

Case-mix-adjusted benchmarking of musculoskeletal outcomes and opioid prescribing in English primary care: a prospective cohort study

HILL, Jonathan C <<http://orcid.org/0000-0001-6246-1409>>, YU, Dahai <<http://orcid.org/0000-0002-8449-7725>>, JORDAN, Kelvin P, BAILEY, James, WATHALL, Simon, THOMPSON, Clare, PARRY, Emma <<http://orcid.org/0000-0003-0278-6898>>, WILKIE, Ross <<http://orcid.org/0000-0003-4825-714X>>, BURGESS, Roanna, BRAYBROOKE, Alex and PEAT, George <<http://orcid.org/0000-0002-9008-0184>>

Available from Sheffield Hallam University Research Archive (SHURA) at:

<https://shura.shu.ac.uk/38069/>

This document is the Published Version [VoR]

Citation:

HILL, Jonathan C, YU, Dahai, JORDAN, Kelvin P, BAILEY, James, WATHALL, Simon, THOMPSON, Clare, PARRY, Emma, WILKIE, Ross, BURGESS, Roanna, BRAYBROOKE, Alex and PEAT, George (2026). Case-mix-adjusted benchmarking of musculoskeletal outcomes and opioid prescribing in English primary care: a prospective cohort study. *BMJ Open*, 16 (10): e118768. [Article]

Copyright and re-use policy

See <http://shura.shu.ac.uk/information.html>

BMJ Open Case-mix-adjusted benchmarking of musculoskeletal outcomes and opioid prescribing in English primary care: a prospective cohort study

Jonathan C Hill ¹, Dahai Yu ¹, Kelvin P Jordan,¹ James Bailey,¹ Simon Wathall,¹ Clare Thompson,¹ Emma Parry ¹, Ross Wilkie ¹, Roanna Burgess,¹ Alex Braybrooke,¹ George Peat ^{1,2}

To cite: Hill JC, Yu D, Jordan KP, *et al.* Case-mix-adjusted benchmarking of musculoskeletal outcomes and opioid prescribing in English primary care: a prospective cohort study. *BMJ Open* 2026;**16**:e118768. doi:10.1136/bmjopen-2026-118768

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2026-118768>).

RW Deceased

Received 24 February 2026
Accepted 21 August 2026



© Author(s) (or their employer(s)) 2026. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

¹School of Medicine, Keele University, Newcastle-under-Lyme, UK

²Centre for Applied Health & Social Care (CARE), Sheffield Hallam University, Sheffield, UK

Correspondence to

Dr Jonathan C Hill;
j.hill@keele.ac.uk

ABSTRACT

Objectives To quantify provider-level variation in musculoskeletal (MSK) health outcomes, consultation experience and opioid prescribing in general practice using a validated case-mix-adjustment model, and to evaluate the feasibility of applying Achievable Benchmark of Care (ABC) and funnel plots to derive performance benchmarks.

Design Prospective observational cohort study.

Setting 30 general practices across 13 primary care networks within a defined geographical region in England, UK (primary care).

Participants 1875 adults aged ≥18 years consulting general practice with non-inflammatory MSK pain between September 2021 and July 2022 who consented to data linkage. Of these, 1317 (70%) completed 3-month follow-up. Patients with inflammatory arthritis were excluded.

Interventions None.

Primary and secondary outcome measures Primary outcome was 3-month change in Musculoskeletal Health Questionnaire (MSK-HQ; range 0–56, higher scores better) score. Secondary outcomes were patient-reported poor consultation experience and opioid prescribing within 14 days of consultation. Multivariable case-mix models included demographic and clinical variables. Provider-level variation was assessed using funnel plots. ABC methodology defined achievable benchmarks based on top-performing providers.

Results Mean unadjusted 3-month MSK-HQ improvement was 4.9 points. After case-mix-adjustment, the weighted mean improvement was 4.8 points, with provider-level adjusted gains ranging from 3.0 to 5.7 points. ABC benchmarks were 5.7 points for MSK-HQ improvement, 4.4% for poor experience and 19.6% for opioid prescribing. Case-mix-adjustment altered provider rankings compared with crude analyses. Funnel plots identified two providers exceeding the 95% control limit for opioid prescribing; no outliers were observed for MSK-HQ.

Conclusions Case-mix-adjusted benchmarking of MSK care in general practice is feasible: combining case-mix models, ABC benchmarks and funnel plots offers a transparent framework for provider comparison. The low baseline response rate, which varied between networks, means the variation observed should be interpreted

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ This prospective cohort study linked patient-reported outcome measures with primary care electronic health records across 30 general practices and 13 primary care networks, enabling provider-level benchmarking in routine care.
- ⇒ Application of a previously validated musculoskeletal-specific case-mix-adjustment model supported fair comparison between providers with differing patient complexity.
- ⇒ Integration of Achievable Benchmark of Care methodology with funnel plots provided transparent identification of variation and potential outliers.
- ⇒ Non-random practice participation and a 14.9% baseline response rate may limit representativeness and introduce selection bias.
- ⇒ Case-mix models calibrated less consistently for consultation experience and opioid prescribing, indicating possible residual confounding and unmeasured practice-level or clinician-level influences.

cautiously rather than as an assessment of individual providers.

Trial registration number ISRCTN18132064.

INTRODUCTION

Musculoskeletal (MSK) conditions are among the most common reasons for primary care consultations, accounting for approximately one in seven general practice contacts in the UK.^{1 2} Evidence suggests wide variability in MSK management across providers and relatively modest improvements in patient outcomes.^{3–5} Despite this, unwarranted variation in MSK outcomes has rarely been systematically examined, and robust methods to identify outlier providers and support case-mix-adjusted performance benchmarking remain underdeveloped.

Audit and feedback improves care quality, particularly where baseline performance is

low or where an area has received limited prior attention, such as MSK care.^{6 7} In primary care, however, MSK-specific approaches to quality reporting and benchmarking remain limited. A critical prerequisite for effective feedback is the establishment of robust methods to generate fair benchmarks and to distinguish true variation in outcomes after appropriate case-mix-adjustment.

