Title: Refinement and validation of a tool for stratifying patients with musculoskeletal pain

Running head: Refinement and validation of the Keele STarT MSK Tool

KM Dunn[1], P Campbell[1], M Lewis[1,2], JC Hill[1], DA van der Windt[1], E Afolabi[1,2], J Protheroe[1], S Wathall[1,2], S Jowett[3], R Oppong[3], CD Mallen[1], EM Hay[1], NE Foster[1,2]

1. Primary Care Centre Versus Arthritis, School of Medicine, Faculty of Medicine and Health Sciences, Keele University, United Kingdom.

2. Keele Clinical Trials Unit (CTU), Faculty of Medicine and Health Sciences, Keele University, United Kingdom.

3. Health Economics Unit, Institute of Applied Health Research, University of Birmingham, United Kingdom.

Corresponding Author: Professor Kate Dunn, Primary Care Centre Versus Arthritis, School of Medicine, Faculty of Medicine & Health Sciences, Keele University, Staffordshire, ST5 5BG, United Kingdom. Tel: +44 (0)1782 734703; Email: k.m.dunn@keele.ac.uk.

Category for which manuscript is being submitted: original article

Funding: This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (https://www.nihr.ac.uk/, grant number: RP-PG-1211-20010 - NEF, KMD, JCH, JP, ML, DAvdW, SJ, CDM, EH), Versus Arthritis (https://www.versusarthritis.org/, grant reference: 20202 – EH, NEF, DAvdW, CDM, ML, KMD) and an NIHR Research Professorship awarded to NEF (NIHR-RP-011-015). CDM is funded by the NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands, the NIHR School for Primary Care Research and an NIHR Research Professorship in General Practice (NIHRRP-2014-04-026). NEF and EH are NIHR Senior Investigators. Funding sources had no role in the design, execution, conduct, data analysis or interpretation, reporting of results, decision to publish or preparation of the manuscript. The views and opinions expressed within this manuscript are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Conflicts of interest: None.

Significance statement: The paper presents the first musculoskeletal pain prognostic stratification tool specifically for use among all primary care patients with the five most common musculoskeletal pain presentations (back, neck, knee, shoulder or multisite pain). The Keele STarT MSK tool identifies groups of musculoskeletal pain patients with clearly different characteristics and prognosis. Using this tool for stratification and treatment matching may be clinically and cost-effective.

ABSTRACT

Background: Patients with musculoskeletal pain in different body sites share common prognostic factors. Using prognosis to stratify and treatment match can be clinically and cost-effective. We aimed to refine and validate the Keele STarT MSK tool for prognostic stratification of musculoskeletal pain patients.

Methods: Tool refinement and validity was tested in a prospective cohort study, and external validity examined in a pilot cluster RCT. Study population comprised 2414 adults visiting UK primary care with back, neck, knee, shoulder or multisite pain returning postal questionnaires (cohort: 1890 (40% response); trial: 524). Cohort baseline questionnaires included a draft tool plus refinement items. Trial baseline questionnaires included the Keele STarT MSK tool. Physical health (SF-36 Physical Component Score (PCS)) and pain intensity were assessed at 2- and 6-months cohort follow-up; pain intensity was measured at 6-months trial follow-up.

Results: The tool was refined by replacing (3), adding (3) and removing (2) items, resulting in a 10-item tool. Model fit (R2) was 0.422 and 0.430 and discrimination (c-statistic) 0.839 and 0.822 for predicting 6-month cohort PCS and pain (respectively). The tool classified 24.9% of cohort participants at low, 41.7% medium and 33.4% high risk, clearly discriminating between subgroups. The tool demonstrated model fit of 0.224 and discrimination 0.73 in trial participants. Multiple imputation confirmed robustness of findings.

Conclusions: The Keele STarT MSK tool demonstrates good validity and acceptable predictive performance, and clearly identifies groups of musculoskeletal pain patients with different characteristics and prognosis. Using prognostic information for stratification and treatment matching may be clinically/cost-effective.

INTRODUCTION

Low back and neck pain are among the leading causes of disability globally, with other musculoskeletal disorders making substantial contributions (Vos et al., 2012). The impact on individuals is considerable, with wider implications such as the heavy burden on healthcare (Jordan et al., 2010; March et al., 2014), and lost productivity for society (Woolf and Pfleger, 2003; Yelin et al., 2019). Musculoskeletal pain is therefore a research priority (GBD 2015 DALYs and HALE Collaborators, 2016; Buchbinder et al., 2018).

Stratified care involves targeting treatments according to patient subgroups, maximising treatment benefit and reducing potential harm or unnecessary interventions (Hingorani et al., 2013). Stratification and targeted treatment is particularly appropriate for patients with musculoskeletal pain due to the wide individual variability in prognosis, plethora of available treatments and the variation in treatment response between patients (Kamper et al., 2015; Linton et al., 2018; Stanton et al., 2010; Patel et al., 2013; Foster et al., 2013). One particular approach in low back pain (LBP), combining a prognostic stratification tool (STarT Back tool (Hill et al., 2008)) with matched treatments, showed significantly better clinical and economic outcomes in the UK (Hill et al., 2011; Foster et al., 2014; Whitehurst et al., 2012; Whitehurst et al., 2015), and is now recommended in LBP guidelines in the UK (National Institute for Health and Clinical Excellence, 2016) and elsewhere (Van Wambeke et al., 2017; New South Wales Agency for Clinical Innovation, 2016; Chenot et al., 2017).

LBP represents approximately 20% of all primary care musculoskeletal consultations, and the other four most common pain presentations are neck, knee, shoulder and multisite pain (Jordan et al., 2010). Systematic reviews have identified various prognostic factors that predict poor outcome across a range of musculoskeletal pain presentations (Mallen et al., 2007; Henschke et al., 2012; Artus et al., 2017), and evidence has shown that a single chronic pain risk score can predict outcomes across pain conditions (Von Korff and Dunn, 2008; Thomas et al., 2008; Muller et al., 2013). Preliminary analyses of a modified STarT Back tool in patients with back, neck, upper limb,

lower limb or multisite pain showed that a single tool was able to predict outcome across pain sites, but required modification as baseline risk varied for the different pain sites (Hill et al., 2016).

The overall aim of this study was to refine and validate a prognostic stratification tool (the Keele STarT MSK tool) for use among all patients consulting in UK primary care with back, neck, knee, shoulder or multisite pain.

METHODS

The Keele Aches and Pains Study (KAPS) was designed to refine and test construct validity of the Keele STarT MSK tool (Campbell et al., 2016). External validation of predictive performance was conducted within a pilot cluster randomised controlled trial (STarT MSK pilot trial).

Refinement and validation

Design and participant recruitment

The KAPS study is a prospective cohort study of patients (18 years and over) consulting a general practitioner (GP) for one (or more) of the five most common musculoskeletal pain presentations – back, neck, knee, shoulder or multisite pain (Jordan et al., 2010). Participants were recruited between July 2014 and February 2015 from 12 general practices in Staffordshire and West Midlands, UK. Relevant Read codes (symptom and diagnostic codes used in UK primary care), entered into electronic records during visits were used to detect potential participants. Consecutive eligible patients were identified through weekly-to-fortnightly electronic record searches. Patients received a postal invitation letter, information sheet, consent form and questionnaire, and were given prepaid return envelopes. Follow-up questionnaires were sent 2- and 6-months later, with reminders if needed (Campbell et al., 2016). All participants continued to receive usual care for their musculoskeletal pain.

Patients were excluded if there were indications of potentially serious underlying pathology (e.g. fracture, infection), urgent care needs (e.g. Cauda Equina Syndrome), if patients were vulnerable (e.g. diagnosed dementia, persistent severe mental health problems, terminal illness, recent trauma or bereavement) and those unable to communicate in English.

