Title

Assessing the Significance of Individual Change in 2 Samples of Patients in Treatment for Low Back Pain Using 5 Different Statistical Indicators

Journal
Journal of Manipulative and Physiological Therapeutics, 44(9)

ISSN
0161-4754

Authors
Hays, Ron D
Slaughter, Mary E
Spritzer, Karen L
et al.

Publication Date
2021-11-01

DOI
10.1016/j.jmpt.2022.03.002

Peer reviewed

Assessing the Significance of Individual Change in 2 Samples of Patients in Treatment for Low Back Pain Using 5 Different Statistical Indicators

Ron D. Hays, PhD,[a] Mary E. Slaughter, PhD,[b] Karen L. Spritzer, BS,[a] and Patricia M. Herman, PhD[b]

ABSTRACT

Objective: The purpose of this study was to estimate the significance of individual change using 5 statistical indicators in 2 samples of patients treated for low back pain.
Methods: This secondary analysis used observational and clinical trial data from 2 samples of patients with low back pain to compare 5 ways of estimating significant individual change on the Impact Stratification Score (ISS) administered at the following 2 time points: 3 months apart in an observational study of 1680 patients undergoing chiropractic care, and 6 weeks apart in a randomized trial of 750 active-duty military personnel with low back pain. The following 5 methods were compared: (1) standard deviation index; (2) standard error of measurement (SEM); (3) standard error of estimate (SEE); (4) standard error of prediction (SEP); and (5) the reliable change index (RCI). The ISS is the sum of the Patient-Reported Outcomes Measurement Information System (PROMIS)-29 v2.1 physical function, pain interference, and pain intensity scores and is scored to have a possible range of 8 (least impact) to 50 (greatest impact).
Results: The amount of change on the ISS needed for significant individual change in both samples was 5 for the SEM and for the SEE and 7 for the SEP and RCI.
Conclusions: The results of the current study provide some preliminary support for use of the SEP or the RCI to identify significant individual change and provide estimated thresholds of individual change that can be used for the ISS. The SEP and RCI estimates of significant change were consistent with retrospective ratings of change of at least *moderately better* in prior research. These 2 were less likely than other methods to classify people with low back pain as responders who have not actually gotten better (false positive). In contrast, the SEM and SEE were less likely to miss real change (false negative). (J Manipulative Physiol Ther 2021;44;699-706)
Key Indexing Terms: *Low Back Pain; Chiropractic; Military Personnel*

INTRODUCTION

Longitudinal studies of healthcare interventions sometimes focus on mean group-level change along with minimally important change estimates as thresholds for change large enough to be of consequence. But there is value and increasing interest in identifying which patients benefit from treatment (which we call "responders").[1,2] Knowing the amount of change that represents a response to treatment can enhance interpretation of clinical trials and observational studies.[3] For clinicians, it is important to know when individual patients have improved or declined.

Responders may be incorrectly identified using average group-level retrospective ratings of change thresholds, such as minimally important change.[4] For example, the U. S. Food and Drug Administration erroneously suggested that the difference in scores between people who reported their condition was the same versus better could be used to identify responders to treatment.[1] This results in an over-optimistic estimate of the number of patients who have benefited from treatment.

Identifying responders is an individual change concept that requires use of individual-level statistics.[5] Using an estimate of minimal group-level change leads to misclassification of patients as responders who may not have changed. In comparison to group change, a much larger change is needed for statistically significant change in an

[a] Department of Medicine, Division of General Internal Medicine & Health Services Research, University of California Los Angeles, Los Angeles, California.
[b] RAND Corporation, Santa Monica, California.
Corresponding author: Ron D. Hays, PhD, 1100 Glendon Avenue, Suite 850; Los Angeles, CA 90024.
(e-mail: *drhays@ucla.edu*).

