system. However, patients may present differently in two separate examinations and testing procedures might differ between examiners. This variability has to be included in a study on the assessment of the reliability of a classification system. Therefore, to avoid overestimation of the inter-rater reliability of testing procedures, raters should assess the patients independently and blinded to each other findings. In general, more

knowledge is needed about the inter-rater reliability of the complete classification procedure.

Was there a significant effect for global perceived effect in favour of classification-based treatment after one year?

The improvement in global perceived effect scores in the classification-based group was not statistically significant (p = 0.14) in our effectiveness analysis (Chapter 6), but reached statistical significance (p = 0.02) in our cost-effectiveness analysis (Chapter 7). This difference was caused by the different statistical analyses that were used. In the effectiveness study, we used longitudinal multilevel analysis with a four level structure. With this procedure we corrected for the fact that outcome measures that are repeatedly measured are correlated within patients and for the fact that the outcomes of patients can cluster (correlate) within therapists, and therapists can cluster under practices. It was not possible to use this sophisticated statistical analysis in the economic evaluation.

Therefore, in the economic evaluation, the hierarchy (e.g. clustering of patients within therapists) existing in our dataset was not taken into account. The longitudinal multilevel analysis is more appropriate for this type of data and provides more accurate information.

Recommendations for future research

- Only a few studies have focused on the inter-rater reliability of the complete classification-based procedure. They also used small study samples and several different algorithms, which makes it difficult to compare and to interpret findings. Consequently, there is a need for further efforts to determine and perhaps to improve the inter-rater reliability of the complete classification-based procedure.

- Ideally, a classification algorithm should be able to classify every patient into (only) one subgroup. In our RCT, using the first part of the algorithm 24% of all patients did not meet any of the subgroups and 16% met more than one subgroup. Our findings are consistent with those of Stanton et al.[29] Therefore, the classification algorithm needs refinement and possibly extension to be able to classify other relevant subgroups.

- Available research supports the influence of key psychological factors (e.g. depression, catastrophizing) on treatment outcome, although there is a lack of evidence how this understanding should be utilized in clinical practice.[30] These data tell us that more research is needed to explore whether incorporating key

psychological factors in a classification-based treatment approach leads to more positive outcomes.[31] There is reason to believe however, that this might be more important for patients with chronic LBP with generally higher prevalence of psychological factors compared to patients with acute or sub-acute LBP.

Conclusion

This first independent broad validation study of Delitto's classification-based treatment approach found that this approach did not generalize to patients with sub-acute and chronic LBP in the Dutch primary physical therapy setting. Only two RCTs have compared this classification-based treatment approach to usual physical therapy care and the results so far are inconclusive. Nevertheless, subgrouping has been identified as a consistent research priority.[32] Therefore, further high-quality research exploring the effectiveness and cost-effectiveness of this classification and comparable approaches is required before firm conclusions can be drawn.

References

1. Hill JC, Vohora K, Dunn KM, et al. Comparing the STarT back screening tool's subgroup allocation of individual patients with that of independent clinical experts. *Clin J Pain* 2010;26:783-7.

2. Grevitt M, Pande K, O'Dowd J, et al. Do first impressions count? A comparison of subjective and psychologic assessment of spinal patients. *Eur Spine J* 1998;7:218-23.

3. Haggman S, Maher CG, Refshauge KM. Screening for symptoms of depression by physical therapists managing low back pain. *Phys Ther* 2004;84:1157-66.

4. Spitzer RL, Williams JB, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *JAMA* 1994;272:1749-56.

5. Manchikanti L, Pampati V, Fellows B, et al. Characteristics of chronic low back pain in patients in an interventional pain management setting: a prospective evaluation. *Pain Physician* 2001;4:131-42.

6. Van der Horst M, Lindeboom R, Lucas C. De aanwezigheid van psychische problemaiek in de eerstelijns fysiotherapiepraktijk: een clusteranalyse. De prognostische waarde van de Vierdimensionale Klachtenlijst (4DKL) (The presence of mental disorders in physiotherapy practice. The prognostic value of the Four Dimensional Symptom Questionnaire [4DSQ]). *Ned Tijdschr Fysiother* 2005;115:106-11.

7. Sackett DL, Hayes RB, Guyatt GH, et al. *Clinical Epidemiology: A basic science for clinical medicine.* 2nd ed. Boston MA: Little, Brown and Company Inc, 1991.

