Original Paper

Responsiveness, reliability, and minimally important and minimal detectable change of three electronic patient reported outcome measures for low back pain: a validation study

Froud R*, PhD – Warwick Medical School, University of Warwick, Coventry, UK and Kristiania University College, Oslo, Norway

Fawkes C, PhD – Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London UK

Foss J, PhD – Department of Computer Science, University of Warwick, UK

Underwood M, MD – Warwick Medical School, University of Warwick, Coventry, UK

Carnes D, PhD – Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK, and Faculty of Health, University of Applied Sciences and the Arts, Western Switzerland

* Corresponding author: R Froud, Warwick Medical School, University of Warwick, Coventry, UK. r.froud@warwick.ac.uk tel. +44(0)2476574221 fax. +44(0)2476 150549

Abstract

Background: The Roland Morris Disability Scale (RMDQ), Visual Analogue Scale of pain intensity (VAS) and Numerical Rating Scale (NRS) are among the most commonly used outcome measures in trials of interventions for low back pain. Their use in paper form is well-established. Few data are available on the metric properties of electronic counterparts.

Objective: To establish responsiveness, minimal important change (MIC) thresholds, reliability, and minimal detectable change (MDC_{95}) for electronic (e) versions of the RMDQ, VAS, and NRS as delivered via iOS app, Android app, and web app.

Methods: We recruited people with low back pain who visited osteopaths. We invited participants to complete the eRMDQ, eVAS, and eNRS at baseline, one-week, and six-weeks, along with a health transition question (TQ) at one and six-weeks. Data from participants reporting recovery were used in responsiveness and MIC analyses, using Receiver Operator Characteristic curves. Data from participants reporting stability were used for analyses of reliability (ICC agreement) and minimal detectable change (MDC_{95}).

Results: We included 442 participants. At one and then six-weeks, ROC AUCs were 0.69 (95%CI 0.59 to 0.80) then 0.67 (0.46 to 0.87) for the eRMDQ; 0.69 (0.58 to 0.80) then 0.74 (0.53 to 0.95) for the eVAS; and 0.73 (0.66 to 0.80) then 0.81 (0.69 to 0.92) for the eNRS. Associated MIC thresholds were estimated as 1 (0 to 2) then 2 (-1 to 5), 13 (9 to 17) then 7 (-12 to 26), and 2 (1 to 3) then 1 (0 to 2) points, respectively. Over one-week in stable and 'about the same' participants ICCs were 0.87 (0.66 to 0.95) and 0.84 (0.73 to 0.91) for the eRMDQ, with MDC_{95} of 4 and 5; 0.31 (-0.25 to 0.71) and 0.61 (0.36 to 0.77) for the eVAS with MDC_{95} of 39 and 34; and 0.52 (0.14 to 0.77) to 0.67 (0.51 to 0.78) with MDC_{95} of 4 and 3 for the eNRS.

Conclusions: The eRMDQ was reliable with borderline adequate responsiveness. The eNRS was responsive with borderline reliability. While the eVAS had adequate responsiveness it did not have an attractive reliability profile. Thus, the eNRS might be preferred over the eVAS for measuring pain intensity. The observed electronic outcome measures' metric properties are within the range of values reported in the literature for their paper counterparts and are adequate for measuring changes in a low back pain population.

Keywords: electronic patient reported outcome measures; validation; responsiveness; reliability; minimally important change; minimal detectable change

Introduction

Low back pain (LBP) is a common and costly problem resulting in a substantial personal, social and economic burden, and is the number one cause of disability globally.[1, 2] LBP is a symptom rather than a specific disease as most LBP is non-specific; *i.e.* where no specific underlying cause has been identified, but where the term lacks formal definition and where definitions in trials have been diverse.[1, 3] The lifetime prevalence of LBP is between 60-84%. [4, 5] The global problem of LBP is getting worse due to aging and increasing population size.[6, 7] The number of clinical trials of interventions for LBP has been increasing, with over 30 trials of interventions for LBP now being published annually.[8] Patient-reported outcome measures (PROMs) in the form of paper questionnaires are typically used in these trials to judge the effectiveness of the health technology under investigation.[8]

Disability and pain are by far the most commonly measured domains in trials of interventions for LBP; each is measured at least twice as often as any other domain.[8] The Visual Analogue Scale (VAS), and the Numerical Rating Scale (NRS) are most commonly used for measuring pain, and the Roland Morris Disability Scale (RMDQ), is most commonly used for measuring disability.[8] These are quasi-continuous measures of pain intensity (VAS, NRS) or functional disability, and for each the relationship between the observed item responses and the unobserved latent variable each is assumed to be consistent with a reflective conceptual framework.[9] There is evidence that paper forms of VAS and NRS have been in use since at least the early to mid 20th century, and the RMDQ has been used since 1983.[10-12]

The validity of a PROM is defined as '*the degree to which an instrument truly measures the construct(s) it purports to measure*'.[13] Several aspects that comprise what we consider to constitute good development and validation of PROMs post-date the introduction of these particular instruments. Validation exercises have been performed retrospectively, results have accrued over time, and endorsement and use of the measures has survived the process.[14-16] Notwithstanding healthy academic debate, it is generally accepted that these outcome measures have reasonable face validity, content validity, and have at times been considered the legacy gold standard for comparison for assessing the criterion/convergent validity of other instruments. [17-19]

Measuring patient/participant change in health status using browser-based technology and mobile device technologies is a natural progression. Digital PROMs and ports of existing paper PROMs to digital media have become known as electronic patient reported outcomes measures (ePROs).[20] When migrating existing paper PROMs into ePROs, there are aspects relating to the metric validity of the instrument that may need to be reassessed. Some aspects of validity are clearly independent of whether the instrument is completed on paper or digitally—for example, the content wording (unless it is culturally/clinically out-of-date), and the extent to which this content is judged to appropriately span the domains of the health construct being measured (*i.e.* content and face validity). However, other aspects of validity that relate directly to measurement performance should not be assumed to be unchanged.

For any instrument that is designed to measure change in a health construct, two properties are particularly relevant: reproducibility (*i.e.* reliability) and responsiveness. Reliability is the extent to which the same results are obtained on repeated measures when no real change in health status has occurred.[21, 22] An analogy using a set of bathroom weighing scales is that it is desirable that the scales show the same weight upon time-standardized daily measurement when there truly is no true change in a person's weight – if this is the case, the scales may be said to be reliable. Conversely, responsiveness is analogous to the scales detecting in important change when one truly exists. As users' physical interactions with ePRO versions of PROMs differs in fundamental respects from paper versions, we suggest that reassessing these two key

change measurement properties is necessary before advocating their widespread use in health research.

In analyses of trials, or evaluations of health interventions, using PROMs to decide when an individual participant has responded, facilitates interpretation of intervention effect.[23] Responder analysis permits the number of improvements to simply be counted and compared by arm using several clear statistics. These are intuitive reporting methods and there is consensus that back pain trials should incorporate these.[23-25] However, to be able to do this it is necessary to know (1) the minimum thresholds considered important to an individual participant – the minimally important change (MIC); and (2) what magnitudes of change can be detected beyond the inherent measurement error of the instrument – the minimal detectable change (MDC).[26, 27] These thresholds may be altered by the change in media from paper to digital, and these thresholds may also be population-specific.[28, 29]

We aimed to determine reliability and responsiveness, MIC and MDC, for electronic versions of the VAS, RMDQ, and NRS as administered to adults with LBP who visit osteopaths, using a web browser, Android or iOS app on their own computers, smart phones, or tablets.

Methods

Recruitment

We recruited participants with LBP from osteopathic clinics in England and Wales. Participants were recruited by osteopaths on our behalf and provided with an enroll code and instructions for installing the iOS or Android app (from the App Store or Google Play) or completing the outcome measures using a web browser.

