RESEARCH ARTICLE

Validity and reliability of the Dutch STarT MSK tool in patients with musculoskeletal pain in primary care physiotherapy

Anke G. van den Broek1,2,3*, Corelien J. J. Kloek2,3, Martijn F. Pisters1,3,4, Cindy Veenhof1,2,3

1 Department of Rehabilitation, Physical Therapy Science and Sports, Rudolf Magnus Institute of Neurosciences, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands, 2 Expertise Center Healthy Urban Living, Research Group Innovation of Human Movement Care, University of Applied Sciences Utrecht, Utrecht, The Netherlands, 3 Center for Physical Therapy Research and Innovation in Primary Care, Julius Health Care Centers, Utrecht, The Netherlands, 4 Research Group Empowering Healthy Behaviour, Fontys University of Applied Sciences, Eindhoven, The Netherlands

* Anke.vandenBroek@han.nl

Abstract

Objective
To evaluate the validity and reliability of the Dutch STarT MSK tool in patients with musculoskeletal pain in primary care physiotherapy.

Methods
Physiotherapists included patients with musculoskeletal pain, aged 18 years or older. Patients completed a questionnaire at baseline and follow-up at 5 days and 3 months, respectively. Construct validity was assessed by comparing scores of STarT MSK items with reference questionnaires. Pearson’s correlation coefficients were calculated to test predefined hypotheses. Test-retest reliability was evaluated by calculating quadratic-weighted kappa coefficients for overall STarT MSK tool scores (range 0–12) and prognostic subgroups (low, medium and high risk). Predictive validity was assessed by calculating relative risk ratios for persisting disability at 3 months.

Results
In total, 142 patients were included in the analysis. At baseline, 74 patients (52.1%) were categorised as low risk, 64 (45.1%) as medium risk and 4 (2.8%) as high risk. For construct validity, nine of the eleven predefined hypotheses were confirmed. For test-retest reliability, kappa coefficients for overall STarT MSK tool scores and prognostic subgroups were 0.71 and 0.65, respectively. For predictive validity, relative risk ratios for persisting disability were 2.19 (95% CI: 1.10–4.38) for the medium-risk group and 7.30 (95% CI: 4.11–12.98) for the high-risk group.
Conclusion

The Dutch STarT MSK tool showed a sufficient to good validity and reliability in patients with musculoskeletal pain in primary care physiotherapy. The sample size for high-risk patients was small (n = 4), which may limit the generalisability of findings for this group. An external validation study with a larger sample of high-risk patients (≥50) is recommended.

Introduction

Musculoskeletal conditions, such as low-back pain, neck pain, osteoarthritis and rheumatoid arthritis, are the most common cause of long-term pain and impaired physical function [1], and have large impact on health-related quality of life [2]. In the Netherlands the point prevalence of musculoskeletal pain is found to be 53.9%, with low-back, shoulder, neck and knee being the most frequently affected pain sites [3]. With the high impact of chronic musculoskeletal conditions on health care and work-related costs [4–8], there is a need for effective and cost-effective treatment options to manage musculoskeletal pain.

Musculoskeletal pain is predominantly managed in primary care, for example by the physiotherapist. Although current evidence shows positive effects of exercise therapy and psychosocial interventions on pain and function in patients with musculoskeletal pain [9, 10], it needs to be considered that every patient is unique and some patients respond better to certain treatments than others [11]. In order to improve effectiveness on clinical outcomes and cost-effectiveness in the treatment of patients with musculoskeletal pain, a stratified care approach is promising [12]. Within stratified care, treatments are matched to patients based on key characteristics such as biomedical and psychosocial risk factors for poor prognosis. To identify modifiable risk factors for poor prognosis at an early stage and, subsequently, to stimulate that the appropriate stratified care will be applied to patients, a valid and reliable risk stratification tool is required.

The Keele STarT Back-Screening Tool (SBT) is an example of a valid and reliable risk stratification tool developed to allocate primary care patients with low-back pain into three prognostic subgroups (low, moderate and high risk of persisting back pain disability), and to apply the appropriate matched treatment [13]. The original SBT was developed in the United Kingdom (UK) and has been translated into several languages, including Dutch [14]. A stratified care approach (use of the SBT and matched treatments) has demonstrated superior clinical and cost outcomes compared to usual non-stratified primary care in patients with low-back pain within the UK [15, 16]. While the SBT focuses primarily on back pain, there is evidence that different regional musculoskeletal pain presentations share common underlying mechanisms and prognostic factors [17–20]. These studies indicate that a comparable risk stratification tool could be useful for patients with a broader range of musculoskeletal pain presentations.

