General section
Review

OPEN

PAIN REPORTS®

# Multidimensional screening for predicting pain problems in adults: a systematic review of screening tools and validation studies

Elke Veirman[a,*], Dimitri M. L. Van Ryckeghem[a,b,c], Annick De Paepe[a], Olivia J. Kirtley[d], Geert Crombez[a]

Downloaded from https://journals.lww.com/painrpts by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3GGzJhRXoJ4AApJOBCZAUXPTcaBLV4J2w8EMY4C1Gg= on 09/16/2019

## Abstract

Screening tools allowing to predict poor pain outcomes are widely used. Often these screening tools contain psychosocial risk factors. This review (1) identifies multidimensional screening tools that include psychosocial risk factors for the development or maintenance of pain, pain-related distress, and pain-related disability across pain problems in adults, (2) evaluates the quality of the validation studies using Prediction model Risk Of Bias ASsessment Tool (PROBAST), and (3) synthesizes methodological concerns. We identified 32 articles, across 42 study samples, validating 7 screening tools. All tools were developed in the context of musculoskeletal pain, most often back pain, and aimed to predict the maintenance of pain or pain-related disability, not pain-related distress. Although more recent studies design, conduct, analyze, and report according to best practices in prognosis research, risk of bias was most often moderate. Common methodological concerns were identified, related to participant selection (eg, mixed populations), predictors (eg, predictors were administered differently to predictors in the development study), outcomes (eg, overlap between predictors and outcomes), sample size and participant flow (eg, unknown or inappropriate handling of missing data), and analysis (eg, wide variety of performance measures). Recommendations for future research are provided.

**Keywords:** Multidimensional screening, Yellow flags, Pain, Risk of bias

## 1. Introduction

Chronic pain is a common experience, with a prevalence of between 10% and 20% in the general adult population.[6,7,34,95,114] Often, chronic pain is disabling and notoriously difficult to treat.[87] At least 2 strategies are possible to face these challenges. First, we can develop new and better medical and psychosocial interventions.[19] Second, we can prevent acute pain from becoming chronic. The latter requires an understanding of how and why acute pain becomes chronic, the identification of individuals at risk, and the timely delivery of preventive actions.[67,126]

Evidence has been accumulating that psychosocial variables are important in the prediction and prevention of chronic pain. First, available experimental and prospective research reveals the role of psychosocial factors in explaining pain, distress, and disability.[57] The roles of learning, emotions, and cognitive factors are well established in laboratory studies,[123] and a number of prospective studies have provided evidence for the role of psychosocial factors in the development and maintenance of pain.[3,60,102] For example, Sobol-Kwapinska et al.[106] reviewed predictors of acute postsurgical pain and found pain catastrophizing, optimism, expectation of pain, neuroticism, anxiety (state and trait), negative affect, and depression to be associated with acute postsurgical pain. Second, contemporary theoretical models have provided insight into how acute pain patients with a particular psychosocial profile may become stuck in a vicious cascade of further pain, distress, and disability.[13,122] Third, evidence is increasing that the timely delivery of cognitive-behavioral interventions can prevent persistent disability.[67]

Taking this evidence into account, Kendall et al.[51] called for the routine assessment of psychosocial factors in people experiencing acute pain. They introduced the concept of "yellow flags" as a method to screen for psychosocial risk factors predicting long-term disability, a concept that has been adopted by a growing number of researchers interested in examining the value of prognostic models.[26,27] This has led to the development of screening tools that include various psychosocial risk factors and

*Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.*

*[a] Department of Experimental Clinical and Health Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium, [b] Institute for Health and Behaviour, INSIDE, Faculty of Language and Literature, Humanities, Arts and Education, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [c] Section Experimental Health Psychology, Clinical Psychological Science, Departments, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands, [d] Center for Contextual Psychiatry, Department of Neurosciences, KU Leuven, Leuven, Belgium*

**Corresponding author. Address: Department of Experimental Clinical and Health Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Henri Dunantlaan 2, B-9000 Gent, Belgium. Tel.: +32 92646392; fax: +32 9 264 64 89. E-mail address: Elke.Veirman@UGent.be (E. Veirman).*

*Copyright © 2019 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of The International Association for the Study of Pain. This is an open access article distributed under the Creative Commons Attribution License 4.0 (CCBY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

*PR9 4 (2019) e775*

*http://dx.doi.org/10.1097/PR9.0000000000000775*

a recommendation for their use in clinical practice (eg, Keele STarT Back Screening Tool [STarT Back][41]; Preventing the Inception of Chronic Pain [PICKUP][115]).

Several reviews have summarized the predictive performance of screening tools.[30,42,49,68,99] For instance, in a meta-analysis of screening tools, Karan et al.[49] showed that screening tools poorly predicted pain, but were acceptable and excellent in predicting disability, and absenteeism, respectively (eg, STarT Back, OMPSQ). This meta-analysis is of high quality and according to the highest standards in the field.[71,109] For that reason, our aim was not to focus upon the actual performance of the screening tools. Available meta-analyses[49,99] have also noted that the methodological quality of studies investigating the predictive performance of screening tools is variable. Nevertheless, these reviews do not provide details of the methodological problems and limitations.

For that reason, our review focuses upon the methodological quality of studies that validate screening tools. First, a detailed analysis and synthesis of the methodological quality of studies is largely missing. Indeed, despite being considered fundamental to guide interpretation of findings, and recommendations for future research and practice,[52] available reviews spend little or no attention to this topic. Second, the methodological quality of the studies in these reviews, typically described as "risk of bias,"[40] was assessed using instruments that were not specifically designed for evaluating the quality of prediction models (eg, Quality in Prognostic Studies tool [QUIPS]).[36,37] Recently, "Prediction model Risk Of Bias ASsessment Tool" (PROBAST), a tool for assessing the risk of bias and applicability of diagnostic and prognostic prediction model studies, has become available and used.[76,127–129]

The aim of this systematic review was 3-fold: (a) to identify available multidimensional screening tools that include psychosocial risk factors for poor pain outcomes (development or maintenance of pain, pain-related distress, and pain-related disability) across pain problems in adults, (b) to evaluate the quality of prospective studies validating these screening tools with up-to-date standards for clinical prediction models, and (c) to synthesize methodological concerns that may bias the predictive performance of these screening tools.

## 2. Methods

### 2.1. Literature search and eligibility criteria

The literature search comprised 4 steps. First, a search was performed for studies published in peer-reviewed journals across relevant electronic databases (MEDLINE, PsychINFO, and Web of Science) using the following terms in the title, key words, or abstract: *screen** AND (*tool* OR *questionnaire*) AND *pain* AND *risk*. Screening of titles, key words, and abstracts allowed identification of screening tools and eligible studies. Second, a list of publications was sent to lead authors in the field of pain research to ask for any other available screening tools of which they were aware. Third, the reference lists of relevant systematic reviews were hand-searched for any articles that were not yielded by our other search methods. Finally, when only the development article for a tool fulfilling the inclusion criteria (see below) was identified in the search, a search was performed for additional articles that fulfilled the inclusion criteria by screening all publications that cited this development article.

The following eligibility criteria were used to identify screening tools for inclusion in this systematic review:

(1) The screening tool is a self-report questionnaire.
(2) The screening tool is multidimensional, containing at least 2 psychosocial risk factors. The report of somatic experiences such as pain, radiation, or other somatic complaints is not considered as psychosocial factors.
(3) The screening tool aims to predict the development (<3 months) or maintenance (≥3 months) of pain, pain-related distress, or pain-related disability.
(4) The screening tool is specifically developed in the context of pain and can target any type of pain (eg, neck pain and low back pain).
(5) The screening tool is a standalone instrument. Therefore, the tool should not consist of a battery of questionnaires, as is often the case for research purposes.
(6) The screening tool is validated in at least 1 independent study, ie, using data that were not used to develop the screening tool.

Six criteria (listed below) were used to select studies for inclusion. Some criteria were included to set a minimum quality (eg, criterion 1), whereas other criteria were applied to narrow the scope of the review (eg, criterion 2).

(1) The study is a full report published in a peer-reviewed scientific journal.
(2) The study includes an adult sample (the average age of the sample was older than 18 years).
(3) At baseline, the study includes patients experiencing no or (sub)acute pain (<3 months), without restriction in the type of pain experienced (eg, musculoskeletal pain, neuropathic pain, and postoperative pain). In line with the development studies of screening tools, we excluded studies involving only patients with chronic pain (≥3 months). Studies involving mixed samples with (sub)acute and chronic pain patients were included. However, when data for separate subsamples were reported, we only included the samples of interest.
(4) The study includes at least 1 screening tool, which is used in its original form. Some differences in translations, item order, and response scale are accepted. Shortened versions are considered different instruments.
(5) The study includes at least one of the following outcomes during outcome assessment (<2 years after baseline assessment): (a) Pain intensity or pain bothersomeness, assessed using a Visual Analogue Scale (VAS), a Numeric Rating Scale (NRS), a verbal rating scale, or a Likert scale; (b) pain-related disability including activity limitations (ie, difficulties in executing a task or an action such as the ability to walk, eat, shower, or dress) and participation restrictions (ie, problems relating to the involvement in life situations such as sick leave or days absent from work or return to work status) according to the International Classification of Functioning, Disability, and Health (ICF) framework.[130] Assessment of these outcomes could be performed with (a subset of questions from) a self-report questionnaire, single questions, or data from existing registration systems; and (c) pain-related distress (eg, anxiety, fear, or low mood), assessed through self-report measures.
(6) The study is a prospective cohort study including patients presenting in primary, secondary, and tertiary health care settings.

Finally, studies were considered ineligible if they aimed to investigate the impact of stratified care (ie, targeted treatment to patient subgroups based on the results of the screening tool) or interventions that specifically targeted psychosocial risk factors (ie, cognitive behavioral therapy) or they consisted of a randomized control trial. We reasoned that the focus of these studies is on the evaluation of a (psychological) therapeutic intervention

and not on the investigation of the predictive value of screening tools.

### 2.2. Data extraction and risk of bias assessment

The assessment of the quality of studies that validated the selected screening tools was based upon a prepublication version of the Prediction model study Risk Of Bias ASsessment Tool (PROBAST) (personal communication, January 2017, Dr. Robert Wolff). The PROBAST has been developed by the Cochrane Prognosis Methods Group using a Delphi process, in which 40 experts in the fields of prediction research and systematic review methodology participated.[129] Its use is recommended by most recent guidelines for performing systematic reviews and meta-analyses of prediction model performance.[16]

Data extraction of eligible validation studies was conducted by E.V. and O.K. following a customized PROBAST template that was created for each of the 5 risk of bias assessment areas: (1) *participant selection*, (2) *predictors*, (3) *outcomes*, (4) *sample size and participant flow*, and (5) *analysis* (details can be retrieved from the authors upon request).[74] Extracted data formed the basis for the risk of bias assessment, where signaling questions across those 5 important areas were rated as *yes*, *probably yes*, *probably no*, *no*, or *no information,* with *yes* indicating the absence of bias and *probably no* or *no* indicating the potential for bias.

For *participant selection*, elements judged were whether appropriate inclusion and exclusion criteria were used and whether patients had a similar state of health at enrollment. For *predictors*, questions considered were whether definition and assessment of predictors were similar across participants, and whether definition and assessment of predictors were similar compared with those of the development model. For *outcomes*, important elements judged were whether a valid outcome was used, whether predictors were excluded from the outcome definition, whether definition and assessment of outcomes were similar across participants, whether definition and assessment of outcomes were similar compared with those of the development model, and whether outcome assessment was blinded to predictor data. For *sample size and participants flow*, elements judged were whether a reasonable number of outcome events were available, whether the time interval between predictor and outcome assessment was appropriate, whether all enrolled participants were included in the analyses, and whether missing data occurred and participants with missing data were handled appropriately. Finally, for *analysis*, evaluated elements focused on whether relevant model performance measures were evaluated. Domains were subsequently rated as *high*, *moderate*, *low*, or *unclear* risk of bias. Risk of bias assessment labels were discussed and assigned upon agreement among team members (G.C., D.V.R., and E.V.).

## 3. Results

### 3.1. Study selection

The study selection process was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA),[71] except for a preregistration of the review. Electronic databases were searched from the earliest record available on September 15, 2016, resulting in 1850 records. After removal of duplicate articles, 2 reviewers (J.C. and E.V.) independently screened a selection of the titles, key words, and abstracts for possible study inclusion. First screening resulted in 187 remaining references.

In the second step, full copies of articles were obtained (E.V.). Full-text reading of these articles resulted in exclusion of several tools for the following reasons (1) not being a screening tool (eg, "Amsterdam Preoperative Anxiety and Information Scale"),[70] (2) the screening tool was not developed in the context of pain (eg, "Distress and Risk Assessment Method"),[65] (3) the screening tool did not assess any psychosocial factors (eg, "London Fibromyalgia Epidemiology Study Screening Questionnaire"),[54] and (4) the screening tool assessed only 1 psychosocial factor (eg, "Fear Avoidance Beliefs Questionnaire").[93]

For 3 potentially eligible screening tools, items were not available in the literature and author contact yielded insufficient access to the tools' items ("Nijmegen Outcome of Lumbar Disc surgery Screening-instrument"[17]; "ABLE Presurgical Assessment Tool"[2]; and "Psychosocial Risk for Occupational Disability Scale"[100]).

Finally, a number of eligible screening tools for which items were available in the literature were not included in the current review as no independent validation studies were retrieved from the electronic database search nor through cited reference search of the development articles of the screening tools (ie, "Absenteeism Screening Questionnaire"[116]; "Back Disability Risk Questionnaire"[103,104]; "Optimal Screening for Prediction of Referral and Outcome cohort yellow flag assessment tool"[58]; "Pain Recovery Inventory of Concerns and Expectations"[105]; "Screening-Instrument zur Feststellung des Bedarfs an medizinisch-beruflich orientierter Rehabilitation"[112]; "Traumatic Injuries Distress Scale"[125]; and "Work and Health Questionnaire"[1]).