Cohort study questionnaires contained the draft Keele STarT MSK tool (identical to the modified STarT Back tool (Hill et al., 2016)). The draft tool comprised items covering pain sites, activity restriction, fear avoidance, catastrophising, anxiety, depression, and pain bothersomeness.

Primary outcome measures

The primary measures for assessing predictive performance were pain intensity (mean of numerical ratings scales for least, average, and current pain over the last 2-weeks (Dunn et al., 2010) and self-reported physical health (SF-36 version 2 physical component summary score (PCS)) (Ware, 2000), assessed at 2- and 6-months follow-up. Minimal data included the SF-12v2, which is a shortened version of the SF-36 version 2 (Ware et al., 1996).

Potential candidate items for refining the Keele STarT MSK tool

Eighteen candidate items were identified for potential addition or replacement within the refined tool and included in cohort study baseline questionnaires. Candidate items were selected based on previous research identifying generic prognostic factors for musculoskeletal pain (Mallen et al., 2007; Henschke et al., 2012; Artus et al., 2017; Von Korff and Dunn, 2008; Campbell et al., 2013; Nicholls et al., 2013), with the predictive value of individual items investigated within existing datasets where possible (Campbell et al., 2013; Dunn & Croft 2005; Hill et al., 2008; Hill et al., 2011; Foster et al., 2014). Items were assessed for suitability, face validity and readability by the research team and a Patient and Public Involvement and Engagement (PPIE) group during a half-day meeting

with 10 PPIE members, supported by two PPIE co-ordinators to ensure autonomy. The PPIE group comprised people with experience of musculoskeletal pain similar to the target population. Item refinements were made, such as wording changes to facilitate simple yes/no response formats. Candidate items covered domains including: vitality/fatigue, comorbidity, coping, sleep problems, previous treatment success, pain interference, pain self-efficacy, pain persistence, pain-related depression, and fear-avoidance. The final choice, wording and format of candidate items was made by the research team based on all available information. The candidate items were included in the questionnaire in the same format and section as the draft Keele STarT MSK tool items.

Additional measures

Measures used to describe the population and differences between risk subgroups included pain duration (Dunn and Croft, 2006), pain spread, pain interference (Amtmann et al., 2010) and bothersomeness (Dunn and Croft, 2005), pain catastrophizing (Harland and Georgieff, 2003), pain self-efficacy (Nicholas, 2007), illness perceptions (Nicholls et al., 2013; Moss-Morris et al., 2002), sleep problems (Jenkins et al., 1988), social support (Krumholz et al., 1998), health-related quality-of-life (EQ-5D-5L) (Herdman et al., 2011), health literacy (Morris et al., 2006), comorbidity, and employment factors (Campbell et al., (2016) contains further details). The STarT Back tool (Hill et al., 2008), for those with LBP, and the short-form Örebro Musculoskeletal Pain Screening Questionnaire (ÖMSPQ) (Linton et al., 2011) were used to assess cross-sectional construct validity.

External validation of the Keele STarT MSK tool

The STarT MSK pilot trial is a pilot cluster randomised controlled trial (RCT) with the same inclusion and exclusion criteria as the KAPS cohort study (Trial registration: ISRCTN 15366334). Recruitment was from eight general practices in the same area, between October 2016 and May 2017 (four

control and four intervention practices). Similar to the cohort, participant identification was based on electronic Read codes, but the trial also included GP point-of-consultation eligibility confirmation using Read code activated computer templates. Patients were invited using similar methods to the cohort study. Baseline questionnaires contained the refined Keele STarT MSK tool. The primary outcome was pain intensity (usual pain over the previous 2-weeks on a 0-10 numerical rating scale) at 6-months; this question has comparable validity to the composite pain intensity measure used in the cohort (Dunn et al., 2010). Control arm patients received usual care for their musculoskeletal pain; care for intervention arm patients was informed by the Keele STarT MSK Tool and matched treatment options (Protheroe et al., 2019).

Sample size

The sample size for the cohort study (1,250 patients at 6-months) was calculated based on the requirement of ≥100 patients per low, medium and high subgroups, anticipating ≥10% of participants in the smallest subgroup (Campbell et al., 2016). Study monitoring indicated lower response rates than estimated, therefore the number of patients identified was increased to ensure sufficient participants with data at 6-months. The STarT MSK pilot trial sample size was based on numbers required to investigate specific pilot trial success criteria (See http://www.isrctn.com/ISRCTN15366334).

Data analysis

There were four specific objectives: to (1) refine the Keele STarT MSK tool based on predictive performance and validity; (2) determine tool risk-strata cut-points based on optimal predictive performance and suitability for matched treatment options; (3) describe tool subgroups and construct validity; and (4) report external validity. The fourth objective was added following protocol

paper publication (Campbell et al., 2016); additional protocol paper objectives relating to qualitative and health economic analyses are or will be reported elsewhere (Saunders et al., 2016).

For analyses with dichotomous outcomes, poor outcome on the PCS was defined as scores <37.17 at 2-months and <39.61 at 6-months, based on lower tertiles from an independent study of UK primary care musculoskeletal pain patients (Salisbury et al., 2013). Poor outcome on pain intensity was defined as scores ≥5 (Von Korff et al., 1992). To provide clinical cut-offs of good / poor outcomes for the two main outcome measures, we pre-defined dichotomies in our protocol paper. The SF-36 PCS was dichotomised at 37.17 and 39.61 at 2- and 6-month follow-up (respectively), based on lower tertiles extracted from a similar cohort (Salisbury et al., 2013), and pain intensity divided at a score of 5 (a score of 5 or more denotes moderate/ severe pain (Von Korff et al., 1992)).

1. Refining the tool

An iterative process was used in tool refinement, considering improvements achieved (predictive, face and construct validity) compared with the draft tool when replacing existing items or adding items, one-by-one. This was carried out during testing (in the cohort study) as initially planned, plus following examination of the refined tool in the trial dataset (when items that were not treatment modifiable were considered).

Predictive performance was determined using linear regression of the association between baseline tool score and PCS and pain intensity at 2- and 6-month cohort follow-up. Performance was assessed based on model fit (R^2) and discrimination (C-statistic, with 95% confidence intervals (CI)) and calibration (calibration slope and Hosmer–Lemeshow test).

Item redundancy and weighting was investigated within multiple linear regression models for estimating PCS and pain intensity at 2- and 6-month cohort follow-up (i.e. 4 models). If items did not add significant predictive performance and/or if average standardized beta weight was small (i.e. <0.05) in most analyses, then the item was deemed statistically redundant. Face validity was

considered in research team decisions about removal of statistically redundant items. Variable weights were applied to reflect the strength of independent associations across non-redundant items (integer weights used to retain scoring simplicity).

2. Determining tool cut-points

The cut-point for identifying the high risk subgroup (versus medium / low risk) was based on classification on the full score most likely to attain positive predictive values and specificity ≥0.8, and positive likelihood ratio ≥5, for predicting PCS and pain intensity at 2- and 6-month cohort follow-up. The cut-point for categorising the low risk subgroup (versus medium / high risk) was based on the classification most likely to achieve negative predictive values and sensitivity ≥0.80, and negative likelihood ratio ≤0.2 (Hayden and Brown, 1999, Grimes and Schulz, 2005, Jaeschke et al., 1994). All decisions about tool cut-points were based on statistical information in the sample overall and within pain sites, plus suitability for matched treatments.

3. Describing the tool subgroups

Proportions classified into low, medium and high risk subgroups were described overall and by pain site. Means (with standard deviation; SD), medians (interquartile range; IQR) or frequencies and percentages (as appropriate) of outcomes were reported at cohort baseline, 2- and 6-months for each risk subgroup. Construct (discriminant) validity was assessed by testing differences between risk subgroups on baseline characteristics using ANOVA for linear tests and chi-square test-for-trend for categorical outcomes. Variations were examined across pain sites.