Recently, a modified, generic version of the SBT was developed for patients with the five most common musculoskeletal pain presentations (i.e., neck, back, shoulder, knee or multisite pain) within the UK [21, 22]. This so-called Keele STarT MSK tool showed a moderate to good predictive ability of the tool’s baseline score for identifying patients who developed persisting disability because of musculoskeletal pain, and subgroup cut-points were comparable across pain sites [22, 23]. Given the promising predictive performance of the STarT MSK tool we believe that the tool could be useful for Dutch clinical practice as well. The initial STarT MSK study that Dunn et al. conducted, included patients who consulted their general practitioner...
(GP) with musculoskeletal pain [22]. In the Netherlands, physiotherapists in particular are involved in the treatment of patients with musculoskeletal pain. A first step toward identifying whether the STarT MSK tool could be useful for Dutch primary care physiotherapy, is to translate the tool into Dutch and evaluate its measurement properties. Therefore, the aim of this study is to evaluate the validity and reliability of the Dutch STarT MSK tool in patients with musculoskeletal pain in primary care physiotherapy.

Methods

Translation of the STarT MSK tool

Prior to the start of this study, we formally translated the original English version of the STarT MSK tool (S1 Appendix) into Dutch using a forward-backward translation method according to the guidelines of Beaton et al. (S2 Appendix) [24]. The Dutch version of the STarT MSK tool is included in S3 Appendix. The tool is designed for patients to self-administer [25].

Design

In this clinimetric study a prospective observational design was used to evaluate the measurement properties of the Dutch STarT MSK tool. Patients were asked to complete baseline and follow-up questionnaires (5 days and 3 months), and received usual care from their physiotherapist. This study was not subject to the Medical Research Involving Human Subjects Act (WMO) and received a ‘non-WMO’ declaration from the Medical Research Ethics Committee of the University Medical Centre Utrecht, The Netherlands (registration number 18–082).

Participants

Physiotherapists. A total of 65 physiotherapists within the authors’ network were invited to participate in this study. Only physiotherapists who were working in primary care and were seeing at least 1 to 2 patients with musculoskeletal pain for a first consultation per week were eligible to participate in this study.

Patients. The inclusion period for patients was February 2018 to May 2018. Patients were eligible for inclusion when (1) they consulted a participating physiotherapist for musculoskeletal pain (i.e., neck, back, shoulder, knee or multisite pain) during the inclusion period, (2) it was the first consultation for the current episode of musculoskeletal pain, (3) they were aged 18 years or older, (4) they were able to read and write in Dutch and (5) they had an email address. Patients were excluded when (1) during the first consultation red flags were found indicating a possible specific underlying pathology (e.g., fracture, infection, tumor, cauda equina) responsible for the musculoskeletal pain, (2) they consulted a physiotherapist for pre- or post-operative rehabilitation related to the musculoskeletal pain presentation, (3) they were diagnosed with inflammatory arthritis, spondyloarthropathy or polymyalgia rheumatica, (4) they experienced pregnancy-related pain problems or (5) they had a cognitive impairment.

Study procedure

The participating physiotherapists received information about the study procedure during a one-hour in-company instruction. Physiotherapists informed eligible patients about the study and screened them on in- and exclusion criteria. Patients who were willing to participate received an information letter from the physiotherapist. After patients had given permission to be contacted, the researcher (AB) received their contact information from the physiotherapist using a secured messenger service. Patients were then e-mailed by the researcher (AB) and received a link to the informed consent form as part of the baseline questionnaire. The link to
the baseline questionnaire was sent within 24 hours after the first consultation with the physio-
therapist. Patients could only start filling-in the questionnaire if informed consent was pro-
vided. When necessary, a reminder was sent to the patient within 3 days. Respectively 5 days
and 3 months after filling in the baseline questionnaire patients received the link to the follow-
up questionnaire.