8. Collie J. Malingering and feigned sickness. New York: Paul B Boeber, 1913.

9. Jones AB, Liewellyn LJ. *Malingering or the simulation of disease.* London: William Heinemann, 1917.

10. Waddell G, McCulloch JA, Kummel E, et al. Nonorganic physical signs in low-back pain. *Spine* 1980;5:117-25.

11. Centeno CJ, Elkins WL, Freeman M. Waddell's signs revisited? Spine 2004;29:1392.

12. Giles LG. Re: Centeno CJ, Elkins WL, Freeman M. Editorial. Waddell's signs revisited? *Spine* 2004;29:1392. *Spine* 2005;30:265-8.

13. Rosomoff HL, Fishbain DA, Goldberg M, et al. Physical findings in patients with chronic intractable benign pain of the neck and/or back. *Pain* 1989;37:279-87.

14. Fishbain DA, Goldberg M, Rosomoff RS, et al. Chronic pain patients and the nonorganic physical sign of nondermatomal sensory abnormalities (NDSA). *Psychosomatics* 1991;32:294-303.

15. Ranney D. A Proposed Neuroanatomical Basis of Waddell's Nonorganic Signs. *Am J Phys Med Rehabil* 2010;89:1036-42.

16. Fishbain DA, Cole B, Cutler RB, et al. A structured evidence-based review on the meaning of nonorganic physical signs: Waddell signs. *Pain Med* 2003;4:141-81.

17. Fritz JM, Delitto A, Erhard RE. Comparison of classification-based physical therapy with therapy based on clinical practice guidelines for patients with acute low back pain: a randomized clinical trial. *Spine* 2003;28:1363-71.

18. Delitto A, Erhard RE, Bowling RW. A treatment-based classification approach to low back syndrome:

identifying and staging patients for conservative treatment. *Phys Ther* 1995;75:470-85.

19. Bekkering GE, Hendriks HJM, Koes BW, et al. KNGF-richtlijn fysiotherapie bij lage-rupijn (National practice guidelines for physical therapy in patients with low back pain). *Ned Tijdschr Fysiother* 2005;115:1-32.

20. Fritz JM, Cleland JA, Childs JD. Subgrouping patients with low back pain: evolution of a classification approach to physical therapy. *J Orthop Sports Phys Ther* 2007;37:290-302.

21. Werneke MW. 'Centralization' and 'directional preference' are not synonymous. *J Orthop Sports Phys Ther* 2009;39:827-8.

22. Brennan GP, Fritz JM, Hunter SJ, et al. Identifying subgroups of patients with acute/subacute 'nonspecific' low back pain: results of a randomized clinical trial. *Spine* 2006;31:623-31.

23. Kent P, Hancock M, Petersen DH, et al. Clinimetrics corner: choosing appropriate study designs for particular questions about treatment subgroups *J Man Manip Ther* 2010;18:147-52.

24. Aina A, May S, Clare H. The centralization phenomenon of spinal symptoms - a systematic review. *Man Ther* 2004;9:134-43.

25. Long A, Donelson R, Fung T. Does it matter which exercise? A randomized control trial of exercise for low back pain. *Spine* 2004;29:2593-602.

26. Werneke MW, Hart DL, Cutrone G, et al. Association between directional preference and centralization in patients with low back pain. *J Orthop Sports Phys Ther* 2011;41:22-31.

27. Fritz JM, George S. The use of a classification approach to identify subgroups of patients with acute low back pain. Interrater reliability and short-term treatment outcomes. *Spine* 2000;25:106-14.

28. Kiesel KB, Underwood FB, Mattacola CG, et al. A comparison of select trunk muscle thickness change between subjects with low back pain classified in the treatment-based classification system and asymptomatic controls. *J Orthop Sports Phys Ther* 2007;37:596-607.

29. Stanton TR, Fritz JM, Hancock MJ, et al. Evaluation of a Treatment-Based Classification Algorithm for Low Back Pain: A Cross-Sectional Study. *Phys Ther* 2011;91:496-509.

30. Nicholas MK, Linton SJ, Watson PJ, et al. Early identification and management of psychological risk factors ('yellow flags') in patients with low back pain: a reappraisal. *Phys Ther* 2011;91:737-53.

31. Billis EV, McCarthy CJ, Oldham JA. Subclassification of low back pain: a cross-country comparison. *Eur Spine J* 2007;16:865-79.

32. Foster NE, Dziedzic KS, Van der Windt DA, et al. Research priorities for non-pharmacological therapies for common musculoskeletal problems: nationally and internationally agreed recommendations. *BMC Musculoskelet Disord* 2009;10:3.