DARREN K. NEELEY, PT, DPT[1] • **STEVEN Z. GEORGE,** PT, PhD, FAPTA[2] • **KATE MINICK,** PT, DPT, PhD[1]
GREG SNOW, PhD[1] • **GERARD BRENNAN,** PT, PhD, FAPTA[1]

Four Variables Were Sufficient for Low Back Pain: Determining Which Patient-Reported Tools Predicted Pain and Disability Improvements

High-quality evidence syntheses support managing low back pain (LBP) with nonpharmacologic interventions.[34,35,43] In US health systems, some patients benefit from receiving physical therapy treatment, yet between 4 and 7 in every 10 patients, they do not achieve a minimal clinically important improvement in back-related disability by the end of an episode of care.[14,18,19] Identifying patients, early in the episode of care, who are less likely to respond to usual physical therapy interventions, would fast-track decisions about alternative clinical care pathways (eg, psychologically informed interventions) and/or referral to other providers.[15,28]

Multiple prognostic tools for managing LBP, including but not limited to the PICKUP tool and Örebro Musculoskeletal Pain Screening Questionnaire, have been well studied.[21,38,44,55] Psychologic and psychosocial measures are recommended in contemporary practice guidelines as prognostic factors for patients with LBP.[2,10] The STarT Back Screening Tool (SBST) is recommended for stratifying patients into high, medium, or low risk for prolonged LBP-related disability.[4,11,17,27,28] One advantage of risk stratification with the SBST is that it facilitates opportunities for tailoring treatment based on the risk level.[12,42]

One limitation of risk stratification is that it does not predict how much improvement in disability or pain one might expect. Broad prognostic categories do not account for predicting improvement at the level of the individual patient.[17,47] A second limitation is that the role of psychosocial factors has been emphasized, whereas other relevant prognostic factors have been neglected. Factors other than psychosocial could play a role in the patient's potential for improvement in disability and pain outcomes.[53] Other prognostic factors that merit investigation for predicting treatment outcomes include socioeconomic status,[30,48] comorbidities,[3,37] and symptoms from other body systems.[23] Including these factors along with established psychosocial risk tools offers an opportunity to further refine and/or improve prediction of disability and pain outcomes.[24]

The purpose was to identify parsimonious predictive models for 30- and 180-day

OBJECTIVE: To predict 30- and 180-day improvements in disability and pain for patients seeking physical therapy care for low back pain (LBP).

DESIGN: Longitudinal cohort.

METHODS: Baseline assessment was completed by 259 patients with chief complaint of LBP, and the assessment includes psychosocial measures (Keele STarT Back Screening [SBST] and the Optimal Screening for Prediction of Referral and Outcome Yellow Flag [OSPRO-YF] tools), the Optimal Screening for Prediction of Referral and Outcome Review of Symptoms (OSPRO-ROS) and the Review of Symptoms Plus (OSPRO-ROS+) tools, the Charlson Comorbidity Index (CCI), the Area Deprivation Index (ADI), and the National Institute of Health Chronic Pain Criteria (NIH-CP). Using the Modified Low Back Disability Questionnaire (MDQ) and the Numeric Pain Rating Scale (NPRS) as primary outcomes, statistical analysis determined multiple sets of predictor variables with similar model performance.

RESULTS: The parsimonious "best model" for prediction of the 180-day MDQ change included 3 predictors (Admit MDQ, NIH-CP, and OSPRO ROS+) because it had the lowest penalized goodness-of-fit statistic (BIC = −35.21) and the highest explained variance (R2 = 0.295). The parsimonious "best model" for 180-day NPRS change included 2 variables (Admit NPRS and OSPRO-ROS+) with the lowest penalized goodness-of-fit statistic (BIC = −18.2) and the highest explained variance (R2 = 0.190).

CONCLUSION: There were many model options with similar statistical performance when using established measures to predict MDQ and NPRS outcomes. A potential variable set for a standard predictive model that balances statistical performance with pragmatic considerations included the OSPRO-ROS+, OSPRO-YF, NIH-CP definition, and admit MDQ and NPRS scores. *J Orthop Sports Phys Ther 2022;52(10):685-693. Epub: 12 August 2022. doi:10.2519/jospt.2022.11018*

KEY WORDS: *back pain, disability, OSPRO, outcomes, pain, prediction, STarT Back*

[1]Intermountain Healthcare, Salt Lake City, UT. [2]Department of Orthopaedic Surgery and Duke Clinical Research Institute, Duke University, Durham, NC. ORCID: Minick, 0000-0003-4683-1937. This study was reviewed and approved by the Intermountain Healthcare IRB committee (IRB No. 1050946). Address correspondence to Darren K. Neeley, Intermountain Healthcare, 81 N Mario Capecchi Dr, Salt Lake City, UT 84113. E-mail: Darren.neeley@imail.org

improvements in disability and pain for patients seeking physical therapy care for LBP. The predictive models we investigated considered established measures from psychosocial and other clinically relevant domains. Our goal was to inform clinical decision-making by providing tailored outcome prediction using the fewest measures necessary to achieve reasonable clinical usefulness.

METHODS

This study included a longitudinal cohort of survey data for 259 participants presenting with a chief complaint of LBP with or without leg pain. Participants were recruited through email contact prior to their initial physical therapy (PT) appointment. At the end of each week, a report was generated that included the names and emails of patients who were scheduled for an initial evaluation, at any one of the 8 participating outpatient PT clinics across the Intermountain Healthcare network. Patients were then sent an institutional review board (IRB)-approved email informing them of the opportunity to participate in the study. Interested patients were invited to follow a secure link to a survey where they completed and signed an informed consent form and the baseline questionnaires. Survey data were collected and managed using REDCap electronic data capture tools hosted at Intermountain Healthcare.[25,26]

All follow-up data were collected electronically through a secure link in an automatically generated email using REDCap sent at 30-, 90-, and 180-days after baseline. The 30- and 180-day Modified Low Back Disability Questionnaire (MDQ) and Numeric Pain Rating Scale (NPRS) scores were primary endpoints in the analysis and reported in the manuscript; the 90-day MDQ and NPRS scores were reported in the **SUPPLEMENTAL FILE.**