In addition to the 27 articles that were considered eligible from the electronic database search, 2 articles[25,50] were identified through cited reference search of the development articles of the screening tools on May 4, 2017, and 3 references[32,53,64] were retrieved by hand-searching of relevant review articles[8,30,42,45,49,59,68,86,88,90,99,107] (O.K.), resulting in a total of 32 references fulfilling the inclusion criteria for the current review. Additional author contact yielded no other tools or studies (see **Figure 1** for a flowchart).

Doubts and disagreements on the inclusion of screening tools and eligible studies were resolved by discussion within the team (G.C., D.V.R., E.V., A.D.P., and O.K.) until consensus was reached. After finalizing the systematic search, all screening tools and development studies were retrieved to extract essential data for the risk of bias assessment. During the screening process, reviewers were not blind to authorship, institution, journal, or results.

### 3.2. Study characteristics: screening tools

The 32 included articles contained 42 study samples. Notably, several articles reported on a similar sample as earlier published articles, whereas other study samples completed multiple screening tools. The articles reported on the validation of 7 screening tools:

(1) Acute Low Back Pain Screening Questionnaire (ALBPSQ; 7 studies)[62]/Örebro Musculoskeletal Pain Screening Questionnaire (OMPSQ; 10 studies)[61]/Örebro Musculoskeletal Screening Questionnaire (OMSQ; 3 studies).[25] The ALBSQ is a 24-item self-report questionnaire aiming to predict poor prognosis—operationalized as accumulated sick leave—in acute and subacute patients presenting with musculoskeletal pain

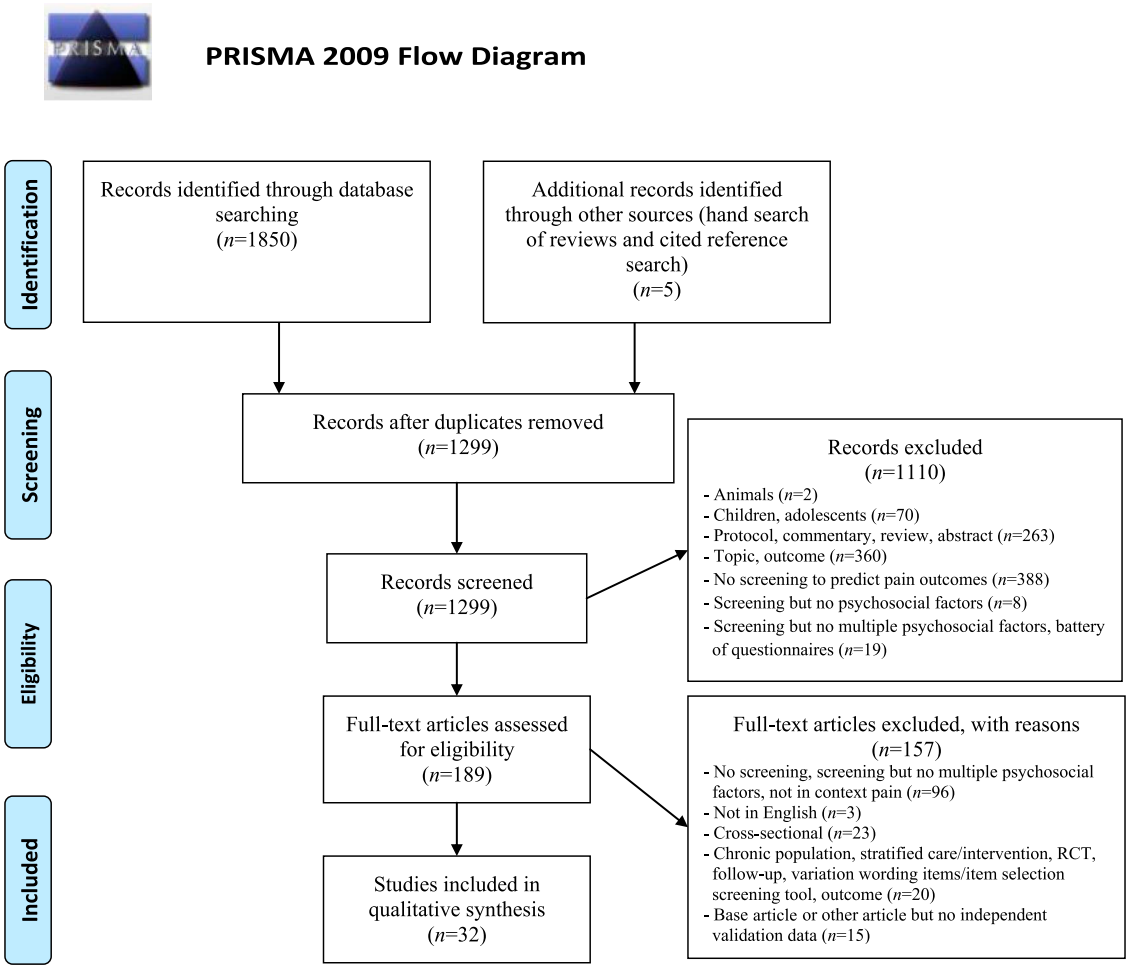


Figure 1. Flow of studies through the review.

(back, neck, and shoulder pain). A few years after its development, it was relabeled as the OMPSQ, including an additional unscored item on employment status. More recently, the OMSQ broadened the focus of the ALBPSQ to general musculoskeletal problems and simplified the questions.

(2) Örebro Musculoskeletal Pain Screening Questionnaire short version (OMPSQs; 2 studies).[63] The OMPSQs is a 10-item self-report questionnaire designed to predict disability—operationalized as sick leave—in workers suffering from musculoskeletal pain (back pain).

(3) Örebro Musculoskeletal Screening Questionnaire short version (OMSQs; 1 study).[23] The OMSQs is a 12-item self-report questionnaire aiming to predict a wide variety of outcomes—including problem severity, functional impairment, absenteeism, long-term absenteeism, cost, and recovery time—in acute and subacute work-injured patients presenting with musculoskeletal pain (whiplash, low back pain).

(4) Heidelberger Kurzfragebogen Rückenschmerz (HKF-R10; 1 study).[79] The HKF-R10 is a 27-item self-report questionnaire developed to predict the likelihood of chronicity in patients with acute low back pain.

(5) Pain Belief Screening Instrument (PBSI; 1 study).[97] The PBSI is a 7-item self-report questionnaire aiming to predict disability in subacute and chronic pain patients with musculoskeletal pain (neck, shoulder, and low back pain).

(6) Keele STarT Back Screening Tool (SBT; 11 studies).[41] The SBT is a 9-item self-report questionnaire developed to predict poor outcome—operationalized as disability—in (sub)acute and chronic primary care patients with nonspecific low back pain.

(7) Preventing the Inception of Chronic Pain (PICKUP; 2 studies).[115] The PICKUP is a 5-item self-report questionnaire aiming to predict the risk of chronic low back pain in patients with acute low back pain.

An overview of the included instruments and more detailed characteristics (as described in the base article) can be found in **Table 1**.

### 3.3. Study characteristics: sources and samples

Studies were conducted between 2000 and 2017.[43,50] The majority of the studies included samples that were collected in Northern European countries (N = 11) or Western European countries (N = 11). A small number of studies collected data from samples outside Europe, including Canada (N = 1), the United States (N = 3), Australia and New Zealand (N = 7), and China (N = 1).

Sex and age of participants differed largely between study samples. The average/median age of participants ranged between 37.7 years and 53.0 years.[21,64] The sex of participants varied from 33.7% female participants to 83.0% female participants.[18,63]

**Table 1**

**Summary of included screening tools.**

| Screening tool | Development study | Summary of instrument | Scoring method | Cutoff scores/subgrouping, follow-up |
|---|---|---|---|---|
| Acute Low Back Pain Screening Questionnaire (ALBPSQ), later renamed as Örebro Musculoskeletal Pain Screening Questionnaire (OMPSQ), and reframed as Örebro Musculoskeletal Screening Questionnaire (OMSQ) | Linton and Hallden,[62] Sweden | 24 items<br>Risk assessment for poor prognosis—operationalized as accumulated sick leave.<br>In acute and subacute patients presenting with musculoskeletal pain (lower back, neck, and shoulder).<br>In primary care setting. | 21 items are scored, covering pain experience (5 items), physical functioning (5 items), coping (1 item), job satisfaction (1 item), anxiety/stress (1 item), depression (1 item), fear-avoidance beliefs (3 items), recovery expectations (2 items), heavy or monotonous work (1 item), and sick leave (1 item).<br>Miscellaneous items relate to age, sex, and nationality. | Cutoff score of 105.<br>6-month follow-up. |
| Örebro Musculoskeletal Pain Screening Questionnaire short (OMPSQs) | Linton et al.,[63] Sweden | 10 items<br>Risk assessment for poor prognosis—operationalized as sick leave.<br>In workers suffering from musculoskeletal pain (back).<br>In occupational health care setting. | 10 items are scored, covering pain experience (2 items), self-perceived function (2 items), distress (2 items), return to work expectancies (2 items), and fear avoidance beliefs (2 items). | Cutoff score of 50.<br>1-year follow-up. |
| Örebro Musculoskeletal Screening Questionnaire short (OMPQs) | Gabel et al.,[23] Australia | 12 items<br>Risk assessment for poor prognosis—operationalized as problem severity, functional status, absenteeism, long-term absenteeism, recovery time, and cost.<br>In acute and subacute workers presenting with musculoskeletal pain (whiplash and low back pain).<br>In primary care setting. | 12 items are scored, covering pain/problem experience (3 items), physical function (2 items), life satisfaction (1 item), depression (1 item), anxiety (1 item), fear-avoidance beliefs (2 items), recovery expectations (1 item), and other (1 item). | No optimal cutoff recommended.<br>6-month follow-up. |
| Heidelberger Short Early Risk Assessment Questionnaire (HKF-R 10) | Neubauer et al.,[79] Germany | 27 items<br>Risk assessment for chronic low back pain<br>In patients with acute low back pain.<br>In primary care setting. | 26 items are scored, covering sociodemographic information (2 items), pain intensity and duration (4 items), efficacy of massage (1 item), depression (5 items), catastrophizing (5 items), and helplessness and hopelessness (9 items).<br>An additional item regarding pain intensity in the past week is present in the measure, but is not included within the total score. | No optimal cutoff recommended.<br>6-month follow-up. |
| Pain Belief Screening Instrument (PBSI) | Sandborgh et al.,[97] Sweden | 7 items.<br>Risk assessment for disability.<br>In subacute and chronic pain patients with musculoskeletal pain (neck, shoulder, and low back)<br>In primary care setting. | 7 items are scored, covering pain intensity (1 item), functional ability (1 item), fear-avoidance (2 items), catastrophizing (1 item), and self-efficacy (2 items). | No optimal cutoff recommended.<br>No follow-up. |
| STarT Back Tool (SBT) | Hill et al.,[41] United Kingdom | 9 items.<br>Risk assessment for pain-related disability.<br>In (sub)acute and chronic patients with nonspecific back pain.<br>In primary care setting. | 9 items are scored, covering bothersomeness of pain (1 item), presence of referred pain (1 item), comorbid pain (1 item), disability (2 items), catastrophizing (1 item), fear (1 item), anxiety (1 item), and depression (1 item). | Stratification of patients in low (overall score 0–3), medium (overall score >3; psychosocial subscale score <4), or high risk (psychosocial subscale scores ≥4) categories of poor clinical outcome, assisting in decision-making about the specific course treatment.<br>6-month follow-up. |
| Predicting the Inception of Chronic Pain (PICKUP) | Traeger et al.,[115] Australia | 5 items.<br>Risk assessment for chronic low back pain.<br>In patients with acute low back pain.<br>In primary care setting. | 5 items are scored online through http://pickuptool.neura.edu.au/, covering pain intensity (1 item), leg pain (1 item), disability compensation (1 item), depression (1 item), and perceived risk (1 item). | Predicted probability risk score in percentage.<br>3-month follow-up. |

Study samples were collected in primary care (83.3%) and secondary care settings (11.9%), and 1 study included a combined sample of participants from primary and secondary care units (4.8%).[92] The terminology used to describe the settings varied, by reference to providers (eg, general practitioner or a physical therapist)[61] and/or type of services (eg, spinal

outpatient clinic).[50] Although some studies detailed the treatment patients received (eg, work conditioning program),[66] others often did not (eg, treated as usual).[81] If information about the use of treatments is reported with insufficient detail, it can potentially bias performance results of the included screening tools because it does not allow researchers to evaluate the impact it might have had on the results.[82,83] Moreover, within the studies that reported on the use of treatments, none of the studies accounted for treatment use.

Most study samples comprised participants with musculoskeletal pain. In particular, patients with back pain were overrepresented. Study samples often also included participants from other populations, such as those experiencing neck pain, pain between the shoulder blades,[124] or multisite pain[24,25] (see **Table 2** for an overview).

### 3.4. Risk of bias assessment of included studies

#### 3.4.1. Participant selection

The majority of the study samples consisted of mixed samples containing both acute and chronic pain patients (59.5%). The remaining samples comprised patients with acute pain (33.3%) or samples for which the type of pain (acute, subacute, or chronic pain) was not clearly described (7.1%) (**Table 2**).

For the PROBAST "participant selection" domain, the majority of the study samples were rated as having a moderate risk of bias (51.1%). Fewer study samples were rated as having low (16.7%) or high risk of bias (19.0%). For the remaining study samples, the risk of bias was rated as unclear (**Table 3**) because the presented information was insufficient to evaluate the appropriateness of the inclusion criteria or the state of health of participants. The reasons for increasing the risk of bias related to the specified inclusion and exclusion criteria and differences in the state of health of participants at enrollment.

##### 3.4.1.1. Inclusion and exclusion criteria

The eligibility criteria were sometimes inappropriate or unclear. For example, some studies did not exclude unemployed participants[14] or did not report information on employment,[38] although the screening tools contained work-related questions. Most studies reported inclusion and exclusion criteria. However, sometimes the criteria had to be retrieved from descriptive information[66] or from a previously published study.[64]

##### 3.4.1.2. Participants' state of health at enrollment

Although most studies aimed to recruit a homogeneous sample, other studies did not. Participants were found not to be in a similar state of health at baseline in cases when studies included patients with (sub)acute and chronic pain in a single sample.[5,14,20–22,25,28,41,43,44,50,53,61,63,64,77,78,80,92,98] For instance, despite George and Beneciuk[28] reported detailed information about their patients with (sub)acute and chronic pain, the analyses were based upon the full sample. For a considerable number of studies, the state of health of the participants had to be derived from descriptive information. For example, Margison and French[66] only reported on average pain duration in weeks. Sometimes, insufficient information was available to conclude whether participants were in a similar state of health at enrolment[63] (**Table 2**).