Paper submitted August 10, 2021; in revised form March 7, 2022; accepted March 7, 2022.
0161-4754

https://doi.org/10.1016/j.jmpt.2022.03.002

individual's score because individual change estimates have larger standard errors.[6] Significant individual change is necessary to classify an individual as a treatment responder.[7]

Although individual-level variation can be estimated by single-case time-series approaches when patient-reported outcomes have been assessed at several time points,[8] most longitudinal studies are limited to a few (eg, 2) time points. The significance of individual change for patient-reported outcomes scores based on 2 assessments can be assessed using at least 5 different methods. Each method compares an individual's change (time 2 − time 1) to the amount that would exceed error. Significance of individual change is usually based on the conventional 2-tailed $P < .05$ threshold and a 1.96 cutoff for each individual-level test statistic.

The 5 methods of estimating significant individual change all include change in the numerator, but each uses a different estimate of "error" in the denominator. The standard deviation index uses the time 1 standard deviation (SD).[9] Another uses the standard error of measurement (SEM) ($SD_1 \ \sqrt{1 - reliability}$).[10] Two other approaches use the standard error of estimation (SEE) ($SD_1 \ \sqrt{reliability(1 - reliability)}$) or the standard error of prediction (SEP) ($SD_1 \ \sqrt{1 - reliability^2}$),[11] and a final method uses the reliable change index (RCI) ($\sqrt{2} \ SEM$).[12] The standard deviation index is limited by ignoring reliability of measurement. The SEM, SEE, SEP, and RCI all include reliability as well as SD. The SEM provides an overall indicator of accuracy of the score. The SEE is designed to be used to set confidence intervals around true scores: true score = mean score + ([reliability] * [observed score − mean]). The SEP is typically used to predict a future score from a past score. The RCI is designed to evaluate differences between 2 scores over time using the standard error of the difference. Limited comparative information about the different approaches is available.

The purpose of this study was to estimate the significance of individual change in pain impact in 2 samples of patients treated for low back pain and compare estimates obtained by 5 different statistical indices of significant change.

Methods

To assess the consistency of results in different applications, we compared results from 2 samples: (1) a 2-wave observational study of patients with chronic low back pain and/or chronic neck pain receiving chiropractic care[13]; and (2) a prospective clinical trial of 750 active-duty military personnel with low back pain.[14]

Data Collection

Ethics. The RAND Human Subjects Protection Committee determined that this secondary analysis was exempt (IRB Number 00000051).

Measures. A National Institutes of Health Pain Consortium research task force proposed an Impact Stratification Score (ISS) for chronic low back pain that is the sum of the PROMIS-29 v2.1 physical function, pain interference, and pain intensity raw scores.[15] The ISS has a possible range of 8 (least impact) to 50 (greatest impact). Physical function (4 items, with response options ranging from *without any difficulty* = 1 to *unable to do* = 5) and pain interference (4 items, with response options ranging from *not at all* = 1 to *very much* = 5) each contribute from 4 to 20 points, and the pain intensity item contributes from 0 to 10 points. The 9-item ISS has promise because of its brevity and focus on core domains associated with low back pain, but there is limited information about the amount of change required to be significant at the individual-level (ie, to identify a responder). Prior work has provided support for the unidimensionality, reliability, and construct validity of the ISS.[16,17]

Sample 1 Design. A multistage systematic stratified sampling with 4 levels was used: regions and/or states, sites (ie, metropolitan areas), providers and/or clinics, and patients. Chiropractic practices were selected in 6 states from major geographical regions of the United States: San Diego, California; Tampa, Florida; Minneapolis, Minnesota; Seneca Falls and/or Upstate, New York; Portland, Oregon; and Dallas, Texas. Patients were recruited by having the front desk staff at each clinic offer every patient who visited the clinic during a 4-week period an iPad-administered prescreening survey to assess initial study inclusion and exclusion criteria.