Tool cross-sectional construct validity was assessed by calculating agreement (percentage agreement and Cohen's kappa, 95% CI) of Keele STarT MSK tool stratification (low risk vs. combined medium and high risk subgroups) versus the two ÖMSPQ risk categories at baseline (using the cut-point proposed by the developers). For patients reporting LBP, baseline agreement of Keele STarT

MSK tool stratification into low, medium, and high risk subgroups versus STarT Back tool subgroups was calculated (Cohen's weighted kappa).

4. External validation

External validation of the Keele STarT MSK tool was examined in STarT MSK pilot trial data. Discriminant and predictive validity was investigated using model fit and discrimination as above. Descriptive analysis of outcomes within risk strata of the final Keele STarT MSK tool were investigated.

Missing data

Percentages of missing data for each variable were determined and patterns of missingness explored. Sensitivity analysis using multiple imputation (40 imputations) was conducted using tool items plus a range of baseline and follow-up variables encompassing the domains measured, via chained equations with predictive mean matching function for numerical variables and logit/ologit functions for categorical variables.

Ethical Approval

Ethical approval for the KAPS study was granted by South East Scotland Research Ethics Committee (14/SS/0083). Approval for the STarT MSK pilot trial was granted by NHS Health Research Authority (16/EM/0257). All participants gave informed consent to provide data.

RESULTS

Refinement and validation sample

4720 patients visited their GP about back, neck, knee, shoulder or multisite pain and were invited to participate in the cohort study. 2057 patients responded (43.6% response), and 1890 consented to

participate (40.2% adjusted response due to incomplete / ineligible questionnaires / refusals). The mean age of participants was 58.3 years (range 18 to 96 years), and 60.6% were female.

Over half of the cohort stated that the pain they visited their GP about (their index pain) was at multiple sites (51.5%). LBP was the next most common (21.6%), followed by knee (18.5%), shoulder (5.4%) and neck pain (3.0%). The mean baseline PCS score was 36.2 (SD 10.1), mean pain intensity was 5.3 (SD 2.4) and 21.7% of the sample reported having had their pain for less than 3-months. Further cohort characteristics are reported in Table 1.

Response at 2- and 6-months was 75.8% (n=1425) and 78.7% (n=1452) respectively. At 2-months and 6-months, mean PCS scores rose to 38.1 and 38.6 respectively (indicating improved physical health); 47.7% (n=560) and 53.4% (n=581) were categorised as having a poor outcome. Mean pain intensity at 2-months fell to 4.4, and 4.1 at 6-months with 45.6% (n=582) and 42.3% (n=482) categorised with a poor pain outcome. Mean pain interference score reduced to 60.1 and 59.1 respectively. In total, 17.8% indicated they were completely recovered or much improved at 2-months and 24.3% at 6 months.

External validation sample

1237 consultations were identified as back, neck, knee, shoulder or multisite pain in the trial. 524 patients (both study arms) returned baseline questionnaires with consent (42.3% response); their mean age was 61.1 years and 60.7% were female. The most common index pain site (coded by the GP) was LBP 29.6%, followed by knee 27.5%, shoulder 23.7%, neck 11.3% and multisite pain 8.0%. From baseline questionnaires, mean pain intensity was 6.2 (SD 2.3); further characteristics are presented in Table 1. Response at 6-months was 91.4% (n=479) and mean pain intensity dropped to 4.1 (SD 2.9) with 42.1% (n=201) categorised as having a poor outcome.

1. Refining the tool

In initial refinement stages, three items were replaced as they improved model fit and discrimination, as well as offering better face validity compared to original items. Model fit for PCS at 2-month follow-up improved to 0.405 and discrimination improved to 0.815, see Table 2. Hosmer-Lemeshow calibration chi square tests were 6.14 (P=0.631) at 2-months and 3.26 (P=0.917) at 6-months with calibration slopes of 1.003 and 1.004, respectively. Improvements were similar for pain intensity. Table 3 describes original and replacement items, with reasons for change.

This revised version was then examined within the external validation sample. Model fit dropped to 0.149 and discrimination fell to 0.69 against pain intensity at 6-months. Due to this drop, and because discrimination fell below the cut-off for acceptable discrimination (0.70), the decision was made to go back to the refinement and validation sample and investigate changes to further improve performance. Three additional items improved tool performance and / or improved face validity without detriment to performance, and were therefore added (Table 3). Item redundancy was subsequently investigated by examining the magnitude of the standardized beta coefficients (averaged across four models for predicting PCS and Pain at 2 and 6 months) and statistical significance for the predictors in the multiple linear regression model. On this basis, three tool items were considered redundant (items on dressing, anxiety and low mood), i.e. these items had average beta coefficients <0.05 and were mostly not statistically significant. In research team discussions, it was agreed to remove two items, but it was felt to be important for face validity to include an item on low mood given its clinical importance for primary care decision-making. Average beta values for all items are shown in Table 3. The beta value for pain intensity was much larger than the other items, so a decision was made to weight the scoring in favour of this item (compared to the others). There was little difference in performance statistics between a model using actual beta values as weights, and a simplified model where weights of 1 were given to all (retained) items except pain intensity, therefore a simplified approach was taken which assigned a weight of 3 to pain intensity and 1 to all the other items.

The final Keele STarT MSK tool comprises 10-items with model fit of 0.422 and discrimination 0.839 for PCS at 6-month cohort follow-up (Table 2). Performance was also perceived to be acceptable across the pain sites. Multiple imputation indicated that tool performance was robust to missing data: model fit was consistently above 0.4 and discrimination was consistently above 0.8, showing similar results as the primary (available case) analysis.

2. Tool cut-points for defining risk subgroups

The cut-points determined to provide the best combination of sensitivity, specificity, predictive values and likelihood ratios, in combination with suitability for matched treatments, overall and across pain sites, were 0-4 for low risk, 5-8 for medium risk, and 9-12 for high risk, on the full scale (see table S1). The cut-point of 0-4 / 5+ ensured a consistently high negative predictive value such that lower risk patients as classified by the tool (scores 0-4) would have over 80% chance of having low pain/ high function (and conversely less than 20% chance of having high pain / low function) and medium/higher tool classification (scores 5+) ensures over 80% of truly poor outcome patients are captured (sensitivity >0.8); whilst keeping specificity as high as possible. The cut-point of 0-8 / 9+ ensured a consistently high positive predictive value such that higher risk patients with scores between 9-12 would have over 80% chance of having high pain/ low function (and conversely less than 20% chance of having low pain / high function) and low/medium tool classification (scores 0-8) ensures over 80% of truly better outcome patients are not given high risk classification (specificity > 0.8); whilst keeping sensitivity as high as possible. These cut-points also provide consistent good performance across individual pain sites (as indicated by summary results in the table S1 footer).

3. Describing the tool subgroups

The proportion of cohort participants classified in low, medium and high risk strata based on the Keele STarT MSK tool were 24.9%, 41.7% and 33.4%, respectively. Characteristics of participants in each risk stratum are shown in Table 4. Overall, discriminant validity is clear: scores consistently demonstrate statistically significant differences between risk subgroups, with increasingly “better”

outcomes reported by participants stratified as medium or low risk. These patterns were still evident when the sample was stratified by pain site.