#### 3.4.2. Predictors

The most frequently used screening tool was the ALBPSQ/OMPSQ/OMSQ. However, we noted that the cutoff score to identify the high risk group varied substantially, ranging between 72 and 147 (**Table 4**).

For the PROBAST "predictors" domain, the majority of the study samples were rated as having a low risk of bias (40.5%). Only a small number of study samples were rated as having moderate (19.0%) or high risk of bias (11.9%). For 28.6% of the study samples, the risk of bias was rated as unclear (**Table 3**) because the presented information was insufficient to evaluate whether differences occurred in the assessment of the screening tools either across participants or compared with the development study. The reasons for increasing the risk of bias related to differences in the assessment of the screening tools across participants and differences in the assessment of screening tools compared with the development study.

##### 3.4.2.1. Definition and assessment of predictors across participants

Study samples that validated the ALBPSQ, the OMPSQ, and the PICKUP—tools that include work-related questions—sometimes did not report information on employment status. This could mean that participants were all employed, or that some of the participants were unemployed, but it was not reported.[50] Furthermore, those studies that did report on employment status did not always administer these tools in a similar way across participants. For example, Hurley et al. instructed participants to fill out ALBPSQ work-related questions as best they could, even when they were unemployed.[43,44] When these questions were left blank, the mean score of the other questions was used as replacement. In the study by Grotle et al.,[33] it is noted that for participants who were unemployed, OMPSQ work-related questions were replaced by the mean score of the other questions.

##### 3.4.2.2. Definition and assessment of predictors compared with the developmental model

Furthermore, across the included studies, significant variation was observed in the applied screening tool cutoff points used to categorize patients. Selective reporting of results based only on cutoff values other than those specified in the original development study for the screening tool, was considered a risk for underestimation or overestimation of the screening tool's predictive accuracy. Moreover, variable use of cutoffs prohibits to estimate the influence of a given setting on the performance at the recommended (original) threshold. For example, for the ALBPSQ, the standard cutoff originates from Linton and Halldén,[62] who used 105 as their cutoff score for detecting poor prognosis in the form of sick leave. Hurley et al.[44] and Vos et al.[124] only reported results using a cutoff of 112 and 72, respectively, for the outcome sick leave. In addition, few studies also treated screening tool scores as continuous without additional reporting of the cutoff values from the screening tool's development study[38] (**Table 4**).

#### 3.4.3. Outcomes

The majority of study samples assessed one or more outcomes related to *pain* (66.7%), activity limitations (54.8%), and participation restrictions (50.0%). In addition, about half of the study samples reported also mixed or composite outcomes (40.5%) (**Table 4**).

For the PROBAST "outcomes" domain, the majority of the study samples were assigned an unclear risk of bias (40.5%), mainly due to insufficient information to evaluate blinding, or a moderate risk of bias (42.9%). None of the study samples were

**Table 2**

**Key study and participant characteristics of included validation studies.**

| Study | Country; setting | Age in years [SD; (IQ-)range] | % female | Pain type | Pain duration | Pain intensity |
|---|---|---|---|---|---|---|
| *ALBPSQ* | | | | | | |
| Hurley et al.[43] | United Kingdom;<br>Primary care<br>Physiotherapy departments and health centers | M = 43.19 [range: 17–77] | 60 | Low back pain | <12 weeks: N = 64<br>>12 weeks: N = 50 | $MPQ_{ALBPSQ \leq 112}$:<br>*Med* = 14.5 [IQ range = 12.2; range = 1–54]<br>$MPQ_{ALBPSQ > 112}$:<br>*Med* = 27.5 [IQ range = 24.5; range = 0–70] |
| Hurley et al.[44] | United Kingdom;<br>Primary care<br>Physiotherapy departments | *Med* = 41.5 [range: 17–77] | 60 | Low back pain | <12 weeks: 56% | MPQ:<br>*Med* = 19.0 [IQ range: 20.0, range: 0–70] |
| Grotle et al.[32] | Norway;<br>Primary care<br>General practitioners, chiropractors, and physical therapists (27% recruited through advertisement) | M = 38.9 [SD = 10.3] | 57 | Low back pain with or without radiation | ≤3 days: N = 41, 34%<br>4–12 days: N = 43, 36%<br>13–20 days: N = 36, 30% | NA |
| Grotle et al.[33] | Norway;<br>Primary care<br>General practitioners, chiropractors, and physical therapists (27% recruited through advertisement) | M = 38.0 [SD = 10.1] | 54 | Low back pain with or without radiation | M = 2.3 weeks [SD = 2.2] | ALBPSQ current pain:<br>M = 6.7 (SD = 1.8)<br>ALBPSQ average pain:<br>M = 3.0 (SD = 2.5) |
| Grotle et al[31] | Norway;<br>Primary care<br>General practitioners, chiropractors, and physical therapists (27% recruited through advertisement) | M = 37.9 [SD = 10.1] | 55 | Low back pain with or without radiation | M = 8.1 days [SD = 6.6] | NRS pain intensity last week:<br>M = 6.7 [SD = 1.8] |
| Heneweer et al.[38] | The Netherlands;<br>Primary care<br>Physical therapists | Recovered*: M = 40.8 [SD = 9.2]<br>Not recovered*: M = 43.1 [SD = 9.1] | 39 | Nonspecific low back pain | Recovered*:<br><4 weeks: N = 20, 64.5%<br>4–6 weeks: N = 9, 29.0%<br>7–12 weeks: N = 2, 6.5%<br>Not recovered*:<br><4 weeks: N = 9, 36.0%<br>4–6 weeks: N = 6, 24.0%<br>7–12 weeks: N = 10, 40.0% | NA |
| Vos et al.[124] | The Netherlands;<br>Primary care<br>General practitioners | Male/female: M = 43.2/38.2 | 64 | Neck pain | M = 2.76 weeks [SD = 3.00] | ALBPSQ current pain:<br>M = 6.5 [SD = 1.75]<br>ALBPSQ average pain:<br>M = 3.78 [SD = 2.76] |
| *OMPSQ* | | | | | | |
| Linton and Boersma[61] | Sweden;<br>Primary care<br>General practitioners and physical therapists | M = 41.1 [range: 22–66] | 48 | Neck and back pain | >24 weeks: 43% | OMPSQ current pain:<br>M = 6.2 [SD = 2.1]<br>OMPSQ average pain:<br>M = 5.1 [SD = 2.2] |
| Dunstan et al.[18] | Australia;<br>Primary care<br>Occupational injury compensation database | [range: 18–65] | 34 | Musculoskeletal pain | NA | NA |
| Margison and French[66]—Derivation sample | Canada;<br>Primary care<br>Private-sector clinics and physiotherapy clinics | M = 41.2 [SD = 10.8] | 41 | Neck, shoulder, upper back, lower back, arm, wrist, and hand, leg, ankle, and foot, and other pain | M = 6.7 weeks [SD = 1.7] | Pain intensity past 3 months:<br>M = 6.8 [SD = 2.0] |

(*continued on next page*)

Table 2 (*continued*)

**Key study and participant characteristics of included validation studies.**

| Study | Country; setting | Age in years [SD; (IQ-)range] | % female | Pain type | Pain duration | Pain intensity |
|---|---|---|---|---|---|---|
| Margison and French[66]—Validation sample | Canada;<br>Primary care<br>Private-sector clinics and physiotherapy clinics | M = 41.5 [SD = 9.8] | 39 | Neck, shoulder, upper back, lower back, arm, wrist, and hand, leg, ankle, and foot, and other pain | M = 6.6 weeks [SD = 1.5] | Pain intensity past 3 months:<br>M = 7.1 [SD = 2.1] |
| Maher and Grotle[64]—Australasian sample | Australia and New Zealand;<br>Primary care<br>Physiotherapy clinics | M = 43.3 [SD = 12.1] | 43 | Nonspecific low back pain | 6–8 weeks: N = 45<br>9–11 weeks: N = 38<br>12 weeks: N = 17 | OMPSQ current pain:<br>M = 5.2 [SD = 1.9] |
| Maher and Grotle[64]—Norwegian sample | Norway;<br>Primary care<br>Doctors and chiropractors | M = 38.7 [SD = 9.7] | 56 | Low back pain | <1 week: N = 50<br>1–2 weeks: N = 21<br>2–3 weeks: N = 29 | OMPSQ current pain:<br>M = 6.8 [SD = 1.8] |
| Gabel et al.[25]—OMPSQ | Australia;<br>Primary care<br>Physiotherapy outpatient clinics | M = 39 [SD = 7; range: 18–58] | 42 | Lower back, lower back and leg, lower back and neck, back, neck, and shoulder pain | M = 4.0 weeks [SD = 8.2]<br>6% chronic | OMPSQ current pain:<br>M = 6.5 [SD = 1.8]<br>OMPSQ average pain:<br>M = 6.2 [SD = 3.0] |
| Gabel et al.[25]—OMSQ | Australia;<br>Primary care<br>Physiotherapy outpatient clinics | M = 39 [SD = 9; range: 18–58] | 43 | Neck/back, arm, leg, both sides, and several areas | M = 4.1 weeks [SD = 8.1]<br>8% chronic | OMSQ intensity acute:<br>M = 6.6 [SD = 1.9]<br>OMSQ severity chronic:<br>M = 5.8 [SD = 2.7] |
| Linton et al.[63] | Sweden;<br>Primary care | M = 48 | 83 | Nonspecific back or neck pain | NA | NA |
| Gabel et al.[24] | Australia;<br>Primary care<br>Physiotherapy centers | M = 38.9 [SD = 10.5; range: 18–65] | 43 | Musculoskeletal pain resulting from work injury (back, neck, upper limbs, lower limbs, and multisite pain) | Item 3 OMSQ: M = 4.1 [SD: 2.9] | OMSQ intensity acute:<br>M = 6.3 [SD = 2.0]<br>OMSQ severity chronic:<br>M = 6.0 [SD = 2.9] |
| Nonclercq and Berquin[81] | Belgium;<br>Secondary care<br>Emergency facility and outpatient clinic | M = 42.2 [SD = 10.7] | 56 | Back pain (lumbar pain, cervical pain, and multisite pain) | <3 weeks: 58% | NA |
| Dagfinrud et al.[14] | Norway;<br>Primary care<br>Manual therapists | M = 44.3 [SD = 14.4; range: 18–81] | 59 | Neck pain and low back pain | 0–2 weeks: 23.4%<br>2–12 weeks: 24.1%<br>3–12 months: 13.9%<br>>1 year: 38.6%. | OMPSQ current pain:<br>M = 6.36 [SD = 3.54] |
| Gabel et al.[23] | Australia;<br>Primary care<br>Physiotherapy centers | M = 39.3 [SD = 9.7] | 43 | General musculoskeletal pain (spine, upper and lower limbs) | NA | NA |
| Law et al.[55] | China;<br>Primary care<br>Physiotherapy outpatient clinics | M = 44.2 [SD = 11.2] | 43 | Nonspecific low back pain | M = 3.0 weeks [SD = 1.8]<br>1–2 weeks: N = 114, 47.3%<br>3–5 weeks: N = 100, 41.5%<br>6–10 weeks: N = 24, 9.9% | NPRS pain intensity:<br>M = 5.8 [SD = 2.1] |
| Riewe et al.[92] | Germany;<br>Primary, secondary care<br>Orthopaedic specialists, rehabilitation facilities, and private physiotherapy practices | M = 43 | 65 | Nonspecific back pain | >1 week: 94%<br>>24 weeks: 15% | OMPSQ current pain:<br>M = 5.5 [SD = 2.1]<br>OMPSQ average pain:<br>M = 4.8 [SD = 2.0] |
| *OMPSQs* | | | | | | |
| Linton et al.[63] | Sweden;<br>Primary care | M = 48 | 83 | Nonspecific back or neck pain | NA | NA |
| Karran et al.[50] | Australia;<br>Secondary care<br>Spinal outpatient clinic | M = 49 [SD = 16] | 49 | Low back pain, with or without leg symptoms | <3 months: 20.9%<br>3–6 months: 33.6%<br>>6 months: 44.6% | NRS pain intensity previous week:<br>M = 7.1 [SD = 2.2] |

(*continued on next page*)

Table 2 (*continued*)