Measurements

Baseline. At baseline (T0), general patient characteristics including age, gender, educa-
tional level, pain site, pain duration and presence of comorbidity were obtained. The prediction
of persisting musculoskeletal-related disability was assessed with the Dutch STarT MSK
tool, consisting of 10 independent items that cover biomedical and psychosocial prognostic
factors (Fig 1). Subgroup cut-points are 0–4 for low risk, 5–8 for medium risk and 9–12 for
high risk, based on an overall score ranging from 0–12 [25]. The average pain in the past week
was measured with the 11-point Numeric Pain Rating Scale (NPRS) [26], ranging from 0 (no
pain) to 10 (worst possible pain). Pain self-efficacy beliefs were assessed with the Pain Self-Effi-
cacy Questionnaire (PSEQ) [27], consisting of 10 items, each scored on a 7-point Likert scale
(0 = not at all confident; 6 = completely confident), with a higher score reflecting stronger self-
efficacy beliefs. Disability was measured with the physical functioning subscale of the 36-Item
Short Form Health Survey (SF-36 PF) [28], consisting of 10 statements with three answer
options varying from 'Yes, limited a lot’ to ’No, not limited at all’. Each item is scored 1 to 3
points and the total score was transformed to a 100-point scale, with a higher score indicating
better physical functioning. Timeline illness perception was assessed with the timeline ques-
tion of the Brief Illness Perception Questionnaire (Brief IPQ) [29], ranging from 'a very short
time’ to ’forever’ on an 11-point scale. Depressive symptoms were measured with the depres-
sion subscale of the Hospital Anxiety and Depression Scale (HADS-D) [30], consisting of 7
items, each scored on a 4-point Likert scale, with a higher score reflecting more depressive
symptoms. Fear of movement was assessed with the shortened version of the Tampa Scale of
Kinesiophobia (TSK-11) [31]. The TSK-11 consists of 11 statements with four answer options
varying from ‘strongly disagree’ to ‘strongly agree’. The total score ranges from 11 to 44 points,
with a higher score reflecting greater fear of movement.

Follow-up. Five days after filling-in the baseline questionnaire patients received a follow-
up questionnaire (T1) consisting of the Dutch STarT MSK tool, the NPRS to assess the average
pain in the past week and the Global Perceived Effect (GPE) scale [32] to measure recovery.
The GPE scale consists of the question "To what extend have your complaints improved since
filling-in the baseline questionnaire?”, scored on a 7-point Likert scale (1 = completely recov-
ered; 7 = worse than ever).

Three months after filling-in the baseline questionnaire patients received a follow-up ques-
tionnaire (T2) consisting of the SF-36 PF to assess disability, the NPRS to assess the average
pain in the past week and the GPE scale to measure recovery.

Main study parameters

The measurement properties construct validity, test-retest reliability and predictive validity
were evaluated according to the Consensus-based Standards for the selection of health Mea-
surement Instruments (COSMIN) checklist [33, 34].

Construct validity. To evaluate construct validity, we compared scores of the separate
STarT MSK items with reference questionnaires. This method was used because the STarT
MSK tool is a formative model [35], which means that each item contributes a part of the con-
struct and together they will give a prognosis for persisting musculoskeletal-related disability.
As recommended by Mokkink et al. [36], hypotheses were formulated a priori about the expected magnitude of correlations between scores on the instrument (i.e., STarT MSK items) and scores on other instruments. In concordance with Bier et al. [14], reference questionnaires were chosen based on the intended domains being measured using validated Dutch questionnaires. Scores on STarT MSK items 7 and 10 were compared with clinical variables [36], since no validated questionnaires were available. Hypotheses were formulated based on the comparability of the domains being measured. A priori we expected a moderate ($r \geq 0.3, < 0.5$) to
high ($r \geq 0.5$) positive correlation between STarT MSK item 1 with the NPRS, item 6 with the timeline question of the Brief IPQ, item 7 with the single-item question on ‘comorbidity’, item 8 with the HADS-D, item 9 with the TSK-11, item 10 with the single-item question on ‘pain duration’; and a moderate ($r \leq -0.3, > -0.5$) to high ($r \leq -0.5$) negative correlation between item 2 with the PSEQ and item 4 with the SF-36 PF [37]. We expected a moderate ($r \geq 0.3, < 0.5$) to high ($r \geq 0.5$) positive correlation between the STarT MSK bothersomeness item 3 with the NPRS and a moderate ($r \leq -0.3, > -0.5$) to high ($r \leq -0.5$) negative correlation between item 3 with the SF-36 PF, as bothersomeness has been associated with pain and disability [38]. Finally, we expected a low positive correlation ($r < 0.3$) between item 5 and the NPRS, as this item focuses on the location of pain and not on pain intensity.

**Test-retest reliability.** To evaluate test-retest reliability, we assessed the agreement between scores of the Dutch STarT MSK tool on baseline and after 5 days. The time interval was considered long enough to prevent for recall bias, given the large number of questionnaires patients had to complete at baseline. Next, 5 days were considered short enough to prevent substantial improvement [39]. A sensitivity analysis was performed in a subset of patients reporting stable musculoskeletal pain symptoms during the test-retest period. In concordance with the criteria proposed by Bier et al. [14, 40], patients were considered stable between T0 and T1 when they scored ‘slightly worsened’, ‘no change’, or ‘slightly improved’ on the GPE, and had a stable pain score (i.e., the same score on the NPRS plus or minus one point compared with T0).