A recommended quality indicator for MSK care is overall health gain measured using the Musculoskeletal Health Questionnaire (MSK-HQ), a validated patient-reported outcome measure designed for use across a wide range of MSK conditions and care settings.^{8–10} The MSK-HQ captures domains including pain, stiffness, physical function, sleep, mood and self-management confidence. Its increasing use in the UK presents an opportunity to develop standardised outcome-based benchmarking in primary care.

Opioid prescribing differs fundamentally from these outcomes. For the MSK-HQ and for patient experience, higher scores are better, whereas for opioids in non-inflammatory MSK pain the guideline-concordant direction is lower: National Institute for Health and Care Excellence (NICE) guidance on chronic pain (NG193)¹¹ advises against initiating opioids for chronic primary pain, and their benefits in most MSK presentations are marginal. Benchmarks for prescribing should therefore be read as descriptions of current practice and as starting points for improvement rather than as clinically endorsed targets.

The Achievable Benchmark of Care (ABC) method derives 'real world' performance targets by calculating the pared mean of the highest-performing providers whose combined patient populations represent at least 10% of the total cohort. Unlike arbitrary thresholds, ABC produces stretch-but-attainable goals anchored in observed practice and has been associated with subsequent improvements in quality performance.^{12–14}

Fair benchmarking also requires adjustment for differences in patient complexity. Case-mix adjustment accounts for variation in patient characteristics such as age, comorbidity and baseline severity, ensuring that provider comparisons reflect genuine differences in care quality rather than differences in patient mix.^{15–17} We have previously developed an MSK-specific case-mix adjustment model for the MSK-HQ^{15 17} but its applicability to other important outcomes, including patient-reported consultation experience and opioid prescribing, remains uncertain. To date, no studies have integrated case-mix adjustment with ABC benchmarking and funnel plot methodology to evaluate variation in MSK care quality and outcomes in routine general practice.

The Multi-level Integrated Data for musculoskeletal health intelligence and ActionS (MIDAS) programme was established to strengthen the use of real-world data in MSK primary care.¹⁸ Using data from the MIDAS-GP cohort (ISRCTN18132064), this study aimed to:

1. Apply a validated case-mix adjustment model to quantify provider-level variation in 3-month MSK health outcomes, patient experience and opioid prescribing.
2. Explore the application of ABC methods and funnel plots to derive benchmarks and identify outlier providers across these indicators.

METHODS

Study design and setting

MIDAS-GP was a prospective cohort study conducted in general practices in North Staffordshire and Stoke-on-Trent, UK, between September 2021 and July 2022. The region included 70 practices organised into 13 primary care networks (PCNs). Practices were eligible if they had compatible clinical systems and text-messaging capability to support patient identification and recruitment. To ensure representation, we aimed to recruit at least one practice from each PCN, including those serving socioeconomically disadvantaged and ethnically diverse populations.

Participants

Patients were eligible if they were aged 18 years or older and had a recorded SNOMED CT (Systematised Nomenclature of Medicine—Clinical Terms) code, the structured clinical terminology used to record diagnoses and clinical activity in National Health Service (NHS) primary care electronic health records, for a painful MSK condition during the recruitment period. To support eligible case identification, a list of SNOMED CT Concept IDs tailored for UK primary care was developed.¹⁹ Patients with codes indicating inflammatory arthritis were excluded.

Practice staff, supported by the National Institute of Health Research (NIHR) Clinical Research Network: West Midlands, conducted weekly or fortnightly electronic health record searches to identify eligible patients. The index consultation was the first eligible MSK consultation recorded during the recruitment period and could relate to either a new or ongoing MSK problem. Recruitment was staggered across practices throughout the study period. Patients with an active mobile phone number recorded in their practice system were invited by SMS to complete an online questionnaire within a few days of their consultation. For patients without a mobile number, paper questionnaires were mailed to their home address. Participants provided informed consent for linkage of baseline and 3-month follow-up survey responses with their primary care electronic health records. The study size was determined pragmatically, with all eligible patients identified during the recruitment period invited to participate; no formal sample size calculation was undertaken.

Outcomes and measures

The primary outcome was 3-month MSK-HQ score, following their index consultation. The MSK-HQ is a validated 14-item patient-reported outcome measure scored from 0 to 56, with higher scores indicating better MSK

health.⁸ The primary analysis focused on provider-level change in case-mix-adjusted MSK-HQ scores between baseline and 3 months.

Secondary outcomes were (1) patient-reported consultation experience, assessed in the immediate postconsultation survey and dichotomised as poor ('fairly poor' or 'very poor') vs not poor. (2) Opioid prescribing within 14 days of the index consultation was extracted from the electronic health records for the full linked cohort (n=1875).

Case-mix variables

Adjustment variables were based on a previously validated MSK-specific case-mix model.¹⁵ These included age, sex and employment status (unemployed or health-related work absence vs other categories); baseline MSK-HQ score; days of moderate physical activity in the prior week; pain duration (<3 months, 3–6 months, 7–12 months, 13 months–3 years, or >3 years); pain site (1, 2, or ≥3 sites); history of MSK surgery (yes/no) and number of prior MSK pain episodes (0, 1, 2–3, 4–9, or ≥10). Comorbidities were assessed from primary care records in the preceding 5 years and categorised as 0, 1, 2, or ≥3. Comorbidity code lists were developed by cross-mapping the NICE multimorbidity indicator, the Charlson and Elixhauser indices, and other established case-mix-adjustment methods.¹⁹ Continuous variables (including baseline MSK-HQ score and days of moderate physical activity) were retained in their original scale. Categorisations for pain duration, number of prior episodes and comorbidity count were prespecified based on the previously validated MSK-specific case-mix model and chosen to reflect clinically meaningful groupings while ensuring adequate sample size within categories.

Statistical analysis

Analyses followed the NHS England methodology for case-mix-adjustment of patient-reported outcome measures.^{20 21} For the primary outcome (3-month MSK-HQ), a four-step analytical approach was used to examine variation across PCNs. First, a generalised linear fixed-effects model was used to estimate associations between baseline covariates (prespecified from our existing validated case-mix model) and 3-month follow-up MSK-HQ scores, with PCN included as a fixed effect. Variables with minimal contribution ($p>0.20$) were removed to avoid overfitting while maintaining consistency with the validated framework. Second, predicted patient-level MSK-HQ scores were calculated from the final model coefficients using the case-mix covariates only. The PCN fixed effect was excluded at the prediction stage, so that between-PCN differences remained in the residuals rather than being absorbed into the predicted values. Third, predicted scores were aggregated to PCN level by calculating the mean difference between observed and predicted scores within each PCN and adding this value to the overall sample mean, generating a case-mix-adjusted mean 3-month MSK-HQ score for each PCN.