We assumed an attrition rate of 30%, and a recovery (*i.e.* participants who state they are much better or completely recovered on a health transition question – below) rate of over 90% in those with acute and sub-acute LBP (*i.e* LBP present for less than three months).[30] Thus, for our responsiveness study, for which we required improved participants, we sought to recruit a minimum of 200 people with acute and sub-acute LBP to ensure at least 50 eligible six-week measurements (see Sample Size section). For people with chronic LBP receiving manual therapy, we assumed the same rate of attrition, but a lower rate of recovery, of 45%.[24] For our test-retest study, we required stable participants who identified as remaining stable over a period of one–week; thus, we sought to recruit 400 chronic patients to find 50 participants self-identifying as stable (*i.e.* reporting 'no change' on a health transition question – below). Participants were invited to complete the electronic versions of outcome measures at baseline, one-week, and six-week follow-up time points.

Software

We used Android and iOS apps, and a web app with an associated form builder that was developed by Clinvivo Ltd, a University of Warwick spin-out company.[31] The apps, which function identically across platforms, permitted PROMs to be typeset and then administered to patients securely on their own devices. Data in transit are encrypted using a secure socket layer (SSL) and data at rest are encrypted using an Rivest-Shamir-Aldeman (RSA) and Advanced Encryption Standard (AES) encryption hybrid. At the end of the study period, data were encrypted using the Open Pretty Good Privacy (PGP) standard and transferred from Clinvivo to researchers. The iOS, Android, and web apps sent data one-way and did not receive or redisplay personal data. The platform presented an electronic version of the instrument and reminded participants to complete outstanding follow-up measurements, as appropriate. Off-line completion in apps was permitted in cases of interrupted connectivity, with submissions occurring upon restoration of connectivity. Reminders, which were received up to twice per

follow-up measurement due, were sent directly to devices for app-enrolled participants, and web-enrolled participants were sent up to two reminder e-mails.

Electronic Versions of Patient Reported Outcome Measures

The Visual Analogue Scale of Pain Intensity (VAS) is a continuous scale running from 0 to 100mm, measuring current pain intensity.[32] It is the most commonly used outcome measure in trials of interventions for non-specific LBP overall.[8] Huskisson is commonly credited with its development in 1974; however, there is evidence that it was being used at least as far back as 1921.[11] Intellectual Property Rights (IPR) are in the public domain and no permissions are required for use, reproductions, or modifications. Completion of the paper scale involves a person marking a line on the scale indicating their level of pain between to anchored scales that typically have wordings of 'No pain' on the left (*i.e.* 0mm) and 'Worst possible' or 'Worst imaginable' pain on the right (*i.e.* 100mm).[33, 34] On paper, the distance of the marked line is then measured from the point of zero pain and reported in mm. In migrating this to an electronic version (eVAS), we implemented a slider that could be dragged into position. We did not force the scale to render at 10cm, so as to allow for resizing to screens of different devices. Thus, we report scores in units rather than mm, where one unit is $1/100^{th}$ of the scale (*i.e.* where the pointer can be set at any one of 101 different positions), as rendered (Figure 1).

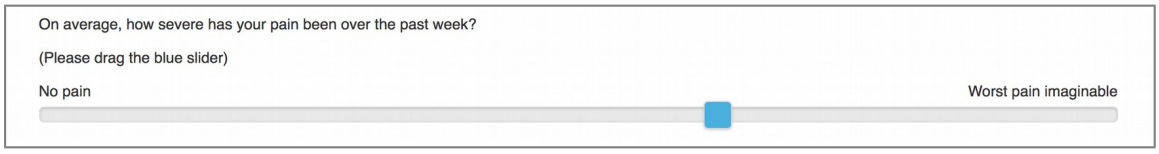

Figure 1: Electronic Visual Analogue Scale (eVAS) for pain intensity. The figure shows a screenshot of the eVAS set to show 63 units of pain intensity

The Roland Morris Disability Questionnaire (RMDQ) is a 24-item questionnaire, measuring functional disability due to back pain, which was developed in the early 1980s.[10] It is the most commonly used outcome measure in trials of interventions for LBP overall.[8] The original paper version of the instrument is well established.[35-38] No permissions are required for its use, reproductions, or modifications.[39] Scores on the RMDQ range from 0 to 24, where higher scores indicate greater disability. Participants are given a statement with which they may indicate agreement by ticking a box. Participants are asked to tick statements that they feel describe them on that day and to leave blank boxes next to statements that they feel do not. The score is then the sum total of checked items. Our electronic (eRMDQ) migration was an exact copy utilising multi-select check-boxes (Figure 2). One year into the research we added a box stating 'None of the above symptoms' for participants to confirm that none of the statements applied to them and to confirm zero scores were genuine and not reflective of a skipped question.

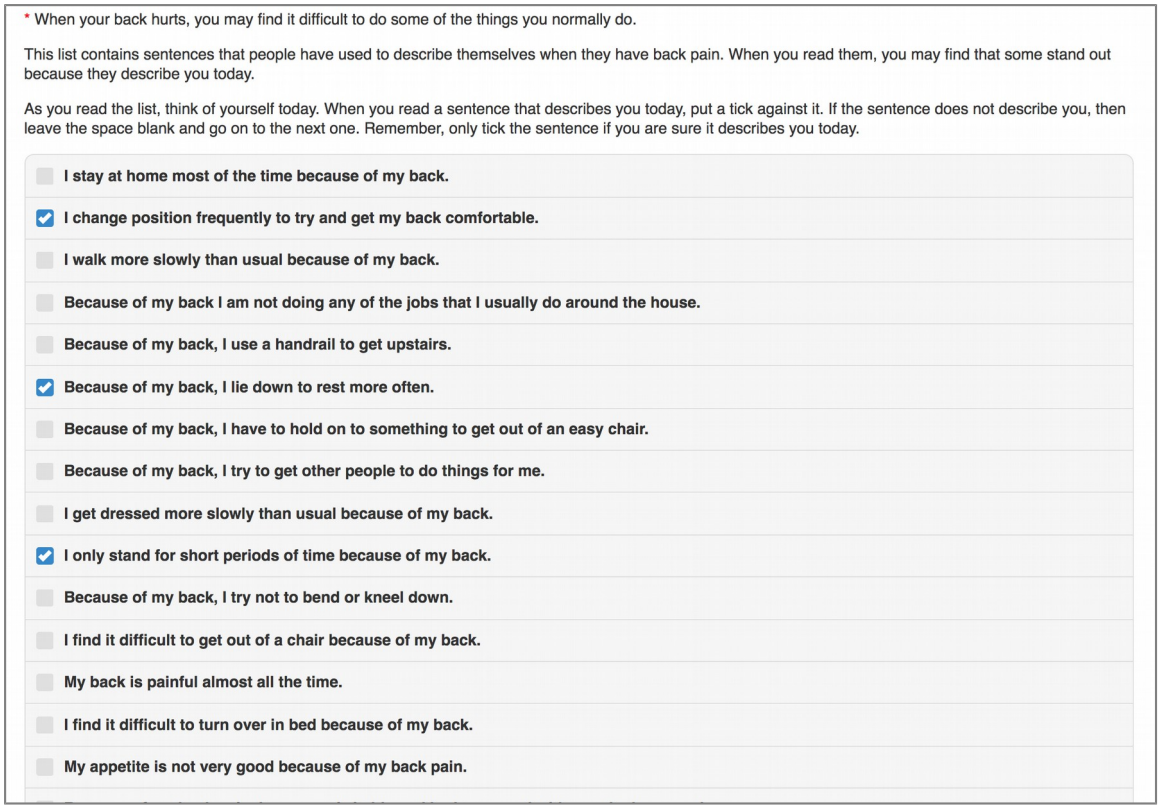

Figure 2: Electronic Roland Morris Disability Scale (eRMDQ). The figure shows a screenshot of part of the eRMDQ showing a part score of 3 units.

The NRS is an 11-point ordinal scale measuring current pain intensity.[40, 41] Validation of the paper version is well established.[41-43] It is the fourth most commonly used outcome in trials of interventions for LBP overall.[8] It is well established with IPR in the public domain. Scores on the NRS range from zero, which typically is anchored '*No pain*', and 10, which typically is anchored '*Worst pain possible*'. Our electronic (eNRS) migration was an exact copy with these anchor wordings (Figure 3). As the range of responses is exhaustive, completion of the scale was required for submission.