Patients who met these criteria were invited to be in the study, and, if they agreed, they were asked to provide their email address and a phone number. Patients invited to the study were emailed a longer screening questionnaire to determine whether they met the study criteria (ie, reported back or neck pain for at least 3 months before seeing the chiropractor and/or stated that their pain was chronic). If they were eligible for the study, patients were then consented and asked additional questions for which they received a $20 gift card. They completed a baseline and 3-month follow-up questionnaire. Participants received a $25 gift card for completing the baseline questionnaire and $25 for completing the 3-month follow-up questionnaire.

Sample 1 was registered as an observational study on ClinicalTrials.gov (ID: NCT03162952). The RAND Human Subjects Protection Committee reviewed and approved the original study, and this secondary analysis of it was deemed exempt (2019−0651-AM02).

Sample 1 participants were clearly receiving chiropractic care for their back or neck pain, but some were also receiving other health care. More detail on this sample is published.[13]

Sample 2 Design. Data were collected in a multi-site clinical trial of active-duty U.S. military personnel. The study was conducted at 3 military treatment facilities: Naval Hospital in Pensacola, Florida; Walter Reed National Military Medical Center in Bethesda, Maryland; and Naval Medical Center in San Diego, California. The trial was preregistered on clinicaltrials.gov (NCT01692275), approved by each participating institution's institutional review board, and oversight was provided by an independent data and safety monitoring committee. Written informed consent was given by all study participants. The detailed protocol and primary results were previously published.[14] Study participants were randomized to either usual medical care or usual medical care plus chiropractic care for low back pain.

Sample 1 Characteristics. Table 1 summarizes characteristics of those who completed the baseline survey (n = 2024) and the subset of 1680 patients with complete data for baseline and the 3-month endpoint survey. The characteristics of the subset of people who completed both the baseline and the endpoint survey are very similar to that of those who completed the baseline survey. The average age of the endpoint sample was 49 years, 74% were female patients, and the majority had a college degree, were non-Hispanic white, worked full-time, and had an annual income of $60,000 or more.

Sample 2 Characteristics. The lower part of Table 1 summarizes the characteristics of those in sample 2 who completed the baseline survey (n = 750) and the subset of 619 patients with complete data on the variables used in the analyses reported in this paper at baseline and 6 weeks. The characteristics of those who completed the 6-week follow-up were very similar to that of the baseline sample. The average age of the analytic sample was 32 years, 23% were female patients, and the majority were non-Hispanic white.

Analysis Plan. We first compare the internal consistency reliability[18] of the ISS in the 2 samples at baseline. Then we provide the mean and range of ISS change scores, skewness, and kurtosis. We compare estimates of the significance of individual change on the ISS from each dataset by the 5 different methods summarized above: (1) standard deviation index, (2) SEM, (3) SEE, (4) SEP, and (5) RCI. For each method, we calculate the amount of individual change in the ISS required to be significant at P <.05. In addition, we estimate agreement between each pair of methods using the kappa statistic.[19] We also report the number of people that improved, stayed the same, or got worse based on each of the statistical indices.

Results

Internal consistency reliability of the ISS at baseline was 0.90 in sample 1 and 0.92 in sample 2. ISS scores at baseline and follow-up are summarized in Table 2. The mean scores at baseline and follow-up suggest mild impact, but the maximum scores represent severe impact.[13] The magnitude of improvement was larger in sample 2 than in sample 1.

The estimates of error (denominators) for the 5 methods in sample 1 (sample 2) were very similar by sample: 7.6 (8.4) for the standard deviation index, 2.4 (2.4) for the SEM, 2.3 (2.3) for the SEE, and 3.3 (3.4) for the SEP and the RCI (Table 3). The amount of change on the ISS needed for statistically significant individual change rounded to the nearest integer was similar by sample but different by method: 15 to 16 for the standard deviation index, 5 for the SEM and SEE, and 7 for the SEP and RCI. As shown in Tables 4 and 5, the percentage of people classified as responders ranged from 1% (standard deviation index in sample 1) to 57% (SEM and SEE in sample 2). The differences between samples in the percentage of people in the 3 categories of change reflect differences in mean ISS change in the 2 studies.