**Key study and participant characteristics of included validation studies.**

| Study | Country; setting | Age in years [SD; (IQ-)range] | % female | Pain type | Pain duration | Pain intensity |
|---|---|---|---|---|---|---|
| *OMSQs* | | | | | | |
| Gabel et al.[23] | Australia;<br>Primary care<br>Physiotherapy centers | M = 39.3 [SD = 9.7] | 43 | General musculoskeletal pain (spine, upper and lower limbs) | NA | NA |
| *STarT Back* | | | | | | |
| Hill et al.[41] | United Kingdom;<br>Primary care<br>General practice | M = 45 [SD = 9.7] | 59 | Nonspecific back pain | <1 month: N = 83, 17%<br>1–3 months: N = 94, 19%<br>4–6 months: N = 77, 15%<br>7 months–3 years: N = 125, 25%<br>>3 years: N = 112, 22% | NRS pain intensity mean (least, average, current):<br>Mild (0–5): N = 325, 65%<br>Moderate (6–7): N = 113, 23%<br>Severe (8–10): N = 54, 11% |
| Fritz et al.[22] | United States;<br>Primary care<br>Outpatient physical therapy clinics | M = 44.3 [SD = 15.8] | 57 | Low back pain | *Med* = 46 days [IQ range: 18.5–147] | NRS initial pain intensity:<br>M = 5.3 [SD = 2.3] |
| Field and Newell[20] | United Kingdom;<br>Primary care<br>Chiropractic clinics | Low risk: M = 45.4 [SD = 15.1]<br>Medium risk: M = 45.9 [SD = 15.0]<br>High risk: M = 45.8 [SD = 14.1] | Low risk: 55<br>Medium risk: 53<br>High risk: 51 | Nonspecific low back pain | <1 month: 56.2%<br>1–3 months: 12.4%<br>>3 months: 31.4% | BQ pain:<br>Low risk: *Med* = 5 (range: 4–7)<br>Medium risk: *Med* = 7 (range: 6–8)<br>High risk: *Med* = 7 (range: 6–9) |
| Beneciuk et al.[5] | United States;<br>Primary care<br>Outpatient physical therapy clinics | M = 41.1 [SD = 13.5] | 61 | Low back pain | *Med* = 90.0 days [IQ range: 30–365]<br>≤14 days: 11.8%<br>15–90 days: 39.2%<br>≥90 days: 49.0% | NRS pain intensity mean (current, best, and worst):<br>M = 5.3 [SD = 2.0] |
| Morsø et al.[77]—UK sample | United Kingdom;<br>Primary care<br>General practices | *Med* = 46.0 [IQ range = 39–53] | 59 | Nonspecific low back pain | <4 weeks: N = 327, 38.2%<br>4–12 weeks: N = 221, 25.8%<br>>12 weeks: N = 285, 33.3% | NRS pain intensity:<br>*Med* = 5 [IQ range: 3–7]<br>Mild (0–5): N = 527, 61.6%<br>Moderate (6–7): N = 196, 22.9%<br>Severe (8–10): N = 127, 14.8% |
| Morsø et al.[77]—Danish sample | Denmark;<br>Primary care<br>General practices and physiotherapy clinics | *Med* = 50.0 [IQ range = 41–59] | 58 | Nonspecific low back pain | <4 weeks: N = 149, 44.2%<br>4–12 weeks: N = 66, 19.6%<br>>12 weeks: N = 122, 36.2% | NRS pain intensity:<br>*Med* = 7 [IQ range: 5–8]<br>Mild (0–5): N = 130, 38.7%<br>Moderate (6–7): N = 98, 29.2%<br>Severe (8–10): N = 108, 32.1% |
| Morsø et al.[78]—Primary care sample | Denmark;<br>Primary care<br>General practices and physiotherapy clinics | M = 52.0 [SD = 15.2] | 57 | Low back pain | <1 month: N = 65, 38.9%<br>1–3 months: N = 39, 23.4%<br>>3 months: N = 63, 37.7% | NRS low back pain intensity:<br>*Med* = 6 (IQ range: 4–7)<br>NRS leg pain intensity:<br>*Med* = 3 (IQ range: 0–6) |
| Morsø et al.[78]—Secondary care sample | Denmark;<br>Secondary care<br>Spine center | M = 52.0 [SD = 14.1] | 54 | Low back pain | <1 month: N = 47, 5.0%<br>1–3 months: N = 139, 14.9%<br>>3 months: N = 746, 80.0% | NRS low back pain intensity:<br>*Med* = 5 (IQ range: 4–7)<br>NRS leg pain intensity:<br>*Med* = 5 (IQ range: 2–7) |
| Foster et al.[21] | United Kingdom;<br>Primary care<br>Family practices | M = 53.0 [SD = 15.0] | 55 | Nonspecific low back pain | <1 month: N = 75, 20%<br>1–3 months: N = 62, 17%<br>3–6 months: N = 75, 20%<br>6 months–3 years: N = 82, 22%<br>>3 years: N = 74, 20% | NRS pain intensity:<br>M = 5.3 [SD: 2.4] |
| George and Beneciuk[28] | United States;<br>Primary care<br>Outpatient physical therapy clinics | Med = 45, M = 43.5 [SD = 12.4] | 65 | Low back pain | ≥90 days: N = 53, 47.7% | NRS pain intensity mean (current, best, and worst):<br>*Med* = 5.3, M = 5.4 [SD: 1.9] |

(*continued on next page*)

Table 2 (*continued*)

**Key study and participant characteristics of included validation studies.**

| Study | Country; setting | Age in years [SD; (IQ-)range] | % female | Pain type | Pain duration | Pain intensity |
|---|---|---|---|---|---|---|
| Newell et al.[80] | United Kingdom;<br>Primary care<br>Chiropractic clinics | M = 47.8 [SD = 13.9] | 57 | Nonspecific low back pain | <1 month: 43.2%<br>1–3 months: 10.0%<br>>3 months: 46.6% | BQ pain:<br>M = 6.4 [SD: 2.0] |
| Kongsted et al.[53] | Denmark;<br>Primary care<br>Chiropractic clinics | M = 43 | 44 | Nonspecific low back pain or lumbar nerve root involvement | 0–2 week: 62%<br>2–4 weeks: 13%<br>1–3 months: 11%<br>>3 months: 14% | NRS low back pain intensity:<br>M = 6.5<br>NRS leg pain intensity:<br>M = 2.4 |
| Karran et al.[50] | Australia;<br>Secondary care<br>Spinal outpatient clinic | M = 49 [SD = 16] | 49 | Low back pain, with or without leg symptoms | <3 months: 20.9%<br>3–6 months: 33.6%<br>>6 months: 44.6% | NRS pain intensity previous week:<br>M = 7.1 [SD = 2.2] |
| *HKF-R10* | | | | | | |
| Riewe et al.[92] | Germany;<br>Primary, secondary care<br>Orthopaedic specialists, rehabilitation facilities, and private physiotherapy practices | NA | 67 | Nonspecific back pain | >8 days: 88% | HKF-R10 pain past week:<br>M = 53.14 [SD = 22.13]<br>HKF-R10 pain past week in best stage:<br>M = 27.85 [SD = 21.14] |
| *PBSI* | | | | | | |
| Sandborgh et al.[98] | Sweden;<br>Primary care<br>Physical therapy departments and occupational health care organization | M = 46 [SD = 11; range: 19–64] | 68 | Musculoskeletal pain | *Med* = 12 months (IQ range: 3–59, range 1–300).<br>Subacute: N = 22, 22%<br>Chronic: N = 131, 78% | NA |
| *PICKUP* | | | | | | |
| Traeger et al.[115] | Australia;<br>Primary care<br>General practitioners, pharmacists, and physiotherapists | M = 45 [SD = 15.8] | 46 | Low back pain with or without leg pain | <2 weeks: N = 1183, 78%<br>2–3 weeks: N = 149, 10%<br>3–4 weeks: N = 77, 5%<br>4–6 weeks: N = 116, 8% | Likert Pain intensity:<br>None: N = 0, 0%<br>Very mild: N = 290, 19%<br>Mild: N = 242, 16%<br>Moderate: N = 565, 37%<br>Severe: N = 346, 23%<br>Very severe: N = 70, 5% |
| Karran et al.[50] | Australia;<br>Secondary care<br>Spinal outpatient clinic | M = 49 [SD = 16] | 49 | Low back pain, with or without leg symptoms | <3 months: 20.9%<br>3–6 months: 33.6%<br>>6 months: 44.6% | NRS pain intensity previous week:<br>M = 7.1 [SD = 2.2] |

* Split by the outcome recovery, which is defined as the patient's individual perception of well-being within the current health state.

BQ, Bournemouth Questionnaire; MPQ, McGill Pain Questionnaire; NA, not available; NRS, Numeric Rating Scale.

**Table 3**

**Methodological quality of included validation studies.**

| Study | Participant selection | Predictors | Outcomes | Sample size and participants flow | Analysis |
|---|---|---|---|---|---|
| *ALBPSQ* | | | | | |
| Hurley et al.[43] | High | High | Unclear | Unclear | Moderate |
| Hurley et al.[44] | High | High | Unclear | Unclear | Moderate |
| Grotle et al.[32] | Moderate | Unclear | Moderate | Unclear | Moderate |
| Grotle et al.[33] | Moderate | High | Unclear | Unclear | Moderate |
| Grotle et al.[31] | Moderate | Unclear | Moderate | Unclear | Moderate |
| Heneweer et al.[38] | Unclear | Unclear | — | Unclear | — |
| Vos et al.[124] | Moderate | High | Unclear | Unclear | Moderate |
| *OMPSQ* | | | | | |
| Linton and Boersma[61] | High | Unclear | Moderate | Unclear | Moderate |
| Dunstan et al.[18] | Low | Unclear | — | Unclear | — |
| Margison and French[66]—Derivation sample | Low | Moderate | Moderate | Unclear | Moderate |
| Margison and French[66]—Validation sample | Low | Moderate | Moderate | Unclear | Moderate |
| Maher and Grotle[64]—Australasian sample | Moderate | Moderate | Unclear | Unclear | Moderate |
| Maher and Grotle[64]—Norwegian sample | Low | Moderate | Unclear | Unclear | Moderate |
| Gabel et al.[25]—OMPSQ | Moderate | Moderate | Moderate | Unclear | Moderate |
| Linton et al.[63] | Unclear | Low | Unclear | Unclear | Moderate |
| Nonclercq and Berquin[81] | Moderate | Unclear | Moderate | Unclear | Moderate |
| Dagfinrud et al.[14] | High | Unclear | Unclear | Unclear | Moderate |
| Law et al.[55] | Moderate | Unclear | Moderate | High | Moderate |
| Riewe et al.[92] | High | High | Moderate | Moderate | Moderate |
| *OMSQ* | | | | | |
| Gabel et al.[25] | Moderate | Moderate | Moderate | Unclear | Moderate |
| Gabel et al.[24] | Low | Moderate | Moderate | Unclear | Moderate |
| Gabel et al.[23] | Low | Moderate | Moderate | Unclear | Moderate |
| *OMPSQs* | | | | | |
| Linton et al.[63] | Unclear | Low | Unclear | Unclear | Moderate |
| Karran et al.[50] | High | Unclear | Unclear | Low | Low |
| *OMSQs* | | | | | |
| Gabel et al.[23] | Low | Unclear | Moderate | Unclear | Moderate |
| *HKF-R10* | | | | | |
| Riewe et al.[92] | Moderate | Low | Moderate | Moderate | Moderate |
| *PBSI* | | | | | |
| Sandborgh et al.[98] | High | Low | Unclear | Low | Moderate |
| *SBT* | | | | | |
| Hill et al.[41] | Moderate | Low | Unclear | Unclear | Moderate |
| Fritz et al.[22] | Moderate | Low | — | Unclear | — |
| Field and Newell[20] | Moderate | Low | — | Unclear | — |
| Beneciuk et al.[5] | Moderate | Low | Moderate | Unclear | Moderate |
| Morsø et al.[77]—UK sample | Moderate | Low | Moderate | Unclear | Moderate |
| Morsø et al.[77]—Danish sample | Moderate | Low | Moderate | Unclear | Moderate |
| Morsø et al.[78]—Primary care sample | Moderate | Low | Unclear | Unclear | Moderate |
| Morsø et al.[78]—Secondary care sample | Moderate | Low | Unclear | Unclear | Moderate |
| Foster et al.[21] | Moderate | Low | — | Unclear | — |
| George and Beneciuk[28] | Moderate | Low | — | High | Moderate |
| Newell et al.[80] | Moderate | Low | — | Unclear | — |
| Kongsted et al.[53] | Moderate | Low | Moderate | Low | Moderate |
| Karran et al.[50] | Moderate | Low | Unclear | Low | Low |
| *PICKUP* | | | | | |
| Traeger et al.[115] | Moderate | Unclear | Unclear | Low | Low |
| Karran et al.[50] | High | Unclear | Unclear | Low | Low |

rated as having low risk of bias. For 16.7% of the study samples, the risk of bias was not rated because no performance measures were reported for the outcomes of interest (**Table 3**). The reasons for increasing the risk of bias related to the validity of the outcome, overlap between predictors and outcomes, differences in the assessment of outcomes across participants, differences in the assessment of outcomes compared with the development study, and blinding.

#### *3.4.3.1. Validity of outcome definition*

Outcome measures that mixed outcome domains were rated as inadequate. Also, composite outcomes that combined outcome measures or outcome domains were considered inadequate.[28] For example, the 10-item modified version of the Oswestry Disability Index contains items that assess activity limitations and participation restrictions.[22] Mixed or composite outcomes have the potential to increase the event rate and thus the statistical power. However, they may be misleading when the outcome domains included in the outcome differ in importance to patients, the number of events in the outcome domains of greater importance is small, and the magnitude of effect differs markedly across the outcome domains.[72]

#### *3.4.3.2. Exclusion of predictors from outcome definition*

Next, overlap between predictor and outcome assessment was frequently observed and considered as problematic. Several