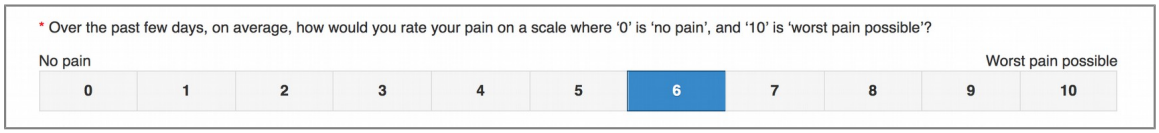

Figure 3: Electronic Numerical Rating Scale (eNRS) for pain intensity. The figure shows a screenshot of the eNRS showing a part score of 6 units.

Participants were also asked to complete (electronically) a health transition question (TQ) at one and six-week follow-up time points. The TQ was a single question with the wording '*Overall, how would you rate the change in your symptoms since beginning this study?*' and where the participant could respond on a seven-point scale: 1) Completely recovered; 2) Much improved; 3) Slightly improved; 4) No change; 5) Slightly worsened; 6) Much worsened; 7) Vastly worsened.[44]

Assessment

We aimed to have 50 completed paired measurements in 'improving' participants for responsiveness assessments and 50 completed test-retest measurements in 'stable' participants. We defined improving participants, *a priori,* as participants who select 'Much improved' or

'Completely recovered' using the TQ. Improving participants' scores were used to assess responsiveness at one and six-weeks. For our test-retest study, we defined stable participants, *a priori,* as those who select 'no change' at one-week, and in the case of having too few observations, a *post hoc* sensitivity analysis including those who selected either 'slightly worsened', 'no change', or 'slightly improved'. This alternative 'about the same' approach to marking stability has been used elsewhere.[45] Allowing one-week is typical in low back pain test-retest studies; clinically, this is close enough for the people with chronic pain to remain stable, but far enough apart that participants cannot easily recall their initial responses. It was anticipated that the chronic population would predominantly contribute participants to the test-retest study, and that improving participants would come from across all chronicity sub-populations.

Statistical Analyses

To measure responsiveness in a way that is consistent with the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) definition, we constructed Receiver Operator Characteristic (ROC) curves for one and six-week data using a dichotomised TQ as the external criterion.[22] The area under the ROC curve (AUC) is then a metric of responsiveness, accepting that the external criterion reasonably includes the construct of interest.[46] The approach has previously been used to quantify responsiveness across all three paper versions of instruments.[47] ROC AUCs of over 0.70 were considered to be adequate.[9, 48] We dichotomized the TQ such that participants responding 'Completely recovered' and 'Much improved' were considered 'improved', and all other responses were considered 'not improved'.

We also used ROC curves and the TQ external criterion for one and six-week data to quantify minimally important change (MIC), which was defined as: "*the smallest [change] in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management*." (See Note 1 of Multimedia Appendix 1) [43, 49] We used a MIC estimator based upon the minimum sums of squares method, which consistently selects the cut-point closest to the top-left corner of ROC space; as required when sensitivity and specificity are valued equally.[50] We calculated confidence intervals for MIC point estimates using bootstrapping.[51]

To estimate reliability, we calculated intra-class correlation coefficients (ICCs).[52, 53] ICC values usually range from 0 to 1.[54] ICC values above 0.75 may be interpreted as excellent agreement, values of 0.40 to 0.75 indicate poor to fair agreement, and values of below 0.40 indicate poor agreement.[55] We calculated the standard error of measurement (SEM).[53] We used this to estimate the minimal detectable change (MDC_{95}) (See Notes 2 to 4 of Multimedia Appendix 1) .[53, 56, 57]

TQs can be highly correlated with follow-up score rather than change.[24, 43, 58] Guyatt *et al* assert that if a TQ is truly measuring change then a correlation between baseline score and the TQ, and follow-up score and the TQ should ideally be present, equal, and opposite.[58] In addition, they suggest that in a linear regression model with follow-up score entered as the initial explanatory variable, the baseline score should explain a significant proportion of the residual variance in the transition rating.[58] We performed Pearson correlations and fitted regression models to explore the degree to which the TQ measured change or simply reflected follow-up status. Log rank tests were used to assess significance of the addition of baseline score.

All analyses were performed using Stata version 14.2 (Statacorp, Texas). The program *rocmic* was used to estimate MIC and the ROC AUC, which for ROC AUC utilises the *lroc* program.[51, 59]

Power and Sample Size

With the notable exception of construct validity, sample sizes in validation studies generally are not calculated based on power to test hypotheses: the estimation of reliability and responsiveness parameters is focused on the extent to which the coefficients describing these parameters approach 1 (which would represent perfect reliability/responsiveness), rather than their difference from zero or some other null value. Generally, a sample size of at least 50 participants is considered adequate for this purpose.[9, 60] Assuming an ICC of 0.7, with 50 participants we would be able to estimate the ICC to within a 95% confidence interval of +/- 0.14. Alternatively, for an ICC of 0.8, we would be able to estimate to within a 95% CI of +/- 0.10. [9] For responsiveness, with 50 participants and assuming an AUC of 0.8, and equal numbers of cases and non-cases, we would be able to estimate AUC to within a 95%CI of +/- 0.12.[61]

As standard errors (SEs) for MIC estimates are not readily calculable, we used bootstrapping to generate standard errors and 95% CIs.[51, 62] Previous simulation work on the paper-based RMDQ in a similar population suggested that 2,500 bootstrap samples was sufficient to ensure standard error convergence.[63] To explore whether this is the case for the eRMDQ (and also whether it is an appropriate number of replications for the eNRS and eVAS) we simulated SEs by randomly sampling *n* observations (with replacement) from our dataset, for an increasing number of *n*; where n is an integer, beginning at 20 and increasing by increments of 20, up to 6,000.[62, 64] We then graphically assessed SE convergence and used the point of convergence to inform the number of bootstrap replications.

Data exclusions, assumptions, and variations

Prior to the addition of the 'none' box we imputed zero scores for all baseline submissions with no eRMDQ boxes ticked, and assumed and imputed a zero score for eRMDQ follow-up scores in the case that the baseline eRMDQ score was greater than zero, and a submission had been made for the follow-up period in question. When the eVAS rendered it did so with the slider in the zero position. In the case of a submission for an untouched eVAS, zero slider zeros were assumed valid. The eNRS is a required response and necessitated a selection for submission.

As part of the basic demographic details we collected, we included a list of presenting complaints, featuring LBP among 15 other common musculoskeletal presentations and the opportunity to report a complaint not listed in a free-text box. The list of complaints was derived from earlier survey work developed as part of a national data collection initiative.[65, 66] We excluded all cases where a participant had not checked the LBP box (data from non-LBP cases were used in unrelated research).

Ethics approval

Ethics approval was obtained from the research ethics committee at Queen Mary University of London (QMERC2014/18).

Results

User Statistics and Demographics

We collected data from 575 people from 30 osteopathic clinics, between July 15, 2014 and May 3, 2017. Of these, 442 (77%) reported LBP as their main complaint. The average submission time for one-week scores was 7.4 (standard deviation (SD)=0.79) days after baseline. The average submission time for six-week scores was 42.5 (SD=0.9) days after baseline. Of the 442 participants, 267 were female (60%); 306 (69%) identified as being in full or part-time employment, five (1%) were long-term sick, 16 (4%) identified as looking after home/family, 87 (20%) were retired, six (1%) were in full-time education, 13 (3%) were unemployed, and 9 (2%) selected other or preferred not to disclose. Figure 4 shows a histogram of patient-reported age at baseline.