Cross-sectional construct validity was demonstrated through the 'moderate' kappa for agreement between stratification by the Keele STarT MSK tool and the ÖMSPQ: 0.49 and 0.48 'moderate' for the two possible Keele STarT MSK tool cut-offs, with overall agreement 76% and 73% (see Table 5[A]). The discordance in the [A] cross-tabulations is due to the Keele STarT MSK tool medium risk, which over-classifies low risk and under-classifies high-risk (when classified alongside low risk), and under-classifies low risk and over-classifies high risk (when classified alongside high-risk). For patients with LBP, the weighted kappa for agreement between stratification by the Keele STarT MSK tool and the STarT Back tool was 0.52 'moderate' and 0.64 'substantial' for linear and quadratic weights respectively; discordance was greater in respect of higher risk categorisation of the Keele STarT MSK tool than the STarT Back tool (i.e. off-diagonal counts being higher on the left-side of the cross-tab than the right-side at a ratio of 171:61) (see Table 5[B]). The observed strong correlation of 0.8 between the numerical scales for Keele STarT MSK tool versus ÖMSPQ and the STarT Back tool (see footer of Table 5) further indicates strong construct validity of the Keele STarT MSK tool.

4. External validation

In the external validation sample, the Keele STarT MSK tool demonstrated model fit of 0.224 and discrimination 0.73 for pain intensity at 6-month follow-up (Table 2). Among all baseline participants, 25.3% were classified at low risk, 50.0% medium and 24.7% high risk. When stratified by pain site, 12.8% (multisite pain) to 33.5% (neck or shoulder pain; combined due to low numbers) of participants were classified at low risk, with 16.5% (neck or shoulder pain) to 33.3% (multisite pain) classified at high risk. Mean 6-month pain intensity was 5.7 in the high risk subgroup, 4.1 in medium risk and 2.3 in the low risk subgroup; all differences between subgroups were statistically significant.

DISCUSSION

We have refined and validated the first musculoskeletal pain prognostic stratification tool specifically for use among primary care patients with the five most common musculoskeletal pain presentations (back, neck, knee, shoulder or multisite pain). The tool clearly and simply allocates patients to subgroups with distinct characteristics and different prognosis, and its performance is acceptable upon external validation.

This study has confirmed that generic prognostic factors can be combined to produce a stratification tool appropriate for use among patients with a range of musculoskeletal pain presentations. While there are no existing tools specifically designed for stratifying all primary care patients with musculoskeletal pain, comparison with the ÖMSPQ (developed to predict time to return-to-work following work-related soft tissue injuries (Linton et al., 2011, Linton and Boersma, 2003)) indicated moderate agreement. The tool also demonstrated substantial agreement with the STarT Back tool (Hill et al., 2008) among patients with LBP. Comparison with other instruments such as the Optimal Screening for Prediction of Referral and Outcome Yellow Flag Tool (developed in a physical therapy setting (Lentz et al., 2016)) or the Graded Chronic Pain Scale-Revised (developed in a general population sample (von Korff et al., 2020)) would also be helpful.

The approach of refinement and validation was robust and comprehensive, utilising information on item selection from systematic reviews of generic prognostic factors, a bespoke prospective cohort to test key aspects including face validity (using information both from clinicians and service users), construct validity (using available prognostic tools), predictive performance (using clinically meaningful outcomes), plus examination of external validity. Refining or updating an existing tool (as done here) is preferable to developing new ones, as updated versions are then based on both original and new data, leading to improved stability and generalisability (Moons et al., 2009). The plans for the cohort study and the refinement and validation of the tool were published (Campbell et al., 2016), although the process was amended following publication in order to include external

validation of predictive performance, which we felt was essential prior to the tool being disseminated and implemented.

There are also some limitations. In the refinement / validation (cohort) study, just over 40% of those invited took part. This could lead to bias if participants were systematically different to non-participants. This may result in differences in the proportions of patients in the risk subgroups, but is less likely to have influenced the refinement of the tool itself, as comparisons are internal within the dataset. There were large differences in the proportion with each pain site between the refinement / validation study and the external validation sample, most notably for multisite pain (51.5% cohort, 8.0% trial). This was predominantly due to the fact that cohort study patients self-reported their pain site, whereas participating GPs recorded index pain sites in the trial. That the tool demonstrated discrimination and prediction in both scenarios is an indicator of robustness, strengthened by the performance of the tool within groups with different pain sites. Performance in the refinement / validation cohort and the external sample was generally similar, indicating good generalisability, although predictive performance, as expected, was lower in the external sample. This may reflect differences in the sample, as although many characteristics of the two samples were similar (e.g. pain intensity 6.2 at baseline and 4.1 at follow-up in both), more people were allocated to high risk in the refinement / validation sample (33.4%), compared to the external validation sample (24.7%). Further external validation in other samples is needed to further examine this. In the external sample, the only identical outcome available was pain intensity, so it was not possible to assess external performance against physical health, or further examine construct validity. The analysis was undertaken using self-report data from postal questionnaires, and further testing with electronic versions may be needed. Our aim was to produce a tool for use among all patients consulting with back, neck, knee, shoulder or multisite pain, regardless of duration of pain, but further analysis among those with shorter or longer pain duration may provide further insights.

We have produced a prognostic stratification tool suitable for use among patients with the five most common musculoskeletal pain presentations in UK primary care (available at https://www.keele.ac.uk/startmsk/). Over 95% of the UK population are registered with a GP, and in light of demonstrated validity and predictive performance, the Keele STarT MSK tool may be broadly generalizable to people seeking healthcare for musculoskeletal pain in other settings, although further studies are needed to confirm this. The tool identifies distinct groups of patients with different prognoses and clearly identifiable characteristics, which can inform treatment decisions. For example, patients classified at low risk generally have a good prognosis, and may only need advice and support to self-manage. Patients classified at high risk have increased likelihood of a poor prognosis, a more complex clinical presentation, often with physical and/or psychosocial comorbidities, and may require more intensive healthcare intervention.

The Keele STarT MSK tool was developed as part of a programme of work investigating stratified primary care for patients with musculoskeletal pain. The main STarT MSK trial is ongoing, therefore it is inappropriate to speculate on whether using the tool, alongside matched treatment options (Protheroe et al., 2019), improves patient outcomes.

This paper reports the refinement and validation of a brief prognostic stratification tool for use among patients with the five most common musculoskeletal pain presentations in primary care: the Keele STarT MSK tool. It has demonstrated strong results in terms of validity and acceptable performance, and clearly identifies groups of patients with different characteristics and prognosis.

ACKNOWLEDGMENTS

The authors would like to thank the STarT MSK Programme Grant Steering Group overseeing the research, the staff from Keele Clinical Trials Unit, in particular Stephanie Tooth, Steff Garvin and Nicola Halliday, the Keele Patient and Public Involvement and Engagement (PPIE) group for their

support and assistance throughout both studies, the wider STarT MSK programme research team, and all the patients and GP practices for their participation. The STarT MSK research team acknowledges the support for recruitment of the National Institute for Health Research Clinical Research Network (NIHR CRN).

AUTHOR CONTRIBUTIONS

KMD, PC, ML, DAvdW, JCH, JP, SW, CDM, EH, SJ and NEF designed the study. NEF, KMD, JCH, JP, ML, DAvdW, SJ, CDM, EH obtained funding for the study. KMD, PC, EA, ML, JCH, SW and NEF collected the data. ML, EA, KMD, JCH, RO and PC were involved in data cleaning and analysis. KMD, PC, ML, JCH, EA, DAvdW and NEF contributed to the interpretation of the results. KMD developed the first draft of the manuscript. All authors were involved in critical revision of the manuscript for important intellectual content and approved the final version of the manuscript.

REFERENCES

Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W. H., Choi, S., Revicki, D., Cella, D., Rothrock, N., Keefe, F., Callahan, L. & Lai, J. S. 2010. Development of a PROMIS item bank to measure pain interference. *Pain,* 150**,** 173-82.

Artus, M., Campbell, P., Mallen, C. D., Dunn, K. M. & Van Der Windt, D. A. 2017. Generic prognostic factors for musculoskeletal pain in primary care: a systematic review. *BMJ Open,* 7**,** e012901.