Finally, funnel plots were used to assess PCN-level variation and identify potential outliers using 95% and 99.8% control limits ($\approx \pm 1.96 \times \text{SE}$ and $\pm 3.09 \times \text{SE}$) around the overall case-mix-adjusted mean (continuous outcome) or prevalence (binary outcomes). For the continuous outcome (3-month MSK-HQ), the SE was calculated as $\text{SE} = \text{SD} / \sqrt{n}$, where SD was taken as the pooled within-PCN SD and n was the number of participants contributing to the PCN estimate. For binary outcomes (poor experience and any opioid prescribing), the SE was calculated using the binomial model, $\text{SE} = \sqrt{[p_0(1-p_0)/n]}$, where p_0 is the overall case-mix-adjusted prevalence for that outcome; estimates were expressed as percentages for plotting.²² Lower control limits for binary outcomes were truncated at 0% (as prevalences cannot be negative).

Stratified funnel plots were constructed to explore socioeconomic inequalities, comparing patients in the most deprived Index of Multiple Deprivation (IMD) quintile (IMDq1) with those in quintiles 2–5. For each stratum, adjusted PCN-level estimates and control limits were derived using stratum-specific sample sizes.

For secondary outcomes (consultation experience and opioid prescribing), generalised logistic fixed-effects models were used to derive case-mix-adjusted PCN-level percentages, following the same four-step approach and again excluding the PCN term at the prediction stage. Funnel plots were again used to identify variation and outliers.

Benchmarks for each outcome were derived using the ABC method, defined as the weighted mean performance of the highest-performing providers whose combined patient populations represented at least 10% of the total cohort (n=1875). For consultation experience and opioid prescribing, this threshold was reached by aggregating the two highest-performing PCNs. For the MSK-HQ, the three highest-performing PCNs were tied on adjusted health gain and were therefore all included.

The primary analysis was restricted to observed data from baseline responders with linked electronic health records. To address missing data arising from loss to follow-up and incomplete covariate information, multi-level multiple imputation (patients nested within PCNs) was conducted with 30 imputed datasets.²³ Separate imputation models were developed for each outcome. Estimates were combined using Rubin's rules under the missing at random assumption. Imputed results were compared with complete-case analyses to assess the robustness of findings to missing data assumptions. All analyses were conducted using Stata MP V.19 (StataCorp). Further details of the statistical analyses are provided in online supplemental file 2.

Patient and public involvement

Patient and public involvement was embedded throughout the MIDAS research programme. A dedicated Patient Advisory Group (PAG), comprising seven standing members, met regularly (monthly) with members of the research team and contributed across the lifecycle of the



programme. Public contributors were involved in shaping research priorities, advising on questionnaire content and acceptability, and informing strategies to minimise exclusion and selective non-response.

PAG members advised on recruitment methods within general practice, including the use of SMS text messaging, online and paper-based questionnaires and the provision of telephone support to maximise accessibility. They advocated strongly for inclusion of practices serving socioeconomically deprived and ethnically diverse populations and contributed to the design of additional outreach approaches (including targeted doorstep surveying in MIDAS-POP). The PAG also helped prioritise exemplar analyses, including those relating to inequalities in care and outcomes.

Public contributors provided input into the interpretation and presentation of findings, emphasising the importance of clear, non-judgemental communication and avoiding stigmatisation when reporting variation or inequality. Their feedback informed the co-development of prototype dashboards and data visualisation tools to ensure outputs were understandable and meaningful to patients as well as clinicians and commissioners.

Public contributors were not involved in data collection itself, nor in the statistical analysis, but contributed to interpretation, dissemination planning and the broader governance of the programme.

RESULTS

A total of 13 447 patients from 30 general practices were invited to participate, including at least one practice from each of the 13 PCNs, and including four research-inexperienced practices serving higher proportions of patients from Black, Asian, Mixed/Multiple and other ethnic minority backgrounds. At baseline, 2008 patients (14.9%) responded, of whom 1875 consented and were successfully linked to their electronic health record (EHR), forming the primary analysis cohort. Of these, 1317 (70%) completed both the immediate post-consultation and 3-month follow-up surveys. Reasons for non-participation (eg, non-response to invitation or failure to complete follow-up) were not systematically recorded. Baseline characteristics are shown in [table 1](#).

Participant characteristics

The 1875 participants had a mean age of 57.7 years, 66% were female and had a mean body mass index (BMI) of 29.2 kg/m² (SD 6.9); most were in employment or retired (83.8%). Ethnic composition was predominantly White, with limited representation of other ethnic groups. Clinically, 13% reported prior MSK-related surgery, 40% had symptoms lasting less than 3 months and 46% reported 10 or more previous episodes of MSK pain. Pain was limited to a single site in 50% of participants, while 29% reported pain at three or more sites ([table 1](#)).

Participant characteristics across PCNs are summarised in online supplemental table 1. Modest between-PCN

Table 1 Characteristics of participants, overall

	Total
N	1875
Age, mean (SD)	57.7 (15.5)
Female	1233 (66)
BMI, mean (SD)	29.2 (6.9)
Ethnicity	
Asian	31 (2)
Black	28 (2)
Mixed/multiple	11 (1)
Other	17 (1)
White	1788 (95)
Employment status	
In paid employment/self-employed/retired/homemaker/voluntary work/student/other	1572 (83.8)
Missing/unable to work due to health/unemployed/seeking work	303 (16.2)
Previous MSK surgery	238 (13)
Duration of symptoms	
<3 months	758 (40)
4–6 months	205 (11)
7–12 months	251 (13)
13 months–3 years	290 (15)
>3 years	371 (20)
Previous episodes	
0	345 (18)
1	129 (7)
2–3	306 (16)
4–9	233 (12)
10+	862 (46)
Pain sites	
1 site	936 (50)
2 sites	390 (21)
≥3 sites	549 (29)
Physical activity (days), mean (SD)	2.2 (2.4)
Comorbidity count	
0	829 (44)
1	577 (31)
2	329 (18)
≥3	140 (7)
Baseline MSK-HQ, mean (SD)	25.3 (10.7)
Month-3 MSK-HQ, mean (SD)	30.2 (12.7)
Poor experience	119 (6)
Opioid analgesic	493 (26)

Numbers are n (%) unless otherwise stated.
BMI, body mass index; MSK, musculoskeletal; MSK-HQ, Musculoskeletal Health Questionnaire.