**Table 4**

**Key predictor, outcome, sample size and participants flow, and analysis characteristics of included validation studies.**

| Study | N at baseline, (follow-up(s); N at follow-up(s); % at final follow-up) | Outcome (assessment, applied cutoff) | Events (N and/or %) | Recommended criterion | Performance measures |
|---|---|---|---|---|---|
| *ALBPSQ* | | | | | |
| Hurley et al.[43] | 118 (at treatment discharge; 118; 100%) | Pain intensity (MGPQ, NA) | NA | 112 | Kendall's τ |
| | | Functional disability (RMDQ, NA) | NA | | Kendall's τ |
| | | Return to work (yes/no) | 29/15 | | Mann–Whitney U tests, sensitivity, and specificity |
| Hurley et al.[44] | 118 (12 months; 90; 76%) | Pain intensity (MGPQ, NA) | NA | 112 | Kendall's τ |
| | | Functional disability (RMDQ, NA) | NA | | Kendall's τ |
| | | Work loss (yes/no) | 14/55 (20.2%/79.7%) | | Mann–Whitney U tests, sensitivity, and specificity |
| Grotle et al.[32] | 123 (1, 3 months; 120; 98%) | Pain intensity (NRS, NA) | NA | 90 | NA |
| | | Disability (RMDQ, >4 on both 1 and 3 months) | 24% | | ORs |
| | | Sickness absence (NA) | 8% at 1 month<br>6% at 3 months | | NA |
| Grotle et al.[33] | 123 (6 and 12 months; 112; 91%) | Pain intensity (NRS, score >2) | NA | 90 (105 for 12 months RMDQ) | Specificity, sensitivity, LRs (−/+), AUC, and ORs (for all outcomes) |
| | | Disability (RMDQ, >4) | NA | | |
| | | Work loss (disability days, >30 days) | NA | | |
| Grotle et al.[31] | 123 (1, 3, 6, 9, and 12 months; 112; 91%) | Pain intensity (NRS, NA) | NA | 112 | NA |
| | | Disability (RMDQ, >4) | 17% at 12 months | | ORs at 12 months |
| | | Sickness absence (disability days, NA) | 12 (11%) at 1 month<br>10 (9%) at 3 months<br>7 (7%) at 6 months<br>7 (8%) at 9 months<br>9 (9%) at 12 months | | NA |
| Heneweer et al.[38] | 66 (2, 4, 8, and 12 weeks; 56; 95%) | Pain intensity (VAS, NA) | NA | Continuous | NA |
| | | Disability (QBPDS, NA) | NA | | NA |
| | | Work absenteeism (yes/no) | 7/49 (87%/13%) at 12 weeks | | NA |
| Vos et al.[124] | 187 (6, 12, 26, and 52 weeks; 180; 96%) | Pain intensity (NRS, NA) | NA | 72 | NA |
| | | Sick leave (>7 days) | 31 (22%) | | Specificity, sensitivity, PPV, NPV, and AUC |
| *OMPSQ* | | | | | |
| Linton and Boersma[61] | 122 (6 months, 107; 88%) | Pain intensity (OMPSQ items, ≥17) | 48% | 90 | Specificity, sensitivity, and Wilks' λ (for all outcomes) |
| | | Function (OMPSQ items, ≥45) | 60% | | |
| | | Sick leave (>0 days, >30 days) | 60%/23%/17% | | |
| Dunstan et al.[18] | 55 (6 months, 55; 100%) | Return to work (yes/no) | 24/31 | Continuous | NA |
| Margison and French[66]—Derivation sample | 200 (200; 100%) | Clinical discharge status (fit/not fit for return to work) | NA | 147 | Sensitivity and FPR |
| Margison and French[66]—Validation sample | 211 (211; 100%) | Clinical discharge status (fit/not fit for return to work) | 195/16 | 147 | Specificity, sensitivity, PPV, and NPV |
| Maher and Grotle[64]—Australasian sample | 133 (6 weeks, 3, 12 months; 133; 100%) | Pain intensity (OMPSQ item, NA) | NA | Continuous | Regression coefficients |
| | | Disability (RMDQ, NA) | NA | | Regression coefficients |
| Maher and Grotle[64]—Norwegian sample | 97 (4 weeks, 3, 12 months; 97; 100%) | Pain intensity (OMPSQ item, NA) | NA | Continuous | Regression coefficients |
| | | Disability (RMDQ, NA) | NA | | Regression coefficients |
| Gabel et al.[25]—OMPSQ | 66 (6 months; 58; 88%) | Problem severity (NRS, >1) | NA | 113 | Specificity, sensitivity, LRs, and AUC (for all outcomes) |
| | | Functional status (SFI, ≥10%) | NA | 113 | |
| | | Absenteeism (PDO, >0 days) | NA | 115 | |
| | | Long-term absenteeism (PDO, >28 days) | NA | 120 | |
| Gabel et al.[25]—OMSQ | 106 (6 months; 97; 92%) | Problem severity (NRS, >1) | NA | 112 | Specificity, sensitivity, LRs, and AUC (for all outcomes) |
| | | Functional status (SFI, ≥10%) | NA | 112 | |

(*continued on next page*)

Table 4 (*continued*)

**Key predictor, outcome, sample size and participants flow, and analysis characteristics of included validation studies.**

| Study | N at baseline, (follow-up(s); N at follow-up(s); % at final follow-up) | Outcome (assessment, applied cutoff) | Events (N and/or %) | Recommended criterion | Performance measures |
|---|---|---|---|---|---|
| | | Absenteeism (PDO, >0 days) | NA | 116 | |
| | | Long-term absenteeism (PDO, >28 days) | NA | 120 | |
| Linton et al. (2011) | 183 (12 months; 183; 100%) | Sick leave (>14 days of work during past 6 months) | 171 | 90 | Specificity, sensitivity, LRs, and AUC |
| Gabel et al. (2012) | 143 (1 month, 6 months; 43; 100%) | Problem severity (NRS, >10%) | NA | 114 | Specificity, sensitivity, and LRs (+) (for all outcomes) |
| | | Functional status (SFI/LLFI, >10%) | NA | | |
| | | Absenteeism (PDO, >0 days) | NA | | |
| | | Long-term absenteeism (PDO, >28 days) | NA | | |
| Nonclercq & Berquin (2012) | 91 (6 months; 73; 80%) | Pain intensity (OMPSQ items, >16) | 34% | Low/high | Specificity, sensitivity, PPV, and AUC (for all outcomes) |
| | | Function (OMPSQ items, <45; ODI, >20%) | 58%; 18% | 75/97 | |
| | | Work absence (OMPSQ item, >6 [scores corresponding to >30 days]) | 37% | 76/86<br>75/106<br>71/106 | |
| Dagfinrud et al. (2013) | 157 (8 weeks; 128; 82%) | Functional limitations (ODI/NDI, NA) | NA | Continuous/105 | Regression coefficients |
| Gabel et al. (2013) | 143 (6 months; 143; 100%) | Problem severity (NRS, >10%) | NA | 126 | Specificity, sensitivity, LRs (+), and *t* tests |
| | | Functional status (PRO, >10%) | NA | | |
| | | Absenteeism (PDO, >0 days) | NA | | |
| | | Long-term absenteeism (PDO, >28 days) | NA | | |
| Law et al. (2013) | 241 (3–4 weeks, 12 months; per outcome: 184, 160, 220, 202; 76%, 66%, 91%, 84%) | Pain intensity (NRS, NA) | NA | 105, 130 | NA |
| | | Functional disability (RMDQ, NA) | NA | | NA |
| | | Return to work (yes/no) | 171/49 at 12 months | | Specificity, sensitivity, AUC, and ORs (both outcomes) |
| | | Sick leave (>30 days) | 88 at 12 months | | |
| Riewe et al. (2016) | 241 (6 months; per outcome: 122, 122, 108; 51%, 51%; 45%) | Pain intensity (OMPSQ items, ≥17) | 61 | 84 | Specificity, sensitivity, PPV, NPV, LRs (+/−), and AUC (for all outcomes) |
| | | Function (OMPSQ items, <45) | 64 | | |
| | | Sick leave (>0 days) | 40 | | |
| *OMPSQs* | | | | | |
| Linton et al.[63] | 183 (12 months; 183; 100%) | Sick leave (>14 days of work during past 6 months) | 171 | 50 | Specificity, sensitivity, LRs, and AUC |
| Karran et al.[50] | 220 (4 months; 195; 89%) | Poor outcome (composite pain/disability NRS, ≥3) | 164 (84%) | Lowest 10th through highest 10th decile of risk | Nagelkerke $R^2$, AUC, calibration plot (for poor outcome), net benefit, post hoc sensitivity analysis (for poor outcome and high pain), and AUC (for all outcomes) |
| | | Pain intensity (NRS, ≥3) | 155 (79%) | | |
| | | Disability (NRS, ≥3) | 159 (82%) | | |
| | | High pain (NRS, ≥5) | 129 (66%) | | |
| | | High disability (NRS, ≥5) | 126 (65%) | | |
| *OMSQs* | | | | | |
| Gabel et al.[23] | 143 (6 months; 143; 100%) | Problem severity (NRS, >10%) | NA | 72 | Specificity, sensitivity, LRs (+), and *t* tests |
| | | Functional status (PRO, >10%) | NA | | |
| | | Absenteeism (PDO, >0 days) | NA | | |
| | | Long-term absenteeism (PDO, >28 days) | NA | | |
| *STarT Back* | | | | | |
| Hill et al.[41] | 500 (6 months; 500, 100%) | Disability (RMDQ, ≥7) | Low risk: 39 (16.7%)<br>Medium risk: 99 (53.2%)<br>High risk: 58 (78.4%) | Low, medium, and high risk groups | Sensitivity, specificity, LRs (+/−), and AUC |
| Fritz et al.[22] | 214 (at each visit; 177, 83%) | Pain intensity (NRS, NA) | NA | Low, medium, and high risk groups | NA |
| | | Disability (DISQ, NA) | | | NA |
| Field and Newell[20] | 404 (14, 30, 90 days; per follow-up per outcome: 218/204, 123/119, 142/136; 54%/50%, 30%/29%, 35%/34%) | Pain (BQ, NA) | NA | Low, medium, and high risk groups | NA |
| | | Total (BQ, NA) | NA | | NA |

(*continued on next page*)

Table 4 (*continued*)

**Key predictor, outcome, sample size and participants flow, and analysis characteristics of included validation studies.**

| Study | N at baseline, (follow-up(s); N at follow-up(s); % at final follow-up) | Outcome (assessment, applied cutoff) | Events (N and/or %) | Recommended criterion | Performance measures |
|---|---|---|---|---|---|
| Beneciuk et al.[5] | 146 (4 weeks, 6 months; 128, 111; 88%, 76%) | Pain intensity (NRS, NA)<br>Disability (RODQ, NA) | NA<br>NA | Continuous | Regression coefficients<br>Regression coefficients |
| Morsø et al.[77]—UK sample | 856 (3 months; 845, 99%) | Pain intensity (NRS, ≥8)<br>Activity limitations (RMDQ, >30)<br>Pain bothersomeness (1 item, severe or very severe) | NA<br>36%<br>NA | Low, medium, and high risk groups | AUC<br>RR, ORs, and AUC<br>AUC |
| Morsø et al.[77]—Danish sample | 344 (3 months, 322, 94%) | Pain intensity (NRS, ≥8)<br>Activity limitations (RMDQ, >30)<br>Pain bothersomeness (1 item, severe or very severe) | NA<br>47%<br>NA | Low, medium, and high risk groups | AUC<br>RR, ORs, and AUC<br>AUC |
| Morsø et al.[78]—Primary care sample | 172 (6 months; 144, 83%) | Pain intensity (NRS, ≥8)<br>Activity limitations (RMDQ, >30) | NA<br>40.2% | Low, medium, and high risk groups | AUC<br>RR, ORs, and AUC |
| Morsø et al.[78]—Secondary care sample | 960 (6 months; 960, 100%) | Pain intensity (NRS, ≥8)<br>Activity limitations (RMDQ, >30) | NA<br>69.0%<br>NA | Low, medium, and high risk groups | AUC<br>RR, ORs, and AUC |
| Foster et al.[21] | 368 (2, 6 months; 254 (69%), 233 (63%) | Pain intensity (NRS, NA)<br>Disability (RMDQ, NA) | NA | Low, medium, and high risk groups | NA<br>NA |
| George and Beneciuk[28] | 146 (6 months; 111, 76%) | Pain intensity (NRS = 0)<br>Disability (RMDQ, ≤ 2)<br>Recovery (NRS = 0 and RMDQ ≤2) | 14 (12.6%)<br>36 (32.4%)<br>14 (12.6%) | Low, medium, and high risk groups | Wilks' λ<br>Wilks' λ<br>Wilks' λ |
| Newell et al.[80] | Initial treatment/2-days post-initial treatment: 749/716 (14, 30, 90 days; per follow-up: 542, 416, 318; 58% | Pain (BQ, NA)<br>Total (BQ, NA) | NA | Low, medium, and high risk groups | NA |
| Kongsted et al.[53] | 859 (2 weeks, 3, 12 months; per follow-up: 710, 676, 636; 83%, 79%, 74%) | Pain intensity (NRS, >0)<br><br><br>Disability (RMDQ, >8) | 92% at 2 weeks<br>60% at 3 months<br>56% at 12 months<br>79% at 2 weeks<br>61% at 3 months<br>57% at 12 months | Low, medium, and high risk groups | LR (+/−), AUC, and $R^2$ |
| Karran et al.[50] | 220 (4 months; 195; 89%) | Poor outcome (composite pain/disability NRS, ≥3)<br>Pain intensity (NRS, ≥3)<br>Disability (NRS, ≥3)<br>High pain intensity (NRS, ≥5)<br>High disability (NRS, ≥5) | 164 (84%)<br><br>155 (79%)<br>159 (82%)<br>129 (66%)<br>126 (65%) | Low, medium, and high risk groups | Nagelkerke $R^2$, AUC, calibration plot (for poor outcome), net benefit, post hoc sensitivity analysis (for poor outcome and high pain), and AUC (for all outcomes) |
| *HKF-R10* | | | | | |
| Riewe et al.[92] | 242 (6 months; 128; 58%) | Pain intensity (HKF-R10 items, ≥30) | 90 | 37 | Specificity, sensitivity, PPV, NPV, LRs (+/−), and AUC |
| *PBSI* | | | | | |
| Sandborgh et al.[98] | 168 (8 months; 146, 85%) | High pain intensity (NRS, ≥5)<br>High disability (PDI, ≥35) | NA<br>33 | Continuous | NA<br>Specificity, sensitivity, and Wilks' λ |

(*continued on next page*)

**Table 4** *(continued)*

**Key predictor, outcome, sample size and participants flow, and analysis characteristics of included validation studies.**

| Study | N at baseline, (follow-up(s); N at follow-up(s); % at final follow-up) | Outcome (assessment, applied cutoff) | Events (N and/or %) | Recommended criterion | Performance measures |
|---|---|---|---|---|---|
| *PICKUP* | | | | | |
| Traeger et al.[115] | 1528 (3 months; per outcome: 1528, 1525, 1504; 100%, 99%, 98%) | Pain intensity (Likert, >2)<br>High pain intensity (Likert, >3)<br>Disability (Likert, ≥2) | 291 (19%)<br>162 (10%)<br>217 (14%) | Calculator | Nagelkerke $R^2$, AUC, calibration plot (intercept/slope), net benefit at incidence rate cutoff, and net number of unnecessary interventions avoided at 30% risk cutoff (for all outcomes) |
| Karran et al.[50] | 220 (4 months; 195; 89%) | Poor outcome (composite pain/disability NRS, ≥3)<br>Pain intensity (NRS, ≥3)<br>Disability (NRS, ≥3)<br>High pain intensity (NRS, ≥5)<br>High disability (NRS, ≥5) | 164 (84%)<br>155 (79%)<br>159 (82%)<br>129 (66%)<br>126 (65%) | Lowest 10th through highest 10th decile of risk | Nagelkerke $R^2$, AUC, calibration plot (for poor outcome), net benefit, post hoc sensitivity analysis (for poor outcome and high pain), and AUC (for all outcomes) |