Buchbinder, R., Van Tulder, M., Oberg, B., Costa, L. M., Woolf, A., Schoene, M. & Croft, P. 2018. Low back pain: a call for action. *Lancet,* 391**,** 2384-2388.

Campbell, P., Foster, N. E., Thomas, E. & Dunn, K. M. 2013. Prognostic indicators of low back pain in primary care: five-year prospective study. *J Pain,* 14**,** 873-883.

Campbell, P., Hill, J. C., Protheroe, J., Afolabi, E. K., Lewis, M., Beardmore, R., Hay, E. M., Mallen, C. D., Bartlam, B., Saunders, B., Van Der Windt, D. A., Jowett, S., Foster, N. E. & Dunn, K. M. 2016. Keele Aches and Pains Study protocol: validity, acceptability, and feasibility of the Keele STarT MSK tool for subgrouping musculoskeletal patients in primary care. *J Pain Res,* 9**,** 807-818.

Chenot, J. F., Greitemann, B., Kladny, B., Petzke, F., Pfingsten, M. & Schorr, S. G. 2017. Non-Specific Low Back Pain. *Dtsch Arztebl Int,* 114**,** 883-890.

Dunn, K. M. & Croft, P. R. 2005. Classification of low back pain in primary care: using "bothersomeness" to identify the most severe cases. *Spine,* 30**,** 1887-1892.

Dunn, K. M. & Croft, P. R. 2006. The importance of symptom duration in determining prognosis. *Pain,* 121**,** 126-132.

Dunn, K. M., Jordan, K. P. & Croft, P. R. 2010. Recall of medication use, self-care activities and pain intensity: a comparison of daily diaries and self-report questionnaires among low back pain patients. *Primary Health Care Research & Development,* 11**,** 93-102.

Foster, N. E., Hill, J. C., O'sullivan, P. & Hancock, M. 2013. Stratified models of care. *Best Practice & Research Clinical Rheumatology,* 27**,** 649-661.

Foster, N. E., Mullis, R., Hill, J. C., Lewis, M., Whitehurst, D. G., Doyle, C., Konstantinou, K., Main, C., Somerville, S., Sowden, G., Wathall, S., Young, J. & Hay, E. M. 2014. Effect of Stratified Care for Low Back Pain in Family Practice (IMPaCT Back): A Prospective Population-Based Sequential Comparison. *Ann Fam Med,* 12**,** 102-111.

GBD 2015 DALYs and HALE Collaborators. 2016. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet,* 388**,** 1603-1658.

Grimes, D. A. & Schulz, K. F. 2005. Refining clinical diagnosis with likelihood ratios. *Lancet,* 365**,** 1500-5.

Harland, N. J. & Georgieff, K. 2003. Development of the Coping Strategies Questionnaire 24, a Clinically Utilitarian Version of the Coping Strategies Questionnaire. *Rehabilitation Psychology,* 48**,** 296-300.

Hayden, S. R. & Brown, M. D. 1999. Likelihood ratio: A powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. *Ann Emerg Med,* 33**,** 575-80.

Henschke, N., Ostelo, R. W., Terwee, C. B. & Van Der Windt, D. A. 2012. Identifying generic predictors of outcome in patients presenting to primary care with non-spinal musculoskeletal pain. *Arthritis Care & Research,* 64**,** 1217-1224.

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., Bonsel, G. & Badia, X. 2011. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res,* 20**,** 1727-36.

Hill, J. C., Afolabi, E. K., Lewis, M., Dunn, K. M., Roddy, E., Van Der Windt, D. A. & Foster, N. E. 2016. Does a modified STarT Back Tool predict outcome with a broader group of musculoskeletal patients than back pain? A secondary analysis of cohort data. *BMJ Open,* 6**,** e012445.

Hill, J. C., Dunn, K. M., Lewis, M., Mullis, R., Main, C. J., Foster, N. E. & Hay, E. M. 2008. A primary care back pain screening tool: Identifying patient subgroups for initial treatment. *Arthritis Rheum,* 59**,** 632-641.

Hill, J. C., Whitehurst, D. G., Lewis, M., Bryan, S., Dunn, K. M., Foster, N. E., Konstantinou, K., Main, C. J., Mason, E., Somerville, S., Sowden, G., Vohora, K. & Hay, E. M. 2011. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet,* 378**,** 1560-1571.

Hingorani, A. D., Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G. & Hemingway, H. 2013. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ,* 346**,** e5793.

Jaeschke, R., Guyatt, G. & Sackett, D. L. 1994. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA,* 271**,** 389-391.

Jenkins, C. D., Stanton, B. A., Niemcryk, S. J. & ROSE, R. M. 1988. A scale for the estimation of sleep problems in clinical research. *J Clin Epidemiol,* 41**,** 313-321.

Jordan, K. P., Kadam, U. T., Hayward, R., Porcheret, M., Young, C. & Croft, P. 2010. Annual consultation prevalence of regional musculoskeletal problems in primary care: an observational study. *BMC Musculoskelet Disord,* 11**,** 144.

Kamper, S. J., Apeldoorn, A. T., Chiarotto, A., Smeets, R. J., Ostelo, R. W., Guzman, J. & Van Tulder, M. W. 2015. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: Cochrane systematic review and meta-analysis. *Bmj,* 350**,** h444.

Krumholz, H. M., Butler, J., Miller, J., Vaccarino, V., Williams, C. S., Mendes De Leon, C. F., Seeman, T. E., Kasl, S. V. & Berkman, L. F. 1998. Prognostic importance of emotional support for elderly patients hospitalized with heart failure. *Circulation,* 97**,** 958-964.

Lentz, T. A., Beneciuk, J. M., Bialosky, J.E., Zeppieri Jr, G., Dai, Y., Wu, S. S., George, S. Z. 2016 Development of a Yellow Flag Assessment Tool for Orthopaedic Physical Therapists: Results From the Optimal Screening for Prediction of Referral and Outcome (OSPRO) Cohort. *J Orthop Sports Phys Ther*, 46, 327-43.

Linton, S. J. & Boersma, K. 2003. Early identification of patients at risk of developing a persistent back problem: the predictive validity of the Örebro musculoskeletal pain questionnaire. *Clin J Pain,* 19**,** 80-86.

Linton, S. J., Nicholas, M. & Macdonald, S. 2011. Development of a short form of the Orebro Musculoskeletal Pain Screening Questionnaire. *Spine,* 36**,** 1891-1895.

Linton, S. J., Nicholas, M. & Shaw, W. 2018. Why wait to address high-risk cases of acute low back pain? A comparison of stepped, stratified, and matched care. *Pain*.

Mallen, C. D., Peat, G., Thomas, E., Dunn, K. M. & Croft, P. R. 2007. Prognostic factors for musculoskeletal pain in primary care: a systematic review. *Br J Gen Pract,* 57**,** 655-661.

March, L., Smith, E. U., Hoy, D. G., Cross, M. J., Sanchez-Riera, L., Blyth, F., Buchbinder, R., Vos, T. & Woolf, A. D. 2014. Burden of disability due to musculoskeletal (MSK) disorders. *Best Pract Res Clin Rheumatol,* 28**,** 353-366.

Moons, K. G., Altman, D. G., Vergouwe, Y. & Royston, P. 2009. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ,* 338**,** b606.

Morris, N. S., Maclean, C. D., Chew, L. D. & Littenberg, B. 2006. The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability. *BMC Fam Pract,* 7**,** 21.

Moss-Morris, R., Weinman, J., Petrie, K., Horne, R., Cameron, L. & Buick, D. 2002. The Revised Illness Perception Questionnaire (IPQ-R). *Psychology & Health,* 17**,** 1-16.

Muller, S., Thomas, E., Dunn, K. M. & Mallen, C. D. 2013. A prognostic approach to defining chronic pain across a range of musculoskeletal pain sites. *Clin J Pain,* 29**,** 411-416.

