Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 59, No. 5, May 15, 2008, pp 632–641
DOI 10.1002/art.23563

ORIGINAL ARTICLE

A Primary Care Back Pain Screening Tool: Identifying Patient Subgroups for Initial Treatment

JONATHAN C. HILL, KATE M. DUNN, MARTYN LEWIS, RICKY MULLIS, CHRIS J. MAIN, NADINE E. FOSTER, AND ELAINE M. HAY

Objective. **To develop and validate a tool that screens for back pain prognostic indicators relevant to initial decision making in primary care.**

Methods. **The setting was UK primary care adults with nonspecific back pain. Constructs that were independent prognostic indicators for persistence were identified from secondary analysis of 2 existing cohorts and published literature. Receiver operating characteristic curve analysis identified single screening questions for relevant constructs. Psychometric properties of the tool, including concurrent and discriminant validity, internal consistency, and repeatability, were assessed within a new development sample (n = 131) and tool score cutoffs were established to enable allocation to 3 subgroups (low, medium, and high risk). Predictive and external validity were evaluated within an independent external sample (n = 500).**

Results. **The tool included 9 items: referred leg pain, comorbid pain, disability (2 items), bothersomeness, catastrophizing, fear, anxiety, and depression. The latter 5 items were identified as a psychosocial subscale. The tool demonstrated good reliability and validity and was acceptable to patients and clinicians. Patients scoring 0–3 were classified as low risk, and those scoring 4 or 5 on a psychosocial subscale were classified as high risk. The remainder were classified as medium risk.**

Conclusion. **We validated a brief screening tool, which is a promising instrument for identifying subgroups of patients to guide the provision of early secondary prevention in primary care. Further work will establish whether allocation to treatment subgroups using the tool, linked with targeting treatment appropriately, improves patient outcomes.**

INTRODUCTION

Back pain is a common reason for visits to general practitioners (GPs), with 6–9% of all adults (approximately one-quarter of those with back pain) consulting GPs annually in the UK (1). Although many back pain episodes are short lived, approximately 60–80% of patients consulting primary care still report problems a year later (2). It is estimated that 85% of these patients have nonspecific back pain with no known specific underlying disease or pathology (3). Diagnosing back pain is problematic in primary care and the lack of a clear biomedical model to legitimize the pain and direct initial treatment decisions can be frustrating for patients and practitioners alike (4,5).

Evidence-based guidelines for nonspecific back pain in primary care highlight the need to consider prognostic clinical indicators (6,7). Although there is evidence that clinicians' global prognostic assessment compares favorably with that of formal epidemiologic prediction rules (8), identifying specific indicators as appropriate targets of primary care treatment has proved problematic (9). Even when clinicians identify negative prognostic indicators, they appear not to tailor patient management accordingly (10). Investigators increasingly advocate effective early secondary prevention of back pain in primary care through better identification of prognostic indicators that require treatment (11–15).

Several back pain classification tools have been developed to aid clinical decision making (16–22), and such

Supported by a program grant from the Arthritis Research Campaign UK (13413) and by the North Staffordshire Primary Care Research Consortium. Mr. Hill is an Arc Lecturer in Physiotherapy awarded by the Arthritis Research Campaign UK. Dr. Foster's work is supported by a Primary Care Career Scientist award from the Department of Health and the National Health Service Research and Development UK.

Jonathan C. Hill, MSc, Kate M. Dunn, PhD, Martyn Lewis, PhD, Ricky Mullis, MSc, Chris J. Main, PhD, Nadine E. Foster, DPhil, Elaine M. Hay, MD: Keele University, Keele, Staffordshire, UK.

Address correspondence to Jonathan C. Hill, MSc, Primary Care Musculoskeletal Research Centre, Keele University, Keele, Staffordshire, ST5 5BG, UK. E-mail: j.hill@cphc.keele.ac.uk.

Submitted for publication June 8, 2007; accepted in revised form December 3, 2007.

tools may improve clinical outcomes when subgrouping guides treatment (23,24). However, the conceptual purposes of these tools vary substantially, with most designed for the occupational setting or to identify patients for specific treatment modalities. There is a need for an adequately validated screening tool that allocates and prioritizes treatment for the entire spectrum of patients with nonspecific low back pain presenting to primary care, on the basis of treatment modifiable indicators, that is brief and quick to score.

Our overall aim was to develop and validate a back pain screening tool to identify prognostic indicators relevant to GP decision making concerning initial treatment options in primary care. The key objectives of the screening tool were 1) to identify patients with potentially treatment-modifiable prognostic indicators using a brief, user-friendly tool, and 2) to validate cutoff scores for subgrouping patients into 1 of 3 a priori initial treatment options in primary care: low risk subgroup (patients with few negative prognostic indicators, suitable for primary care management according to best-practice guidelines [e.g., analgesia, advice, and education]), medium risk subgroup (patients with an unfavorable prognosis with high levels of physical prognostic indicators, appropriate for physiotherapy), and high risk subgroup (patients with a very unfavorable prognosis, with consistently high levels across psychosocial prognostic indicators, appropriate for management by a combination of physical and cognitive–behavioral approaches).

PATIENTS AND METHODS

There were 3 consecutive steps to the development and validation of the Subgroups for Targeted Treatment (STarT) Back Screening Tool: 1) selecting items for inclusion, 2) validating psychometric properties and identifying cutoff scores for subgroup allocation, and 3) independent external validation. The study received ethical approval from the North Staffordshire Local Research Ethics Committee.

Step 1: selecting items for inclusion in the tool. First, we identified prognostic constructs that could potentially be modified by treatment options in primary care and were therefore relevant to clinical decision making. A secondary analysis of 1 randomized controlled trial (n = 402) (25) and 1 prospective cohort study (n = 739) (1) was undertaken. Poor outcome was defined as 12-month followup Roland-Morris Disability Questionnaire (RMDQ) (26) scores above the median, as this was available in both data sets and standard cutoffs were not available in the literature. Crude odds ratios with 95% confidence intervals (95% CIs) were calculated for associations between prognostic indicators and outcome. Statistically significant indicators ($P < 0.05$) were entered into forward stepwise binary logistic regression analysis to identify independent predictors of outcome within each data set. In parallel, the primary care back pain literature was reviewed to further identify relevant prognostic indicators. Articles were identified through a search of PubMed, databases held by authors, and scrutiny of reference lists, using the search terms "low back pain" and "primary care," with either "randomized controlled trials" or "prospective cohort studies." This was not intended to be a systematic review of the literature, but we aimed to identify important indicators for low back pain prognosis in primary care. A list of constructs for potential inclusion was compiled from both sources. A clinical advisory panel (consisting of primary care back pain specialists including GPs, physiotherapists, osteopaths, pain management nurses, patient representatives, and the study team) reviewed the list of identified constructs and excluded those considered nonmodifiable or rare, or that were likely to be inappropriate targets for primary care intervention. The final constructs were chosen through discussion within the expert panel using information on strength, independence, consistency of association with outcome, and perceived face validity.