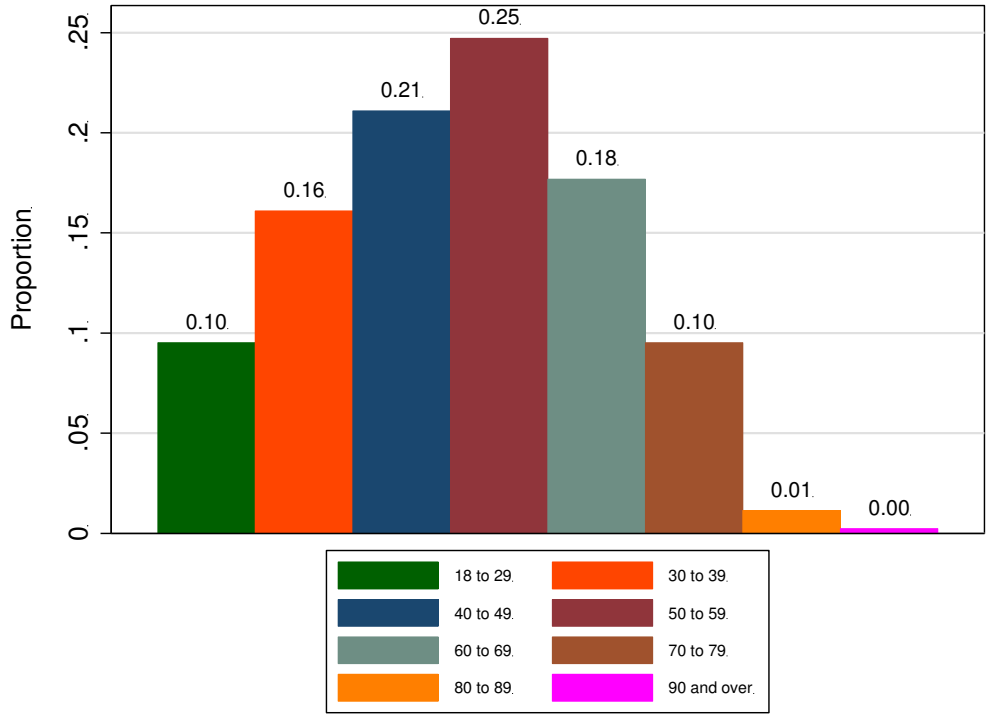

Figure 4: Histogram of patient age. The figure shows a histogram of patient age at baseline

We collected baseline eNRS data from 442 participants, and eVAS and eRMDQ data from 247 participants. One-week data were collected from 187 participants and 97 participants respectively; and six-week data were collected from 86 participants and 40 participants respectively. Figure 5 shows the incidence of recovery in these groups. There was one missing data point for eNRS at baseline (0.2%) and one-week (0.5%) for which we were unable to confirm cause. Table 1 summarizes ePRO submissions scores and cumulative recovery using median and inter-quartile range (IQR). Change scores (not shown) more closely followed normal distributions.

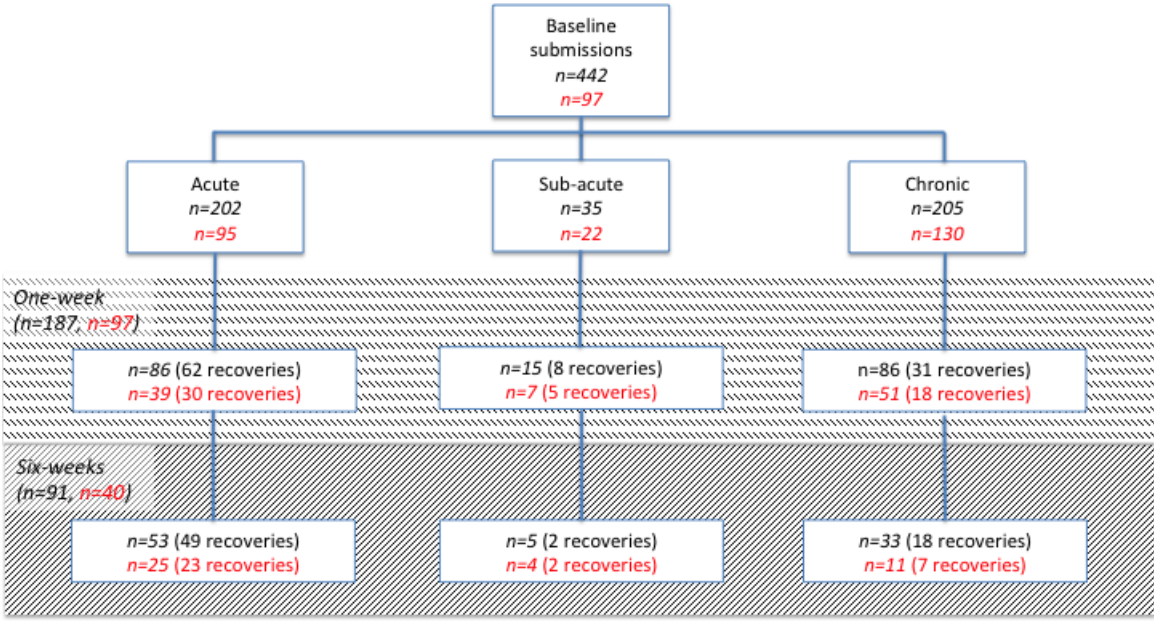

Figure 5: Flow chart showing completion rates at one and six-weeks, chronicity status, and the incidence of self-reported recovery using the health transition question, for participants who also completed eNRS, and eRMDQ and eVAS measurement (red)

Table 1. Baseline, one-week, and six-week scores across the whole sample

	Baseline Median (IQR)	***n***	**One-week Median (IQR)**	***n***	**Six-week Median (IQR)**	***n***
eRMDQ						
	4 (6)	247	2 (6)	97	2 (3.5)	40
eVAS						
	41 (32)	247	24 (19)	97	19 (19)	40
eNRS						
	5 (4)	441	3 (3.0)	186	2 (2)	91
			n (%)		***n (%)***	
TQ (recovery)	N/A	N/A	101 (54)	187	69 (76)	91
Cumulative recovery (TQ)	N/A	N/A	101 (23)	442[a]	170 (38)	442[a]

[a] *i.e.* as a proportion of all baseline participants

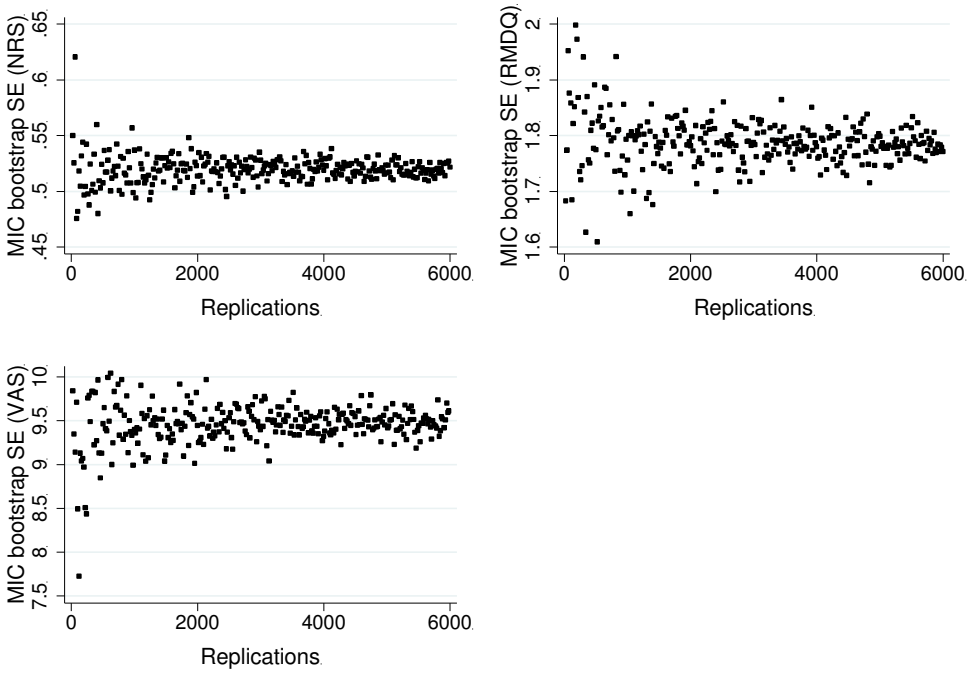

Figure 6: MIC standard error convergence. A table of graphs showing MIC bootstrap standard error (SE) convergence from simulations with increasing replication numbers

The addition of baseline score generally explained a significant proportion of the variance in the TQ over and above follow-up score. The TQ correlated with follow-up score but not with baseline score. Comprehensive results for the Guyatt analyses on the TQ's performance in measuring change are listed in Note 2 of Multimedia Appendix 1.