National Institute for Health and Clinical Excellence. 2016. Low back pain and sciatica in over 16s: assessment and management. NICE guideline.

New South Wales Agency for Clinical Innovation. 2016. Management of people with acute low back pain: model of care. Chatswood, Australia.

Nicholas, M. K. 2007. The pain self-efficacy questionnaire: Taking pain into account. *Eur J Pain,* 11**,** 153-163.

Nicholls, E. E., Hill, S. & Foster, N. E. 2013. Musculoskeletal pain illness perceptions: factor structure of the Illness Perceptions Questionnaire-Revised. *Psychol Health,* 28**,** 84-102.

Patel, S., Friede, T., Froud, R., Evans, D. W. & Underwood, M. 2013. Systematic review of randomized controlled trials of clinical prediction rules for physical therapy in low back pain. *Spine (Phila Pa 1976),* 38**,** 762-9.

Protheroe, J., Saunders, B., Bartlam, B., Dunn, K. M., Cooper, V., Campbell, P., Hill, J. C., Tooth, S., Mallen, C. D., Hay, E. M. & Foster, N. E. 2019. Matching treatment options for risk sub-groups in musculoskeletal pain: a consensus groups study. *BMC Musculoskelet Disord,* 20**,** 271.

Salisbury, C., Montgomery, A. A., Hollinghurst, S., Hopper, C., Bishop, A., Franchini, A., Kaur, S., Coast, J., Hall, J., Grove, S. & Foster, N. E. 2013. Effectiveness of PhysioDirect telephone assessment and advice services for patients with musculoskeletal problems: pragmatic randomised controlled trial. *Bmj,* 346**,** f43.

Saunders, B., Bartlam, B., Foster, N. E., Hill, J. C., Cooper, V. & Protheroe, J. 2016. General Practitioners' and patients' perceptions towards stratified care: a theory informed investigation. *BMC Fam Pract,* 17**,** 125.

Stanton, T. R., Hancock, M. J., Maher, C. G. & Koes, B. W. 2010. Critical appraisal of clinical prediction rules that aim to optimize treatment selection for musculoskeletal conditions. *Phys Ther,* 90**,** 843-54.

Thomas, E., Dunn, K. M., Mallen, C. D. & Peat, G. 2008. A prognostic approach to chronic pain: application to knee pain in older adults. *Pain,* 139**,** 389-397.

Van Wambeke, P., Desomer, A., Ailliet, L., Berquin, A., Demoulin, C., Dewachter, J., Dolphens, M., Forget, P., Fraselle, V., Hans, G., Hoste, D., Mahieu, G., Michielsen, J., Nielens, H., Orban, T., Parlevliet, T., Simons, E., Tobbackx, Y., Van Schaeybroeck, P., Van Zundert, J., Vanderstraeten, J., Vlaeyen, J. W. & Jonckheer, P. 2017. Summary: Low back pain and radicular pain: assessment and management. KCE report 287Cs. Good Clinical Practice (GCP) Brussels.

Von Korff, M. & Dunn, K. M. 2008. Chronic pain reconsidered. *Pain,* 138**,** 267-276.

Von Korff, M., Ormel, J., Keefe, F. J. & Dworkin, S. F. 1992. Grading the severity of chronic pain. *Pain,* 50**,** 133-149.

Von Korff, M., DeBar, L. L., Krebs, E. E., Kerns, R. D., Deyo, R. A., Keefe, F. J. 2020. Graded chronic pain scale revised: mild, bothersome, and high-impact chronic pain. *Pain*, 161, 651-661.

Vos, T., Flaxman, A. D., Naghavi, M., Lozano, R., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J. A., Abdalla, S., Aboyans, V., Abraham, J., Ackerman, I., Aggarwal, R., Ahn, S. Y., Ali, M. K., Alvarado, M., Anderson, H. R., Anderson, L. M., Andrews, K. G., Atkinson, C., Baddour, L. M., Bahalim, A. N., Barker-Collo, S., Barrero, L. H., Bartels, D. H., Basanez, M. G., Baxter, A., Bell, M. L., Benjamin, E. J., Bennett, D., Bernabe, E., Bhalla, K., Bhandari, B., Bikbov, B., Bin, A. A., Birbeck, G., Black, J. A., Blencowe, H., Blore, J. D., Blyth, F., Bolliger, I., Bonaventure, A., Boufous, S., Bourne, R., Boussinesq, M., Braithwaite, T., Brayne, C., Bridgett, L., Brooker, S., Brooks, P., Brugha, T. S., Bryan-Hancock, C., Bucello, C., Buchbinder, R., Buckle, G., Budke, C. M., Burch, M., Burney, P., Burstein, R., Calabria, B., Campbell, B., Canter, C. E., Carabin, H., Carapetis, J., Carmona, L., Cella, C., Charlson, F., Chen, H., Cheng, A. T., Chou, D., Chugh, S. S., Coffeng, L. E., Colan, S. D., Colquhoun, S., Colson, K. E., Condon, J., Connor, M. D., Cooper, L. T., Corriere, M., Cortinovis, M., De Vaccaro, K. C., Couser, W., Cowie, B. C., Criqui, M. H., Cross, M., Dabhadkar, K. C., Dahiya, M., Dahodwala, N., Msere-Derry, J., Danaei, G., Davis, A., De, L. D., Degenhardt, L., Dellavalle, R., Delossantos, A., Denenberg, J., Derrett, S., Des, J., D.C., Dharmaratne, S. D., Dherani, M., et al. 2012. Years lived with disability (YLDs) for 1160

sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet,* 380**,** 2163-2196.

Ware, J., Jr., Kosinski, M. & Keller, S. D. 1996. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care,* 34**,** 220-33.

Ware, J. E., JR. 2000. SF-36 Health Survey Update. *Spine,* 25**,** 3130-3139.

Whitehurst, D. G., Bryan, S., Lewis, M., Hay, E. M., Mullis, R. & Foster, N. E. 2015. Implementing stratified primary care management for low back pain: cost-utility analysis alongside a prospective, population-based, sequential comparison study. *Spine (Phila Pa 1976),* 40**,** 405-14.

Whitehurst, D. G., Bryan, S., Lewis, M., Hill, J. & Hay, E. M. 2012. Exploring the cost-utility of stratified primary care management for low back pain compared with current best practice within risk-defined subgroups. *Ann Rheum Dis,* 71**,** 1796-802.

Woolf, A. D. & Pfleger, B. 2003. Burden of major musculoskeletal conditions. *Bulletin of the World Health Organisation,* 81**,** 646-656.

Yelin, E., Weinstein, S. & King, T. 2019. An update on the burden of musculoskeletal diseases in the U.S. *Semin Arthritis Rheum*.