Predictors

Psychosocial predictors included the SBST[28] and the total score for the Optimal Screening for Prediction of Referral and Outcome Yellow Flag (OSPRO-YF) tool.[6] The SBST has a total of 8 yes/no questions and a ninth question with a range of responses from "not at all" to "extremely." Items included on the SBST are referred to as leg pain, comorbid pain, disability, loathsomeness, catastrophizing, fear, anxiety, and depression. The scoring of the SBST tool includes 1 point for an answer of yes on questions 1-8 and an additional point for an answer of "very much" or "extremely" on question 9, for a total range of 0-9. A higher score is correlated with high psychosocial distress and poorer prognosis.[28] The reliability, validity, and predictive capabilities of the SBST in primary care and physical therapy settings have been well established.[4,32,51,52]

The OSPRO-YF includes 3 psychosocial domains (ie, negative mood, negative coping, and positive affect/coping). It differs from the SBST by allowing for a range in responses using a 10-point Likert scale and including psychosocial constructs, such as self-efficacy, that are not measured in the SBST. We used the 10-item scale, which has acceptable test-retest reliability and internal consistency and factorial, convergent, and know-groups validity.[6,46] We calculated the simple summary score,[24] with a higher score indicating higher psychosocial involvement, and a total range of 3-53.

A measure of contributing symptoms from other body systems was collected using the Optimal Screening for Prediction of Referral and Outcome Review of Symptoms (OSPRO-ROS) and the Review of Symptoms Plus (OSPRO-ROS+) tool.[23,24] Together, these measures cover 23 yes/no questions. Scoring is 1 point for every "yes" response so a higher score is more indicative of other body systems contributing to the patient's current pain problem (total range of 0-23). The predictive validity for the OSRPO-ROS+ tool has been established for persistent musculoskeletal pain 12 months after a physical therapy episode.[5]

Two additional measures were collected automatically through electronic medical record data at the time of participant initial assessment in PT: a measure of comorbidities present using the Charlson Comorbidity Index (CCI),[7] and a measure of socioeconomic status using the Area Deprivation Index (ADI).[33,49] The CCI is a tool that assigns a point value based on a patient's current comorbidities. The tool is scored from 0 to 33 with a higher score indicating more comorbidities. In a recent review of the clinometric properties of the CCI, the inter-rater reliability was excellent with extremely high agreement between self-report and medical charts. Additionally, the CCI had concurrent validity with several other prognostic scales or to result in concordant predictions and clinometric sensitivity in a variety of medical conditions.[36] The ADI is a standardized measurement of social determinants of health based upon United States census data with the intent of characterizing a patient's socioeconomic condition given their neighborhood. The ADI includes factors for the domains of income, education, employment, and housing quality and is frequently reported in quintiles, with 1 indicating the least deprived neighborhoods and 5, the most deprived.[33]

Chronic pain state was determined using the National Institute of Health Chronic Pain (NIH-CP) criteria.[13] These predictive variables align with recent recommendations for predicting LBP outcome.[24] The NIH-CP criteria includes 2 questions to define chronic LBP, classifying it by its impact (defined by pain intensity, pain interference, and physical function).[13] We classified individuals as having chronic LBP or not (a binary measure) at initial physical therapy assessment. Despite the recommendation to use the NIH-CP criteria in clinical practice and research, there are no published studies that have used this variable to predict clinical outcomes.

Outcomes

Primary outcomes were a measure of patient-reported functional disability and patient-reported pain using the MDQ[20] and the NPRS.[9] The MDQ is a modified version of the Oswestry Disability Index

and has 10 questions related to the impact of LBP on physical function. All 10 items are scored on a 0-5 scale and summed, for a total score of 0-50, which is then reported as a percentage of the total sum score (0-100%). Higher ODI scores represent higher self-report of disability. The MDQ has higher levels of test-retest reliability and responsiveness compared with the Quebec Low Back Disability Scale.[20] The NPRS is a numeric scale for pain intensity ranging from 0-10. Patients can indicate their level of pain intensity from 0, being no pain to 10 being the worst imagined pain. The total range of scores for the NPRS is from 0 to 10, with acceptable levels of reliability, validity, and responsiveness reported in the literature.[1,8,29]

Analysis

The leaps package[54] for the R statistical software[45] was used to find the "best fit" linear regression models for each possible number of predictors to compare competing models. The exhaustive search method was used for best subsets regression and is equivalent to fitting every possible combination of candidate variables into the prediction models. This is preferable to methods like forward or backward selection by showing multiple similar models that could be used interchangeably. The primary outcomes (one outcome for each set of model fit) were the 30- and 180-day change scores in MDQ and NPRS (subtract baseline score from corresponding 30- or 180-day score) and were considered dependent variables in our analysis. Candidate predictor variables were measured at admission (baseline) and all predictors were considered in the models. The same approach was taken with the 90-day scores reported in the **SUPPLEMENTAL FILE**.

Regression models were compared using the Bayesian Information Criterion (BIC), which measures goodness of fit penalized by model size. For the BIC criterion, lower values indicate better fitting models. Linear regression R^2 values were also used to assess model performance. For the R^2 criterion, larger values indicate better fit but there is no penalty for number of predictors. We planned to identify multiple sets of predictor variables that give similar quality of predictions, and in recommending "best" models placed a premium on parsimonious models (ie, fewest predictor variables for model performance).

RESULTS

FIGURE 1 provides a summary of recruitment, and **TABLE 1** provides a descriptive summary of this cohort of patients seeking initial consultation for LBP (n = 259). The follow-up rate was 85% (n = 220) at 30 days, 82% (n = 212) at 90 days, and 78% (n = 201) at 180 days. Data from the 30- and 180-day outcomes are reported in the paper; the 90-day outcomes are reported in the **SUPPLEMENTAL FILE**.