**Predictive validity.** To evaluate predictive validity, we assessed the ability of the Dutch STarT MSK tool to predict persisting disability at 3 months. Persisting disability was defined as a SF-36 PF score equal to or below the baseline median [13], as standard cutoffs for the SF-36 PF were not available in the literature. We used the SF-36 PF score as poor outcome, because the SF-36 is an assessment tool with good validity regarding its 8 subscales [28, 41]. It was hypothesised that patients with lower baseline scores (i.e., low-risk group compared to medium- or high-risk group and medium-risk group compared to high-risk group) will have better outcomes on the SF-36 PF. As advised by Hill et al. [42], a sensitivity analysis was performed using different subgroup cut-points on the Dutch STarT MSK tool (i.e., low risk 0–3, medium risk 4–7, high risk 8–12; and low risk 0–4, medium risk 5–7, high risk 8–12).

**Sample size**

To evaluate construct validity, test-retest reliability and predictive validity, a minimal sample size of 50 patients is advised and a sample size of 100 patients is adequate according to the COSMIN checklist [39, 43]. In this study, we aimed for an adequate sample size of at least 100 patients [43].

**Statistical analysis**

Statistical analysis was performed using IBM SPSS Statistics version 25.0 (Armork, New York, USA). Descriptive statistics were calculated for baseline characteristics of the study population, with continuous variables presented using mean and standard deviations. Categorical and nominal/dichotomous data were presented as proportions for each category.

**Construct validity.** Pearson’s correlations were calculated between specific items of the Dutch STarT MSK tool and their reference questionnaires. The construct validity was defined as good if at least 75% of the a priori hypotheses could be confirmed [39].

**Test-retest reliability.** The quadratic-weighted kappa was calculated for overall scores of the Dutch STarT MSK tool and prognostic subgroups, with $\kappa \leq 0$ indicating poor agreement,
κ = 0.01–0.20 slight agreement, κ = 0.21–0.40 fair agreement, κ = 0.41–0.60 moderate agreement, κ = 0.61–0.80 substantial agreement and κ = 0.81–1.00 almost perfect agreement [44].

**Predictive validity.** To examine calibration of the STarT MSK tool, we calculated the proportion of patients with persisting disability at 3 months in each risk group. Relative risk ratios (RRs) were calculated for medium risk and high risk, each compared with low risk, in their ability to predict persisting disability at 3 months.

**Results**

In total, 44 physiotherapists were willing to participate and instructed. A number of 22 physiotherapists from 11 primary care physiotherapy clinics actually recruited patients. Of the participating physiotherapists, the majority was specialised as manual therapist or sports physiotherapist. A total of 167 patients were recruited, of whom 146 patients were included (Fig 2). Four patients were excluded from the analysis, because they did not fully complete the

![Flowchart for inclusion of patients](https://doi.org/10.1371/journal.pone.0248616.g002)
baseline questionnaire. For the follow-up questionnaires at T1 and T2, a 100% and 96% follow-up rate were achieved, respectively.

On average, the time between the first consultation of physiotherapy and completing the baseline questionnaire was 2 days. Baseline characteristics of the study population are presented in Table 1. The mean age of study participants was 48.2 ± 15.6 years (range 18–81), and 57.0% were female. The study population consisted of 142 patients with musculoskeletal pain, of whom 44 patients (31.0%) had back pain, 35 (24.7%) shoulder pain, 28 (19.7%) multisite pain, 25 (17.6%) knee pain and 10 (7.0%) neck pain. Of the patients with multisite pain, the majority (57.1%) reported neck and shoulder pain. At baseline, 74 patients (52.1%) were categorised as low risk, 64 (45.1%) as medium risk and 4 (2.8%) as high risk. Age and gender were comparable across risk groups. For each increase in risk profile, there was a corresponding increase in pain intensity, disability, timeline illness perception, depression and fear of movement.

Table 1. Baseline characteristics of the study population†.