Because brevity was important, individual tool items were selected from multi-item instrument constructs identified above. Where validated single questions existed, these were used. For constructs without validated single screening questions, we used receiver operating characteristic (ROC) curves (27) to select optimal individual screening items to identify patients above the median on the full questionnaires (28). The best-performing individual items were discussed and agreed on by the panel, and the item response categories and wording were standardized. The panel categorized items as broadly physical or psychosocial to facilitate linkage with the appropriate a priori treatment subgroups, and to identify items for a psychosocial subscale. Patient acceptability of the screening tool was assessed using feedback from a small sample (n = 12) of patients with back pain consulting primary care.

Step 2A: psychometric testing in the development sample. Psychometric properties of the screening tool, including discriminant validity, internal consistency, and repeatability, were assessed within a new primary care cohort of patients with nonspecific low back pain: the development sample.

Participants were recruited from 8 general practices in North Staffordshire and Central Cheshire, UK. All consecutive patients with low back pain ages 18–59 years were invited to participate. Patients were identified using previously tested computerized electronic Read codes for nonspecific low back pain (1) entered by GPs at the time of consultation. Participants were recruited over a 5-week period (January to February 2005). Sample size calculations were based on previous survey responses in similar patient populations; an initial sample of 200 patients was identified to provide sufficient responses to evaluate validity based on a minimal 10:1 ratio of patients to variables (29).

Measurement and data collection was performed by postal self-completion questionnaire, including age, sex, employment status, and days off work due to back pain. Disability was measured using the RMDQ; pain intensity was measured using the mean of three 11-point numerical rating scales for least, average (over previous 2 weeks), and current pain (30); back pain bothersomeness (over previous 2 weeks) was measured using a single question (1);

duration of back pain was measured through recall of the last pain-free month (31); and fear avoidance beliefs were measured using the Tampa Scale of Kinesiophobia (TSK) (32). Questionnaires also included the Pain Catastrophizing Scale (PCS) (33) and a validated primary care depression screen: the Patient Health Questionnaire-2 (PHQ-2) (34). The STarT Back Screening Tool was included with selected items together on a single page.

Data were analyzed to assess discriminant validity using ROC curves, and by calculating the area under the curve (AUC) for overall screening tool scores and a subscale using psychosocial items alone, against baseline cases on relevant reference standards. Reference standard multi-item instruments were dichotomized to provide cases and noncases using established cutoffs from the primary care literature where available. The definitions for reference standard cases were very or extremely bothersome back pain (1), catastrophizing (PCS score $\geq$20) (35), fear (TSK score $\geq$41) (36), and depression (PHQ-2 score $\geq$2) (34). A median score was used to dichotomize back pain disability (using the reference standard RMDQ score $\geq$7) because an established cutoff was not available from the primary care literature (28). A single item identified patients reporting referred leg pain from their back. AUCs for overall and subscale tool scores were compared to determine which method best discriminated patients according to physical and psychosocial reference standards. Strength of discrimination was classified according to the following descriptors: 0.7–<0.8 indicated acceptable discrimination, 0.8–<0.9 indicated excellent discrimination, and $\geq$0.9 indicated outstanding discrimination (37).

A confirmatory factor analysis was performed to determine whether psychosocial screening items provided a homogeneous single subscale. Item redundancy and internal consistency were investigated by calculating Cronbach's alpha (38,39) for overall and subscale screening tool scores (poor internal consistency was defined as $\alpha < 0.70$, and item redundancy was defined as $\alpha > 0.90$) (38). Floor and ceiling effects were considered present if >15% of respondents achieved the highest/lowest possible tool scores (29). To investigate repeatability (test–retest reliability) of the tool scores, we calculated the quadratic weighted Cohen's kappa (38) for overall scores and subscale scores. The test–retest sample comprised baseline responders who completed a second questionnaire sent 2 weeks after initial baseline questionnaires. Repeatability was further assessed in a subset of these patients who reported stable back pain symptoms during this 2-week period.

Step 2B: screening tool cutoff scores for treatment subgroups. A primary objective of the tool was to provide subgroups to facilitate treatment decision making. We therefore used development sample data to derive scoring rules to allocate patients to 3 a priori treatment subgroups (as defined in the Introduction). First, we defined the low risk subgroup as patients who were not cases based on relevant prognostic reference standards (described in step 1). To identify cutoffs, we used the ROC curves produced in step 2A for overall tool scores against reference standard cases together with average sensitivity and specificity values for each potential cutoff score. The cutoff for low risk was selected using the screening tool overall score that most consistently discriminated between reference standard cases and noncases (i.e., with the highest average sensitivity and specificity).

To determine the optimal method to discriminate between the remaining patients eligible for the medium and high risk subgroups, we used ROC curves of the screening tool psychosocial subscale scores against psychosocial reference standards, including a combined variable to estimate pain-related psychosocial distress. Psychosocial distress was defined as patients who were consistently cases across psychosocial reference standards (defined in step 2A for bothersomeness, fear, catastrophizing, and depression). Average sensitivity and specificity were calculated for each potential cutoff score. Particular consideration was given to a cutoff with high specificity, as physiotherapy has been shown to be effective in modifying moderate levels of distress (40,41), whereas cognitive–behavioral approaches may be detrimental in nondistressed patients (42). The 3 treatment subgroup proportions were then calculated.

Step 3: external and predictive validity testing using an independent external sample. An independent external sample was used to investigate the external validity of subgroup cutoffs and to test the predictive validity of the screening tool.