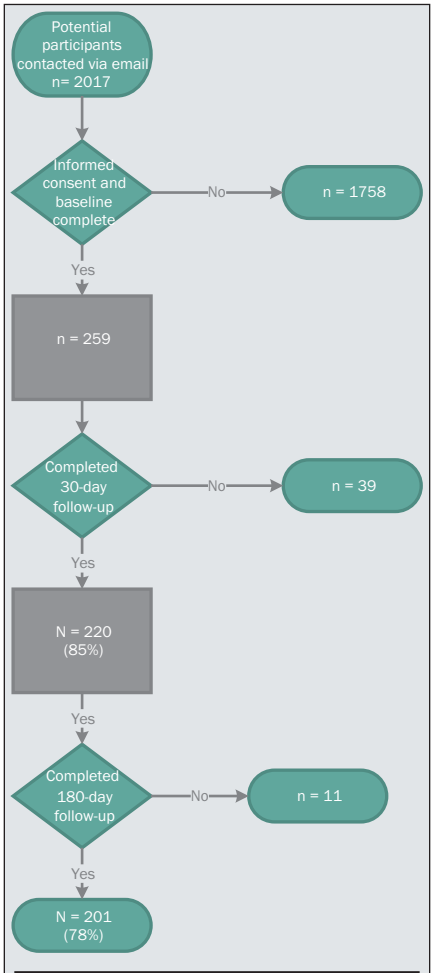

FIGURE 1. Participant recruitment and retention flow.

The focus of the study was the use of psychosocial and baseline characteristics. We looked at age and sex as predictors, and the best subsets regression models did on occasion include them as predictors, with age included more often than sex. However, they were not included in any of the best fitting parsimonious models (ie, those with 4 or fewer items), so they were not included in associated graphs and tables.

Individual Predictors for 30-Day Improvement Scores

Multivariate regression models including all predictor variables were first investigated for MDQ and NPRS improvement at 30 days (**TABLE 2**). In these models, the variables with the highest association with MDQ improvement, based on regression coefficients in **TABLE 2**, included the baseline MDQ score, the SBST, the OSPRO-YF, and the CCI. Variables with the highest association with NPRS improvement, based on regression coefficients in **TABLE 2**, included baseline NPRS score, the ROS score, the CCI score, and the OSPRO-YF. Please see **TABLE 2** for details.

Model Selection for 30-Day Improvement Scores

Predictive models were then explored for improvement in 30-day MDQ (**TABLE 3**) and NPRS (**TABLE 4**) scores. Variable combinations between 2-8 predictors were considered. For these models, the lowest BIC and highest R^2 combinations were for predictive models with 2-4 variables.

For the MDQ, the parsimonious predictive model that was selected as a "best model" included 2 variables (Admit MDQ and OSPRO-ROS+) with the lowest penalized goodness-of-fit statistic (BIC = −21.5) and the highest explained variance (R^2 = 0.21). Other predictive models for the MDQ reported in **TABLE 3** were considered "candidate models." These models included 2-4 variables based on BIC (ranging from −19.6 to −14.4) and R^2 values (ranging from 0.20 to 0.22).

TABLE 1 Baseline Descriptive Summary of Cohort Demographics[a]

	Age	Female	+ NIH-CP	Visits	MDQ	NPRS	CCI	ADI	ROS	ROS+	YF	SBST
Baseline	46.9 (16.47)	168 [65%]	142 [56%]	3.6 (2.88)	28.3 (15.96)	4.4 (2018)	0.9 (0.040)	91.5 (19.88)	3.8 (2.15)	2.5 (2.18)	20.6 (7.53)	4.1 (2.36)
30-day follow-up					19.5 (17.59)	3.6 (2.25)			2.2 (2.14)	1.3 (1.93)	15.5 (9.64)	2.5 (2.41)
180-day follow-up					16.7 (15.47)	3.0 (2.43)						

Abbreviations: ADI, mean total score on the Area Deprivation Index; Baseline, results at the time of the initial physical therapy assessment; CCI, mean total score on the Charlson Comorbidity Index; MDQ, mean score on the Modified Low Back Disability Questionnaire; + NIH-CP, percentage of patients who met the National Institute of Health Chronic Pain criteria; NPRS, mean score on the Numeric Pain Rating Scale; ROS, mean total score on the Optimal Screening for Predication of Referral and Outcome Review of Symptoms tool; ROS+, mean total score on the Optimal Screening for Predication of Referral and Outcome Review of Symptoms Plus tool; SBST, mean total score on the STarT Back Screening Tool; YF, mean total score on the Optimal Screening for Predication of Referral and Outcome Yellow Flag tool.
[a]*Mean (standard deviation).*

TABLE 2 30- and 180-Day MDQ and NPRS Improvement (Linear Regression Results)[a]

	30-Day MDQ Change	30-Day NPRS Change	180-Day MDQ Change	180-Day NPRS Change
R2	0.192	0.213	0.258	0.176
BIC	1283	687	1288	723
Baseline MDQ	0.40 (0.092)	−0.02 (0.014)[b]	0.52 (0.093)	−0.01 (0.016)[b]
Baseline pain	0.70 (0.623)	0.64 (0.096)	−0.32 (0.633)	0.62 (0.107)
ADI	−0.06 (0.049)[b]	0.00 (0.008)[b]	−0.03 (0.050)[b]	−0.00 (0.008)[b]
CCI	−3.35 (3.340)	0.18 (0.513)	−0.80 (3.394)	0.93 (0.574)
OSPRO-ROS	−0.36 (0.613)	−0.02 (0.094)[b]	−0.14 (0.623)	−0.09 (0.105)
OSPRO-ROS+	−1.15 (0.703)	−0.10 (0.108)	−1.38 (0.715)	−0.18 (0.121)
NIH-CP	−3.04 (1.941)	−0.15 (0.298)	−5.19 (1.972)	−0.16 (0.333)
OSPRO-YF	−0.03 (0.194)	−0.04 (0.030)[b]	0.16 (0.197)	−0.02 (0.033)[b]
SBST	0.04 (0.567)	−0.02 (0.087)	−0.10 (0.576)	−0.01 (0.097)