<table>
<thead>
<tr>
<th></th>
<th>Total  (n = 142)</th>
<th>Low risk  (n = 74)</th>
<th>Medium risk  (n = 64)</th>
<th>High risk  (n = 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>81 (57.0)</td>
<td>40 (54.1)</td>
<td>39 (60.9)</td>
<td>2 (50.0)</td>
</tr>
<tr>
<td>Age in years</td>
<td>48.2 ± 15.6</td>
<td>47.9 ± 16.8</td>
<td>47.9 ± 14.5</td>
<td>56.8 ± 8.2</td>
</tr>
<tr>
<td>Pain site</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neck</td>
<td>10 (7.0)</td>
<td>6 (8.1)</td>
<td>3 (4.7)</td>
<td>1 (25.0)</td>
</tr>
<tr>
<td>Back</td>
<td>44 (31.0)</td>
<td>22 (29.7)</td>
<td>21 (32.8)</td>
<td>1 (25.0)</td>
</tr>
<tr>
<td>Shoulder</td>
<td>35 (24.7)</td>
<td>23 (31.1)</td>
<td>12 (18.8)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Knee</td>
<td>25 (17.6)</td>
<td>12 (16.2)</td>
<td>12 (18.8)</td>
<td>1 (25.0)</td>
</tr>
<tr>
<td>Multisite</td>
<td>28 (19.7)</td>
<td>11 (14.9)</td>
<td>16 (25.0)</td>
<td>1 (25.0)</td>
</tr>
<tr>
<td>Episode duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;3 months</td>
<td>72 (50.7)</td>
<td>46 (62.2)</td>
<td>25 (39.1)</td>
<td>1 (25.0)</td>
</tr>
<tr>
<td>3 to 6 months</td>
<td>26 (18.3)</td>
<td>16 (21.6)</td>
<td>9 (14.1)</td>
<td>1 (25.0)</td>
</tr>
<tr>
<td>6 months or longer</td>
<td>44 (31.0)</td>
<td>12 (16.2)</td>
<td>30 (46.9)</td>
<td>2 (50.0)</td>
</tr>
<tr>
<td>Presence of comorbidity</td>
<td>59 (41.5)</td>
<td>28 (37.8)</td>
<td>29 (45.3)</td>
<td>2 (50.0)</td>
</tr>
<tr>
<td>Dutch STarT MSK tool score¶</td>
<td>4.2 ± 2.3</td>
<td>2.4 ± 1.2</td>
<td>6.0 ± 1.0</td>
<td>9.5 ± 0.6</td>
</tr>
<tr>
<td>Dutch STarT MSK tool risk profile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>74 (52.1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>64 (45.1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>4 (2.8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain intensity¶</td>
<td>5.4 ± 2.0</td>
<td>4.5 ± 2.0</td>
<td>6.2 ± 1.6</td>
<td>8.0 ± 0.0</td>
</tr>
<tr>
<td>Mild (0–5)</td>
<td>62 (43.7)</td>
<td>47 (63.5)</td>
<td>15 (23.4)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Moderate (6–7)</td>
<td>60 (42.3)</td>
<td>21 (28.4)</td>
<td>39 (60.9)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Severe (8–10)</td>
<td>20 (14.1)</td>
<td>6 (8.1)</td>
<td>10 (15.6)</td>
<td>4 (100.0)</td>
</tr>
<tr>
<td>Pain self-efficacy¶</td>
<td>49.4 ± 9.6</td>
<td>53.5 ± 7.3</td>
<td>45.6 ± 9.4</td>
<td>34.0 ± 9.9</td>
</tr>
<tr>
<td>Disability¶</td>
<td>75.1 ± 19.5</td>
<td>83.0 ± 13.7</td>
<td>67.2 ± 21.4</td>
<td>56.3 ± 19.7</td>
</tr>
<tr>
<td>Timeline illness perception¶</td>
<td>4.2 ± 2.9</td>
<td>3.1 ± 2.2</td>
<td>5.3 ± 3.0</td>
<td>7.5 ± 2.1</td>
</tr>
<tr>
<td>Depression¶</td>
<td>2.3 ± 2.5</td>
<td>1.3 ± 1.8</td>
<td>3.1 ± 2.4</td>
<td>7.3 ± 4.0</td>
</tr>
<tr>
<td>Fear of movement¶</td>
<td>20.0 ± 5.6</td>
<td>18.5 ± 5.1</td>
<td>21.2 ± 5.4</td>
<td>28.3 ± 7.2</td>
</tr>
</tbody>
</table>

† Values are numbers (percentage) or mean ± standard deviation.
¶ Dutch STarT MSK tool score (0–12); Pain intensity (0–10); Pain self-efficacy (0–60); Disability (0–100); Timeline illness perception (0–10); Depression (0–21); Fear of movement (11–44).

https://doi.org/10.1371/journal.pone.0248616.t001
Construct validity
The highest correlations were found between item 1 with the NPRS and item 10 with the single-item question on ‘pain duration’ (Table 2). The correlations between item 2 with the PSEQ and item 9 with the TSK-11 were lower than hypothesised. Of the predefined hypotheses, 81.8% were confirmed which indicates a good construct validity.

Test-retest reliability
In total, 142 patients (100%) completed the Dutch STarT MSK tool at T0 and T1, of whom 77 patients were regarded as stable. On average, the time between T0 and T1 was 7 days. The quadratic-weighted kappa coefficients for the overall STarT MSK tool scores (range 0–12) and prognostic subgroups (low, medium and high risk) were 0.71 and 0.65, respectively, indicating substantial agreement (Table 3). Distribution of prognostic subgroups was skewed due to the very low prevalence of patients at high risk. Test-retest reliability increased to 0.75 for the overall tool scores and decreased to 0.60 for prognostic subgroups, when agreement was calculated in the subset of 77 patients reporting stable musculoskeletal pain symptoms. Of the clinically stable patients, 23.4% were categorised differently within approximately 7 days.