Participants were recruited using recruitment methods identical to those used in the development sample, although 8 different general practices in North Staffordshire and Central Cheshire, UK, were used. Sample participants were those recruited to an ongoing prospective cohort study of primary care patients with low back pain (43). The first 500 participants who returned baseline and 6-month followup questionnaires were included (September 2004 to January 2006). The external and development samples were both assembled as part of an ongoing program on low back pain in primary care, and the use of the independent sample to externally validate the screening tool was an a priori stated objective of the study.

Measurement and data collection procedures were identical to those of the development sample except that PCS and PHQ-2 questionnaires were replaced with the Coping Strategies Questionnaire catastrophizing subscale (44) and the Hospital Anxiety and Depression Scale (45). The use of these alternative reference standards ensured that results were generalizable to these prognostic constructs, rather than to specific measurement instruments. The independent external sample questionnaire was designed before the final completion of step 1 of the screening tool development, when prognostic constructs had been selected but not their individual items. This meant that some of the tool items were included within the context of their multi-item instruments, rather than in their final 1-page format. However, tool scores for the external sample were calculated using the single items identified in step 1 to maintain consistency; where necessary, category responses were collapsed to reflect the categories used in the screening tool.

Table 1. Odds ratios for the association between baseline indicators and poor outcome (RMDQ higher than median values) at 12-month followup*

Indicator†	Cohort study (RMDQ median 5) (n = 410)		RCT (RMDQ median 3) (n = 329)	
	Crude	Stepwise	Crude	Stepwise
Anxiety/distress	3.23 (2.20–4.73)	Not significant	1.57 (1.01–2.44)	Not significant
Bothersomeness	2.36 (1.62–3.42)	Not significant	No data	No data
Catastrophizing beliefs/perceived risk of not recovering	14.54 (7.98–26.4)	7.63 (3.69–15.7)	1.77 (1.13–2.75)	Not significant
Coping strategies	1.45 (0.98–2.11)	Not applicable	1.83 (1.18–2.84)	1.66 (1.02–2.72)
Depression	3.66 (2.48–5.38)	Not significant	2.02 (1.28–3.17)	Not significant
Disability	6.68 (4.45–10.0)	2.28 (1.36–3.83)	2.47 (1.58–3.86)	2.02 (1.22–3.33)
Duration	2.74 (1.87–4.02)	1.90 (1.15–3.16)	1.20 (0.72–1.99)	Not applicable
Educational status	1.78 (1.23–2.57)	Not significant	1.24 (0.80–1.92)	Not applicable
Fear avoidance behavior/beliefs	3.14 (1.96–5.00)	Not significant	2.06 (1.32–3.23)	1.97 (1.19–3.27)
Female sex	1.08 (0.74–1.56)	Not applicable	0.75 (0.48–1.16)	Not applicable
History of back pain	1.26 (0.75–2.11)	Not applicable	1.69 (1.01–2.82)	Not significant
Job dissatisfaction	3.23 (2.17–4.77)	Not significant	1.27 (0.82–1.98)	Not applicable
Pain elsewhere	3.70 (2.37–5.77)	2.19 (1.23–3.87)	2.44 (1.26–4.71)	2.72 (1.31–5.66)
Pain intensity	3.16 (2.15–4.64)	Not significant	1.85 (1.19–2.88)	Not significant
Pain radiating to the leg/sciatica	2.67 (1.79–3.98)	Not significant	1.31 (0.82–2.09)	Not applicable
Self-rated health	4.95 (3.29–7.42)	2.29 (1.34–3.91)	1.52 (0.98–2.36)	Not applicable
Unemployment	9.79 (5.52–17.3)	3.21 (1.56–6.58)	1.79 (1.08–2.95)	Not significant
Work absence	4.83 (2.94–7.94)	Not significant	1.61 (0.92–2.83)	Not applicable

* Values are the odds ratio (95% confidence interval). Not applicable is used where variables were not included in the stepwise analysis because they were not significant in the crude analysis. RMDQ = Roland-Morris Disability Questionnaire; RCT = randomized controlled trial.
† Independent variables were dichotomized for statistical analysis: the median value was used as a cutoff for numerical scales.

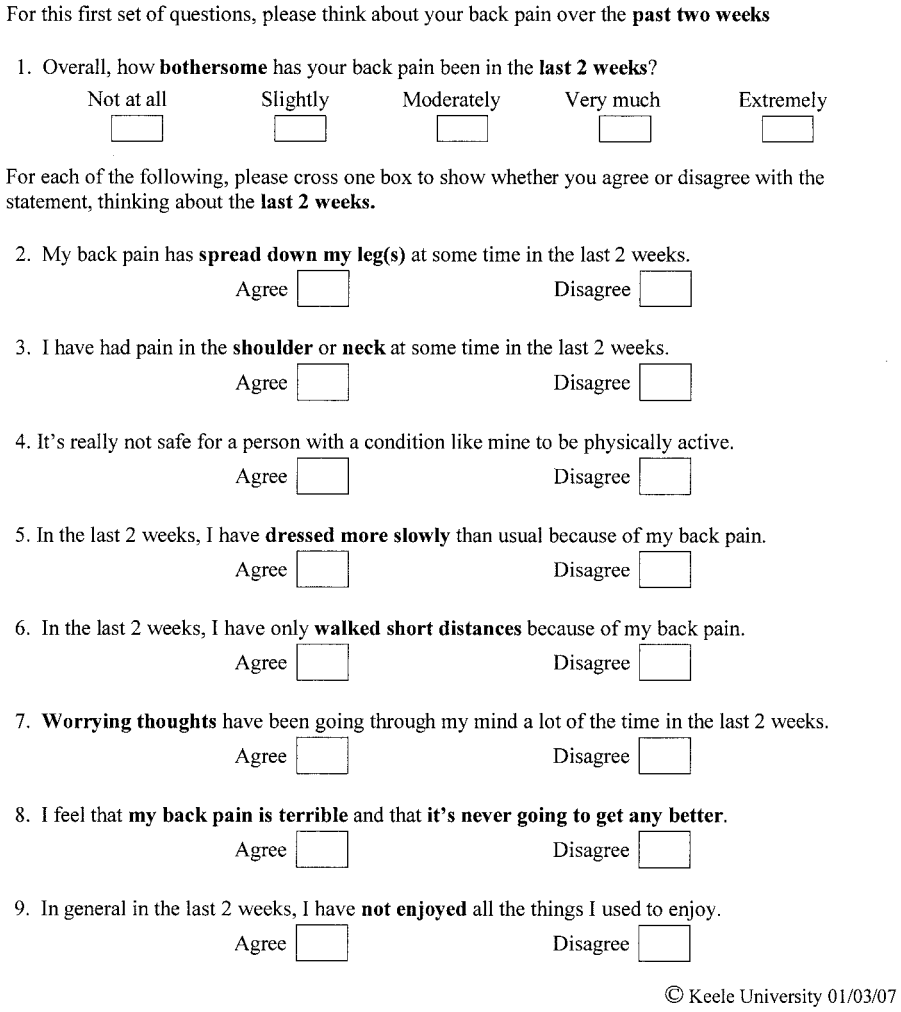
Statistical analysis. Tool scores were computed for baseline data and the subgroup proportions in the external sample were calculated. To investigate external validity, we tested whether the tool's discriminative abilities, according to reference standards, decreased in the independent external sample compared with the development sample. This was performed by calculating AUCs for the tool overall scores against baseline reference standard cases for disability (RMDQ ≥7) and the tool psychosocial subscales scores against bothersomeness (very or extremely) and fear (TSK ≥41).