Abbreviations: ADI, Area Deprivation Index; BIC, Bayesian Information Criterion; Baseline MDQ, Modified Low Back Disability Questionnaire at initial physical therapy assessment; Baseline pain, numeric pain rating at initial physical therapy assessment; CCI, Charlson Comorbidity Index; NIH-CP, National Institute of Health Chronic Pain criteria; OSPRO-ROS, Optimal Screening for Predication of Referral and Outcome Review of Symptoms tool; OSPRO-ROS+, Optimal Screening for Prediction of Referral and Outcome Review of Symptoms Plus tool; OSPRO-YF, Optimal Screening for Predication of Referral and Outcome Yellow Flag tool; R2, linear regression R2 values; SBST, STarT Back Screening Tool.
[a]*Predictor raw coefficient value (standard error).*
[b]*Indicates P<0.05 for coefficient.*

For the NPRS, the parsimonious predictive model that was selected as a "best model" included 2 variables (Admit NPRS and OSPRO-YF) with the lowest penalized goodness-of-fit statistic (BIC = −28.0) and the highest explained variance (R^2 = 0.238). Other predictive models for the NPRS reported in **TABLE 4** were considered "candidate models." These models included 2-4 variables based on BIC (ranging from −26.0 to −20.4) and R^2 values (ranging from 0.228 to 0.250).

Individual Predictors for 180-Day Improvement Scores

TABLE 2 also reports the multivariate regression models for MDQ and NPRS improvement at 180 days. In regression models for the MDQ, predictor variables included the baseline MDQ, the ROS+, and chronic LBP state, and for the NPRS, predictor variables included the baseline pain, CCI, ROS+, ROS, and OSPRO-YF for 180-day NPRS change.

Model Selection for 180-Day Improvement Scores

Multiple predictive models were explored for 180-day improvement in MDQ scores (**TABLE 3**) and NPRS (**TABLE 4**). Variable combinations between 2-8 predictors and model fit were again considered. The lowest BIC and highest R^2 combinations were associated with predictive models with 2-4 variables. For the MDQ, the parsimonious predictive model that was selected as a "best model" included 3 predictors (Admit MDQ, NIH-CP, and OSPRO ROS+) because it had the lowest penalized goodness-of-fit statistic (BIC = −35.21) and the highest explained variance (R^2 = 0.295). Other predictive models for MDQ reported in **TABLE 3** were considered "candidate models." These models included 2-4 variables based on BIC (ranging from −31.16 to −29.04) and R^2 values (ranging from 0.267 to 0.296)

For the NPRS, the parsimonious predictive model selected as "best model" included 2 predictors (Admit NPRS and OSPRO-ROS+) because it had the lowest penalized goodness-of-fit statistic (BIC = −18.2) and the highest explained variance

TABLE 3 Predictor Subsets: 30-Day (30) and 180-Day (180) MDQ Change Scores[a]

No. of variables	MDQ	NPRS	NIH-CP	CCI	ADI	OSPRO-ROS	OSPRO-ROS+	OSPRO-YF	SBST	BIC	R2
30-Day MDQ Improvement Scores[b]											
2	0.43						−1.10			−21.5	0.206
2	0.36				−0.07					−19.6	0.196
2	0.40					−0.71				−19.5	0.196
3	0.44		−2.96				−1.22			−18.9	0.218
3	0.43				−0.06		−1.06			−18.1	0.214
3	0.42			−3.03			−1.08			−17.3	0.211
4	0.44		−2.97		−0.06		−1.17			−15.5	0.227
4	0.40	0.54	−3.20				−1.40			−14.6	0.222
4	0.43		−2.79	−2.49			−1.19			−14.4	0.221
180-Day MDQ Improvement Scores[c]											
2	0.43		−4.70							−33.9	0.266
2	0.50						−1.22			−32.4	0.259
2	0.51	−0.93								−30.4	0.249
3	0.52		−5.35				−1.44			−35.2	0.295
3	0.48		−4.87			−0.80				−31.2	0.276
3	0.50	−0.80	−4.46							−30.7	0.274
4	0.50		−5.38				−1.56	0.11		−30.6	0.296
4	0.54	−0.35	−5.19				−1.32			−30.5	0.296
4	0.52		−5.35		−0.02		−1.42			−30.4	0.296

Abbreviations: ADI, Area Deprivation Index; BIC, Bayesian Information Criterion; CCI, Charlson Comorbidity Index; MDQ, Modified Low Back Disability Questionnaire; NIH-CP, National Institute of Health Chronic Pain criteria; NPRS, Numeric Pain Rating Scale; OSPRO-ROS, Optimal Screening for Prediction of Referral and Outcome Review of Symptoms tool; OSPRO-ROS+, Optimal Screening for Prediction of Referral and Outcome Review of Symptoms Plus tool; OSPRO-YF, Optimal Screening for Prediction of Referral and Outcome Yellow Flag tool; R2, linear regression R2 values; SBST, STarT Back Screening Tool.

[a] *Raw coefficient values are for baseline data.*

[b] *Dark gray shading indicates important variables in "best model" and "candidate models" for 30-day MDQ change scores.*

[c] *Light gray shading indicates important variables in "best model" and "candidate models" for 180-day MDQ change scores.*

(R^2 = 0.190). Other predictive models for the NPRS reported in **TABLE 4** were considered "candidate models." These models included 2-4 variables based on BIC (ranging from −16.8 to −12.8) and R^2 values (ranging from 0.183 to 0.213).

DISCUSSION

We used established measures to identify parsimonious models that predicted 30- and 180-day improvement in disability and pain for patients seeking physical therapy for LBP. The analysis meant we could recommend "best parsimonious model" options using BIC and R^2 criteria. Recommended models explained 18%-21% and 27%-30% variance for 180-day change in pain intensity and disability scores respectively while including 2-4 of the established measures. However, there were many other candidate models that performed similarly statistically to the recommended models.