Table 2. Pearson’s correlation between the Dutch STarT MSK tool and reference questionnaires.

<table>
<thead>
<tr>
<th>STarT MSK</th>
<th>Reference questionnaire</th>
<th>A priori</th>
<th>r</th>
<th>Correlation</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>NPRS</td>
<td>r ≥ 0.3</td>
<td>0.82</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 2</td>
<td>PSEQ</td>
<td>r ≤ -0.3</td>
<td>-0.22</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>Item 3</td>
<td>NPRS</td>
<td>r ≥ 0.3</td>
<td>0.48</td>
<td>Moderate</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 3</td>
<td>SF-36 PF</td>
<td>r ≤ -0.3</td>
<td>-0.42</td>
<td>Moderate</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 4</td>
<td>SF-36 PF</td>
<td>r ≤ -0.3</td>
<td>-0.47</td>
<td>Moderate</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 5</td>
<td>NPRS</td>
<td>r &lt; 0.3</td>
<td>0.13</td>
<td>Low</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 6</td>
<td>Timeline Brief IPQ</td>
<td>r ≥ 0.3</td>
<td>0.57</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 7</td>
<td>Single item ‘comorbidity’</td>
<td>r ≥ 0.3</td>
<td>0.34</td>
<td>Moderate</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 8</td>
<td>HADS-D</td>
<td>r ≥ 0.3</td>
<td>0.39</td>
<td>Moderate</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 9</td>
<td>TSK-11</td>
<td>r ≥ 0.3</td>
<td>0.29</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>Item 10</td>
<td>Single item ‘pain duration’</td>
<td>r ≥ 0.3</td>
<td>0.80</td>
<td>High</td>
<td>Yes</td>
</tr>
</tbody>
</table>

r = Pearson’s correlation.
NPRS = Numeric Pain Rating Scale, PSEQ = Pain Self-Efficacy Questionnaire. SF-36 PF = Physical Functioning subscale of the 36-item Short Form Health Survey, IPQ = Illness Perception Questionnaire, HADS-D = depression subscale of the Hospital Anxiety and Depression Scale, TSK-11 = 11-item Tampa Scale of Kinesiophobia.

Table 3. Quadratic-weighted kappa for the Dutch STarT MSK tool.

<table>
<thead>
<tr>
<th></th>
<th>Test-retest sample (n = 142)</th>
<th>Subset of patients reporting stable symptoms (n = 77)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall tool scores</td>
<td>0.71 (0.62–0.79)</td>
<td>0.75 (0.66–0.83)</td>
</tr>
<tr>
<td>Prognostic subgroups</td>
<td>0.65 (0.54–0.76)</td>
<td>0.60 (0.45–0.76)</td>
</tr>
</tbody>
</table>

* Values are represented as quadratic-weighted kappa coefficients with 95% confidence intervals.
κ ≤ 0: poor agreement, κ = 0.01–0.20: slight agreement, κ = 0.21–0.40: fair agreement, κ = 0.41–0.60: moderate agreement, κ = 0.61–0.80: substantial agreement, κ = 0.81–1.00: almost perfect agreement.
Predictive validity

In total, 136 patients (96%) completed the follow-up questionnaire at T2, of whom 73 patients were classified as low risk at baseline, 60 as medium risk, and 3 as high risk for persisting disability because of musculoskeletal pain. The mean (SD) SF-36 PF score at follow-up was 88.2 (14.4). The proportion of patients with persisting disability (SF-36 PF ≤ 80.0) was 13.7% in the low-risk group, 30.0% in the medium-risk group, and 100.0% in the high-risk group (Table 4). The RRs for persisting disability were 2.19 (95% CI: 1.10–4.38) for the medium-risk group and 7.30 (95% CI: 4.11–12.98) for the high-risk group, both compared to the low-risk group. The sensitivity analysis using different subgroup cut-points did not result in a substantially larger sample of high-risk patients and revealed less sufficient predictive performance of the Dutch STarT MSK tool (S1 Table).

Discussions

The aim of this study was to evaluate the validity and reliability of the Dutch STarT MSK tool in patients with musculoskeletal pain in primary care physiotherapy. The results showed a good construct validity, as 81.8% of the predefined hypotheses were confirmed. Test-retest reliability was substantial for the overall tool scores and prognostic subgroups in the test-retest sample, with quadratic-weighted kappa coefficients of 0.71 and 0.65, respectively. Test-retest reliability remained substantial for the overall scores and decreased slightly to a kappa of 0.60 for prognostic subgroups in the subset of patients reporting stable symptoms between T0 and T1. The predictive validity was sufficient, with RRs for persisting disability of 2.19 (95% CI: 1.10–4.38) for the medium-risk group and 7.30 (95% CI: 4.11–12.98) for the high-risk group, both compared to the low-risk group. The sample size for high-risk patients was small, therefore results for this group should be interpreted with caution.