The predictive validity of the tool subgroup cutoffs (low/medium and medium/high) was assessed by calculating sensitivity, specificity, and negative and positive likelihood ratios (LRs) for subgroup cutoffs against 6-month disability outcome (RMDQ ≥7). The LRs discriminate between good/poor outcome according to the reference standards: higher positive LRs and lower negative LRs indicate better discrimination. In addition, to evaluate the influence of nonmodifiable patient characteristics on the tool's predictive abilities, calculations were presented for subgroups of patients stratified by age (dichotomized at the median), sex, and episode duration (5 categories from <1 month to >3 years).

RESULTS

Step 1: items selected for inclusion in the tool. The independent prognostic indicators identified from the secondary analyses of the randomized controlled trial (n = 402) and prospective cohort study (n = 739) are presented in Table 1. A final list of 9 screening items covering 8 constructs was selected for inclusion in the tool: bothersomeness, referred leg pain, comorbid pain, disability, catastrophizing, fear, anxiety, and depression. Two disability (RMDQ) items were selected to achieve an appropriate level of sensitivity. Items that were identified in the search or analysis but were excluded on discussion due to being rare, nonmodifiable, of perceived low face validity, or inappropriate for primary care intervention included episode duration, educational status, height, age, sex, history of back pain, family history of back pain, frequent consultation, self-rated health, obesity, treatment expectations, and workers compensation status. The screening tool is presented on a single sheet and is suitable for self-completion (see Figure 1). Bothersomeness responses are recorded as positive for very much or extremely bothersome back pain. For all other items, responses are recorded as positive if the person agrees with the statement. Overall tool scores are produced by summing positive items (items 1–9). The psychosocial subscale score is a sum of bothersomeness, fear, catastrophizing, anxiety, and depression items (items 1, 4, 7, 8, and 9). The patient acceptability sample (n = 12) reported that the tool was acceptable, quick, and simple to complete, and the clinical advisory panel reported it was acceptable and simple to score.

Step 2A: development sample psychometric testing. Of 244 patients identified, 131 (54%) returned the questionnaire and 107 (82%) agreed to further contact. Baseline development sample characteristics are presented in Table



For this first set of questions, please think about your back pain over the **past two weeks**

1. Overall, how **bothersome** has your back pain been in the **last 2 weeks**?

Not at all ☐ Slightly ☐ Moderately ☐ Very much ☐ Extremely ☐

For each of the following, please cross one box to show whether you agree or disagree with the statement, thinking about the **last 2 weeks.**

2. My back pain has **spread down my leg(s)** at some time in the last 2 weeks.

Agree ☐ Disagree ☐

3. I have had pain in the **shoulder** or **neck** at some time in the last 2 weeks.

Agree ☐ Disagree ☐

4. It's really not safe for a person with a condition like mine to be physically active.

Agree ☐ Disagree ☐

5. In the last 2 weeks, I have **dressed more slowly** than usual because of my back pain.

Agree ☐ Disagree ☐

6. In the last 2 weeks, I have only **walked short distances** because of my back pain.

Agree ☐ Disagree ☐

7. **Worrying thoughts** have been going through my mind a lot of the time in the last 2 weeks.

Agree ☐ Disagree ☐

8. I feel that **my back pain is terrible** and that **it's never going to get any better**.

Agree ☐ Disagree ☐

9. In general in the last 2 weeks, I have **not enjoyed** all the things I used to enjoy.

Agree ☐ Disagree ☐

Figure 1. Subgroups for Targeted Treatment (STarT) Back Screening Tool. Item 1 is scored as positive if "very much" or "extremely" bothered is marked. Items 2–9 are positive if "agree" is marked. Psychosocial subscale items are 1, 4, 7, 8, and 9. Patients are allocated to the high risk group if the psychosocial subscale score is ≥4. The remaining patients are allocated to the low risk group if the overall tool score is <4 and to the medium risk group if the overall tool score is ≥4.

2. We posted 74 test–retest questionnaires, of which 53 (72%) were returned; 23 patients reported that their symptoms remained stable within the 2-week period.

The discriminant validity of the screening tool is presented in Table 3, with AUCs for overall and subscale tool scores against reference standard cases ranging from 0.73 (acceptable) for referred leg pain to 0.92 (outstanding) for disability. The results demonstrated that overall tool scores best discriminated self-report physical reference standards (e.g., disability and referred leg pain), whereas psychosocial subscale scores best discriminated psychosocial reference standards (e.g., catastrophizing, fear, and depression) and therefore allocation to the high risk subgroup.

Factor analysis confirmed that the psychosocial subscale formed a single dimension. Cronbach's alpha was 0.79 for overall tool scores and 0.74 for the 5 psychosocial items, indicating that no items were redundant. Of the respondents, 10.8% had tool scores of 0 and 5.4% had tool scores of 9, demonstrating that floor and ceiling effects were not present.

The test–retest reliability quadratic weighted kappa scores for the overall tool scores and psychosocial subscale scores were 0.73 (95% CI 0.57–0.84) and 0.69 (95% CI 0.51–0.81), respectively, signifying substantial reliability (46). Test–retest reliability increased to 0.79 (95% CI 0.73–0.95) and 0.76 (95% CI 0.52–0.89), respectively, when agreement was calculated using the 23 patients reporting stable symptoms.