Baseline Scores Are an Important Part of Each Prediction Model

The variation observed in the number and type of variables included in the predictive models suggests there are various implementation options for forecasting 30- and 180-day outcomes. One common feature across all models was the importance of baseline MDQ and NPRS scores, which are typically the strongest predictor of 30- and 180-day improvements. This is not a surprising finding, but it is worth reinforcing the predictive value of baseline scores for frameworks where predicting patient outcomes is the goal.[24] Our results converge with other analyses indicating that psychosocial risk stratification measures have less predictive value when admit scores are included in the model.[16] A novel finding of our study was that the NIH-CP definition for chronic LBP[13] and the OSPRO-ROS+ tool[22,23] were consistently included in parsimonious prediction models for 30- and 180-day disability outcome.

Potential for Clinical Application

Collectively, our results suggest 3 important findings: (1) there was only a modest amount of variance explained by these models; (2) for these outcomes, there appears to be an upper limit of accuracy for 4 predictors when using these established measures; and (3) there is notable flexibility in which established measures

TABLE 4 Predictor Subsets: 30-day (30) and 180-day (180) NPRS Change Scores[a]

No. of variables	MDQ	NPRS	NIH-CP	CCI	ADI	OSPRO-ROS	OSPRO-ROS+	OSPRO-YF	SBST	BIC	R2
30-Day NPRS Improvement Scores[b]											
2		0.54						−0.07		−28.0	0.238
2	−0.03	0.58								−26.0	0.228
2		0.52					−0.21			−23.3	0.215
3	−0.02	0.61						−0.05		−25.4	0.250
3		0.58					−0.12	−0.06		−24.5	0.245
3		0.57				−0.08		−0.07		−23.9	0.242
4	−0.02	0.63					−0.10	−0.04		−21.5	0.255
4	−0.02	0.62				−0.06		−0.05		−20.9	0.252
4	−0.02	0.60		0.17				−0.05		−20.4	0.250
180-Day NPRS Improvement Scores[c]											
2		0.54					−0.27			−18.2	0.190
2		0.51				−0.22				−16.8	0.183
2	−0.03	0.54								−15.0	0.173
3		0.53		1.04			−0.27			−16.8	0.208
3	−0.02	0.61					−0.23			−15.2	0.200
3		0.57				−0.13	−0.20			−14.9	0.199
4	−0.02	0.60		0.95			−0.23			−13.2	0.215
4		0.56		1.05			−0.22	−0.03		−12.9	0.214
4		0.56		0.95		−0.11	−0.21			−12.8	0.213

Abbreviations: ADI, Area Deprivation Index; BIC, Bayesian Information Criterion; CCI, Charlson Comorbidity Index; MDQ, Modified Low Back Disability Questionnaire; NIH-CP, National Institute of Health Chronic Pain criteria; NPRS, Numeric Pain Rating Scale; OSPRO-ROS, Optimal Screening for Prediction of Referral and Outcome Review of Symptoms tool; OSPRO-ROS+, Optimal Screening for Prediction of Referral and Outcome Review of Symptoms Plus tool; OSPRO-YF, Optimal Screening for Prediction of Referral and Outcome Yellow Flag tool; R2, linear regression R2 values; SBST, STarT BACK Screening Tool.

[a] *Raw coefficient values are for baseline data.*

[b] *Dark grey shading indicates important variables in "best model" and "candidate models" for 30-day NPRS change scores.*

[c] *Light grey shading indicates important variables in "best model" and "candidate models" for 180-day NPRS change scores.*

(eg, CCI, SBST, OSPRO, etc) can be used in models predicting these outcomes. Clinically, this flexibility could be advantageous for implementation efforts that want to support decision-making at the patient level. Our results support healthcare delivery systems, with established measures in their standard clinical workflow, which may have some predictive value. The flexibility means different health systems might use different measures to generate outcome predictions with similar predictive accuracy.

Proposed Standard Prediction Model

There are advantages to implementing a single prediction model that would enable prediction of both 30- and 180-day MDQ and NPRS improvement scores. We found statistical evidence to recommend using a standard predictive model. The foundation for the model would include the baseline MDQ and NPRS scores as these were the strongest individual predictors in the corresponding 30- and 180-day change scores. The third variable to consider in a standard model would be the OSPRO-ROS+ tool as it consistently contributed to models predicting 30- and 180-day improvement scores.

Two other variables worth considering in a standard predictive model are the NIH-CP status and/or the OSPRO-YF tool(s). The NIH-CP variable contributed to all models predicting 180-day MDQ improvement, whereas the OSPRO-YF variable contributed to all models predicting 30-day NPRS improvement. Based on our findings, a standard prediction model that could be implemented and tested clinically for further refinement would include the OSPRO-ROS+, NIH-CP status, and/or OSPRO-YF tool in addition to the baseline MDQ and NPRS scores. The findings for the OSPRO tools converge with results from a musculoskeletal pain cohort in which the OSPRO-YF tool predicted 12-month function and pain outcomes,[23] whereas the ROS+ predicted persistent pain state at 12 months.[5] This study is the first prospective validation of these OSPRO tools, but future research is necessary to validate the accuracy of the proposed standardized model and its ability to support clinical decision-making.

SBST or OSPRO-YF as Part of the Model?

Psychosocial factors are an important part of clinical decision-making, so these measures may still need to be incorporated

for intervention type and/or care process (eg, psychologically informed physical therapy,[31,40] multidisciplinary care, consultation with other healthcare providers, etc). In this analysis, candidate models for predicting short-term MDQ improvement included admit MDQ, NIH-CP, and/or CCI, and either the SBST or OSPRO-YF tool. Neither the SBST nor the OSPRO-YF were significant factors in predicting 180-day outcomes. From a clinical standpoint, it may be prudent to include yellow flag assessment tools in a predictive framework, as there already exist treatment pathways that are related to these assessments, including but not limited to psychologically informed PT interventions.[31,40]

The decision as to which yellow flag tool to implement is not aided by statistical findings and instead should be determined by what is readily available and meets the needs of the patient population. While the OSPRO-YF tool performed slightly better in some models, there was no clear statistical advantage over the SBST. This means clinical use of the OSPRO-YF or SBST should be made based on what is needed to support decision-making. If risk stratification and applying matched treatment is the primary goal, use the SBST.[41] If there is interest in characterizing the number of yellow flags across positive coping, negative coping, and negative mood measures to support subsequent clinical actions, use the OSPRO-YF tool.[50]