The risk group distribution in our study was different from the distribution in the UK validation study [22] and STarT MSK pilot cluster Randomised Controlled Trial [23], in which our cohort had a shift toward low risk at the expense of high risk. In the present study, only 4 patients (2.8%) were categorised as high risk, which is significantly lower compared to the UK cohorts (respectively 19.0% and 13.3%). Although the low proportion of high-risk patients might suggest that the Dutch STarT MSK tool is not adequately able to distinguish between medium- and high-risk subgroups, several factors might explain the skewed distribution found in our study. First, we included patients who consulted their physiotherapist for musculoskeletal pain, while in the UK study patients consulted their GP. Most of the participating physiotherapists were specialised as manual therapist or sports physiotherapist. Consequently, it is possible that the more severe psychosomatic cases and potentially high-risk patients were
missed, since they possibly visit GPs or specialised psychosomatic physiotherapists. Second, different cultures and health care systems might have contributed to the discrepancy in high-risk patients found between cohorts. Finally, we considered that the original STarT MSK subgroup cut-points might not be appropriate for our clinical population; however the sensitivity analysis with different subgroup cut-points did not result in a substantial larger sample of high-risk patients and revealed less sufficient predictive performance of the tool.

Although the test-retest reliability found in our study suggests that the Dutch STarT MSK tool is able to classify patients into the same prognostic subgroup over time, further research may need to focus on the timing of stratification. We found a quadratic weighted kappa coefficient of 0.65 for prognostic subgroups, indicating substantial agreement. The kappa coefficient decreased to 0.60, indicating moderate agreement, when reliability was calculated using clinically stable patients. It seems more appropriate to consider the test-reliability of prognostic subgroups as substantial since distribution of prognostic subgroups was skewed due to the very low prevalence of patients at high risk, which increases chance agreement and reduces the value of kappa accordingly [45]. Additional analyses showed that 23.4% of the clinically stable patients were categorised differently within approximately 7 days. This finding is in line with a previous finding that changes in SBT categorisation might occur in the first few days after start of initial treatment [46]. Newell et al. found that one third of patients switched SBT risk groups within the 2 days between the initial stratification and after the first treatment. Especially in acute high-risk patients, psychosocial risk factors might be addressed during the first consultation [47], which could influence the results [14]. For example, specific concerns such as the likelihood of a serious underlying pathology and unhelpful beliefs and behaviors such as fear of movement can be addressed during the primary consultation. Despite the test-retest reliability was at an acceptable level, it has to be acknowledged that changes in risk categorisation after the first consultation of physiotherapy could potentially interfere the test-retest reliability. Future research should critically evaluate the most optimal timing of stratification (i.e., at initial assessment or within a few days after first treatment) for successfully predicting persisting musculoskeletal-related disability.

In our study patients received usual care from their physiotherapist. This methodology was chosen, because it most closely resembles how the tool will be used in physiotherapy practice and it seems not ethical to refrain patients from physiotherapy treatment during a 3-month period. Physiotherapy treatment might have influenced the predictive validity of the Dutch STarT MSK tool. In a study regarding changes in SBT categorisation during routine physiotherapy care, 81.8% of the high-risk, 76.0% of the medium-risk and 11.3% of the low-risk patients were categorised differently within a few weeks [48]. In a Danish SBT validation study, difference in risk prediction was observed between the Danish and original UK cohort, in which exposure to physiotherapy treatment was considered to be a confounder [49]. Morso et al. found that patients in the physiotherapy treatment group had a substantially lower risk of persisting disability than the GP group. Therefore, in our study, it is possible that the predictive validity of the Dutch STarT MSK tool might be reduced because of the natural history of the musculoskeletal pain condition being modified by physiotherapy treatment [49]. Despite the possible influence of physiotherapy treatment, the predictive validity showed statistically significant results. However, the confidence intervals of medium- and high risk were wide and show minor overlap, which may indicate a lack of power. The small sample size of high-risk patients and 100% of these patients having persisting disability at 3 months, may have led to an overestimation of the RR for the high-risk group.