Table 2 Variation in MSK-HQ outcomes across PCNs: crude versus case-mix-adjusted

PCN	Crude				Case-mix-adjusted				
	Baseline, mean (SD) MSK-HQ	Mean MSK-HQ 3-month follow-up score	MSK-HQ health gain	Rank	Predicted MSK-HQ 3-month follow-up score	Difference between actual and predicted MSK-HQ score	Adjusted MSK-HQ 3-month score (mean 3m MSK-HQ (30.2) + difference)	Adjusted Gain (from baseline mean of 25.3)	Adjusted rank
1	22.2 (11.0)	27.2	5.0	4	26.4	0.8	31.0	5.7	1
2	22.4 (10.3)	26.4	4.0	9	26.6	-0.2	30.0	4.7	8
3	23.2 (10.3)	26.3	3.1	13	27.3	-1.0	29.2	3.9	12
4	26.1 (11.3)	30.4	4.3	7	30.8	-0.4	29.8	4.5	9
5	24.8 (10.0)	30.6	5.8	1	30.5	0.1	30.3	5.0	5
6	29.0 (11.6)	32.2	3.2	12	32.7	-0.5	29.7	4.4	11
7	27.3 (11.2)	32.2	4.9	6	32.1	0.1	30.3	5.0	5
8	25.6 (11.0)	31.4	5.8	1	30.6	0.8	31.0	5.7	1
9	26.9 (10.8)	31.9	5.0	4	31.1	0.8	31.0	5.7	1
10	24.3 (9.2)	28.4	4.1	8	28.4	0.0	30.2	4.9	7
11	25.1 (10.4)	29.1	4.0	9	29.5	-0.4	29.8	4.5	9
12	27.0 (11.4)	30.8	3.8	11	32.7	-1.9	28.3	3.0	13
13	25.0 (9.9)	30.5	5.5	3	29.9	0.6	30.8	5.5	4

Adjusted mean MSK-HQ 3-month follow up scores were then calculated by taking the average MSK-HQ 3-month follow up score for the whole dataset (30.2) and adding it to the mean difference between actual and predicted scores at PCN level to create an adjusted mean MSK-HQ 3-month follow up score per PCN.

Adjusted MSK-HQ health gain was then calculated as the adjusted mean MSK-HQ 3-month follow up score for each PCN and taking away the mean baseline MSK-HQ score for the dataset (25.3) to create the adjusted MSK-HQ health gain per PCN.

The weighted mean MSK-HQ gain across these PCNs was approximately 4.8 points.

The ABC benchmark weighted mean was 5.7, as all three highest-performing PCNs had the same adjusted MSK-HQ gain of 5.7 points.

ABC, Achievable Benchmark of Care; MSK-HQ, Musculoskeletal Health Questionnaire; PCN, primary care network.

variation included mean age ranging from approximately 50 to 66 years and female representation from 56% to 72%. Mean BMI ranged from 28.4 to 31.4 kg/m². Employment rates varied from 70% to 91%. The proportion reporting three or more pain sites ranged from 22% to 38%. Self-reported physical activity was stable across PCNs (mean 1.8–2.5 days/week).

Primary outcome: MSK-HQ

Across the cohort, mean baseline MSK-HQ score was 25.3 (SD 10.7), increasing to 30.2 (SD 12.7) at 3 months, corresponding to an unadjusted health gain of 4.9 points (table 2). Baseline scores ranged from 22.2 to 29.0 across PCNs, and unadjusted 3-month scores ranged from 26.3 to 32.2. Crude health gains ranged from 3.1 to 5.8 points.

Predictors of 3-month MSK-HQ score are shown in table 3. Older age was associated with slightly better outcomes, whereas unemployment or health-related work absence was associated with poorer outcomes. Multiple pain sites, longer symptom duration, frequent prior episodes and higher comorbidity burden were associated with worse outcomes. Baseline MSK-HQ score was the strongest predictor of follow-up health status. Sex and physical activity were not retained in the final model. Calibration plots demonstrated generally good model fit (online supplemental figure 1),

and the complete-case analysis of the MSK-HQ outcome produced estimates very close to those from the imputed analysis (online supplemental table 2).

After case-mix-adjustment, between-PCN variation narrowed (table 2). Adjusted 3-month MSK-HQ scores ranged from 28.3 to 31.0 and adjusted health gains from 3.0 to 5.7 points, with a weighted mean gain of 4.8 points. Several PCNs changed rank following adjustment, indicating confounding by patient complexity in crude comparisons. For example, PCN5 fell from joint highest to joint fifth, whereas PCN1 rose from joint fourth to joint highest. PCNs 6 and 12 remained among the lowest performers.

The ABC benchmark for adjusted MSK-HQ gain was 5.7 points, achieved by three PCNs. These PCNs also exceeded the published minimum clinically important difference of 5.5 points.⁸

Funnel plots (figure 1) identified no statistical outliers for MSK-HQ, with all PCNs within the 95% and 99.8% control limits. However, stratified analyses (online supplemental figure 2) demonstrated a socioeconomic gradient, with lower adjusted MSK-HQ scores among patients in the most deprived quintile (mean 26.9) compared with less deprived groups (mean 31.3).