OSPRO-ROS+ and NIH-CP May Have Predictive Value

Some other clinically relevant findings included the OSPRO-ROS+ tool and the NIH-CP criteria both showing up repeatedly in many of the 30- and 180-day outcome prediction models. The OSPRO-ROS+ is intended as a comprehensive red flag screening tool, but from our analysis, it appears to also be an important variable in predicting 30- and 180-day pain and MDQ improvement. Interestingly, this finding for the OSPRO-ROS+ tool converges with it being a predictor of musculoskeletal pain in an analysis of another physical therapy cohort.[5] The NIH-CP variable being included in these models was expected, but other variables such as the ADI and the CCI did not have the impact on forecasting models that we expected. This could be because the recruitment area for participants in our study did not have significant social determinants of health disparities and our population cohort was relatively healthy with minimal comorbidities.

Limitations

We were unable to do separate analyses for patients who met the NIH-CP criterion. Originally, we had planned to recruit sufficient participants to allow for separate analyses. However, due to SARS-CoV-2 changes in research protocols, we had to end enrollment in our study sooner than anticipated. We modified our original plan by incorporating the NIH-CP status in the predictive models. Another limitation is an absence of measures to capture constructs that could explain individual variance beyond the measures used in our predictive models. These additional constructs would involve measures other than self-report and could include biological markers, functional imaging (ie, cortical activity), and/or patient behaviors.

Participants represented 1 geographical area of a relatively healthy population. Accordingly, our findings may not generalize to other geographical areas or to populations with higher comorbidity rates. Finally, because we aimed to determine the extent to which prediction tools could provide prognostic value, the type and quantity of physical therapy interventions provided to study participants was left to the discretion of the treating physical therapist, and treatment parameters were not included in the statistical analysis. Future research should include validation of model performance in an independent cohort by comparing predicted to actual patient outcomes.[39] Furthermore, whether clinical implementation of these predictive models better guides clinical pathway options could also be a topic for future research.

CONCLUSION

The derived predictive models typically included 2-4 variables for predicting 30- and 180-day MDQ and NPRS improvement scores. While "best fit" parsimonious models could be identified, there were many other model options with similar statistical performance and the overall variance explained was modest. A suggested variable set for a standard predictive model that balances statistical performance with pragmatic considerations included OSPRO-ROS+, OSPRO-YF, and the NIH-CP definition, in addition to baseline MDQ and NPRS scores. ◉

KEY POINTS

FINDINGS: Baseline ratings were the strongest predictors of 30- and 180-day improvements in back pain. Our recommended models for forecasting pain and disability improvement explained 18%-21% and 27%-30% of variance in 180-day outcomes, respectively.

IMPLICATIONS: There is notable flexibility in which measures can be used in models predicting pain and disability outcomes. Four variables were often the upper limit for model accuracy. A model consisting of baseline pain and disability ratings, NIH task force definition of chronic LBP, OSPRO-YF, and Review of Systems + tools was suggested as a standard model to consider for clinical implementation.

CAUTION: The purpose of this study was to determine the extent to which and what number of existing tools could provide prognostic value; the type and quantity of physical therapy interventions were not controlled or included in the analysis. There are other tools that have been studied for predicting outcomes in LBP that were not part of this study. Future research is needed to validate the performance of the suggested model in an independent cohort by

comparing predicted to actual patient outcomes.

STUDY DETAILS

AUTHOR CONTRIBUTIONS: Darren Neeley, Steve George, Kate Minick, and Gerard Brennan conceived and designed the study. Darren Neeley and Kate Minick acquired the data for the study. Data analysis was performed by Greg Snow. All authors participated in data interpretation, drafting, reading, revising, and approving of the final manuscript.

DATA SHARING: Aggregate, de-identified data that underlie the results reports in this article are available upon request to the corresponding author for use in meta-analyses or other reasonable studies, given a methodologically sound proposal.

PATIENT AND PUBLIC INVOLVEMENT: Staff from Intermountain Rehabilitation Services supported and facilitated data collection for this study. Patient involvement included consent to use of de-identified patient reported outcomes, screening tools, and electronic medical record data for study analysis.

ACKNOWLEDGMENTS: *We want to thank Intermountain Healthcare Rehabilitation Research assistants Brendan Digan and Erna Serezlic for their help with facilitating follow-up survey data collection and responding to participant questions throughout the duration of the study. Both research assistants were compensated for their work in this study.*