Evaluation Outcomes

Graphically, SE convergence appeared to be asymptotically complete at around 5,000 bootstrap replications (Figure 6); thus 5,000 replications were used to generate confidence intervals for the MIC estimates in Table 2. Responsiveness point estimates (Table 2) were borderline adequate (AUC≈0.7) or above adequate for all instruments and time points. The AUC confidence interval for the RMDQ at six-weeks spanned the null value. (Table 2)

Table 2 Responsiveness and minimally important change, by instrument and one-week and six-week follow-up time periods

	ROC AUC	95% CI	*n*	MIC	95% CI
				Points/eVAS units (% of baseline score)	
eRMDQ – 1W					
	0.69	0.59 to 0.80	97	1 (19%)	0 to 2
eRMDQ – 6W					
	0.67	0.46 to 0.87	40	2 (38%)	-1 to 5
eVAS – 1W					
	0.69	0.58 to 0.80	93	13 (32%)	9 to 17
eVAS – 6W					
	0.74	0.53 to 0.95	40	7 (17%)	-12 to 26
eNRS – 1W					
	0.73	0.66 to 0.80	185	2 (43%)	1 to 3
eNRS – 6W					
	0.81	0.69 to 0.92	91	1 (21%)	0 to 2

Table 3. Intra-class correlation coefficients from test-retest study in a per protocol stable sample and a pseudo-stable sample, and associated minimal detectable change thresholds

	n	$ICC_{agreement}$	95% CI	MDC_{95}
				Points/eVAS units
eRMDQ – per protocol				
	15	0.87	0.66 to 0.95	4
eRMDQ – allowing slight change				
	43	0.84	0.73 to 0.91	5
eVAS – per protocol				
	15	0.31	-0.25 to 0.71	39
eVAS – allowing slight change				
	43	0.61	0.36 to 0.77	34
eNRS – per protocol	22	0.52	0.14 to 0.77	4
eNRS - allowing slight change	83	0.67	0.51 to 0.78	3

Using 'no change' as a criterion for judging stability, we did not achieve our *a priori* threshold of 50 test-retest data points for comparison across any of the instruments. Of the 23 people who said they had 'no change' at one-week, 15 (65%) had chronic pain. Allowing 'slightly improved' and 'slightly worsened' to count as stable, allowed us to achieve this threshold for the eNRS only. Of 84 people who said they had 'no or slight change' at one-week, 53 (63%) had chronic pain. Notwithstanding the lack of data, the eRMDQ reliability (agreement) was excellent using either analysis, with CIs spanning fair to excellent in both analyses (Table 3). For the eVAS per protocol analysis, the agreement was fair with CIs spanning poor to fair, and in the sensitivity analysis the agreement was poor to fair with a CI range spanning poor to fair (Table 3). For the eNRS per protocol analysis the agreement was poor to fair with a CI spanning poor to excellent, and for the sensitivity analysis agreement was fair with a CI spanning poor to fair to excellent (Table 3).

Discussion

Principal Results

The results suggest that the eRMDQ had borderline adequate responsiveness levels and excellent reliability. Conversely, the eNRS had relatively good responsiveness at six-weeks but borderline adequate reliability. The eNRS outperformed the eVAS, which had adequate responsiveness but relatively poor reliability. As test-retest numbers were few, eVAS confidence intervals spanned poor to excellent and thus further investigation is warranted. While exploring use by age was not a specific study objective, we note the results indicate encouraging use by older people from this population.

Comparison with Prior Work

Across acute and chronic back pain populations there has been like-for-like evaluation (*i.e.* using similar and directly comparable methods) of the properties of paper versions of the outcome

measures explored. ROC AUC for the RMDQ ranges from 0.64 to 0.93. [45, 47, 67-75] ROC AUC for the NRS ranges from 0.67 to 0.93.[41, 42, 47, 67, 75, 76] ROC AUC for the VAS ranges from 0.71 to 0.93.[47, 72, 77-79] Our results are within these ranges at six-weeks for all but our VAS estimate at one-week, where our point-estimate approaches the lower border of the range. Our eVAS data are nevertheless consistent with the range (*i.e* insofar as the upper CI overlaps). As estimates of ROC AUC for the VAS are fewer in the literature, which might explain why the range of reported results is narrower than it is for the RMDQ and NRS.

MIC thresholds for RMDQ has ranged between 1.5 and 5;[21, 24, 35, 67, 68, 72, 75, 80-83] for the NRS between 1.5 and 4;[41-43, 67, 75, 81, 84] and for the VAS between 15 and 28mm.[72] Our absolute MIC thresholds are comparable, but are towards the lower side of this range. MIC estimates are known to increase with baseline severity and relatively low baseline scores likely explain our relatively low thresholds.[68, 75, 81, 84] However, MIC thresholds in our results, expressed as percentage change from baseline, averages 28%, which is consistent with Ostelo *et al*'s suggestion (following a review of MIC and MDC literature) of using an improvement of between 20 and 30% of baseline score for the RMDQ, NRS and the VAS as a MIC threshold.[29] We emphasise that the MIC thresholds relate to the degree of change may be considered important for an individual, and not what degree of difference may be considered important at a population-level. [27, 85, 86] We note that negative confidence intervals imply consistency of the data with the true MIC thresholds in the opposite direction. This is likely an artifact of low power and we suggest inflated sample sizes below for future studies based on the bootstrapped standard error observations.

Reported ICC estimates for the RMDQ have ranged from 0.42 to 0.95;[45, 67, 81, 87] for the NRS, from 0.92 to 0.98;[67, 81] and an estimate for the VAS of 0.71 has been reported.[88] Our results are within the ranges reported but our ICC point-estimate for the eVAS is lower than the reported paper VAS estimate. It is conceivable that rendering the eVAS slider in a zero position might lead to additional variance in the case that the outcome is overlooked (*i.e.* leading to a comparatively lower ICC) and future research might explore whether a 'touch to confirm zero' design is acceptable to users. We also note that some of the ICC values in the literature ranges may have been derived from ICCs for consistency rather than agreement; this is a practice that is known to exist (although it is not always clear which approach is used) and known to overestimate reliability.[53]

MDC_{95} estimates reported (or, in the case of the NRS only, either reported or where the MDC_{95} can be readily calculated from reported SEMs) have ranged from 5.0 to 12.1 for the RMDQ;[21, 24, 35, 45, 56, 67, 81, 83] from 2.4 to 11 (*i.e.* the full width of the scale) for the NRS;[41, 45, 67, 81, 84] and from 21.0 to 33.5 for the VAS.[79, 88, 89] Our estimates are slightly better than average for the RMDQ, towards the lower end of the range for the NRS, and comparable to the available estimates for the VAS.

In terms of comparison to studies assessing these instruments as ePROs, Bird *et al* conducted a test-retest study among 22 healthy adults of the VAS administered on iPad and found ICCs of 0.90 (0.82 to 0.95) as compared to 0.96 (0.92 to 0.98) in a paper version that participants completed simultaneously.[90] It is difficult to compare the results with this study, as the time between test and retest was less than 30 minutes. A much shorter period between test and retest might be appropriate in some populations (*e.g.* where change in acute pain must be measured over short spaces of time). In these cases, participants may be more prone to panel conditioning; where second response is affected by recall of the first response.[91] For back pain, most interventions focus on chronic pain and longer time periods. When exploring reliability of LBP outcome measures, a one-week gap between test and retest is typical. Bijur *et al* and Gallagher *et al* have similarly used small time-frames between tests on a paper-based VAS in acute pain populations and demonstrate similarly high ICCs of 0.97 (0.96 to 0.98) and 0.99 (0.989 to 0.992) respectively.[92, 93] Also of relevance, but again not directly comparable, is

work by Bishop *et al*, who administered the RMDQ on paper and on-line and constructed limits of agreement and explored differences between the instruments demonstrating equivalence with a score difference of only 0.03 points and a Bland-Altman range of -2.77 to 2.83.[94]

Finally, we note that the distribution of the user age of the health outcomes app in this population appears to be higher than the age of health app users.[95]

Implications

None of the results in this study is materially different from those that have been observed in population-similar studies of paper counterparts that are methodologically alike. There is thus some suggestion that the ePROs under evaluation are a suitable substitute for PROMs for measuring change in LBP. The eNRS outperformed the eVAS in terms of both responsiveness and reliability. As such we suggest the eNRS might be preferred the eVAS for the measurement of LBP pain intensity, but we caution that subsequent confirmatory research is warranted.