Step 2B: screening tool cutoff scores for treatment subgroups derived using the development sample. ROC analysis and average sensitivity and specificity for each potential tool cutoff score are presented in Figure 2. Tool overall scores of 0–3 identified the low risk subgroup, as scores ≥4 provided the greatest average sensitivity and specificity to identify cases (i.e., 86% for disability and 76% for referred leg pain). A score ≥4 for psychosocial

Table 2. Baseline characteristics of the development and independent external samples*

Baseline characteristics	Development sample (n = 131)	External sample (n = 500)
Female sex	77 (60)	293 (59)
Age, mean ± SD years	44 ± 10.0	45 ± 9.7
Currently employed	95 (73)	370 (75)
Reason for unemployment		
Not working due to back	17 (13)	49 (10)
Looking after home	7 (5)	28 (6)
Retired	0 (0)	15 (3)
Student	2 (2)	3 (1)
Not working other	9 (7)	23 (5)
Days off work in past 6/12		
No time off work	31 (24)	144 (29)
<7 days	23 (18)	91 (18)
1–4 weeks	18 (14)	88 (18)
1–3 months	15 (12)	28 (6)
>3 months	6 (5)	21 (4)
Episode duration		
Less than 1 month	32 (25)	83 (17)
1 to 3 months	19 (15)	94 (19)
4 to 6 months	15 (12)	77 (15)
7 months to 3 years	29 (22)	125 (25)
More than 3 years	34 (26)	112 (22)
Screening tool score, mean ± SD		
Overall	4.33 ± 2.6	3.83 ± 2.3
Psychosocial subscale	2.17 ± 1.6	1.82 ± 1.5
Pain intensity		
Mild (0–5)	85 (65)	325 (65)
Moderate (6–7)	29 (22)	113 (23)
Severe (8–10)	17 (13)	54 (11)
Disability (RMDQ), mean ± SD	8.6 ± 6.6	9.1 ± 5.9
Referred leg pain	75 (57)	303 (61)
Comorbid pain in neck/shoulder	63 (48)	276 (55)
Very or extremely bothered by back	56 (43)	276 (55)
Fear (TSK), mean ± SD	40.9 ± 7.5	39.5 ± 6.9
Catastrophizing, mean ± SD		
PCS	18.9 ± 12.4	–
CSQ catastrophizing subscale	–	10 ± 7.9
Anxiety (HADS subscale), mean ± SD	–	8.2 ± 4.5
Depression (HADS subscale), mean ± SD	–	6.7 ± 4.3

* Values are the number (percentage) unless otherwise indicated. RMDQ = Roland-Morris Disability Questionnaire; TSK = Tampa Scale of Kinesiophobia; PCS = Pain Catastrophizing Scale; CSQ = Coping Strategies Questionnaire; HADS = Hospital Anxiety and Depression Scale.

Table 3. Discriminant validity in the development sample: area under the receiver operating characteristic curve (AUC) for screening tool overall scores and psychosocial subscale scores against reference standard cases at baseline*

Reference standards	Case definition	Overall tool scores, AUC (95% CI)	Psychosocial subscale scores, AUC (95% CI)
Disability	RMDQ ≥7	0.92 (0.88–0.97)	0.90 (0.85–0.93)
Referred leg pain	Yes	0.84 (0.77–0.91)	0.73 (0.64–0.81)
Bothersomeness	Very or extremely	0.92 (0.88–0.97)	0.92 (0.88–0.97)
Catastrophizing	PCS ≥20	0.79 (0.71–0.87)	0.83 (0.76–0.90)
Fear	TSK ≥41	0.79 (0.71–0.87)	0.81 (0.74–0.99)
Depression	PHQ-2 = 2	0.74 (0.65–0.82)	0.76 (0.68–0.84)

* 95% CI = 95% confidence interval; PHQ-2 = Patient Health Questionnaire-2; see Table 2 for additional definitions.