SEM and SEE (methods 2 and 3) yielded very similar estimates of the amount of change needed to be statistically significant, and the SEP and the RCI (methods 4 and 5) produced essentially the same estimates. Agreement in classifying individuals as getting worse, staying the same, or getting better (unweighted kappa) was perfect (1.00) between methods 2 and 3 and between methods 4 and 5 in both samples. The kappa coefficient comparing methods 2 and 3 to methods 4 and 5 was 0.69 in sample 1 and 0.80 in sample 2. Kappa was only 0.09 to 0.16 between method 1 and the other 4 methods in sample 1 and 0.37 to 0.51 in sample 2.

Discussion

Identifying how many individuals significantly improve (ie, responders) provides important supplementary information beyond group mean change about the effects of treatment options. The amount of change that represents a response to treatment is important for clinicians to use in assessing the trajectory of their patients and allows researchers to identify predictors of response to treatment.

In this study, we found that the amount of change needed for individual significance on the standard deviation index was much greater than any of the other indices. It is worth noting that we applied the conventional 2-tailed cutoff to the standard deviation index and the other indices, but some have mentioned 1-tailed tests for the standard deviation index.[9] Nonetheless, the standard deviation index is problematic because its denominator does not reflect any information about the reliability of measurement. The amount of individual-level change required for the other 4 indices to be significant ranged from 5 to 7 points on the ISS. The SEM and SEE estimates of error were like one another and

Table 1. *Characteristics of the Samples*

Sample 1	Baseline Survey (n = 2024)	Baseline and 3-Month Survey (n = 1680)
Age	Mean = 49 (range: 21-95)	Mean = 49 (range: 21-95)
Age 50+	50%	50%
Female (%)	72%	73%
Education		
Less than HS	0.3%	0.3%
HS degree/GED	7%	7%
Some college	37%	36%
BA or higher	56%	57%
Race and ethnicity		
Hispanic	5%	5%
Non-Hispanic		
White	88%	89%
Asian	2%	2%
African American	2%	2%
American Indian/Pacific Islander/Other	3%	3%
Working full time	59%	60%
Gross income		
Income <$ 10K	2%	2%
$10K ≤=income ≤ $60K	36%	37%
$60K ≤=income ≤ $100K	30%	30%
Income ≥ $100K	32%	32%
Sample 2	**Baseline Survey (n-750)**	**Baseline and 6-Weeks (n=619)**
Age	Mean = 31 (range: 18-50)	Mean = 32 (range: 18-50)
Age 50+	0%	0%
Female (%)	23%	23%
Race and ethnicity		
Hispanic	16%	16%
Non-Hispanic		
White	58%	57%
Asian	4%	4%
African American	19%	20%
American Indian/Pacific Islander/other	3%	3%

BA, Bachelor's degree; *GED,* General Educational Development; *HS,* high school.

Table 2. *Impact Stratification Score (ISS) at Baseline to 3 Months Later in Sample 1 (n = 1680) and 6 Weeks Later in Sample 2 (n = 619)*

Time	Minimum	Mean	Median	Maximum	SD	Skewness	Kurtosis
Sample 1							
Baseline	8	19.2	18	48	7.6	0.86	0.39
3-months	8	18.1	16	47	7.7	1.02	0.71
Change	−30	−1.2	−1	24	5.6	−0.29	2.38
Sample 2							
Baseline	9	24.0	23	49	8.4	0.41	−0.62
3-months	8	19.1	17	47	8.8	0.79	−0.02
Change	−40	−5.0	−4	25	9.2	−0.53	0.92

Possible range of the ISS is 8 (least impact) to 50 (greatest impact).
ISS, Impact Stratification Score; *SD,* standard deviation.

smaller than those of the SEP and RCI. The SEP and RCI estimates of error reflect error over time (ie, predict a future score and difference in scores over time), not just measurement error (ie, score accuracy and true scores).