REFERENCES

1. Alghadir AH, Anwer S, Iqbal A, Iqbal ZA. Test-retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. *J Pain Res*. 2018;11:851-856. https://doi.org/10.2147/JPR.S158847
2. Artus M, Campbell P, Mallen CD, Dunn KM, van der Windt DAW. Generic prognostic factors for musculoskeletal pain in primary care: a systematic review. *BMJ Open*. 2017;7:e012901. https://doi.org/10.1136/bmjopen-2016-012901
3. Badley EM, Millstone DB, Perruccio AV. Back pain and co-occurring conditions: findings from a nationally representative sample. *Spine*. 2018;43:E935-E941. https://doi.org/10.1097/BRS.0000000000002590
4. Beneciuk JM, Bishop MD, Fritz JM, et al. The STarT back screening tool and individual psychological measures: evaluation of prognostic capabilities for low back pain clinical outcomes in outpatient physical therapy settings. *Phys Ther*. 2013;93:321-333. https://doi.org/10.2522/ptj.20120207
5. Beneciuk JM, Lentz TA, He Y, Wu SS, George SZ. Prediction of persistent musculoskeletal pain at 12 months: a secondary analysis of the optimal screening for prediction of referral and outcome (OSPRO) validation cohort study. *Phys Ther*. 2018;98:290-301. https://doi.org/10.1093/ptj/pzy021
6. Butera KA, George SZ, Lentz TA. Psychometric evaluation of the optimal screening for prediction of referral and outcome yellow flag (OSPRO-YF) tool: factor structure, reliability, and validity. *J Pain*. 2020;21:557-569. https://doi.org/10.1016/j.jpain.2019.09.003
7. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *J Clin Epidemiol*. 1994;47:1245-1251. https://doi.org/10.1016/0895-4356(94)90129-5
8. Chien CW, Bagraith KS, Khan A, Deen M, Strong J. Comparative responsiveness of verbal and numerical rating scales to measure pain intensity in patients with chronic pain. *J Pain*. 2013;14:1653-1662. https://doi.org/10.1016/j.jpain.2013.08.006
9. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine*. 2005;30:1331-1334. https://doi.org/10.1097/01.brs.0000164099.92112.29
10. Coronado RA, Brintz CE, McKernan LC, et al. Psychologically informed physical therapy for musculoskeletal pain: current approaches, implications, and future directions from recent randomized trials. *Pain Rep*. 2020;5:e847. https://doi.org/10.1097/PR9.0000000000000847
11. Crofford LJ. Psychological aspects of chronic musculoskeletal pain. *Best Pract Res Clin Rheumatol*. 2015;29:147-155. https://doi.org/10.1016/j.berh.2015.04.027
12. Delitto A, Patterson CG, Stevans JM, et al. Study protocol for targeted interventions to prevent chronic low back pain in high-risk patients: a multi-site pragmatic cluster randomized controlled trial (TARGET Trial). *Contemp Clin Trials*. 2019;82:66-76. https://doi.org/10.1016/j.cct.2019.05.010
13. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH task force on research standards for chronic low back pain. *J Pain*. 2014;15:569-585. https://doi.org/10.1016/j.jpain.2014.03.005
14. Eleswarapu AS, Divi SN, Dirschl DR, Mok JM, Stout C, Lee MJ. How effective is physical therapy for common low back pain diagnoses?: a multivariate analysis of 4597 patients. *Spine*. 2016;41:1325-1329. https://doi.org/10.1097/BRS.0000000000001506
15. Foster NE, Mullis R, Hill JC, et al. Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. *Ann Fam Med*. 2014;12:102-111. https://doi.org/10.1370/afm.1625
16. Foster NE, Thomas E, Bishop A, Dunn KM, Main CJ. Distinctiveness of psychological obstacles to recovery in low back pain patients in primary care. *Pain*. 2010;148:398-406. https://doi.org/10.1016/j.pain.2009.11.002
17. Fritz JM, Beneciuk JM, George SZ. Relationship between categorization with the STarT Back Screening Tool and prognosis for people receiving physical therapy for low back pain. *Phys Ther*. 2011;91:722-732. https://doi.org/10.2522/ptj.20100109
18. Fritz JM, Cleland JA, Speckman M, Brennan GP, Hunter SJ. Physical therapy for acute low back pain: associations with subsequent healthcare costs. *Spine*. 2008;33:1800-1805. https://doi.org/10.1097/BRS.0b013e31817bd853
19. Fritz JM, Hunter SJ, Tracy DM, Brennan GP. Utilization and clinical outcomes of outpatient physical therapy for medicare beneficiaries with musculoskeletal conditions. *Phys Ther*. 2011;91:330-345. https://doi.org/10.2522/ptj.20090290
20. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. *Phys Ther*. 2001;81:776-788. https://doi.org/10.1093/ptj/81.2.776
21. Fuhro FF, Fagundes FRC, Manzoni ACT, Costa LOP, Cabral CMN. Örebro musculoskeletal pain screening questionnaire short-form and STarT back screening tool: correlation and agreement analysis. *Spine*. 2016;41:E931-E936. https://doi.org/10.1097/BRS.0000000000001415
22. George SZ, Beneciuk JM, Bialosky JE, et al. Development of a review-of-systems screening tool for orthopaedic physical therapists: results from the Optimal Screening for Prediction of Referral and Outcome (OSPRO) Cohort. *J Orthop Sports Phys Ther*. 2015;45:512-526. https://doi.org/10.2519/jospt.2015.5900
23. George SZ, Beneciuk JM, Lentz TA, et al. Optimal Screening for Prediction of Referral and Outcome (OSPRO) for musculoskeletal pain conditions: results from the validation cohort. *J Orthop Sports Phys Ther*. 2018;48:460-475. https://doi.org/10.2519/jospt.2018.7811
24. George SZ, Lentz TA, Beneciuk JM, Bhavsar NA, Mundt JM, Boissoneault J. Framework for improving outcome prediction for acute to chronic low back pain transitions. *Pain Rep*. 2020;5:e809. https://doi.org/10.1097/PR9.0000000000000809
25. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. https://doi.org/10.1016/j.jbi.2019.103208