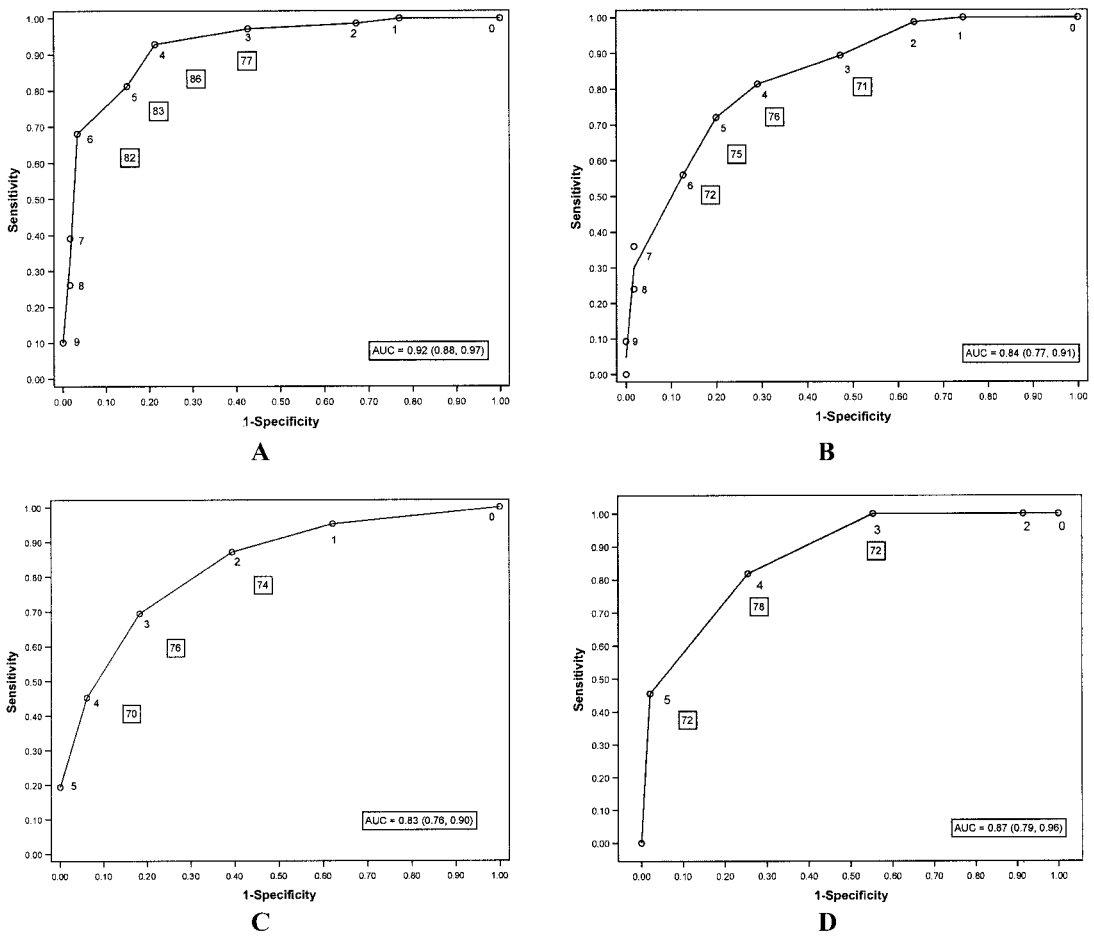


Figure 2. Scoring cutoffs for subgroup allocation derived using the development sample. Receiver operating characteristic (ROC) curves for overall tool scores against reference standard cases for **A,** disability (Roland-Morris Disability Questionnaire score ≥7) and **B,** referred leg pain ("yes"), and for tool psychosocial subscale scores against cases for **C,** catastrophizing (Pain Catastrophizing Scale score ≥20) and **D,** psychosocial distress defined using all 4 psychosocial reference standards. The numbers in the boxes are the average sensitivity and specificity. Numbers in parentheses are the 95% confidence interval. AUC = area under the ROC curve.

subscale items was determined for the high risk subgroup. Overall tool scores >3 but <4 on the psychosocial subscale allocated patients to the medium risk subgroup. These cutoffs produced a distribution of 52 patients (40%) in the low risk subgroup, 45 (35%) in the medium risk subgroup, and 33 (25%) in the high risk subgroup.

Step 3: testing for external and predictive validity with an independent external sample. *Independent external sample.* Participants (n = 500) were similar to the development sample and their characteristics are presented in Table 2. The distribution of screening tool treatment subgroup cutoffs when applied to the independent external sample was as follows: 234 (47%) were low risk, 186 (38%) were medium risk, and 74 (15%) were high risk.

External validity. AUCs for tool overall scores against baseline reference standard cases for disability and psychosocial subscale scores against baseline bothersomeness and fear did not fall substantially between development and external samples. AUCs (95% CIs) for disability, bothersomeness, and fear were 0.92 (0.88–0.97), 0.92 (0.88–0.97), and 0.81 (0.74–0.89), respectively, in the development sample and 0.90 (0.88–0.93), 0.89 (0.86–0.91), and 0.79 (0.75–0.83), respectively, in the external sample.

Predictive validity of screening tool. Sensitivity, specificity, and negative and positive LRs for tool subgroup cutoffs against 6-month disability outcome are presented in Table 4. There was a small influence of age and sex on the tool's predictive abilities, with older persons and men having higher positive LRs and lower negative LRs signifying better discrimination of outcome. There was a stronger influence observed from episode duration, particularly among patients who reported having back pain for 1–6 months. At the 6-month followup, 39 (16.7%) of the 234 patients in the low risk group had a poor disability outcome (RMDQ ≥7), 99 (53.2%) of the 186 patients in the medium risk group had a poor outcome, and 58 (78.4%) of the 74 patients in the high risk group had a poor outcome.

DISCUSSION

We have developed and validated a simple, brief, and practical way to subgroup patients with nonspecific low

Table 4. Sensitivity, specificity, and likelihood ratios (LRs) for screening tool subgroup cutoffs (low/medium and medium/high) to predict poor disability outcome at 6 months (Roland-Morris Disability Questionnaire ≥7) within the external sample data*

Stratification	Subgroup cutoffs	Sensitivity, %	Specificity, %	Neg. LR (95% CI)	Pos. LR (95% CI)
Total sample	L vs. M/H	80.1	65.4	0.30 (0.23–0.40)	2.32 (1.96–2.76)
	L/M vs. H	39.6	94.6	0.74 (0.67–0.81)	5.51 (3.30–9.28)
Age ≤46.5 years†	L vs. M/H	75.7	65.8	0.37 (0.25–0.52)	2.21 (1.73–2.86)
	L/M vs. H	28.2	93.2	0.77 (0.67–0.87)	4.11 (2.14–8.00)
Age >46.5 years†	L vs. M/H	84.9	65.1	0.23 (0.14–0.37)	2.44 (1.94–3.10)
	L/M vs. H	31.2	96.1	0.72 (0.61–0.81)	7.90 (3.52–17.99)
Female	L vs. M/H	78.0	63.6	0.35 (0.24–0.48)	2.14 (1.72–2.70)
	L/M vs. H	25.2	94.4	0.79 (0.70–0.87)	4.54 (2.30–9.08)
Male	L vs. M/H	84.1	67.6	0.24 (0.13–0.40)	2.60 (2.01–3.41)
	L/M vs. H	37.7	94.9	0.66 (0.53–0.77)	7.39 (3.44–15.81)
<1 month duration	L vs. M/H	78.0	65.0	0.34 (0.17–0.62)	2.25 (1.50–3.51)
	L/M vs. H	31.3	91.8	0.75 (0.56–0.93)	3.82 (1.39–10.79)
1–3 months' duration	L vs. M/H	81.0	69.9	0.27 (0.11–0.59)	2.69 (1.75–4.04)
	L/M vs. H	14.3	98.6	0.87 (0.67–0.99)	7.43 (1.11–50.04)
4–6 months' duration	L vs. M/H	80.0	73.2	0.27 (0.11–0.58)	2.99 (1.83–4.89)
	L/M vs. H	35.0	98.1	0.66 (0.44–0.84)	19.60 (3.36–117.71)
7 months' to 3 years' duration	L vs. M/H	71.2	59.7	0.48 (0.30–0.75)	1.77 (1.28–2.49)
	L/M vs. H	21.2	90.3	0.87 (0.72–1.01)	2.18 (0.93–5.12)
>3 years' duration	L vs. M/H	86.8	54.5	0.24 (0.13–0.46)	1.91 (1.42–2.76)
	L/M vs. H	36.8	93.2	0.68 (0.55–0.82)	5.39 (1.91–16.24)

* Results are provided for the total sample and stratified by age, sex, and episode duration. 95% CI = 95% confidence interval; L = low risk; M = medium risk; H = high risk.
† Age 46.5 years = median age.

back pain in primary care. The new STarT Back Screening Tool identifies potentially modifiable prognostic indicators that may be appropriate targets for primary care interventions.