Limitations

The principal limitation is that in several cases we had small sample sizes. We had intended to recruit sufficient numbers to have at least 50 people for each assessment, in-line with recommendations, but we failed to meet these targets as we underestimated the incidence of stability.[9] There were high rates of improvement in people receiving treatment and this is a hazard of nesting a test-retest design within a protocol where participants are receiving routine clinical treatment. This was of consequence in the eRMDQ responsiveness analysis, where the data are consistent with a null population parameter and thus six-week responsiveness of the eRMDQ requires confirmation in a larger sample. Having too few data has greater implications for the test-retest assessment of the VAS where the confidence intervals span coefficient values that can be interpreted at their extremes as either poor or excellent. It is less of an issue for the eRMDQ, as while the numbers are low and lower at one and six-weeks respectively, the stronger signal combined with boundary proximity leads to narrower and more useful CIs.

It is not ideal that we permitted slightly worse and slightly improved to indicate stability in our test-retest. Although, we note a similar approach has been observed previously.[45] Further, this was a *post hoc* decision taken in light of having too few observations to use our more stringent *a priori* criteria of only those reporting 'no change'. The results using our *a priori* approach, but with few observations, are offered as sensitivity analyses that may provide useful comparison.

Having relatively few observations also meant that we were unable to explore differences by platform, *i.e.* iOS, Android, and, web browser, or to explore MIC as a function of baseline score (*e.g.* stratifying by number in category of severity), or separately by chronicity, which may have been useful and allowed us to explore any differences in these metrics by chronicity. Thus, our focus here is pragmatic and results are generalizable to the population of adults with LBP who consult osteopaths, notwithstanding chronicity.

We recorded in our database only the summed eRMDQ score, rather than individual responses. Had we retained detail of individual response profiles of the eRMDQ we could have also calculated internal consistency (as well as aspects of modern test theory; *e.g.* Rasch analysis to examine item performance, or factor analysis to explore data dimensionality). Whereas COSMIN conflate internal consistency with reliability in their taxonomy,[22, 96] we consider internal consistency to be an indication of the unidimensionality of a scale and of item redundancy, rather than the degree to which a scale is free from measurement error. As such, and with respect to the reliability definition, we preferred to consider it separately. We had not immediately considered that completion media type might affect internal consistency or item functioning of a scale. On reflection however, we think that it is conceivable that presenting the

scale digitally may alter the way patients respond. Additionally, there may be self-selection effects of those more familiar with digital media joining the study and this may be a factor that could be confounded with how a person responds.

It is not ideal that our TQ correlates with follow-up score but not with baseline score. This is emerging to be the case generally and is not something particular to evaluating electronic outcome measures.[24, 43, 58] This emergence in our view raises the more general question of whether it is appropriate to use TQs at all to evaluate change in outcome measures. Apart from being overly driven by follow-up score, the assumption that the TQ is sufficiently driven by the same latent construct as the PROM, to the extent that it may be considered a gold standard, may be unrealistic. We have previously explored what people think about when they complete the TQ and what they think about when the complete the paper RMDQ version, and we found discordance.[97] Pain appears to be a greater driver of the TQ and the wording of the TQ (*i.e.* attempting to place focus specifically on function or an explicit domain) does not appear to matter. In the current study, we used the term 'symptoms'. However, in the case that the suggestion arising from our previous research is incorrect, then using a generic wording in the TQ might have the advantage of not favoring any one ePRO over another, but the disadvantage of disassociating the TQ from any specific latent health construct. Use of a generically worded TQ would then introduce some information bias; for example, if people systematically attend more to a particular domain upon reading the word 'symptoms'. We caution that the logic of this typically taken approach of using one outcome measure as a proxy gold standard of recovery, and then using this proxy to judge domain-specific responsiveness and MIC thresholds in another, may be questionable where there is domain mismatch.

There was a small amount of missing data at baseline and one-week (one person in each case) which should have been impossibility as a selection on the eNRS was a required response. We are uncertain of the cause of this but we suspect that this might have been due to use of an obscure and/or obsolete browser.

This research was conducted solely in private care and people who pay to see osteopaths may differ from those attending publically funded health care, as is more routinely the case in health services research. We note a lower than typical baseline severity (as compared to clinical trials) and thus some caution is indicated before generalizing to typical trial populations. Finally, our focus here was on the most commonly used domains and outcome measures in trials. The VAS is most commonly used overall (pain), RMDQ second most common (disability), and the NRS the fourth most commonly used (pain). We did not include the third most commonly used outcome, the Oswestry Disability Questionnaire, which also measures disability.[8] Unlike the VAS and NRS, which are both single-item instruments, including two full disability questionnaires risked being unduly burdensome for participants. Qualitative work suggests that participants would prefer to spend only 5-10 minutes completing ePROs.[98, 99] Including a direct comparison with paper versions would have permitted direct exploration of criterion validity; however, this approach would likely have been affected by panel condition and further added to participant burden.

Recommendations for future research

Sampling stable participants from people receiving routine clinical treatment allows the nesting of a test-retest design makes for an efficient design. However, it produces some challenges for achieving sufficient recruitment over a realistic time period. It assumes that the TQ classification of unchanged is valid. As data suggest that TQ is driven more by follow-up state than change, the approach has some limitations. It would be scientifically preferable that test-retest studies are conducted within untreated populations. However, this has both ethical and practical implications. When planning to nest a test-retest design within any treatment-containing protocol, based on rates observed in this study (using the lower eNRS 'no chance' incidence) we

recommend planning for a study that is around three times larger, *i.e.* seeking around 1,200 people to obtain 50 stable participants. For study of responsiveness alone, around 250 participants should be sufficient to achieve 50 improvements at six-weeks. The most extreme MIC threshold we estimated was 7mm (-12 to 26) for the VAS at 6 weeks. This is lower than has been noted in studies of paper counterparts. Assuming the point estimate is representative of the population parameter, around 300 participants would be required to power a study to confirm the finding.

Retaining data at item level in future studies will permit more sophisticated analytics. There may need to be a cultural change as we transition from paper to digital measurement. The ability to more easily retain greater data resolution is a clear advantage of digital measurement and one that would be sensible to exploit. Further advantages in terms of cost, logistics, form validation, reminders, time logging, environmental factors, and reach, are undeniable and, in our view, make electronic health measurement very attractive. More generally, routine outcome measurement in clinical practice will permit learning healthcare systems and so should be a shared goal by stakeholders across healthcare.[100, 101] To achieve this, greater collaboration may be needed between clinicians, informatics, and policy makers. We also encourage further metric testing of electronic versions these and other legacy PROMs as so that results may inform health services researchers' and clinicians' choices of measure.

Conclusion

Each of the electronic outcome measures have metric properties that do not materially differ from values reported in the literature for their paper counterparts. A possible exception may the reliability of the eVAS, for which there is insufficient existing research to make useful comparisons to the paper version. The eRMDQ is adequate for measuring back-related disability and the eNRS is adequate for measuring pain intensity. The eNRS should be preferred over the eVAS for the measurement of pain intensity.

Acknowledgements

RF and MU conceived of the study, applied for, and were awarded, the funding to do the study. CF undertook the day-to-day management of the study and submitted documents for consideration by the Queen Mary University of London ethics committee. JF and RF were responsible for the data management for the study. RF performed all the analyses. All authors commented on and approved the manuscript. Part of RF and JF's time on the study was funded by the Warwick Impact Fund, which administers the same HEIF5 grant. The study was sponsored by University of Warwick but conducted from Queen Mary University of London that sponsored the remainder of CF's PhD research. Neither the sponsor nor the funder had any involvement in the study design, analysis, or reporting of results.