26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377-381. https://doi.org/10.1016/j.jbi.2008.08.010
27. Hill JC, Dunn KM, Lewis M, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. *Arthritis Rheum*. 2008;59:632-641. https://doi.org/10.1002/art.23563
28. Hill JC, Whitehurst DGT, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet*. 2011;378:1560-1571. https://doi.org/10.1016/S0140-6736(11)60937-9
29. Hjermstad MJ, Fayers PM, Haugen DF, et al. Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: a systematic literature review. *J Pain Symptom Manage*. 2011;41:1073-1093. https://doi.org/10.1016/j.jpainsymman.2010.08.016
30. Jöud A, Petersson IF, Jordan KP, Löfvendahl S, Grahn B, Englund M. Socioeconomic status and the risk for being diagnosed with spondyloarthritis and chronic pain: a nested case-control study. *Rheumatol Int*. 2014;34:1291-1298. https://doi.org/10.1007/s00296-014-3039-6
31. Keefe FJ, Main CJ, George SZ. Advancing psychologically informed practice for patients with persistent musculoskeletal pain: promise, pitfalls, and solutions. *Phys Ther*. 2018;98:398-407. https://doi.org/10.1093/ptj/pzy024
32. Kendell M, Beales D, O'Sullivan P, Rabey M, Hill J, Smith A. The predictive ability of the STarT Back Tool was limited in people with chronic low back pain: a prospective cohort study. *J Physiother*. 2018;64:107-113. https://doi.org/10.1016/j.jphys.2018.02.009
33. Knighton AJ, Savitz L, Belnap T, Stephenson B, VanDerslice J. Introduction of an area deprivation index measuring patient socioeconomic status in an integrated health system: implications for population health. *EGEMS (Wash DC)*. 2016;4:1238. https://doi.org/10.13063/2327-9214.1238
34. Koes BW, van Tulder M, Lin CWC, Macedo LG, McAuley J, Maher C. An updated overview of clinical guidelines for the management of non-specific low back pain in primary care. *Eur Spine J*. 2010;19:2075-2094. https://doi.org/10.1007/s00586-010-1502-y
35. Koes BW, van Tulder MW, Ostelo R, Kim Burton A, Waddell G. Clinical guidelines for the management of low back pain in primary care: an international comparison. *Spine*. 2001;26:2504-2513. https://doi.org/10.1097/00007632-200111150-00022
36. Kraus KH, Turrentine MA, Johnson GS. Multimeric analysis of von Willebrand factor before and after desmopressin acetate (DDAVP) administration intravenously and subcutaneously in male beagle dogs. *Am J Vet Res*. 1987;48:1376-1379.
37. Leopoldino AAO, Megale RZ, Diz JBM, et al. Influence of the number and severity of comorbidities in the course of acute non-specific low back pain in older adults: longitudinal results from the Back Complaints in the Elders (BACE-Brazil). *Age Ageing*. 2019;49:96-101. https://doi.org/10.1093/ageing/afz134
38. Linton SJ, Nicholas M, MacDonald S. Development of a short form of the Örebro Musculoskeletal Pain Screening Questionnaire. *Spine*. 2011;36:1891-1895. https://doi.org/10.1097/BRS.0b013e3181f8f775
39. Lutz AD, Brooks JM, Chapman CG, Shanley E, Stout CE, Thigpen CA. Risk adjustment of the modified low back pain disability questionnaire and neck disability index to benchmark physical therapist performance: analysis from an outcomes registry. *Phys Ther*. 2020;100:609-620. https://doi.org/10.1093/ptj/pzaa019
40. Main CJ, George SZ. Psychologically informed practice for management of low back pain: future directions in practice and research. *Phys Ther*. 2011;91:820-824. https://doi.org/10.2522/ptj.20110060
41. Main CJ, Sowden G, Hill JC, Watson PJ, Hay EM. Integrating physical and psychological approaches to treatment in low back pain: the development and content of the STarT Back trial's "high-risk" intervention (StarT Back; ISRCTN 37113406). *Physiotherapy*. 2012;98:110-116. https://doi.org/10.1016/j.physio.2011.03.003
42. Murphy SE, Blake C, Power CK, Fullen BM. The effectiveness of a stratified group intervention using the STarTBack screening tool in patients with LBP - a non randomised controlled trial. *BMC Musculoskelet Disord*. 2013;14:342. https://doi.org/10.1186/1471-2474-14-342
43. Oliveira CB, Maher CG, Pinto RZ, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *Eur Spine J*. 2018;27:2791-2803. https://doi.org/10.1007/s00586-018-5673-2
44. Opsommer E, Rivier G, Crombez G, Hilfiker R. The predictive value of subsets of the Örebro Musculoskeletal Pain Screening Questionnaire for return to work in chronic low back pain. *Eur J Phys Rehabil Med*. 2017;53:359-365. https://doi.org/10.23736/S1973-9087.17.04398-2
45. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org/
46. Razmjou H, Palinkas V, Robarts S, Kennedy D. Psychometric properties of the OSPRO-YF screening tool in patients with shoulder pathology. *Physiother Can*. 2021;73:26-36. https://doi.org/10.3138/ptc-2019-0046
47. Rodeghero JR, Cook CE, Cleland JA, Mintken PE. Risk stratification of patients with low back pain seen in physical therapy practice. *Man Ther*. 2015;20:855-860. https://doi.org/10.1016/j.math.2015.04.007
48. Sheth MM, Morris BJ, Laughlin MS, Elkousy HA, Edwards TB. Lower socioeconomic status is associated with worse preoperative function, pain, and increased opioid use in patients with primary glenohumeral osteoarthritis. *J Am Acad Orthop Surg*. 2020;28:287-292. https://doi.org/10.5435/JAAOS-D-19-00490
49. Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. *Am J Public Health*. 2003;93:1137-1143. https://doi.org/10.2105/AJPH.93.7.1137
50. Stearns ZR, Carvalho ML, Beneciuk JM, Lentz TA. Screening for yellow flags in orthopaedic physical therapy: a clinical framework. *J Orthop Sports Phys Ther*. 2021;51:459-469. https://doi.org/10.2519/jospt.2021.10570
51. Stevans JM, Delitto A, Khoja SS, et al. Risk factors associated with transition from acute to chronic low back pain in US patients seeking primary care. *JAMA Netw Open*. 2021;4:e2037371. https://doi.org/10.1001/jamanetworkopen.2020.37371
52. Suri P, Delaney K, Rundell SD, Cherkin DC. Predictive validity of the STarT back tool for risk of persistent disabling back pain in a U.S. primary care setting. *Arch Phys Med Rehabil*. 2018;99:1533-1539.e2. https://doi.org/10.1016/j.apmr.2018.02.016
53. Tagliaferri SD, Miller CT, Owen PJ, et al. Domains of chronic low back pain and assessing treatment effectiveness: a clinical perspective. *Pain Pract*. 2020;20:211-225. https://doi.org/10.1111/papr.12846
54. Thomas Lumley based on Fortran code by Alan Miller. *Leaps: Regression Subset Selection*; 2020. https://CRAN.R-project.org/package=leaps
55. Traeger AC, Henschke N, Hübscher M, et al. Estimating the risk of chronic pain: Development and validation of a prognostic model (PICKUP) for patients with acute low back pain. *PLOS Med*. 2016;13:e1002019. https://doi.org/10.1371/journal.pmed.1002019