Identifying patient subgroups has been referred to as "the Holy Grail" of back pain by the Cochrane Back Review Group (47) and prognostic assessment is highlighted in the European guidelines for low back pain in primary care (6). This is the first instrument specifically developed and validated for use in primary care in the UK and it specifically addresses identified research priorities (48). The tool was designed for specific screening purposes and is therefore quick to complete and score. Our approach combined statistical and clinical methods, both in the tool's conceptual purposes and development. A further strength was the use of 4 large data sets including 3 primary care samples (2 cohorts and 1 randomized controlled trial) to develop the tool and a fourth cohort to evaluate external validity, which optimizes generalizability.

There are a number of potential limitations to our study. First, a formal comprehensive systematic review or consensus technique was not used in item selection, although the broad experience of the study team and clinical advisory panel and the wide range of methods used ensured that our methods were robust. Second, this study was conducted within the context of UK primary care, and therefore some clinical examination findings, such as imaging, were not included because they are not part of usual practice in this setting. Third, for practical reasons the tool was not included in a 1-page format in the external sample. However, all of the individual items were included in the context of full construct measures within the questionnaire. Fourth, there was a lack of a single reference standard (criterion validity) against which to compare the performance of the tool. Research in The Netherlands has compared back pain prognostic instruments against clinicians' subjective opinion (8), and we therefore plan a similar study. A further limitation relates to possible nonresponse bias, which may have inflated the proportion of patients classified as appropriate for physiotherapy (medium and high risk subgroups) if patients with longer-term or more severe problems were more likely to respond to the questionnaire (38). In the extreme circumstance of all nonresponders being low risk, 70% of patients would be classified as low risk, but we believe the real situation is likely to lie in the middle, with referral to physiotherapy being appropriate for 30–50% of back pain patients. This is slightly higher than current UK estimates (49) but similar to The Netherlands (50). While nonresponse bias might affect the proportions of subgroup allocation, it is unlikely to affect data presented on the validity of the tool (38).

Recent back pain intervention trials in primary care demonstrated only small differences in outcome when compared with active controls (51). One possible explanation is that interventions have not effectively targeted modifiable risk factors (52). Therefore, at the beginning of the tool development, 3 subgroups were conceptually defined to identify a low risk group, high physical risk group, and high psychosocial risk group to facilitate future targeting of treatment. However, the predictive validity study confirmed that in fact these were low, medium, and high risk subgroups across a broad range of clinical outcomes and therefore these new terms were given to the sub-

groups. The fact that allocation to the high risk subgroup is driven by the psychosocial subscale highlights the importance of psychosocial prognostic indicators among patients with low back pain.

Decision making for referral to physiotherapy is currently inconsistent. The STarT Back Screening Tool may help to provide a more systematic approach by reassuring clinicians that important modifiable indicators have not been missed. The influence of episode duration on the tool's predictive performance indicates that there may be an important window of opportunity for screening that is within 1–6 months. This complements the tool's primary conceptual purpose to provide subgroups for initial treatment of modifiable risk factors. Specificity for allocation to the high risk subgroup was purposefully set high, because cognitive–behavioral approaches may be detrimental for nondistressed patients (42) and conservative physiotherapy may be effective in modifying moderate levels of pain-related distress (40,41).

Further work is now needed to establish whether improved clinical outcomes are demonstrated by implementing the STarT Back Screening Tool and linking this with appropriate targeted treatment in primary care. We are currently evaluating this in a randomized controlled trial.

ACKNOWLEDGMENTS

The authors would like to thank the administrative and health informatics staff at Keele University's Primary Care Musculoskeletal Research Centre and the Keele General Practice Partnership. The authors also thank the members of the expert clinical advisory group, the independent monitoring committee, staff and patients of the 8 participating general practices, and the wider members of the research team who were involved with the study.

AUTHOR CONTRIBUTIONS

Mr. Hill had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study design. Hill, Dunn, Lewis, Mullis, Main, Foster, Hay.
Acquisition of data. Hill, Dunn, Lewis, Foster, Hay.
Analysis and interpretation of data. Hill, Dunn, Lewis, Main, Foster, Hay.
Manuscript preparation. Hill, Dunn, Lewis, Mullis, Main, Foster, Hay.
Statistical analysis. Hill, Dunn, Lewis, Main, Hay.