The similar estimates for the SEM and the SEE and for the SEP and the RCI occurred because the reliability of the ISS was high (0.90 in sample 1 and 0.92 in sample 2). Differences between the 2 methods in both pairs would increase as the reliability of measurement decreases. Future studies would be useful to document differences in these methods for measures with lower levels of reliability, but if the reliability meets the 0.90 threshold for individual-level assessment,[20] then the results reported here in terms of similarities between these pairs of methods will apply.

Which threshold should be used to indicate significant individual change in the ISS: a 5-point change or a 7-point change? It has been suggested that retrospective perceptions of change (individuals' reports of "meaningful" change) should also be examined.[1] The optimal cut-point on the ISS was 7 points based on improvement defined by the patient's rating of change in back pain (*moderately better, much better,* or *completely gone*) in a prior analysis of sample 2.[7] This 7-point change is consistent with the significance of individual change in the ISS based on either the SEP or the RCI in the current study. Another study using a different method[21] reported 7.5 points as the optimal ISS cut point using data from 223 patients of a Dutch spine clinic who reported pain in their lower back and/or leg for more than 12 weeks.[22] Improved was defined as "much improved" or "completely improved" (versus "extremely worsened," "much worsened," "little worsened," "unchanged," or "little improved").

Table 3. *Amount of Change in Impact Stratification Score (ISS) Representing Significant Individual Change in Sample 1 and in Sample 2*

Method to Estimate Individual Change on the Impact Stratification Score	Each Method's Estimate of Individual Error	Amount of Individual Change Required to Be Statistically Significant at Individual Level
(1) Standard deviation index (SD_1)	7.6 (8.4)	14.9 (16.5)
(2) Standard error of measurement [SEM] ($SD_1 \sqrt{1 - reliability}$)	2.4 (2.4)	4.7 (4.7)
(3) Standard error of estimation (SEE) ($SD_1 \sqrt{reliability}\ (1 - reliability)$)	2.3 (2.3)	4.5 (4.5)
(4) Standard error of prediction (SEP) ($SD_1 \sqrt{1 - reliability^2}$)	3.3 (3.4)	6.5 (6.7)
(5) Reliable change index (RCI) ($\sqrt{2}\ SEM$)	3.3 (3.4)	6.5 (6.7)

Sample 1 results are given followed by sample 2 results in parentheses.
ISS, Impact Stratification Score; *RCI,* reliable change index; *SD,* standard deviation; *SD_1*, Standard deviation at time 1 (baseline); *SEE,* standard error of estimation; *SEM,* standard error of measurement; *SEP,* standard error of prediction.

Table 4. *Number of Those Who Improved, Stayed the Same, or Got Worse for 5 Different Indices of Significant Change in Sample 1*

Method to Estimate Individual Change	Improved	Stayed Same	Got Worse
Standard deviation index (SD_1)	24 (1%)	1645 (98%)	11 (0.7%)
Standard error of measurement (SEM)	375 (22%)	1106 (66%)	199 (12%)
Standard error of estimate (SEE)	375 (22%)	1106 (66%)	199 (12%)
Standard error of prediction (SEP)	243 (14%)	1331 (79%)	106 (6%)
Reliable Change Index (RCI)	243 (14%)	1331 (79%)	106 (6%)

Row percentages are shown.
RCI, reliable change index; SD_1standard deviation at time 1 (baseline); *SEE,* standard error of estimation; *SEM,* standard error of measurement; *SEP,* standard error of prediction.

When choosing the threshold to indicate significant individual change on the ISS, requiring perceived change that is at least moderately better is less likely to result in classifying people as responders who have not clearly gotten better (false positive). Using that logic, one might prefer the SEP or the RCI. Another argument for a larger threshold comes from a recent study showing that the constant SEM used for all respondents can be overly optimistic and that when the standard error varies appropriately by estimated score, fewer people are classified as responders.[23] On the other hand, the risk of this threshold is that it may be too conservative and not include those who are likely to have changed (ie, false negative).