To our knowledge, the STarT MSK tool is the only screening tool originally developed to allocate primary care patients with musculoskeletal pain into prognostic subgroups and subsequently stratify for the appropriate matched treatment. The prediction of persisting
musculoskeletal-related disability is based on the presence of biomedical and psychosocial prognostic factors at initial assessment. Beneciuk et al. concluded that reliance only on initial (SBT) risk categorisation may have limitations due to changes that might occur following routine physical therapy, and suggested that repeated assessment of risk categorisation can potentially improve prognosis for long-term low-back pain related disability [48]. Given the rapid development of “big data”, this raises the question whether other screening tools, incorporating big data analysis such as machine learning algorithms, may be more adequate to use in targeting treatment to individual patients. The application of machine learning in musculoskeletal physiotherapy is upcoming [50]. However, despite the possible advantages of big data and machine learning algorithms, there are still several challenges that needs to be addressed before further application in health care [51–53]. For now, the Dutch STaRt MSK tool seems an easy to use screening tool to start with in physiotherapy practice.

The strength of this study is that we evaluated the measurement properties of the Dutch STaRt MSK tool in primary care physiotherapy. In the Netherlands, physiotherapists in particular are involved in the treatment of patients with musculoskeletal pain and therefore are likely the primary users of the tool. Another strength is that we achieved an appropriate sample size (>100 patients) and respectively 100% and 96% follow-up rate. The present study has some limitations. First, the setting in which patients were included might have resulted in the low proportion of high-risk patients and limited generalisability of findings. Second, for feasibility and ethical reasons, patients received usual care from their physiotherapists. Effective physiotherapy treatment could have influenced the predictive validity of the Dutch STaRt MSK tool. Third, the sample size for high-risk patients was too small to calculate performance characteristics (i.e., sensitivity and specificity) in order to examine discrimination of the Dutch STaRt MSK tool [54]. Finally, we followed the method of Bier et al. [14] and used the guidelines of Cohen [37] as cut-off points for defining the strength of the expected correlations between the STaRt MSK items and reference questionnaires. It has to be acknowledged that these cut-off points are arbitrary and there are no widely accepted criteria for defining the strength of a relationship [55]. For the interpretation of complex abstract phenomena, lower correlations are often used as evidence of relationships [56] and therefore the cut-off point of 0.3 seems appropriate for the psychosocial variables. However, for some biomedical variables, a higher cut-off point might have been more appropriate.

This validation study is a first step toward identifying whether the Dutch STaRt MSK tool could be useful for physiotherapy practice. According to the psychometric analysis, the tool showed a sufficient to good validity and reliability in patients with musculoskeletal pain in primary care physiotherapy. Based on these results, the tool can be implemented in Dutch physiotherapy practice, in which it can be used to predict persisting disability at an early stage. Because of the small sample size for high-risk patients (n = 4), generalisability of findings may be limited. An external validation study with a larger sample of high-risk patients (≥50) is recommended. Furthermore, the Dutch STaRt MSK tool is not just a prognostic screening tool, but can be used to stratify patients for the appropriate matched treatment as well. Matched treatment packages are available for patients with low-back pain [47]. In addition, recently, primary care matching treatment options for patients with the five most common musculoskeletal pain presentations were proposed in a consensus groups study [57]. Future research should focus on whether these proposed matching treatment options are applicable within the Netherlands and particularly the primary care physiotherapy setting. Subsequently, further research is needed to determine whether the entire stratified care approach (use of the STaRt MSK tool and matched/targeted treatments) can be effectively used in primary care physiotherapy in the Netherlands.
Supporting information

S1 Appendix. The Keele STarT MSK tool© (including scoring method). (PDF)

S2 Appendix. Forward-backward translation method Dutch STarT MSK tool. (PDF)

S3 Appendix. The Dutch version of the STarT MSK tool (including scoring method). (PDF)

S1 File. The Keele STarT MSK tool. (PDF)

S1 Table. Sensitivity analysis predictive validity using different subgroup cut-points. (PDF)

Author Contributions


Data curation: Anke G. van den Broek, Corelien J. J. Kloek.


Investigation: Anke G. van den Broek.


Project administration: Anke G. van den Broek, Corelien J. J. Kloek.

Resources: Cindy Veenhof.

Supervision: Corelien J. J. Kloek, Martijn F. Pisters, Cindy Veenhof.


Visualization: Anke G. van den Broek.

Writing – original draft: Anke G. van den Broek, Corelien J. J. Kloek.

Writing – review & editing: Anke G. van den Broek, Corelien J. J. Kloek, Martijn F. Pisters, Cindy Veenhof.

References


52. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC medical informatics and decision making. 2018; 18(1):139. Epub 2018/12/31. https://doi.org/10.1186/s12911-018-0719-2 PMID: 30594159; PubMed Central PMCID: PMC6311005 Not applicable. COMPETING INTERESTS: MP and JB are members of the editorial board of BMC Medical Informatics and Decision Making. The authors declare that they have no other competing interests. PUBLISHER’S NOTE: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