Table 1: Baseline characteristics of the KAPS cohort study (internal validation and refinement) and STarT MSK pilot trial (external validation) populations

	Cohort study (n=1890)	External validation (n=524)
Age, mean (SD)	58.3 (16.1)	61.1 (14.8)
Female, n (%)	1145 (60.6%)	318 (60.7%)
Index pain site, n (%)[1]		
Knee	349 (18.5%)	144 (27.5%)
Neck	57 (3.0%)	59 (11.3%)
Back	408 (21.6%)	155 (29.6%)
Shoulder	103 (5.4%)	124 (23.7%)
Multisite	973 (51.5%)	42 (8.0%)
Live alone, n (%)	394 (21.0%)	87 (16.6%)
Employed, n (%)	747 (41.1%)	234 (46.0%)
Time off work in last 6 months, n (%)	318 (16.8%)	66 (12.6%)
Pain @ consultation, mean (SD)	n/a	6.3 (2.0)
Pain intensity, mean (SD)		
Mean of least, average, & current pain	5.3 (2.4)	-
Usual pain	6.2 (2.5)	6.2 (2.3)
Duration: How long since no pain, n (%)		
< 3 months	403 (21.7%)	136 (26.0%)
3-6 months	225 (12.1%)	77 (14.7%)
7-12 months	212 (11.4%)	89 (17.0%)
1-5 years	521 (27.6%)	128 (24.4%)
6+ years	500 (26.5%)	94 (17.9%)
SF-36 Component Scales , mean (SD) (n=116 missing)		
Physical	36.2 (10.1)	n/a
Mental	43.6 (13.2)	n/a
PROMIS Pain interference, mean (SD) (n=46 missing)	62.1 (8.1)	n/a
Pain self-efficacy, mean (SD), (n=31 missing)	37.2 (16.1)	n/a
Catastrophising, mean (SD), (n=13 missing)	9.7 (8.9)	n/a
Long-term medical conditions, n (%)		
Diabetes	217 (11.5%)	54 (10.3%)
Breathing problems / COPD / Asthma	334 (17.7%)	92 (17.6%)
Heart problems or high blood pressure	579 (30.7%)	171 (32.6%)
Chronic fatigue, ME, fibromyalgia, WP	84 (4.5%)	43 (8.2%)
Anxiety, depression, stress	446 (23.6%)	100 (19.1%)
Other	495 (26.2%)	129 (24.6%)
Health Literacy problems, n (%)		
Never / rarely	1555 (82.3%)	472 (91.4%)
Sometimes / often / always	325 (17.3%)	44 (8.5%)
EQ-5D-5L, mean (SD)	0.56 (0.27)	0.56 (0.24)

[1] Self-reported in cohort study, coded by GP in external validation sample

Table 2: Model fit and discrimination of the draft, revised and finals versions of the Keele STarT MSK Tool

			SF-36 Physical Component Summary Score (PCS)		Pain intensity	
Tool version	**Dataset**	**Outcome point**	**Model fit (R^2)**	**Discrimination (c-statistic, 95% CI)**	**Model fit (R^2)**	**Discrimination (c-statistic, 95% CI)**
Draft Keele STarT MSK Tool	Cohort	2 months	0.337	0.791 (0.766, 0.817)	0.330	0.799 (0.774,0.823)
		6 months	0.334	0.804 (0.778, 0.830)	0.321	0.785 (0.758, 0.812)
Revised Keele STarT MSK tool	Cohort	2 months	0.405	0.815 (0.792, 0.839)	0.341	0.801 (0.778, 0.825)
		6 months	0.389	0.817 (0.793, 0.842)	0.331	0.783 (0.757, 0.809)
Revised Keele STarT MSK tool	External validation	6 months	n/a	n/a	0.149	0.685 (0.636, 0.735)
Final Keele STarT MSK tool	Cohort	2 months	0.423	0.818 (0.794, 0.842)	0.429	0.838 (0.816, 0.860)
		6 months	0.422	0.839 (0.816, 0.863)	0.430	0.822 (0.799, 0.846)
Final Keele STarT MSK tool	External validation	6 months	n/a	n/a	0.224	0.725 (0.679, 0.772)

Additional information on performance of the Final Keele STarT MSK tool in the cohort stratified according to pain site:
PCS at 2 months – R^2=.305 (AUC=.801) [Neck/Shoulder]; R^2=.404 (AUC=.803) [Back]; R^2=.369 (AUC=.798) [Knee]; R^2=.424 (AUC=.814) [Multi-site pain].
PCS at 6 months – R^2=.289 (AUC=.763) [Neck/Shoulder]; R^2=.436 (AUC=.833) [Back]; R^2=.351 (AUC=.816) [Knee]; R^2=.404 (AUC=.839) [Multi-site pain].
Pain at 2 months – R^2=.191 (AUC=.749) [Neck/Shoulder]; R^2=.382 (AUC=.831) [Back]; R^2=.435 (AUC=.831) [Knee]; R^2=.424 (AUC=.827) [Multi-site pain].
Pain at 6 months – R^2=.266 (AUC=.824) [Neck/Shoulder]; R^2=.410 (AUC=.822) [Back]; R^2=.344 (AUC=.794) [Knee]; R^2=.411 (AUC=.801) [Multi-site pain].

Table 3. Item changes made during refinement of the Keele STarT MSK Tool

Draft Keele STarT MSK Tool item	Reason the change was suggested	Final Keele STarT MSK Tool Item*
In the last two weeks, has your most painful area been in your hand/wrist/elbow or shoulder?	Draft tool item not applicable to all pain sites.	Do you have any other important health problems?
Do you feel that pain is terrible and it's never going to get better (yes to both)?	Draft tool item was reported to be difficult to interpret as it was two questions in one.	Do you think your pain condition will last a long time?
In the last 2 weeks, have you had pain in more than one part of your body?	Draft tool item was considered to be too inclusive and a more specific question eliciting information about more severe pain may be preferred.	Have you had troublesome joint or muscle pain in more than one part of your body?
n/a	Item added to improve model fit and discrimination of Final tool.	On average, how intense was your pain [where 0 is "no pain" and 10 is "pain as bad as it could be"]?
n/a	Item added to improve model fit and discrimination of Final tool.	Have you had your current pain problem for 6 months or more?
n/a	Item added to improve model fit and discrimination of Final tool.	Do you often feel unsure about how to manage your pain condition?

*The average beta values for the items (original and additional) were (in order of magnitude): 0.31 'average pain intensity'; 0.18 'pain duration'; 0.18 'walk short distances'; 0.13 'other important health problems'; 0.11 'pain condition will last a long time'; 0.08 'pain in more than one part of the body'; 0.08 'bothersomeness'; 0.07 'unsafe to be physically active'; 0.04 'dress more slowly'; 0.02 'stopped enjoying things'; 0.00 'worrying thoughts about pain'.

Table 4: Characteristics and outcomes in the cohort study (refinement and validation) population, overall and within subgroups defined by the draft and final Keele STarT MSK tools