Table 3 Coefficients of model predicting 3-month MSK-HQ score with adjustment of case-mix variables

Variables	Regression coefficients (95% CI)
Age, years	0.04 (0.00 to 0.07)
Female	Not selected
Employment status	
In paid employment/self-employed/retired/homemaker/voluntary work/student/other	Reference
Missing/unable to work due to health/unemployed/seeking work	-2.38 (-3.78 to -0.98)
Pain site	
1 site	Reference
2 sites	0.14 (-1.07 to 1.34)
≥3 sites	-1.16 (-2.38 to 0.05)
Duration of symptoms	
<3 months	Reference
4–6 months	-4.76 (-6.53 to -2.99)
7–12 months	-4.48 (-5.93 to -3.03)
13 months–3 years	-3.74 (-5.43 to -2.05)
Over 3 years	-5.19 (-6.69 to -3.68)
Previous MSK surgery	-0.90 (-2.22 to 0.43)
Comorbidity count	
0	Reference
1	-0.61 (-1.74 to 0.52)
2	-0.93 (-2.27 to 0.40)
≥3	-1.93 (-3.84 to -0.02)
Days of moderate physical activity in previous week	Not selected
Previous episodes	
0	Reference
1	0.01 (-2.05 to 2.08)
2–3	-3.06 (-4.71 to -1.40)
4–9	-4.43 (-6.25 to -2.61)
10+	-5.04 (-6.64 to -3.43)
Baseline MSK-HQ	0.69 (0.64 to 0.74)
Constant	17.67 (15.01 to 20.33)

MSK, musculoskeletal; MSK-HQ, Musculoskeletal Health Questionnaire.

Secondary outcomes: consultation experience and opioid prescribing

Overall, 6% of participants reported a poor consultation experience and 26% received an opioid prescription within 14 days (table 1). Poor experience was more common among younger participants (OR per year, 0.96; 95% CI, 0.95 to 0.97). Higher odds of opioid prescribing were observed among participants who were unemployed or unable to work due to health, those with pain at three or more sites and those with any recorded comorbidity, with a graded increase across categories (one comorbidity OR 1.55, 95% CI 1.21 to 1.99; two 1.75, 1.31 to 2.34;

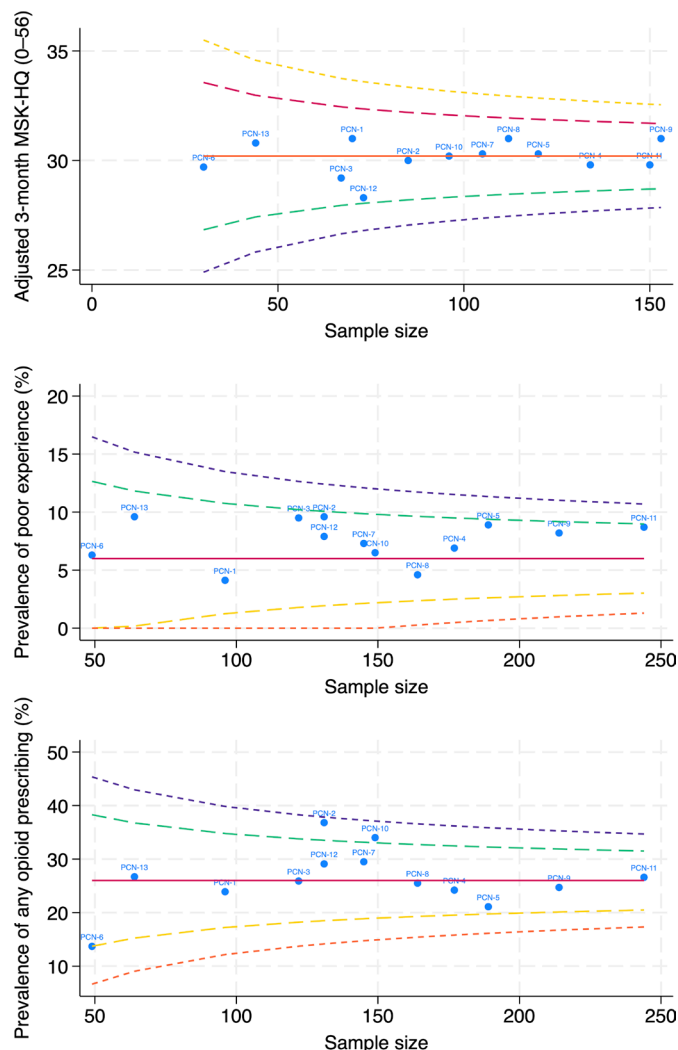


Figure 1 Funnel plots for adjusted outcomes: 3-month Musculoskeletal Health Questionnaire (MSK-HQ) score, poor experience and opioid prescribing within 14 days. This figure presents funnel plots for each primary care network (PCN) across three adjusted 3-month outcomes: (top) mean MSK-HQ score, (middle) prevalence of poor patient experience and (bottom) prevalence of any opioid prescription within 14 days of the index consultation. For each panel, the central horizontal line represents the overall case-mix-adjusted mean (continuous outcome) or prevalence (binary outcomes) across all PCNs. Curves show 95% and 99.8% control limits ($\pm 1.96 \times \text{SE}$ and $\pm 3.09 \times \text{SE}$) that narrow with increasing sample size (n). For MSK-HQ, SE was calculated as pooled within-PCN SD/ $\sqrt{n}$. For binary outcomes, SE was calculated using the binomial model, $\sqrt{[p_0(1-p_0)/n]}$, where p_0 is the overall case-mix-adjusted prevalence; estimates were expressed as percentages for plotting, and lower control limits were truncated at 0%. Points represent PCN-level case-mix-adjusted estimates, with n denoting the number of participants contributing to each PCN estimate. ABC benchmark thresholds (top-performing PCNs) are reported in online supplemental tables 4 and 5 for descriptive benchmarking and are not shown on the funnel plots, which display statistical control limits around the overall adjusted mean/prevalence.