AUC, area under the curve; BQ, Bournemouth Questionnaire; DISQ, 10-item modified version of the Oswestry Low Back Pain Disability Questionnaire; FNR, false negative rate; FPR, false positive rate; LLFI, Lower Limb Functional Index; LRs, likelihood ratios; MCID, minimal clinically important difference; MGPQ, McGill Pain Questionnaire; NA, not available; NDI, Neck Disability Index; NPV, negative predicted value; NRS, Numeric Rating Scale; ODI, Oswestry Disability Index; ORs, odds ratios; PDI, Pain Disability Index; PDO, paid days off; PPV, positive predicted value; PRO, patient-reported outcome; QBPDS, Quebec Back Pain Disability Scale; RMDQ, Roland–Morris Disability Questionnaire; RODQ; Revised Oswestry Disability Questionnaire; RR, relative risk; SFI, Spine Functional Index; VAS, Visual Analogue Scale.

studies used items of the investigated screening tool, measured at follow-up, as primary outcome. For instance, Linton and Boersma[61] used the OMPSQ in its entirety during the outcome assessment, selecting the items on pain, activity limitations, and sick leave. Studies also often included outcomes that showed overlap with domains assessed by the screening tool items. In the study by Grotle et al.,[32] both the activity items of the ALBPSQ and the items of the Roland–Morris Disability Questionnaire (RMDQ) outcome measure address activity limitations. This overlap may lead to overestimation of the predictive performance of the screening tool.[91,117]

#### *3.4.3.3. Definition and assessment of outcomes across participants*

For all studies, outcomes were defined and determined in a similar way across participants. However, they were not always defined and determined similarly to those in the development studies. Indeed, although different outcomes most probably have different predictors, a number of studies targeted outcome domains (eg, pain intensity through OMPSQ items and activity limitations through the RMDQ and not participation restrictions through accumulated sick leave)[64] which differed from the development study. Other studies focused on similar outcome domains, but used other measures (eg, activity limitations through a NRS and not the RMDQ due to the large amount of missing data).[50]

#### *3.4.3.4. Definition and assessment of outcomes compared with the developmental model*

In addition, some studies focused on similar outcome domains and used the same outcome measures as the development study, but used different cutoff points for the outcome measures from those used in the development study. For example, large differences were observed for sick leave. Vos et al.[124] defined long-term sick leave as >7 days off work, while Linton and Hallden[62] initially defined long-term sick leave as being sick listed for >30 days (**Table 4**).

#### *3.4.3.5. Determination of outcomes without knowledge of predictor information*

Information on blinding was most often not reported, which could either mean that the outcome assessment was not blinded or that it was blinded but not described. In cases where studies reported on blinding of outcome assessment, researchers usually applied blinding.[24]

### *3.4.4. Sample size and participant flow*

There was a huge difference between sample sizes of the validation studies. Sample sizes varied considerably at follow-up, ranging from <100 participants,[18,25,38,43,44,64,81] over 500 to 1000 participants,[41,53,77,78,80] to >1500 participants.[115] Also, the number of outcome events differed largely between studies ranging from 14 to 291. The most frequently observed time intervals were 3, 6, and 12 months[92] (see **Table 4** for an overview).

Few studies were rated as having low (16.7%), moderate (2.4%), or high (4.8%) risk of bias for the PROBAST "sample size and participants flow" domain. The majority of studies were assigned an unclear risk of bias (76.2%; **Table 3**) because insufficient information was presented to evaluate the number of

outcome events, the inclusion of enrolled participants, or the occurrence and handling of missing data. The reasons for increasing the risk of bias related to the number of outcome events, the time interval between the assessment of the screening tools and the outcome assessment, dropout, and missing data.

#### 3.4.4.1. Number of outcome events

The number of events (ie, the number of individuals with the outcome event) was not reported in a large number of studies[5,14,21–25,33,38,64,66,80] and considered inappropriate in 5 studies.[28,31,44,63,81] These studies reported <20 events, raising the issue of overfitting (ie, the probability of an event is typically underestimated in low-risk patients and overestimated in high-risk patients).[4,85]

#### 3.4.4.2. Time interval between predictor assessment and outcome determination

Studies sometimes performed multiple follow-ups, reporting results on the predictive validity for one or only a selection of follow-ups (eg, follow-ups at 2- and 4-week intervals until discharge or study completion at 6 months, report of results for 6-month follow-up).[24] Time between screening and outcome assessment was considered inappropriate when results only reported on follow-ups of <3 months, as chronic pain is defined as pain ≥3 months (eg, six weeks).[66] Follow-ups >12 months were also considered inappropriate, as people's (mental) health status changes during the follow-up period and the baseline information becomes increasingly less accurate as time passes (none of the studies). In addition, follow-ups that varied across participants (eg, at treatment discharge, dependent on the number of therapy treatments)[43] were deemed inappropriate. Surprisingly, most studies did not present any theoretical considerations underpinning the choice of a specific follow-up timeframe (**Table 4**).

#### 3.4.4.3. Inclusion of enrolled participants in analysis

Dropout attrition is often poorly reported or presented in a way that prevents readers from being able to fully understand the risk of attrition bias. Studies often limit themselves to reporting the dropout rate. We considered dropout as inappropriate when >20%[96] of the participants were lost at follow-up.[5,20,21,28,44,53,80,92] However, dropout can occur for a number of reasons that may lead to differential dropout, such as motivation (participants lost interest), mobility (participants moved and are no longer able to continue participation), morbidity (participants experience illness preventing their participation), or mortality (participants die before study completion). For example, a low psychosocial risk group may lose more unmotivated participants—that in turn may have different outcomes due to being unmotivated—than a high psychosocial risk assessment group, and this differential dropout may lead to differences in outcomes measured among the remaining participants. Reasons for dropout are, however, rarely specified among the included studies. Furthermore, although characteristics of dropout (ie, baseline characteristics: eg, age, sex, pain intensity, and pain duration) should be available to examine whether systematic differences exist between those who completed a study and those who dropped out,[36] only few studies reported on the differences between completers and noncompleters.[5,28,44,50,53,80,81,98] Of these studies, some provided a detailed tabulation of the characteristics and statistical comparison,[50] whereas other studies only reported the characteristics for which differences were found.[5] Further, numerous studies do not mention whether differences were examined, which could either mean that differences were examined for all or some baseline characteristics but none were found, or no differences were tested.[55]

#### 3.4.4.4. Handling of missing data

Finally, studies did often not report on missing values or how they were or would have been handled,[78] which could either mean that there were no missing data or that missing data were present but not described. Missing values were considered inappropriately handled when complete-case analysis was applied.[92] They were judged as appropriately handled when multiple imputation was used.[74] For example, Karran et al.[50] used Little's Missing Completely at Random test to determine whether values were missing completely at random and used a maximization algorithm to impute missing values.

### 3.4.5. Analyses

Statistics of reported performance measures for pain and related outcomes varied widely. Many studies report sensitivity and specificity of screening tools,[61] whereas others included further details, reporting area under the curve using receiver operating characteristics analyses.[53] Wilk's lambda for discriminative validity is also reported in some studies,[28,98] as are the odds ratios from logistic regression analyses[33] (see **Table 4** for an overview).

For the PROBAST "analyses" domain, the majority of study samples were assigned a moderate risk of bias (76%), and only a few study samples were rated as low risk of bias (9.5%). For 14.3% of the study samples, no risk of bias labels was assigned because no performance measures were reported for the outcomes of interest. The reason for increasing the risk of bias related to the poor use of the performance measures.

#### 3.4.5.1. Evaluation of relevant model performance measures

Statistical analyses were found appropriate when they reflected both calibration (ie, agreement between predicted and observed event rates) and discrimination (ie, the screening tool's ability to distinguish between patients developing and not developing the outcome of interest) components of predictive validity for pain and related outcomes.[74] This was only the case in 2 studies.[50,115] These studies also reported more recently introduced performance measures (eg, net benefit). Moreover, not all studies reported performance measures for pain and related outcomes despite assessing those outcomes. Some studies reported on the course of particular pain and related outcomes. For example, Grotle et al.[31] reported the course of pain intensity, disability, and sickness absence from baseline across follow-ups, but reported no information on the predictive validity of the ALBPSQ for those outcomes, except for disability where odds ratios were provided. Other studies reported differences in mean scores on the screening tool for particular outcomes, used change scores for particular outcomes, or reported on composite outcomes. For example, Dunstan et al.[18] reported differences in mean ALBPSQ scores between those who did and did not return to work. Dagfinrud et al.[14] assessed functional limitations at baseline and

follow-up; however, the predictive validity of the OMPSQ was examined for functional improvement, and the categorization of those that were improved and those that were not was based on change scores. Finally, George and Beneciuk[28] assessed pain intensity and disability; yet, discriminative validity was only examined for recovery, a composite pain intensity and disability outcome. Still others assessed pain and related outcomes, but only reported performance measures related to outcomes that were not within the scope of the current review. For instance, Heneweer et al.[38] assessed pain intensity, disability, work absenteeism, and self-reported recovery, but only reported area under the curve values for the ALBPSQ total and subscale scores in predicting recovery or nonrecovery at final follow-up (**Table 4**).

## 4. General discussion

This review (1) identified multidimensional screening tools that assess psychosocial risk factors for poor pain outcomes, (2) appraised the quality of the evidence in prospective studies validating these tools, and (3) synthesized common methodological concerns in these validation studies.

Seven screening tools were identified, all developed for use in primary care settings to predict chronic pain (HKF-R10, PICKUP) or chronic disability (ALBPSQ/OMPSQ, OMPSQs, OMSQs, PBSI, and STarT Back) in patients with back pain. Notably, we found no tools for the prediction of pain-related distress, a key indicator of health, or for the prediction of acute pain onset, including postoperative pain. These appear to be significant gaps in the literature.[101]

We assessed the quality of the evidence of 32 studies including 42 study samples aiming to validate the predictive value of identified screening tools. Overall, studies showed a moderate risk of bias, which varied largely from domain to domain. Here, we discuss the most notable methodological problems.

Most screening tools were developed to predict the chronification of pain problems, except for the SBT and the PBSI, which were developed to support decision-making for a wide range of patients with pain conditions, regardless of pain duration.[41,97] It is reasonable to expect that validation studies include similar patient populations as those from the development study. Surprisingly, this was often not the case. Indeed, although most tools were developed to be used in patients with acute pain, a substantial number of these validation study samples included also patients with chronic pain. This is concerning for several reasons. First, these studies do not address the same key question as the development study. It may also well be that risk factors developing chronic pain are different from predictors for the maintenance of chronic pain. Second, it is likely that the recovery rate of chronic pain is less than the one of acute pain.[39] Therefore, the presence of chronic patients with chronic pain may (at least partly) account for the apparently high performance in predicting poor pain outcomes. This complicates interpretation of results and may result in an underestimation or overestimation of the predictive value of the screening tools. There is a need to define the inclusion criteria for participants in a more clear and restrictive way and to align these with the original purpose of the screening tools.

The success of initial studies revealing the value of psychosocial risk factors in predicting chronic pain problems has boosted research in this area. However, some of the original studies were designed with specific (clinical) groups in mind. An example is the ALBPSQ, which was designed to target a working population. Some items that are directly related to work (eg, "If you take into consideration your work routines, management, salary, promotion possibilities, and workmates, how satisfied are you with your job?") are therefore inapplicable to a nonworking population. The authors have addressed this problem in various ways. Some replaced the missing scores for those items by the mean for nonworking patients.[33] Others asked patients to fill out those questions related to either current paid or unpaid work.[43,44] Likewise, screening tools were developed for patients with musculoskeletal, in particular back pain, but studies have also investigated the value of the tools in other patient groups (eg, neck pain).[124] Sometimes, items have been adapted accordingly and/or left out. There is a lack of evidence, however, to suggest that these changes are appropriate for the populations in question.

All studies agree that screening tools need to predict poor pain outcome. However, there is less agreement about what exactly poor outcome means. Indeed, a gold standard for poor outcome is lacking. The constructs addressed and the measures and cutoffs used vary largely between studies. For some, poor outcome simply means pain, for others not being able to work, or difficulties in performing physical activities. However, different outcomes most probably also have different predictors. The broad use of the umbrella term "disability" brings additional complications. Indeed, in pain research, "disability" may indicate difficulties in performing particular physical activities (eg, ability to walk, eat, shower, or dress) but also problems related to social role functioning (eg, sick leave, days absent from work, or return to work status). According to the International Classification of Functioning (ICF),[130] these are 2 different constructs, ie, activity limitations and participation restrictions, which should not be confused. The lack of a gold standard may also explain the inconsistency in criteria used across studies. For instance, Morsø et al. defined poor pain outcome as a score greater than 7 on an 11-point NRS,[77,78] whereas George and Beneciuk[28] defined it as a score greater than 0. It is obvious that the patients defined as recovered differ between these studies. The use of an agreed-upon set of outcome measures may provide a solution.[10,11,89] In doing so, we also recommend the selection of measures that are readily applicable to different contexts—occupational and non-occupational settings—and to different pain problems. Such measures already exist, but are underused (eg, Patient-Reported Outcomes Measurement Information System, PROMIS,[9,118,119] available at www.healthmeasures.net).

Some of the identified screening tools were developed to screen for psychosocial risk factors ("yellow flags"), or, at least, are presented as such in studies. Some cautionary notes are warranted. First, all screening tools also include items that could be categorized otherwise (eg, pain duration and disability compensation). Second, screening tools often contain items that could equally well be the primary outcomes (pain intensity, disability, and days off work). Although this may be less of a problem when simply aiming to predict, it is premature to explain the predictive power of these instruments in terms of psychosocial processes. Indeed, given that it is generally known that the best predictor of events in the future is their occurrence in the present or past, it remains to be investigated whether the predictive validity of screening tools is due to the overlap between predictor and outcome.[91,117] To address this problem, one may examine whether tools are able to predict outcomes, beyond the predictive power of baseline pain and pain-related disability.