Limitations and Future Research

Although the analyses were conducted on 2 large samples, results could differ in other samples with different levels of change over time. It is important to conduct similar analyses in other samples that vary in the distribution of ISS change scores. In this study, the effect size of change in sample 1, the observational sample, was −0.16 using the SD at baseline and −0.21 using the SD of change. In sample 2, the clinical trial sample, the effect size of change was −0.59 using the SD at baseline and −0.54 using the SD of change. The consistency of our results over this range of change is an indication that this study's results can be used as estimates of individual change for the ISS in other studies. However, the robustness of these results still needs to be examined in samples with other amounts of change. In addition, this study only included a single patient-reported measure, the ISS. Research with other measures, including those with lower levels of reliability, would be informative. Finally, simulation studies could be helpful to evaluate the different methods evaluated here. It is also worth noting that some have suggested focusing on likely change and partitioning individuals into those who almost certainly changed, quite likely changed, and probably stayed the same.[24,25]

Future research to compare the performance of different methods of estimating individual change with different measures and samples will provide further information about the generalizability of the study results.

Conclusions

The results of the current study provide some preliminary support for use of the SEP or the RCI to identify

Table 5. *Number of Those Who Improved, Stayed the Same, or Got Worse for 5 Different Indices of Significant Change in Sample 2*

Method to Estimate Individual Change	Improved	Stayed Same	Got Worse
Standard deviation index (SD_1)	182 (29%)	408 (66%)	29 (5%)
Standard error of measurement (SEM)	355 (57%)	150 (24%)	114 (18%)
Standard error of estimate (SEE)	355 (57%)	150 (24%)	114 (18%)
Standard error of prediction (SEP)	310 (50%)	224 (36%)	85 (14%)
Reliable Change Index (RCI)	310 (50%)	224 (36%)	85 (14%)

Row percentages are shown.
RCI, reliable change index; SD_1standard deviation at time 1 (baseline); *SEE,* standard error of estimation; *SEM,* standard error of measurement; *SEP,* standard error of prediction.

significant individual change and provide estimated thresholds of individual change that can be used for the ISS. The standard deviation index does not incorporate reliability of measurement and yields much different results than the other methods.

Funding Sources and Conflicts of Interest

This study was funded by the National Center for Complementary and Integrative Health (NCCIH). Grant No. 1R01AT010402-01A1. NCCIH had no role in the design; data collection, analysis, or interpretation; or writing of this manuscript. No other conflicts of interest were reported for this study.

Contributorship Information

Concept development (provided idea for the research): R.D.H., P.M.H.

Design (planned the methods to generate the results): R.D.H.

Supervision (provided oversight, responsible for organization and implementation, writing of the manuscript): R.D.H.

Data collection/processing (responsible for experiments, patient management, organization, or reporting data): M.S., K.S.

Analysis/interpretation (responsible for statistical analysis, evaluation, and presentation of the results): R.D.H., M.S.

Literature search (performed the literature search): R.D.H.

Writing (responsible for writing a substantive part of the manuscript): R.D.H., P.M.H.

Critical review (revised manuscript for intellectual content, this does not relate to spelling and grammar checking): R.D.H., M.S., K.S., P.M.H.

Practical Applications

- We estimate the significance of individual change using 5 statistical indicators in 2 samples of patients treated for low back pain.
- We found that about 5 to 7 points on the Impact Stratification Scale measure of pain impact represent a significant individual change.
- These findings suggest that researchers, clinicians, and decision-makers can identify responders to treatment for low back pain on the Impact Stratification Scale using this guidance.