Table 4 Association of case-mix predictors with adjusted poor experience and opioid prescribing

Variables	Poor experience	Any opioid prescription
Age, years	0.96 (0.95 to 0.97)	Not selected
Female	0.77 (0.52 to 1.13)	Not selected
Employment status		
In paid employment/self-employed/retired/homemaker/voluntary work/student/other	Not selected	Reference
Missing/unable to work due to health/unemployed/seeking work	Not selected	1.38 (1.05 to 1.81)
Pain site		
1 site	Reference	Reference
2 sites	0.68 (0.39 to 1.21)	0.95 (0.72 to 1.26)
≥3 sites	1.44 (0.97 to 2.16)	1.34 (1.05 to 1.71)
Duration of symptoms		
<3 months	Not selected	Reference
4–6 months	Not selected	0.94 (0.65 to 1.35)
7–12 months	Not selected	0.62 (0.43 to 0.90)
13 months–3 years	Not selected	1.03 (0.74 to 1.45)
Over 3 years	Not selected	1.36 (0.98 to 1.88)
Previous MSK surgery	Not selected	Not selected
Comorbidity count		
0	Reference	Reference
1	1.26 (0.82 to 1.92)	1.55 (1.21 to 1.99)
2	0.80 (0.43 to 1.48)	1.75 (1.31 to 2.34)
≥3	1.92 (1.00 to 3.69)	1.96 (1.33 to 2.89)
Days of moderate physical activity in previous week	Not selected	Not selected
Previous episodes		
0	Not selected	Reference
1	Not selected	1.58 (0.98 to 2.56)
2–3	Not selected	1.43 (0.98 to 2.09)
4–9	Not selected	1.55 (1.03 to 2.32)
10+	Not selected	1.23 (0.85 to 1.78)
Constant	0.54 (0.26 to 1.12)	0.22 (0.15 to 0.30)

Values are adjusted odds ratios (95% CIs) from the imputed main analysis. 'Not selected' indicates variables not retained in the final model. MSK, musculoskeletal.

three or more 1.96, 1.33 to 2.89) (table 4). Estimates for the secondary outcome models were more sensitive to the handling of missing data. In the complete-case analysis (online supplemental table 3), days of moderate physical activity was retained in both models whereas previous episodes was retained in neither, and several associations differed in magnitude, most notably for comorbidity count and number of pain sites in the poor experience model. Symptom duration did not show a consistent positive association with opioid prescribing, while previous episodes were retained in the opioid model, with evidence of increased odds for some categories (eg, 4–9 previous episodes), although associations were not consistent across all episode categories.

After case-mix adjustment, PCN-level variation remained (online supplemental tables 4 and 5). Adjusted rates of poor experience ranged from 4.1% to 9.6%, and opioid prescribing from 13.7% to 36.8%. The ABC benchmarks were 4.4% for poor experience and 19.6% for opioid prescribing. Funnel plots showed no outliers for poor experience, but PCN2 and PCN10 exceeded the 95% control limit for high opioid prescribing.

Stratified analyses of inequalities

Stratified funnel plots (online supplemental figures 2–4) revealed consistent socioeconomic gradients. In the most deprived group, mean adjusted MSK-HQ scores were lower than in less deprived groups. Within PCN3 and PCN8, outcomes in deprived populations fell below

the lower 95% control limit, whereas PCN7 and PCN4 exceeded the upper 99.8% limit, indicating high performance in deprived communities.

Adjusted prevalence of poor consultation experience was higher in the most deprived group (7.4%) than in less deprived groups (4.6%), with a similar pattern for opioid prescribing (31.0% vs 26.0%). Calibration plots (online supplemental figures 5 and 6) showed variable model fit across PCNs, suggesting that unmeasured practice-level or clinician-level factors may contribute to residual variation, particularly for prescribing and experience outcomes.

DISCUSSION

Principal findings

In this large prospective cohort of adults consulting primary care with non-inflammatory MSK pain, we identified provider-level variation in 3-month MSK health gain, consultation experience and opioid prescribing after applying a validated case-mix-adjustment model. The mean adjusted MSK-HQ improvement was 4.8 points, and adjustment materially altered PCN-level rankings compared with crude analyses, reinforcing the importance of accounting for patient complexity in performance comparisons. Approximately twofold variation was observed in consultation experience, and opioid prescribing varied considerably across providers. Using the ABC method, we derived a stretch benchmark of 5.7 points for MSK-HQ improvement, achieved by three PCNs, and applied funnel plots to distinguish expected variation from potential outliers. Together, these findings suggest that integrating case-mix adjustment, ABC thresholds and funnel plots provides a feasible framework for benchmarking MSK care. Socioeconomic stratification further showed consistently poorer outcomes and higher prescribing in more deprived populations, underscoring the importance of equity-aware performance reporting.

Strengths and limitations

This study used a well-characterised prospective cohort linked to electronic health records, enabling robust ascertainment of outcomes and prescribing. Regional PCN coverage enhanced representativeness, including participation by less research-active practices. Application of a validated MSK-specific case-mix model alongside ABC benchmarking and funnel plots supported transparent performance assessment. Several limitations merit consideration. Non-random practice participation and modest patient survey response rates may have introduced selection bias. The low baseline response rate (14.9%) raises the possibility of non-response bias. The direction of this bias is uncertain: patients who were most unwell or heavily medicated may have been less likely to respond, but those with more severe or persistent MSK symptoms may also have been more motivated to participate because the questionnaire was more relevant to them. The observed estimates and socioeconomic gradients

should therefore be interpreted cautiously. The baseline response rate also varied between networks, from 9.9% to 20.1% (online supplemental table 1). Because our comparisons are between networks, this differential non-response is a greater threat to validity than the overall response rate alone, since it could produce apparent between-network differences that reflect who responded rather than the care they received. Response to follow-up among those recruited was more complete, ranging from 54.9% to 82.9% at 3 months, so the principal concern relates to entry into the cohort rather than attrition from it. However, PCN coverage was broad and multiple imputation was used to mitigate bias arising from missing data. Limited ethnic diversity restricts generalisability to more heterogeneous populations. Follow-up was limited to 3 months, which may underestimate longer-term recovery trajectories. Although calibration was strong for MSK-HQ, it was weaker for consultation experience and opioid prescribing, and these two models were also less stable between the imputed and complete-case analyses. Together this suggests residual confounding or unmeasured practice- or clinician-level influences, and the case-mix models for experience and prescribing should be treated as provisional rather than ready for benchmarking use. Our prescribing outcome was a binary measure of any opioid prescription within 14 days of the index consultation, and we could not distinguish incident from continuation prescribing, opioid strength, or duration of use; the prescribing benchmark should not therefore be read as reflecting new prescribing decisions alone. The 3-month window also cannot capture longer-term opioid outcomes such as escalation, repeat prescribing or dependence, which are central to opioid-related harm. Future work with larger practice-level sample sizes should examine variation both between practices and within PCNs, as the current analysis was designed primarily for PCN-level benchmarking.