Most studies are not in line with the current guidelines for reporting measures of performance.[110,111] In fact, there is a large disparity in reported performance measures. Many studies reported conventional performance measures, often reporting either calibration (ie, how close predictions are to observed

outcomes) or discrimination (ie, screening tool's ability to correctly distinguish the 2 outcome classifications of event vs nonevent). However, the reporting of both performance measures is crucial. Furthermore, most studies do not consider the clinical consequences of decisions made using a screening tool. Therefore, there is the implicit assumption that false-positive (ie, patient being treated unnecessarily) and false-negative (ie, patient not getting a treatment that (s)he would benefit from) predictions are equally harmful (ie, equally weighted). More recent studies[50,115] do consider the relative harms or benefits of these alternative clinical outcomes. They apply novel performance measures such as net benefit (ie, the expected utility of a decision to treat patients at some threshold, compared with a decision based on an alternative policy such as treating nobody)[75,110,111,120,121] (see also www.decisioncurveanalysis.org).

An assessment of the risk of bias was not possible in a considerable number of studies because of incomplete reporting. A balanced evaluation of the risk of bias of studies may be impeded due to nontransparent reporting. An increased quality of reporting was observed over time, but there is still room for improvement and there is a need for guidance. The "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis" (TRIPOD) statement is particularly helpful and provides guidance for the reporting of studies that develop, validate, or update prediction models[12,73] (available atwww.tripod-statement.org). We encourage researchers to follow its recommendations. Equally important are the availability of study protocols and the availability of data sets. Protocol registration, either through publications, or through open science applications, may reduce the impact of publication bias.[84] A large number of validation studies in our review reported significant results; yet, only 2 studies mentioned a protocol.[50,115] Protocol registration may also reduce reporting bias.[40] It is common practice to measure several outcomes, but the lack of a readily accessible research protocol makes these studies vulnerable to selective reporting of analyses that "worked."[35] Another possibility is to make data sets open, ie, available to all researchers.[115] Available data sets provide the opportunity to conduct secondary analyses that may be informed by advances in theory and scientific standards in the field.

There are some limitations to our review. First, we used a strict search strategy. We excluded batteries of questionnaires and tools that were not originally developed in the context of pain. This may have resulted in missing instruments that are potentially valuable. For example, the Amsterdam Preoperative Anxiety and Information Scale (APAIS) was originally developed to evaluate patient's preoperative anxiety and need for preoperative information regarding the scheduled surgery and anesthesia.[70] Subsequently, this tool was used to predict postoperative pain.[46,48] Second, we focused upon multidimensional screening tools. Otherwise, one may make use of unidimensional questionnaires assessing single psychosocial risk factors to investigate the predictive power of unique psychosocial variables (eg, Pain Catastrophizing Scale[113] and Tampa Scale for Kinesiophobia[69]) for poor pain outcomes. For screening purposes, however, one should aim to minimize the burden of filling out questionnaires for participants. The use of large questionnaire batteries should therefore be avoided. Third, this research field is quickly evolving, with new validation studies appearing at a fast pace. Since our search, new instruments have been validated in an independent study. For instance, the Optimal Screening for Prediction of Referral and Outcome cohort yellow flag assessment tool was developed in a cross-sectional cohort in 2016.[58] Recently, a validation study was published.[29] Fourth, clinical prediction modelling is a dynamic and evolving field[15,47,56,94,108–111] (see also progress-partnership.org). One should keep in mind that the present review is an exploratory mapping of this rapidly evolving field. Assessment of the quality evidence in the included studies was based upon a prepublication version of the PROBAST. This version did not yet provide a guideline for scoring the questions. We constructed, therefore, our own coding system. Now, PROBAST has been published, with some minor changes from the prepublication version of the PROBAST (eg, the signaling questions of the domain "Sample size and participants flow" are now included in the domain Outcomes and the domain Analysis).[76,128] Despite this minor changes, the resulting mapping fulfills the primary goal of providing an entry point to reduce risk of bias in this field. Fifth, we did not perform a meta-analysis. Several meta-analyses are available that synthesize the predictive value of screening tools. They indicate that (1) the predictive value of these screening is highly variable depending on the pain outcome of interest (eg, pain and disability) and (2) substantial heterogeneity between studies exist.[49,99] Taking into account methodological differences and quality criteria is therefore crucial to further our understanding of the predictive value of screening tools. Our insights have the potential to improve research in this area and decision-making based on this research.

## Disclosures

The authors have no conflict of interest to declare.

Preparation of this article was supported by funding from the European Union's Horizon 2020 research and innovation program (Grant 633491).

**Article history:**
Received 11 March 2019
Received in revised form 11 June 2019
Accepted 26 June 2019

## References

[1] Abegglen S, Hoffmann-Richter U, Schade V, Znoj HJ. Work and Health Questionnaire (WHQ): a screening tool for identifying injured workers at risk for a complicated rehabilitation. J Occup Rehabil 2016;27:268–83.
[2] Althof JE, Beasley BD. Psychosocial management of the foot and ankle surgery patient. Clin Podiatr Med Surg 2003;20:199–211.
[3] Andersen JH, Haah JP, Frost P. Risk factors for more severe regional musculoskeletal symptoms: a two-year prospective study of a general working population. Arthritis Rheum 2007;56:1355–64.
[4] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 2017;26:796–808.
[5] Beneciuk JM, Bishop MD, Fritz JM, Robinson ME, Asal NR, Nisenzon AN, George SZ. The STarT back screening tool and individual psychological measures: evaluation of prognostic capabilities for low back pain clinical outcomes in outpatient physical therapy settings. Phys Ther 2013;93:321–33.
[6] Blyth MF, March ML, Nicholas KM, Cousins JM. Chronic pain, work performance and litigation. PAIN 2003;103:41–7.
[7] 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.
[8] Broadbent E, Wilkes C, Koschwanez H, Weinman J, Norton S, Petrie KJ. A systematic review and meta-analysis of the Brief Illness Perception Questionnaire. Psychol Health 2015;30:1361–85.
[9] Cella D, Riley W, Stone A. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol 2010;63:1179–94.

[10] Chiarotto A, Boers M, Deyo RA, Buchbinder R, Corbin TP, Costa L, Foster NE, Grotle M, Koes BW, Kovacs FM, Lin CC, Maher CG, Pearson AM, Peul WC, Schoene ML, Turk DC, van Tulder MW, Terwee CB, Ostelo RW. Core outcome measurement instruments for clinical trials in nonspecific low back pain. PAIN 2018;159: 481–95.
[11] Chiarotto A, Ostelo RW, Turk DC, Buchbinder R, Boers M. Core outcome sets for research and clinical practice. Braz J Phys Ther 2017; 21:77–84.
[12] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Inter Med 2015;162:55–63.
[13] Crombez G, Eccleston C, Van Damme S, Vlaeyen JWS, Karoly P. Fear-avoidance model of chronic pain: the next generation. Clin J Pain 2012; 28:475–83.
[14] Dagfinrud H, Storheim K, Magnussen LH, Odegaard T, Hoftaniska I, Larsen LG, Ringstad PO, Hatlebrekke F, Grotle M. The predictive validity of the Örebro Musculoskeletal Pain Questionnaire and the clinicians' prognostic assessment following manual therapy treatment of patients with LBP and neck pain. Man Ther 2013;18:124–9.
[15] Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015; 68:279–89.
[16] Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, Riley RD, Moons KGM. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460.
[17] den Boer JJ, Oostendorp RA, Evers AW, Beems T, Borm GF, Munneke M. The development of a screening instrument to select patients at risk of residual complaints after lumbar disc surgery. Eur J Phys Rehabil Med 2010;46:497–503.
[18] Dunstan DA, Covic T, Tyson GA, Lennie IG. Does the Orebro Musculoskeletal Pain Questionnaire predict outcomes following a work-related compensable injury? Int J Rehabil Res 2005;28:369–70.
[19] Eccleston C, Crombez G. Advancing psychological therapies for chronic pain. F1000Res 2017;6:461.
[20] Field J, Newell D. Relationship between STarT Back Screening Tool and prognosis for low back pain patients receiving spinal manipulative therapy. Chiropr Man Therap 2012;20:17.
[21] Foster NE, Mullis R, Hill JC, Lewis M, Whitehurst DG, Doyle C, Konstantinou K, Main C, Somerville S, Sowden G, Wathall S, Young J, Hay EM; IMPaCT Back Study team. Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. Ann Fam Med 2014;12:102–11.
[22] Fritz JM, Benecuik JM, George SZ. Relationship between categorization with the STarT back screening tool and prognosis for people receiving physical therapy for low back pain. Phys Ther 2011;91:722–32.
[23] Gabel CP, Burkett B, Melloh M. The shortened Örebro Musculoskeletal Screening Questionnaire: evaluation in a work-injured population. Man Ther 2013;18:378–85.
[24] Gabel CP, Melloh M, Burkett B, Osborne J, Yelland M. The Örebro Musculoskeletal Screening Questionnaire: validation of a modified primary care musculoskeletal screening tool in an acute work injured population. Man Ther 2012;17:554–65.
[25] Gabel CP, Melloh M, Yelland M, Burkett B, Roiko A. Predictive ability of a modified Örebro Musculoskeletal Pain Questionnaire in an acute/subacute low back pain working population. Eur Spine J 2011;20: 449–57.
[26] Gatchel RJ, Polatin PB, Kinney RK. Predicting outcome of chronic back pain using clinical predictors of psychopathology: a prospective analysis. Health Psychol 1995;14:415–20.
[27] Gatchel J, Polatin P, Mayer T. The dominant role of psychosocial risk factors in the development of chronic low back pain disability. Spine 1996;20:2702–9.
[28] George SZ, Beneciuk JM. Psychological predictors of recovery from low back pain: a prospective study. BMC Musculoskelet Disord 2015;16: 49.
[29] George SZ, Beneciuk JM, Lentz TA, Wu SS, Dai Y, Bialosky JE, Zeppieri G Jr. Optimal screening for prediction of referral and outcome (OSPRO) for musculoskeletal pain conditions: results from the validation cohort. J Orthop Sports Phys Ther 2018;48:460–75.
[30] Gray H, Adefolarin AT, Howe TE. A systematic review of instruments for the assessment of work-related psychosocial factors (Blue Flags) in individuals with non-specific low back pain. Man Ther 2011;16:531–43.
[31] Grotle M, Brox JI, Glomsrod B, Lonn JH, Vollestad NK. Prognostic factors in first-time care seekers due to acute low back pain. Eur J Pain 2007;11:290–8.
[32] Grotle M, Brox JI, Veierød MB, Glomsrød B, Lønn JH, Vøllestad NK. Clinical course and prognostic factors in acute low back pain. Patients consulting primary care for the first time. Spine 2005;30:976–82.
[33] Grotle M, Vollestad NK, Brox JI. Screening for yellow flags in first-time acute low back pain: reliability and validity of a Norwegian version of the acute low back pain screening Questionnaire. Clin J Pain 2006;2: 458–67.
[34] Gureje O, Von Korff M, Simon GE, Gater R. Persistent pain and well-being: a World Health Organization study in primary care. JAMA 1998; 280:147–51.
[35] Hahn S, Williamson PR, Hutton JL. Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to a local research ethics committee. J Eval Clin Pract 2002;8:353–9.
[36] Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 2006;44:427–37.
[37] Hayden JA, Van Der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013; 158:280–6.
[38] Heneweer H, Aufdemkampe G, van Tulder MW, Kiers H, Stappaerts KH, Vanhees L. Psychosocial variables in patients with (sub)acute low back pain: an inception cohort in primary care physical therapy in the Netherlands. Spine 2007;32:586–92.
[39] Henschke N, Maher CG, Refshauge KM, Herbert RD, Cumming RG, Bleasel J, York J, Das A, McAuely JH. Prognosis in patients with recent onset low back pain in Australian primary care: inception cohort study. BMJ 2008;337:1–7.
[40] Higgins JPT, Altman DG, Sterne JAC, editors. Chapter 8: assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available at: www.cochrane-handbook.org.Accessed January 26, 2017.
[41] Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, Hay EM. A primary care back pain screening tool: identifying patient subgroups for initial treatment. Arthritis Rheum 2008;59:632–41.
[42] Hockings RL, McAuley JH, Maher CG. A systematic review of the predictive ability of the Orebro Musculoskeletal Pain Questionnaire. Spine 2008;33:E494–500.
[43] Hurley DA, Dusoir TE, McDonough SM, Moore AP, Linton SJ, Baxter GD. Biopsychosocial screening questionnaire for patients with low back pain: preliminary report of utility in physiotherapy practice in Northern Ireland. Clin J Pain 2000;16:214–28.
[44] Hurley DA, Dusoir TE, McDonough SM, Moore AP, Baxter GD. How effective is the acute low back pain screening questionnaire for predicting 1-year follow-up in patients with low back pain? Clin J Pain 2001;17:256–63.
[45] Iles RA, Davidson M, Taylor NF. Psychosocial predictors of failure to return to work in non-chronic non-specific low back pain: a systematic review. J Occup Environ Med 2008;65:507–17.
[46] Janssen KJM, Kalkman CJ, Grobbee D, Bonsel GJ, Moons KGM, Vergouwe Y. The risk of severe postoperative pain: modification and validation of a clinical prediction rule. Anesth Analg 2008;107:1330–9.
[47] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
[48] Kalkman CJ, Visser K, Moen J, Bonsel GJ, Grobbee DE, Moons KG. Preoperative prediction of severe postoperative pain. PAIN 2003;105: 415–23.
[49] Karran EL, McAuley JH, Traeger AC, Hillier SL, Grabherr L, Russek LN, Moseley GL. Can screening instruments accurately determine poor outcome risk in adults with recent onset low back pain? A systematic review and meta-analysis. BMC Med 2017;15:13.
[50] Karran EL, Traeger AC, McAuley JH, Hillier SL, Yau Y, Moseley GL. The value of prognostic screening for patients with low back pain in secondary care. J Pain 2017;18:673–86.
[51] Kendall NAS, Linton SJ, Main CJ. Guide to assessing psychosocial yellow flags in acute low back pain: Risk factors for long term disability and work loss. Wellington: Accident Rehabilitation and Compensation Insurance Corporation of New Zealand and the National Health Committee, 1997.
[52] Khan KS, Kunz R, Kleijnen J, Antes G. Five steps to conducting a systematic review. J R Soc Med 2003;96:118–21.
[53] Kongsted A, Andersen CH, Hansen MM, Hestbaek L. Prediction of outcome in patients with low back pain—a prospective cohort study comparing clinicians' predictions with those of the Start Back Tool. Man Ther 2016;21:120–7.
[54] Lang K, Alexander IM, Simon J, Sussman M, Lin I, Menzin J, Friedman M, Dutwin D, Bushmakin AG, Thrift-Perry M, Altomare C, Hsu MA. The impact of multimorbidity on quality of life among midlife women: findings