			----------Draft Keele STarT MSK Tool--------			----------Final Keele STarT MSK Tool--------		
		All	High risk	Medium risk	Low risk	High risk	Medium risk	Low risk
SF-36 PCS, mean (SD)	Baseline	36.4 (10.1)	29.7 (7.6)	34.0 (8.5)	44.3 (8.2)	28.4 (7.3)	36.8 (8.1)	45.8 (8.0)
	2 months	38.1 (11.2)	31.3 (9.4)	36.3 (10.2)	45.5 (9.0)	29.7 (8.7)	38.7 (9.5)	47.1 (8.6)
	6 months	38.6 (11.4)	32.0 (10.1)	36.0 (10.6)	46.1 (8.7)	30.2 (8.8)	39.0 (10.0)	48.0 (8.2)
➢ 'Poor'[†], n (%)	Baseline	647 (40.0%)	379 (71.0%)	214 (44.7%)	54 (9.0%)	399 (73.8%)	227 (33.7%)	27 (6.6%)
	2 months	560 (47.7%)	273 (75.0%)	204 (56.8%)	83 (18.4%)	295 (80.2%)	233 (46.6%)	42 (13.4%)
	6 months	581 (53.4%)	256 (80.3%)	227 (66.4%)	98 (22.9%)	287 (87.5%)	257 (53.7%)	43 (15.1%)
Pain intensity, mean (SD)	Baseline	5.3 (2.4)	7.0 (1.8)	5.6 (1.7)	3.4 (1.9)	7.2 (1.6)	5.3 (1.7)	2.8 (1.6)
	2 months	4.4 (2.7)	6.2 (2.4)	4.5 (2.4)	2.7 (2.1)	6.4 (2.2)	4.2 (2.3)	2.2 (1.9)
	6 months	4.1 (2.8)	5.8 (2.6)	4.4 (2.5)	2.4 (2.1)	6.2 (2.3)	4.0 (2.4)	1.9 (1.9)
➢ 'Poor'[††], n (%)	Baseline	1009 (59.7%)	507 (88.5%)	344 (68.4%)	158 (25.7%)	525 (92.6%)	447 (63.2%)	51 (12.1%)
	2 months	582 (45.6%)	311 (76.4%)	184 (47.7%)	87 (18.0%)	333 (80.4%)	229 (42.6%)	34 (10.3%)
	6 months	482 (42.3%)	238 (69.8%)	170 (47.9%)	74 (16.7%)	263 (75.1%)	200 (40.7%)	33 (11.1%)
Promis pain interference scale, mean (SD)	Baseline	62.1 (8.1)	69.0 (5.2)	62.9 (5.2)	54.8 (6.5)	68.8 (4.9)	61.9 (5.6)	53.6 (6.6)
	2 months	60.1 (8.4)	66.1 (6.3)	60.4 (7.1)	54.2 (7.0)	66.5 (5.8)	59.7 (6.5)	52.9 (7.1)
	6 months	59.1 (9.0)	65.1 (7.4)	60.3 (7.6)	52.5 (7.3)	65.9 (6.7)	58.4 (7.3)	51.3 (7.4)
EQ5D-L, mean (SD)	Baseline	.56 (.27)	.35 (.28)	.58 (.20)	.75 (.13)	.33 (.26)	.62 (.18)	.78 (.11)
	2 months	.61 (.26)	.43 (.28)	.61 (.20)	.76 (.15)	.40 (.27)	.66 (.18)	.79 (.13)
	6 months	.62 (.26)	.44 (.30)	.61 (.22)	.78 (.14)	.42 (.28)	.66 (.19)	.81 (.15)
Pain Self Efficacy Questionnaire , mean (SD)	Baseline	37.2 (16.1)	24.3 (13.9)	37.0 (12.7)	50.0 (9.7)	24.3 (13.6)	39.3 (12.7)	51.6 (8.8)
	6 months	39.9 (16.1)	28.4 (16.0)	38.2 (14.0)	50.6 (10.2)	27.0 (14.5)	42.1 (13.2)	52.3 (10.0)

			----------Draft Keele STarT MSK Tool--------			----------Final Keele STarT MSK Tool--------		
		All	High risk	Medium risk	Low risk	High risk	Medium risk	Low risk
SF-36 Mental Component Score, mean (SD)	Baseline	43.6 (13.2)	34.8 (12.3)	44.6 (11.7)	51.4 (10.2)	35.1 (12.3)	45.4 (11.7)	52.4 (9.1)
	2 months	47.6 (12.3)	39.7 (13.3)	48.5 (11.1)	53.4 (8.4)	39.5 (12.9)	49.2 (10.8)	54.6 (7.0)
	6 months	47.7 (11.9)	40.4 (12.9)	47.5 (11.5)	53.6 (7.9)	40.2 (13.0)	49.2 (10.4)	54.1 (7.5)
Pain Catastrophising, mean (SD)	Baseline	9.7 (8.9)	16.4 (9.2)	9.1 (7.2)	4.0 (5.0)	16.3 (9.2)	8.3 (6.8)	3.4 (4.7)
	6 months	7.8 (8.4)	13.7 (9.5)	8.0 (7.6)	3.0 (4.4)	13.8 (9.3)	6.9 (7.1)	2.4 (4.0)
Sleep problems, n (%)	Baseline	1193 (63.1%)	461 (80.7%)	344 (68.8%)	269 (43.6%)	464 (82.1%)	449 (63.7%)	161 (38.4%)
	2 months	793 (56.7%)	300 (74.6%)	226 (59.8%)	185 (38.2%)	315 (77.6%)	301 (56.8%)	103 (31.1%)
	6 months	675 (54.3%)	243 (71.7%)	219 (62.2%)	144 (32.7%)	266 (76.4%)	261 (53.5%)	82 (28.0%)
Vigorous physical activity, n (%)	Baseline	710 (38.0%)	116 (20.4%)	186 (37.4%)	345 (56.0%)	106 (18.9%)	296 (42.0%)	252 (59.9%)
Global change “Much improved”, n (%)	2 months	249 (17.8%)	29 (7.4%)	56 (14.7%)	145 (30.1%)	28 (6.9%)	73 (13.7%)	122 (37.2%)
	6 months	353 (24.3%)	56 (13.3%)	69 (17.3%)	202 (41.0%)	38 (9.0%)	118 (21.0%)	167 (50.1%)

High versus Medium versus Low risk statistical testing: Between subgroup differences for all summary measures across each time-point (baseline, 2 and 6 months) for both Tool versions were statistically significant ($p<0.001$) through one-way ANOVA tests with linear contrast and non-parametric Jonckheere's trend (J-T) tests (for numerical outcomes) and by chi-square test-for-trend (for categorical outcomes).

† ‘Poor’ rating according to pre-assigned cut-points for the PCS of: <33.02 (baseline); <37.17 (2 months); <39.61 (6 months); †† ‘Poor’ rating for Pain across all time-points of ≥5.

Table 5: Construct Validity for Final Keele STarT MSK tool versus (A) ÖMSPQ and (B) STarT Back classifications

A.	ÖMSPQ		
	Lower risk#	**Higher risk#**	**Total**
(i) STarT MSK Tool			
Low risk	383 (22.6%)	39 (2.3%)	422 (24.9%)
Medium/High risk	366 (21.6%)	909 (53.6%)	1275 (75.1%)
Total	749 (44.1%)	948 (55.9%)	1697*
Kappa (95% CI)	0.49 (0.45, 0.54) [P<0.001]	Agreement: Observed = 76.1% Expected = 53.0%	
(ii) STarT MSK Tool			
Low/Medium risk	710 (41.8%)	420 (24.7%)	1130 (66.6%)
High risk	39 (2.3%)	528 (31.1%)	567 (33.4%)
Total	749 (44.1)	948 (55.9%)	1697*
Kappa (95% CI)	0.48 (0.44, 0.52) [P<0.001]	Agreement: Observed = 73.0% Expected = 48.1%	* 193 missing STarT MSK tool or ÖMSPQ scores

B.	STarT Back tool			
	Low risk	**Medium risk**	**High risk**	**Total**
STarT MSK Tool				
Low risk	105 (18.0%)	15 (2.6%)	2 (0.3%)	122 (20.9%)
Medium risk	77 (13.2%)	116 (19.9%)	44 (7.5%)	237 (40.6%)
High risk	6 (1.0%)	88 (15.1%)	131 (22.4%)	225 (38.5%)
Total	188 (32.2%)	219 (37.5%)	177 (30.3%)	584**
$Kappa_w$ (95% CI)^	0.52 (0.45, 0.58) [P<0.001]	Agreement: Observed = 79.5% Expected = 57.4%		** 151 missing STarT MSK tool or STarT Back tool scores
$Kappa_w$ (95% CI)^^	0.64 (0.56, 0.70) [P<0.001]	Agreement: Observed = 89.0% Expected = 69.4%		

Percentages within the table are % of the grand total.
Categorisation based on a cut-off of <50 (lower risk) and ≥50 (higher risk).
* Total study population ** Sub-population of participants with back pain (n, 735); back pain only (n, 408) and back pain as part of multisite pain (n, 327)).
^ Weighted-Kappa (linear weights) ^^ Weighted-Kappa (quadratic weights)
For the numerical scales the overall Pearson correlation coefficient between STarTMSK and OMSPQ was r=0.80; correlations by pain area were r=0.76 [Neck], r=0.80 [Back], r=0.71 [Shoulder], r=0.80 [Knee] and r=0.77 [Multisite]. Pearson's correlation between STarT MSK and STarT Back for the back pain subpopulation was r=0.80.