Comparison with other studies

Comparable benchmarking initiatives exist internationally. In the USA, the Center for Medicare & Medicaid Services quality programmes²⁴ and Patient-Reported Outcomes Measurement Information System (PROMIS)-based systems²⁵ incorporate provider-level performance measurement using patient-reported and process indicators. While structurally different, these initiatives share methodological principles with our approach including case-mix adjustment, data-driven benchmarking and funnel plots for comparative reporting. European benchmarking studies, including analyses from the STarT MSK trial, have demonstrated larger health gains under trial conditions. The smaller gains observed in MIDAS likely reflect routine practice constraints and post-pandemic access pressures, emphasising the importance of evaluating outcomes in real-world settings. Process-based benchmarking using ABC methods has also been applied to low back pain in Flemish general practice.²⁶ Our higher opioid benchmark reinforces the need for further

deprescribing initiatives in England while acknowledging contextual differences. The Flemish emphasis on continuous recalibration of pared-mean benchmarks aligns closely with our approach.

Meaning and implications

These findings operationalise prior UK recommendations advocating the integration of patient-reported outcome measures (PROMs), patient-reported experience measures (PREMs), case-mix adjustment and outlier detection methods for fair national benchmarking.²⁷ By combining these elements this study provides a single methodological framework for routine MSK audit and feedback in primary care. Socioeconomic stratification showed a consistent pattern across all three indicators. Patients in the most deprived quintile had lower adjusted MSK-HQ scores than those in less deprived groups and also had higher adjusted prevalence of poor consultation experience and opioid prescribing. These findings suggest that deprivation is associated not only with higher prescribing but also with poorer MSK outcomes and care experience. A benchmarking approach that targets only statistical outliers risks missing this broader equity gradient. Quality improvement based on these benchmarks should therefore be paired with an explicit equity lens and with investment in accessible non-pharmacological alternatives, particularly in more deprived communities. Routine identification of outliers and transparent benchmarking may therefore support targeted quality improvement while monitoring equity. Embedding case-mix-adjusted benchmarking within practice-facing dashboards offers a practical route to strengthen audit and feedback in primary care MSK services.

Case-mix adjustment is intended only to enable fair comparison between networks serving different populations; it is not a statement that higher prescribing in more vulnerable groups is clinically appropriate. Deprivation was deliberately not used as an adjustment variable but was examined through stratified analyses, so that socioeconomic inequities in prescribing remain visible rather than being adjusted away. The higher adjusted prescribing we observed in patients unable to work and in those with multiple comorbidities, groups at greater risk of opioid-related harm, is clinically concerning in its own right, and adjustment should not be used to explain away inequitable or potentially harmful prescribing. The ABC benchmark of 19.6% for opioid prescribing should likewise be read as a stretch-but-attainable description of current best practice within this sample and not as a clinically optimal or endorsed level of prescribing.

Unanswered questions and future research

Future work should refine case-mix models for consultation experience and prescribing outcomes, develop scalable EHR-based adjustment approaches to reduce reliance on self-reported variables and evaluate whether dashboard-enabled feedback leads to measurable

improvements in patient outcomes and prescribing behaviour.

CONCLUSIONS

Case-mix-adjusted benchmarking of MSK outcomes and prescribing in English primary care is feasible. Across 13 PCNs, average 3-month MSK-HQ improvement was modest and variation persisted for consultation experience and opioid prescribing after adjustment. We present this as a proof of principle for the method rather than as an assessment of the participating networks: baseline response rates varied from 9.9% to 20.1% between networks so we cannot exclude the possibility that some of the variation observed reflects differences in who responded rather than differences in care, and we would not support the use of these estimates for performance management in their current form. Establishing whether these differences are real will require higher and more even response rates, which are more likely to be achieved by collecting the MSK-HQ routinely at the point of care than through separate research invitations. Future priorities include refining models for experience and prescribing, developing scalable EHR-based adjustment tools and evaluating whether benchmarking feedback improves patient outcomes.

Acknowledgements The authors wish to acknowledge the invaluable contributions of Gerri Mulcahy and colleagues from the NIHR Clinical Research Network: West Midlands for their support with practice engagement and study delivery. We are grateful to the practice managers, general practitioners and administrative staff at participating practices for their collaboration in patient identification and recruitment. We also thank the team at MJog (by Livi) for their support with text-messaging infrastructure. We further acknowledge the contribution of Sarah Lawton, Steff Garvin, Jo Smith, Sarah Lewis, Rachael Heath, Jacqui Carter and the administrative support staff at Keele Clinical Trials Unit, who played a key role in the design and implementation of practice-based patient recruitment processes for MIDAS-GP. We are especially grateful to the patients who participated in the MIDAS-GP study.

Contributors JCH conceived the study, led protocol development, obtained ethical approval, designed the analytic approach and drafted the manuscript. DY conducted the statistical analyses and contributed to interpretation of the data. KPJ provided senior statistical oversight. JB led data management. SW developed and implemented the electronic health record search and extraction processes. CT coordinated the study, managed ethics processes and oversaw study delivery. EP contributed clinical expertise as a GP. RW and RB contributed to study design, interpretation and public health and clinical academic perspectives. AB contributed to development of code lists and benchmarking methodology. GP was chief investigator of the MIDAS programme and provided overall senior oversight. All authors contributed to interpretation of findings, critically revised the manuscript for important intellectual content, approved the final version and agree to be accountable for the work. JCH is the guarantor. The authors used an AI-based language model (ChatGPT, OpenAI) to assist with editing for clarity, grammar and word count reduction. The AI tool was not used for data analysis, interpretation of results, generation of tables or figures, or development of scientific conclusions. All content was reviewed, verified and approved by the authors, who take full responsibility for the integrity and accuracy of the manuscript.

Funding This work was supported by the Nuffield Foundation's Oliver Bird Fund and Arthritis UK (grant number OBF/43990).