Conflicts of Interest

RF, MU, and JF are directors and shareholders of Clinvivo Ltd, the University of Warwick spin-out company that provided the software for data collection in this study. The HEIF5 grant that paid for the development of IP that has been licensed to Clinvivo and has been used in this study and also paid for high-street vouchers that were used as incentives to recruit participants into the study. RF and DC are non-practising osteopaths, CF is a practising osteopath.

Abbreviations

AES: Advanced Encryption Standard
AUC: area under the curve
CI: confidence interval
COSMIN: Consensus-Based Standards for the Selection of Health Measurement Instruments.

e: electronic; e.g. eRMDQ is the electronic version of the paper RMDQ
JMIR: Journal of Medical Internet Research
ICC: Intra-class correlation coefficient
IPR: Intellectual Property Rights
IQR: Inter-quartile range
PGP: Pretty Good Privacy
RCT: randomized controlled trial
RMDQ: Roland Morris Disability Questionnaire
ROC: receiver operator characteristic
RSA: Rivest-Shamir-Aldeman
MIC: minimally important change
MDC: minimal detectable change
NRS: Numerical Rating Scale
SE: standard error
SSL: secure socket layer
TQ: transition question
VAS: Visual Analogue Scale

Multimedia Appendix

Multimedia Appendix 1: Technical appendix – technical notes and extended technical results for Guyatt analyses.

References

1. Hartvigsen J, Hancock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, Hoy D, Karppinen J, Pransky G, Sieper J, Smeets RJ, Underwood M, Lancet Low Back Pain

Series Working G. What low back pain is and why we need to pay attention. Lancet. 2018 Mar 20. PMID: 29573870. doi: 10.1016/S0140-6736(18)30480-X.
2. Froud R, Patterson S, Eldridge S, Seale C, Pincus T, Rajendran D, Fossum C, Underwood M. A systematic review and meta-synthesis of the impact of low back pain on people's lives. BMC Musculoskelet Disord. 2014 Feb 21;15:50. PMID: 24559519. doi: 10.1186/1471-2474-15-50.
3. Amundsen PA, Evans DW, Rajendran D, Bright P, Bjorkli T, Eldridge S, Buchbinder R, Underwood M, Froud R. Inclusion and exclusion criteria used in non-specific low back pain trials: a review of randomised controlled trials published between 2006 and 2012. BMC musculoskeletal disorders. 2018 Apr 12;19(1):113. PMID: 29650015. doi: 10.1186/s12891-018-2034-6.
4. Airaksinen O, Brox JI, Cedraschi C, Hildebrandt J, Klaber-Moffett J, Kovacs F, Mannion AF, Reis S, Staal JB, Ursin H, Zanoli G, Pain CBWGoGfCLB. Chapter 4. European guidelines for the management of chronic nonspecific low back pain. Eur Spine J. 2006 Mar;15 Suppl 2:S192-300. PMID: 16550448. doi: 10.1007/s00586-006-1072-1.
5. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. Lancet. 2017 Feb 18;389(10070):736-47. PMID: 27745712. doi: 10.1016/S0140-6736(16)30970-9.
6. Disease GBD, Injury I, Prevalence C. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017 Sep 16;390(10100):1211-59. PMID: 28919117. doi: 10.1016/S0140-6736(17)32154-2.
7. Clark S, Horton R. Low back pain: a major global challenge. Lancet. 2018 Mar 20. PMID: 29573869. doi: 10.1016/S0140-6736(18)30725-6.
8. Froud R, Patel S, Rajendran D, Bright P, Bjorkli T, Buchbinder R, Eldridge S, Underwood M. A Systematic Review of Outcome Measures Use, Analytical Approaches, Reporting Methods, and Publication Volume by Year in Low Back Pain Trials Published between 1980 and 2012: Respice, adspice, et prospice. PLoS One. 2016;11(10):e0164573. PMID: 27776141. doi: 10.1371/journal.pone.0164573.
9. de Vet H, Terwee C, Mokkink L, Knol D. Measurement in Medicine. Cambridge: Cambridge University Press; 2011.
10. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine. 1983 Mar;8(2):141-4. PMID: 6222486.
11. Hayes M, Patterson D. Experimental development of the graphic rating method. Psychol Bull. 18. 1921;Psychol Bull(98).
12. Leavell H. Contributions of the Social Sciences to the Solution of Health Problems. N Engl J Med. 1952 (247):885-97.
13. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul;63(7):737-45. PMID: 20494804. doi: 10.1016/j.jclinepi.2010.02.006.
14. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes B, Malmivaara A, Roland M, Von Korff M, Waddell G. Outcome measures for low back pain research.

A proposal for standardized use. Spine. 1998 Sep 15;23(18):2003-13. PMID: 9779535.
15. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N, Carr DB, Chandler J, Cowan P, Dionne R, Galer BS, Hertz S, Jadad AR, Kramer LD, Manning DC, Martin S, Mc Cormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robbins W, Robinson JP, Rothman M, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Wernicke J, Witter J. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005;113(1-2):9-19. PMID: 2005-01238-001.
16. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, Carrino JA, Chou R, Cook K, DeLitto A, Goertz C, Khalsa P, Loeser J, Mackey S, Panagis J, Rainville J, Tosteson T, Turk D, Von Korff M, Weiner DK. Report of the NIH Task Force on Research Standards for Chronic Low Back Pain. Spine J. 2014 Aug 01;14(8):1375-91. PMID: 24950669. doi: 10.1016/j.spinee.2014.05.002.
17. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, Williams JI. The Quebec Back Pain Disability Scale. Measurement properties. Spine (Phila Pa 1976). 1995 Feb 1;20(3):341-52. PMID: 7732471.
18. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain. J Pain. 2004 Mar;5(2):133-7. PMID: 15042521. doi: 10.1016/j.jpain.2003.12.005.
19. Nishiwaki M, Takayama M, Yajima H, Nasu M, Kong J, Takakura N. The Japanese Version of the Massachusetts General Hospital Acupuncture Sensation Scale: A Validation Study. Evid Based Complement Alternat Med. 2017;2017:7093967. PMID: 28676831. doi: 10.1155/2017/7093967.
20. Wallwiener M, Matthies L, Simoes E, Keilmann L, Hartkopf AD, Sokolov AN, Walter CB, Sickenberger N, Wallwiener S, Feisst M, Gass P, Fasching PA, Lux MP, Wallwiener D, Taran FA, Rom J, Schneeweiss A, Graf J, Brucker SY. Reliability of an e-PRO Tool of EORTC QLQ-C30 for Measurement of Health-Related Quality of Life in Patients With Breast Cancer: Prospective Randomized Trial. J Med Internet Res. 2017 Sep 14;19(9):e322. PMID: 28912116. doi: 10.2196/jmir.8210.
21. de Vet HC. Reproducibility and responsiveness of evaluative outcome measures. International journal of technology assessment in health care. 2001;17(4):479-87.
22. Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, Bouter L, de Vet H. COSMIN checklist manual version 9. Amsterdam: EMGO, 2012.
23. Froud R, Underwood M, Carnes D, Eldridge S. Clinicians' perceptions of reporting methods for back pain trials: a qualitative study. Br J Gen Pract. 2012 Mar;62(596):e151-9. PMID: 22429424. doi: 10.3399/bjgp12X630034.
24. Froud R, Eldridge S, Lall R, Underwood M. Estimating the number needed to treat from continuous outcomes in randomised controlled trials: methodological challenges and worked example using data from the UK Back Pain Exercise and Manipulation (BEAM) trial. BMC Med Res Methodol. 2009 Jun 11;9:35. PMID: 19519911. doi: 10.1186/1471-2288-9-35.
25. Froud R, Eldridge S, Kovacs F, Breen A, Bolton J, Dunn K, Fritz J, Keller A, Kent P, Lauridsen HH, Ostelo R, Pincus T, van Tulder M, Vogel S, Underwood M. Reporting outcomes of back pain trials: a modified Delphi study. European journal of pain (London, England). 2011 Nov;15(10):1068-74. PMID: 21596600. doi: 10.1016/j.ejpain.2011.04.015.