from a U.S. nationally representative survey. J Womens Health 2015;24:374–83.
[55] Law RKY, Lee EWC, Law SW, Chan BKB, Chen PP, Szeto GPY. The predictive validity of OMPQ on the rehabilitation outcomes for patients with acute and subacute non-specific LBP in a Chinese population. J Occup Rehabil 2013;23:361–70.
[56] Lee YH, Bang H, Kim DJ. How to establish clinical prediction models. Endocrinol Metab 2016;31:38–44.
[57] Leeuw M, Goossens MEJB, Linton SJ, Crombez G, Boersma K, Vlaeyen JW. The fear-avoidance model of musculoskeletal pain: current state of scientific evidence. J Behav Med 2007;30:77–94.
[58] Lentz TA, Beneciuk JM, Bialosky JE, Zeppieri G Jr, Dai Y, Wu SS, George SZ. Development of a yellow flag assessment tool for orthopaedic physical therapists: results from the optimal screening for prediction of referral and outcome (OSPRO). J Orthop Sports Phys Ther 2016;5:327–43.
[59] Leysen M, Nijs J, Meeus M, Wilgen CP, Struyf F, Vermandel A, Kuppens K, Roussel N. Clinimetric properties of illness perception questionnaire revised (IPQ-R) and brief illness perception questionnaire (Brief IPQ) in patients with musculoskeletal disorders: a systematic review. Man Ther 2014;20:10–17.
[60] Linton SJ. A review of psychological risk factors in back and neck pain. Spine (Phila Pa 1976) 2000;25:1148–56.
[61] Linton SJ, Boersma K. Early identification of patients at risk of developing a persistent back problem: the predictive validity of the Orebro Musculoskeletal Pain Questionnaire. Clin J Pain 2003;19:80–6.
[62] Linton SJ, Hallden K. Can we screen for problematic back pain? A screening questionnaire for predicting outcome in acute and subacute back pain. Clin J Pain 1998;14:209–15.
[63] Linton SJ, Nicolas M, MacDonald S. Development of a short form of the Örebro musculoskeletal pain screening Questionnaire. Spine 2011;36:1891–5.
[64] Maher CG, Grotle M. Evaluation of the predictive validity of the Orebro Musculoskeletal Pain Screening Questionnaire. Clin J Pain 2009;25:666–70.
[65] Main C, Wood P, Hollis S, Spanswick CC, Waddell G. The Distress and Risk Assessment Method. A simple patient classification to identify distress and evaluate the risk of poor outcome. Spine 1992;17:42–52.
[66] Margison DA, French DJ. Predicting treatment failure in the subacute injury phase using the Orebro Musculoskeletal Pain Questionnaire: an observational prospective study in a workers' compensation system. J Occup Environ Med 2007;49:59–67.
[67] Marhold C, Linton SJ, Melin L. A cognitive–behavioral return-to-work program: effects on pain patients with a history of long-term versus short-term sick leave. PAIN 2001;91:155–63.
[68] Melloh M, Elfering A, Egli Presland C, Roeder C, Barz T, Rolli Salathé C, Tamcan O, Mueller U, Theis JC. Identification of prognostic factors for chronicity in patients with low back pain: a review of screening instruments. Int Orthop 2009;33:301–13.
[69] Miller RP, Kori SH, Todd DD. The Tampa Scale: a measure of kinisophobia. Clin J Pain 1991;7:51–2.
[70] Moerman N, van Dam FS, Muller MJ, Oosting H. The Amsterdam preoperative anxiety and information scale (APAIS). Anesth Analg 1996;82:445–51.
[71] Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
[72] Montori VM, Permanyer-Miralda G, Ferreira-González I, Busse JW, Pacheco-Huergo V, Bryant D, Alonso J, Akl EA, Domingo-Salvany A, Mills E, Wu P, Schünemann HJ, Jaeschke R, Guyatt GH. Validity of composite end points in clinical trials. BMJ 2005;330:594–6.
[73] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent reporting of a multivariable prediction model for individual prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–W73.
[74] Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, Reitsma JB, Collins GS. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11:e1001744.
[75] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691–8.
[76] Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1–W33.
[77] Morsø L, Kent P, Albert HB, Hill JC, Kongsted A, Manniche C. The predictive and external validity of the STarT Back Tool in Danish primary care. Eur Spine J 2013;22:1859–67.
[78] Morsø L, Kent P, Manniche C, Albert HB. The predictive ability of the STarT Back Screening Tool in a Danish secondary care setting. Eur Spine J 2014;23:120–8.
[79] Neubauer E, Junge A, Pirron P, Seemann H, Schiltenwolf M. HKF-R 10—screening for predicting chronicity in acute low back pain (LBP): a prospective clinical trial. Eur J Pain 2006;10:559–66.
[80] Newell D, Field J, Pollard D. Using the STarT back tool: does timing of stratification matter? Man Ther 2015;20:533–9.
[81] Nonclercq O, Berquin A. Predicting chronicity in acute back pain: validation of a French translation of the Orebro Musculoskeletal Pain Screening Questionnaire. Ann Phys Rehabil Med 2012;55:263–78.
[82] Pajouheshnia R, Damen JAAG, Groenwold R, Moons KM, Peelen LM. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. Diagn Progn Res 2017;1:15.
[83] Pajouheshnia R, Peelen LM, Moons K, Reitsma JB, Groenwold R. Accounting for treatment use when validating a prognostic model: a simulation study. BMC Med Res Methodol 2017;17:103.
[84] Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, Perel P, Steyerberg EW, Schroter S, Altman DG, Hemingway H; PROGRESS Group. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med 2014;11:e1001671.
[85] 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.
[86] Pengel LHM, Herbert RD, Maher CG, Refshauge KM. Acute low back pain: systematic review of its prognosis. BMJ 2003;327:323.
[87] Phillips CJ. The cost and burden of chronic pain. Rev Pain 2009;3:2–5.
[88] Pincus T. A systematic review of psychological factors as predictors of chronicity/disability in prospective cohorts of low back pain. Spine 2002;27:E109–20.
[89] Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med 2016;374:504–6.
[90] Ramond A, Bouton C, Richard I, Roquelaure Y, Baufreton C, Legrand E, Huez JF. Psychosocial risk factors for chronic low back pain in primary care—a systematic review. J Fam Pract 2011;28:12–21.
[91] Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: assessing methodological quality. In JJ Deeks, PM Bossuyt, C Gatsonis, editors. Cochrane handbook of systematic reviews of diagnostic test accuracy, version 1.0.0. The Cochrane Collaboration, 2009. Available at: http://srdta.cochrane.org.Accessed January 26, 2017.
[92] Riewe E, Neubauer E, Pfeifer AC, Schiltenwolf M. Predicting persistent back symptoms by psychosocial risk factors: validity criteria for the ÖMPSQ and the HKF-r 10 in Germany. PLoS One 2016;11:e0158850.
[93] Rodeghero JR, Cook CE, Cleland JA, Mintken PE. Risk stratification of patients with low back pain seen in physical therapy practice. Man Ther 2015;20:855–60.
[94] Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009;338:1373–7.
[95] Saastamoinen P, Leino-Arjas P, Laaksonen M, Lahelma E. Socio-economic differences in the prevalence of acute, chronic and disabling chronic pain among ageing employees. PAIN 2005;114:364–71.
[96] Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. Edinburgh: Churchill Livingstone, 2000.
[97] Sandborgh M, Lindberg P, Denison E. Pain belief screening instrument: development and preliminary validation of a screening instrument for disabling persistent pain. J Rehabil Med 2007;39:461–6.
[98] Sandborgh M, Lindberg P, Denison E. The Pain Belief Screening Instrument (PBSI): predictive validity for disability status in persistent musculoskeletal pain. Disabil Rehabil 2008;30:1123–30.
[99] Sattelmayer M, Lorenz T, Röder C, Hilfiker R. Predictive value of the Acute Low Back Pain Screening Questionnaire and the Örebro Musculoskeletal Pain Screening Questionnaire for persisting problems. Eur Spine J 2011;21(suppl 6):S773–84.
[100] Schultz IZ, Crook J, Berkowitz J, Milner R, Meloche GR. Predicting return to work after low back injury using the psychosocial risk for occupational disability instrument: a validation study. J Occup Rehabil 2005;15:365–76.
[101] Scott W, McCracken L. Psychological assessment to identify patients at risk of postsurgical pain: the need for theory and pragmatism. Br J Anaesth 2016;117:546–8.

[102] Shahidi B, Curran-Everett D, Maluf KS. Psychosocial, physical, and neurophysiological risk factors for chronic neck pain: a prospective inception cohort study. J Pain 2015;16:1288–99.
[103] Shaw WS, Chin EH, Nelson CC, Reme SE, Woiszwillo MJ, Verma SK. What circumstances prompt a workplace discussion in medical evaluations for back pain? J Occup Rehabil 2013;23:125–34.
[104] Shaw WT, Pransky GS, Patterson WB, Winters T. Early disability risk factors for low back pain assessed at outpatient occupational health clinics. Spine 2005;30:572–80.
[105] Shaw WS, Reme SE, Pransky G, Woiszwillo MJ, Steenstra IA, Linton SJ. The pain recovery inventory of concerns and expectations a psychosocial screening instrument to identify intervention needs among patients at elevated risk of back disability. J Occup Environ Med 2013;55:885–94.
[106] Sobol-Kwapinska M, Bąbel P, Plotek W, Stelcer B. Psychological correlates of acute postsurgical pain: a systematic review and meta-analysis. Eur J Pain 2016;20:1573–86.
[107] Steenstra I, Verbeek J, Heymans M, Bongers P. Prognostic factors for duration of sick leave in patients sick listed with acute low back pain: a systematic review of the literature. Occup Environ Med 2005;62:851–60.
[108] Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer, 2009.
[109] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG; PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10:e1001381.
[110] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31.
[111] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38.
[112] Streibelt M, Bethge M. Prospective cohort analysis of the predictive validity of a screening instrument for severe restrictions of work ability in patients with musculoskeletal disorders. Am J Phys Med Rehabil 2015;94:617–26.
[113] Sullivan M, Bishop SR, Pivik J. The pain catastrophizing scale: development and validation. Psychol Assess 1995;7:524–32.
[114] Toth C, Lander J, Wiebe S. The prevalence and impact of chronic pain with neuropathic pain symptoms in the general population. Pain Med 2009;10:918–29.
[115] Traeger AC, Henschke N, Hübscher M, Williams CM, Kamper SJ, Maher CG, Moseley GL, McAuley JH. Estimating the risk of chronic pain: development and validation of a prognostic model (PICKUP) for patients with acute low back pain. PLoS Med 2016;13:e1002019.
[116] Truchon M, Schmouth ME, Cote D, Fillion L, Rossignol M, Durand MJ. Absenteeism screening questionnaire (ASQ): a new tool for predicting long-term absenteeism among workers with low back pain. J Occup Rehabil 2012;22:27–50.
[117] Tu YK, Gilthorpe MS. Revisiting the relation between change and initial value: a review and evaluation. Stat Med 2007;26:443–57.
[118] Tucker CA, Cieza A, Riley AW, Stucki G, Lai JS, Bedirhan Ustun T, Kostanjsek N, Riley W, Cella D, Forrest CB. Concept analysis of the patient reported outcomes measurement information system (PROMIS®) and the International classification of functioning, disability and Health (ICF). Qual Life Res 2014;23:1677.
[119] Tucker CA, Escorpizo R, Cieza A, Lai JS, Stucki G, Ustun TB, Kostanjsek N, Cella D, Forrest CB. Mapping the content of the patient-reported outcomes measurement information system (PROMIS®) using the International classification of functioning, Health and disability. Qual Life Res 2014;23:2431–8.
[120] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
[121] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6.
[122] Vlaeyen JWS, Linton SJ. Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art. PAIN 2000;85: 317–32.
[123] Vlaeyen JWS, Morley S, Crombez G. The experimental analysis of the interruptive, interfering, and identity-distorting effects of chronic pain. Behav Res Ther 2016;86:23–34.
[124] Vos CJ, Verhagen AP, Koes BW. The ability of the acute low back pain screening Questionnaire to predict sick leave in patients with acute neck pain. J Manipulative Physiol Ther 2009;32:178–83.
[125] Walton DM, Krebs D, Moulden D, Wade P, Levesque L, Elliott J, MacDermid JC. The traumatic Injuries distress scale: a new tool that quantifies distress and Has predictive validity with patient-reported outcomes. J Orthop Sports Phys Ther 2016;46:920–8.
[126] Wilson JMG, Jungner G. Principles and practice of screening for disease. Geneva: WHO, 1968. Available at: http://www.who.int/bulletin/volumes/86/4/07-050112BP.pdf. Accessed January 26, 2017.
[127] Wingbermuhle RW, van Trijffel E, Nelissen PM, Koes B, Verhagen AP. Few promising multivariable prognostic models exist for recovery of people with non-specific neck pain in musculoskeletal primary care: a systematic review. J Physiother 2018;64:16–23.
[128] Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51–8.
[129] Wolff R, Whiting P, Mallet S, Riley R, Westwood M, Kleijnen K, Mallet S. PROBAST—a risk-of-bias tool for prediction-modelling studies. Abstracts of the global evidence summit, Cape Town, South Africa. Cochrane Database Syst Rev 2017;9(suppl 1).
[130] World Health Organization. International classification of functioning, disability and health (ICF) Geneva: World Health Organization; 2001.