Disclaimer The funders had no role in the design of the study; collection, analysis and interpretation of data; writing of the manuscript; or the decision to submit the article for publication.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Consent obtained directly from patient(s)

Ethics approval This study involves human participants and was approved by Yorkshire & The Humber – Leeds West Research Ethics Committee (21/YH/0178). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data are available upon reasonable request. Keele University is a member of the UK Reproducibility Network and is committed to the principles of the UK Concordat on Open Research Data. The School of Medicine and Keele Clinical Trials Unit make data available to bona fide researchers upon reasonable request via open or restricted access through a strictly controlled access procedure. In the first instance, data requests and enquiries should be directed to medicine.datasharing@keele.ac.uk.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <https://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Jonathan C Hill <https://orcid.org/0000-0001-6246-1409>

Dahai Yu <https://orcid.org/0000-0002-8449-7725>

Emma Parry <https://orcid.org/0000-0003-0278-6898>

Ross Wilkie <https://orcid.org/0000-0003-4825-714X>

George Peat <https://orcid.org/0000-0002-9008-0184>

REFERENCES

- 1 Arthritis V. The state of musculoskeletal health 2023. 2023. Available: <https://versusarthritis.org/media/duybjusg/versus-arthritis-state-msk-musculoskeletal-health-2023pdf.pdf> [Accessed 26 Aug 2025].
- 2 Jordan KP, Kadam UT, Hayward R, et al. Annual consultation prevalence of regional musculoskeletal problems in primary care: an observational study. *BMC Musculoskelet Disord* 2010;11:144.
- 3 Peat GM, Hill JC, Yu D, et al. Local variation in musculoskeletal pain consultation rates in primary care: findings from an ecologic study in Staffordshire. *Prim Health Care Res Dev* 2025;26:e52.
- 4 Cieza A, Causey K, Kamenov K, et al. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2021;396:2006–17.
- 5 Lin I, Wiles L, Waller R, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. *Br J Sports Med* 2020;54:79–86.
- 6 Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev* 2012;2012:CD000259.
- 7 Colquhoun HL, Carroll K, Eva KW, et al. Advancing the literature on designing audit and feedback interventions: identifying theory-informed hypotheses. *Implement Sci* 2017;12:117.
- 8 Hill JC, Kang S, Benedetto E, et al. Development and initial cohort validation of the Arthritis Research UK Musculoskeletal Health Questionnaire (MSK-HQ) for use across musculoskeletal care pathways. *BMJ Open* 2016;6:e012331.
- 9 Scott DrlC, McCray DrG, Lancaster ProfG, et al. Validation of the Musculoskeletal Health Questionnaire (MSK-HQ) in primary care patients with musculoskeletal pain. *Semin Arthritis Rheum* 2020;50:813–20.
- 10 Arthritis UK. Health intelligence. Available: <https://www.arthritis-uk.org/policy-and-data/health-intelligence/> [Accessed 27 Jan 2026].
- 11 National Institute for Health and Care Excellence. Chronic pain (primary and secondary) in over 16s: assessment of all chronic pain and management of chronic primary pain. NICE guideline NG193. 2021. Available: <https://www.nice.org.uk/guidance/ng193> [Accessed 18 Aug 2026].
- 12 Kiefe CI, Weissman NW, Allison JJ, et al. Identifying achievable benchmarks of care: concepts and methodology. *Int J Qual Health Care* 1998;10:443–7.
- 13 Hayashino Y, Suzuki H, Yamazaki K, et al. A cluster randomized trial on the effect of a multifaceted intervention improved the technical quality of diabetes care by primary care physicians: The Japan Diabetes Outcome Intervention Trial-2 (J-DOIT2). *Diabet Med* 2016;33:599–608.
- 14 Kiefe CI, Allison JJ, Williams OD, et al. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. *JAMA* 2001;285:2871–9.
- 15 Burgess R, Lewis M, Hill JC. Musculoskeletal case-mix adjustment in a UK primary/community care cohort: Testing musculoskeletal models to make recommendations in this setting. *Musculoskelet Sci Pract* 2021;56:102455.
- 16 Groenewegen PP, Spreeuwenberg P, Leyland AH, et al. Case-mix adjustments for patient reported experience and outcome measures in primary care: an empirical approach to identify patient characteristics as case-mix adjusters based on a secondary analysis of an international survey among patients and their general practitioners in 34 countries. *J Patient Rep Outcomes* 2023;7:127.
- 17 Burgess R, Lewis M, Hill JC. Benchmarking quality of care using patient reported outcome measure data for patients presenting with musculoskeletal conditions in primary care GP practices. *Musculoskeletal Care* 2023;21:713–22.
- 18 Keele University. MIDAS—multi-level integrated data for musculoskeletal health intelligence and actions. Keele University website. Available: <https://www.keele.ac.uk/midas/> [Accessed 26 Aug 2025].
- 19 Peat G, Bailey J, Jordan KP, et al. Multi-level integrated data for musculoskeletal health intelligence and actions (MIDAS). OSF. Available: <https://osf.io/e542w/> [Accessed 26 Aug 2025].
- 20 Department of Health. Patient Reported Outcome Measures (PROMs) in England: the case-mix adjustment methodology. 2012. Available: https://assets.publishing.service.gov.uk/media/5a7c36f440f0b67d0b11fa1e/dh_133449.pdf [Accessed 26 Aug 2025].
- 21 Nuttall D, Parkin D, Devlin N. Inter-provider comparison of patient-reported outcomes: developing an adjustment to account for differences in patient case mix. *Health Econ* 2015;24:41–54.
- 22 Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005;24:1185–202.
- 23 Austin PC, White IR, Lee DS, et al. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can J Cardiol* 2021;37:1322–31.
- 24 Centers for Medicare & Medicaid Services. Quality payment program: MIPS. Available: <https://qpp.cms.gov> [Accessed 27 Jan 2026].
- 25 Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care* 2007;45:S3–11.
- 26 Paridaens R, Vaes B, Van den Bulck S, et al. Benchmarks for low back pain in general practice in Flanders: electronic audit of INTEG0. *BMC Prim Care* 2024;25:431.
- 27 Burgess R, Lewis M, Hill JC. Benchmarking community/primary care musculoskeletal services: A narrative review and recommendation. *Musculoskeletal Care* 2023;21:148–58.