References

1. U.S. Food and Drug Administration. Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. *Fed Regist*. 2009;74(235):65132-65133.
2. Hays RD, Spritzer KL, Sherbourne CD, Ryan GW, Coulter ID. Group and individual-level change on health-related quality of life in chiropractic patients with chronic low back or neck pain. *Spine (Phila Pa 1976)*. 2019;44(9):647-651.
3. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain*. 2008;9(2):105-121.
4. Hays RD, Peipert JD. Minimally important differences do not identify responders to treatment. *JOJ Scin*. 2018;1(1):555552.
5. Hurst H, Bolton J. Assessing the clinical significance of change scores recorded on subjective outcome measures. *J Manipulative Physiol Ther*. 2004;27(1):26-35.
6. Hays RD, Brodsky M, Johnston MF, Spritzer KL, Hui KK. Evaluating the statistical significance of health-related quality-of-life change in individual patients. *Eval Health Prof*. 2005;28(2):160-171.
7. Hays RD, Peipert JD. Between-group minimally important change versus individual treatment responders. *Qual Life Res*. 2021;30(10):2765-2772.
8. Borckardt JJ, Nash MR, Murphy MD, Moore M, Shaw D, O'Neil P. Clinical practice as natural laboratory for psychotherapy research: a guide to case-based time-series analysis. *Am Psychol*. 2008;63(2):77-95.
9. Duff K. Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. *Arch Clin Neuropsychol*. 2012;27(3):248-261.
10. Ware Jr JE, Bayliss MS, Rogers WH, Kosinski M, Tarlov AR. Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service systems. Results from the Medical Outcomes Study. *JAMA*. 1996;276(13):1039-1047.
11. McManus IC. The misinterpretation of the standard error of measurement in medical education: a primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Med Teach*. 2012;34(7):569-576.
12. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991;59(1):12-19.
13. Herman PM, Kommareddi M, Sorbero ME, et al. Characteristics of chiropractic patients being treated for chronic low back and neck pain. *J Manipulative Physiol Ther*. 2018;41(6):445-455.
14. Goertz CM, Long CR, Vining RD, et al. Assessment of chiropractic treatment for active duty, U.S. military personnel with low back pain: study protocol for a randomized controlled trial. *Trials*. 2016;17:70.
15. Deyo RA, Ramsey K, Buckley DI, et al. Performance of a Patient Reported Outcomes Measurement Information System (PROMIS) short form in older adults with chronic musculoskeletal pain. *Pain Med*. 2016;17(2):314-324.
16. Hays RD, Orlando Edelen M, Rodriguez A, Herman P. Support for the reliability and validity of the National Institutes of Health impact stratification score in a sample of active-duty U.S. military personnel with low back pain. *Pain Med*. 2021;22(10):2185-2190.
17. Rodriguez A, Edelen MO, Herman P, Hays RD. Unpacking the impact of chronic pain as measured by the Impact Stratification Score. *J Patient Rep Outcomes*. 2021;5(Suppl 1):P16.

18. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334.
19. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
20. Nunnally J. *Psychometric Theory*. 2nd ed. New York, NY: McGraw-Hill; 1978.
21. Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of pythagoras. theoretical considerations and an example application of change in health status. [published correction appears in PLoS One. 2015;10(3):e0120967] *PLoS One*. 2014;9(12):e114468.
22. Dutmer AL, Reneman MF, Schiphorst Preuper HR, Wolff AP, Speijer BL, Soer R. The NIH minimal dataset for chronic low back pain: responsiveness and minimal clinically important change. *Spine (Phila Pa 1976)*. 2019;44(20):E1211-E1218.
23. Hays RD, Spritzer KL, Reise SP. Using item response theory to identify responders to treatment: examples with the Patient-Reported Outcomes Measurement Information System (PROMIS®) Physical Function Scale and Emotional Distress Composite. *Psychometrika*. 2021;86(3):781-792.
24. Donaldson G. Patient-reported outcomes and the mandate of measurement. *Qual Life Res*. 2008;17(10):1303-1313.
25. de Vries RM, Meijer RR, van Bruggen V, Morey RD. Improving the analysis of routine outcome measurement data: what a Bayesian approach can do for you. *Int J Methods Psychiatr Res*. 2016;25(3):155-167.