REFERENCES

1. Dunn KM, Croft PR. Classification of low back pain in primary care: using "bothersomeness" to identify the most severe cases. Spine 2005;30:1887–92.
2. Croft PR, Macfarlane GJ, Papageorgiou AC, Thomas E, Silman AJ. The outcome of low back pain in general practice: a prospective study. BMJ 1998;316:1356–9.
3. Deyo RA, Weinstein JN. Low back pain. N Engl J Med 2001; 344:363–70.
4. Glenton C. Chronic back pain sufferers: striving for the sick role. Soc Sci Med 2003;57:2243–52.
5. Lillrank A. Back pain and the resolution of diagnostic uncertainty in illness narratives. Soc Sci Med 2003;57:1045–54.
6. Van Tulder M, Becker A, Bekkering T, Breen A, del Real MT, Hutchinson A, et al, and the COST B13 Working Group on Guidelines for the Management of Acute Low Back Pain in Primary Care. Chapter 3: European guidelines for the management of acute nonspecific low back pain in primary care. Eur Spine J 2006;15 Suppl 2:S169–91.
7. Royal College of General Practitioners. Clinical guidelines for the management of acute low back pain. 2nd ed. London: Royal College of General Practitioners; 1999.
8. Jellema P, van der Windt DA, van der Horst HE, Stalman WA, Bouter LM. Prediction of an unfavourable course of low back pain in general practice: comparison of four instruments. Br J Gen Pract 2007;57:15–22.
9. Main CJ, Williams AC. Musculoskeletal pain. BMJ 2002;325: 534–7.
10. Bishop A, Foster NE. Do physical therapists in the United Kingdom recognize psychosocial factors in patients with acute low back pain? Spine 2005;30:1316–22.
11. Boersma K, Linton SJ. Screening to identify patients at risk: profiles of psychological risk factors for early intervention. Clin J Pain 2005;21:38–43.
12. Morley S, Vlaeyen JW. Epilogue to the special topic series. Clin J Pain 2005;21:69–72.
13. Jellema P, van der Horst HE, Vlaeyen JW, Stalman WA, Bouter LM, van der Windt DA. Predictors of outcome in patients with (sub)acute low back pain differ across treatment groups. Spine 2006;31:1699–705.
14. Koes BW, van Tulder MW, Thomas S. Diagnosis and treatment of low back pain. BMJ 2006;332:1430–4.
15. Turk DC. The potential of treatment matching for subgroups of patients with chronic pain: lumping versus splitting. Clin J Pain 2005;21:44–55.
16. Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, et al. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. Ann Intern Med 2004;141: 920–8.
17. Hicks GE, Fritz JM, Delitto A, McGill SM. Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. Arch Phys Med Rehabil 2005;86:1753–62.
18. Dionne CE, Bourbonnais R, Fremont P, Rossignol M, Stock SR, Larocque I. A clinical return-to-work rule for patients with back pain. CMAJ 2005;172:1559–67.
19. Truchon M, Cote D. Predictive validity of the Chronic Pain Coping Inventory in subacute low back pain. Pain 2005;116: 205–12.
20. Duijts SF, Kant IJ, Landeweerd JA, Swaen GM. Prediction of sickness absence: development of a screening instrument. Occup Environ Med 2006;63:564–9.
21. 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.
22. Denison E, Asenlof P, Sandborgh M, Lindberg P. Musculoskeletal pain in primary health care: subgroups based on pain intensity, disability, self-efficacy, and fear-avoidance variables. J Pain 2007;8:67–74.
23. Brennan GP, Fritz JM, Hunter SJ, Thackeray A, Delitto A, Erhard RE. Identifying subgroups of patients with acute/subacute "nonspecific" low back pain: results of a randomized clinical trial. Spine 2006;31:623–31.
24. Fritz JM, Brennan GP, Clifford SN, Hunter SJ, Thackeray A. An examination of the reliability of a classification algorithm for subgrouping patients with low back pain. Spine 2006;31: 77–82.
25. Hay EM, Mullis R, Lewis M, Vohora K, Main CJ, Watson P, et al. Comparison of physical treatments versus a brief pain-management programme for back pain in primary care: a randomised clinical trial in physiotherapy practice. Lancet 2005;365:2024–30.
26. Roland M, Morris R. A study of the natural history of back pain. I. Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8:141–4.
27. Murphy JM, Berwick DM, Weinstein MC, Borus JF, Budman SH, Klerman GL. Performance of screening and diagnostic

tests: application of receiver operating characteristic analysis. Arch Gen Psychiatry 1987;44:550–5.
28. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006;332:1080.
29. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
30. Dunn KM. Epidemiology of low back pain in primary care: a cohort study of consulters. PhD Thesis. Keele: Keele University; 2004.
31. Dunn KM, Croft PR. The importance of symptom duration in determining prognosis. Pain 2006;121:126–32.
32. Kori SH, Miller RP, Todd DD. Kinisophobia: a new view of chronic pain behavior. Pain Manag 1990;3:35–43.
33. Sullivan MJ, Bishop SR, Pivik J. The Pain Catastrophizing Scale: development and validation. Psychol Assess 1995;7: 524–32.
34. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003;41:1284–92.
35. Sullivan MJ, Stanish WD. Psychologically based occupational rehabilitation: the Pain-Disability Prevention Program. Clin J Pain 2003;19:97–104.
36. Nederhand MJ, Ijzerman MJ, Hermens HJ, Turk DC, Zilvold G. Predictive value of fear avoidance in developing chronic neck pain disability: consequences for clinical decision making. Arch Phys Med Rehabil 2004;85:496–501.
37. Hosmer DW, Lemeshaw S. Applied logistic regression. 2nd ed. New York: Wiley; 2000.
38. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. New York: Oxford University; 2003.
39. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess 2003; 80:217–22.
40. Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, van der Heijden GJ, van Geel AC, et al. Active rehabilitation for chronic low back pain: cognitive-behavioral, physical, or both? First direct post-treatment results from a randomized controlled trial. BMC Musculoskelet Disord 2006;7:5.
41. Klaber Moffett JA, Carr J, Howarth E. High fear-avoiders of physical activity benefit from an exercise program for patients with back pain. Spine 2004;29:1167–72.
42. George SZ, Fritz JM, Bialosky JE, Donald DA. The effect of a fear-avoidance-based physical therapy intervention for patients with acute low back pain: results of a randomized clinical trial. Spine 2003;28:2551–60.
43. Foster NE, Bishop A, Thomas E, Main C, Horne R, Weinman J, et al. Illness perceptions of low back pain patients in primary care: what are they, do they change, and are they associated with outcome? Pain 2008. E-pub ahead of print.
44. Rosenstiel AK, Keefe FJ. The use of coping strategies in chronic low back pain patients: relationship to patient characteristics and current adjustment. Pain 1983;17:33–44.
45. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361–70.
46. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): L. Erlbaum Associates; 1998.
47. Bouter LM, Pennick V, Bombardier C, and the Editorial Board of the Back Review Group. Cochrane Back Review Group. Spine 2003;28:1215–8.
48. Borkan JM, Koes B, Reis S, Cherkin DC. A report from the Second International Forum for Primary Care Research on Low Back Pain: reexamining priorities. Spine 1998;23: 1992–6.
49. Mason V, and the Office of Population Censuses and Surveys, Social Survey Division. The prevalence of back pain in Great Britain. London: Her Majesty's Stationery Office; 1994.
50. Picavet HS, Schouten JS. Musculoskeletal pain in The Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. Pain 2003;102:167–78.
51. Macfarlane GJ, Jones GT, Hannaford PC. Managing low back pain presenting to primary care: where do we go from here? Pain 2006;122:219–22.
52. Van der Windt D, Hay E, Jellema P, Main C. Psychosocial interventions for low back pain in primary care: lessons learned from recent trials. Spine 2008;33:81–9.