26. de Vet H, Terwee C, Ostelo R, Beckerman H, Knol D, Bouter L. Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes,. 2006;4(54).
27. Froud R, Underwood M, Eldridge S. Improving the reporting and interpretation of clinical trial outcomes. Br J Gen Pract. 2012 Oct;62(603):e729-31. PMID: 23265234. doi: 10.3399/bjgp12X657008.
28. Henschke N, van Enst A, Froud R, Ostelo RW. Responder analyses in randomised controlled trials for chronic low back pain: an overview of currently used methods. Eur Spine J. 2014 Apr;23(4):772-8. PMID: 24419902. doi: 10.1007/s00586-013-3155-0.
29. Ostelo RWJG, Deyo RA, Stratford P, Waddell G, Croft PP, Von Korff M, Bouter LM, de Vet HC. Interpreting Change Scores for Pain and Functional Status in Low Back Pain: Towards International Consensus Regarding Minimal Important Change. Spine. 2008;33(1):90-4.
30. van Tulder M, Becker A, Bekkering T, Breen A, del Real MT, Hutchinson A, Koes B, Laerum E, Malmivaara A, Care CBWGoGftMoALBPiP. Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care. Eur Spine J. 2006 Mar;15 Suppl 2:S169-91. PMID: 16550447. doi: 10.1007/s00586-006-1071-2.
31. Clinvivo. Clinvivo Ltd Website. 2018 [cited 2018 Jan 12]; Available from: https://clinvivo.com/ archived at http://www.webcitation.org/6wPydLMrk.
32. Huskisson E. Measurement of Pain. The Lancet. 1974 (ii):1127.
33. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J. 2003 Feb;12(1):12-20. PMID: 12592542.
34. Scott JH, EC. Vertical or horizontal visual analogue scales. Ann Rheum Dis. 1979;38(560).
35. Ostelo RW, de Vet HC, Knol DL, van den Brandt PA. 24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. Journal of clinical epidemiology. 2004 Mar;57(3):268-76. PMID: 15066687.
36. Dunn KM, Cherkin DC. The Roland-Morris Disability Questionnaire. Spine. 2007 Jan 15;32(2):287. PMID: 17224833.
37. Roland M, Morris R. A study of the natural history of low-back pain. Part II: development of guidelines for trials of treatment in primary care. Spine. 1983 Mar;8(2):145-50. PMID: 6222487.
38. Beurskens A, de Vet H, Koke A. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain. 1996;65:71-6.
39. Roland M. Roland Morris Disability Questionnaire. 1983 [cited 2018 Jan 12]; Available from: www.rmdq.org archived at http://www.webcitation.org/6wPzCvX6M.
40. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. Ann Rheum Dis. 1978 Aug;37(4):378-81. PMID: 686873.
41. Childs J, Piva S, Fritz J. Responsiveness of the Numeric Pain Rating Scale in Patients with Low Back Pain. Spine. 2005;30(11).
42. Farrar JT, Young JP, Jr., LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain. 2001 Nov;94(2):149-58. PMID: 11690728.

43. de Vet H, Ostelo R, Terwee C, van der Roer N, Knol D, Beckerman H, Boers M, Bouter L. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Quality of life research. 2007;16:131-42.
44. Lauridsen HH, Hartvigsen J, Korsholm L, Grunnet-Nilsson N, Manniche C. Choice of external criteria in back pain research: Does it matter? Recommendations based on analysis of responsiveness. Pain. 2007 Feb 1. PMID: 17276006.
45. Davidson M, Keating J. A comparison of five low back disability questionnaires: Reliability and responsiveness. Physical therapy. 2002;82(1).
46. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis. 1986;39(11):897-906. PMID: 2947907.
47. Grotle M, Brox JI, Vollestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine. 2004 Nov 1;29(21):E492-501. PMID: 15507789.
48. Terwee C. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of clinical Epidemiology. 2007;60:34-42.
49. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Controlled clinical trials. 1989 Dec;10(4):407-15. PMID: 2691207.
50. 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. PLoS One. 2014;9(12):e114468. PMID: 25474472. doi: 10.1371/journal.pone.0114468.
51. ROCMIC: Stata module to estimate minimally important change (MIC) thresholds for continuous clinical outcome measures using ROC curves [database on the Internet]. 2004. Available from: <https://ideas.repec.org/c/boc/bocode/s457052.html>.
52. Shrout PF, J. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological bulletin. 1979;86(2):420-8.
53. de Vet HC, Terwee C, Knol DL, Bouter L. When to use agreement versus reliability measures. Journal of Clinical Epidemiology. 2006;59:1033-9.
54. Sitgreaves R. Review of Intraclass correlation and the analysis of variance by E. A. Haggard. Journal of the American Statistical Association. 1960;55:384-5.
55. Fleiss JL, editor. The design and analysis of clinical experiments. New York: Wiley; 1986.
56. Stratford PW. Using the Roland-Morris questionnaire to make decisions about patients. Physiotherapy Canada. 1996;48:107-10.
57. Ostelo RW, de Vet H. Clinically important outcomes in low back pain. Best Practice & Research Clinical Rheumatology. 2005;19(4):593.
58. Guyatt G, Norman G, Juniper E, Grifith L. A critical look at transition ratings. J Clin Epidemiol 2002;55(9):900-8.
59. Tilford JM, Roberson PK, Fiser DH. Using lfit and lroc to evaluate mortality prediction models. . Stata Technical Bulletin 1995;5:77-81.
60. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1991.

61. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29-36. PMID: 7063747. doi: 10.1148/radiology.143.1.7063747.
62. Efron B, Tibshirani R. Bootstrap measures for standard errors, confidence intervals, and other mea- sures of statistical accuracy. Statistical Science. 1986 (1):54-77.
63. Froud R. Improving interpretation of patient-reported outcomes in low back pain trials: Queen Mary University of London; 2010.
64. Gould W, Pitblado J. ACCUM: Stata module. on-line2001 [cited 2018 January 12]; Available from: http://www.stata.com/support/faqs/stat/reps.html archived at http://www.webcitation.org/6wPzQpaXc.
65. Fawkes CA, Leach CM, Mathias S, Moore AP. Development of a data collection tool to profile osteopathic practice: use of a nominal group technique to enhance clinician involvement. . Manual Therapy. 2014;19(2):119-24.
66. Fawkes CA, Leach CM, Mathias S, Moore AP. A profile of osteopathic care in private practices in the United Kingdom: a national pilot using standardised data collection. . Manual Therapy. 2014;19(2):125-30.
67. Maughan EF, Lewis JS. Outcome measures in chronic low back pain. Eur Spine J. 2010 Sep;19(9):1484-94. PMID: 20397032. doi: 10.1007/s00586-010-1353-6.
68. Stratford PW. Sensitivity to Change of the Roland-Morris Back Pain Questionnaire: Part 1. Phys Ther. 1998;78(11).
69. Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine. 1995 May 1;20(9):1017-28. PMID: 7631231.
70. Frost H, Lamb SE, Stewart-Brown S. Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index v2.1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. Spine (Phila Pa 1976). 2008 Oct 15;33(22):2450-7; discussion 8. PMID: 18824951. doi: 10.1097/BRS.0b013e31818916fd.
71. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. Phys Ther. 1994 Jun;74(6):528-33. PMID: 8197239.
72. Mannion AF, Junge A, Grob D, Dvorak J, Fairbank JC. Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery. Eur Spine J. 2006 Jan;15(1):66-73. PMID: 15856340. doi: 10.1007/s00586-004-0816-z.
73. Coelho RA, Siqueira FB, Ferreira PH, Ferreira ML. Responsiveness of the Brazilian-Portuguese version of the Oswestry Disability Index in subjects with low back pain. Eur Spine J. 2008 Aug;17(8):1101-6. PMID: 18512083. doi: 10.1007/s00586-008-0690-1.
74. Brouwer S, Kuijer W, Dijkstra PU, Goeken LN, Groothoff JW, Geertzen JH. Reliability and stability of the Roland Morris Disability Questionnaire: intra class correlation and limits of agreement. Disability and rehabilitation. 2004 Feb 4;26(3):162-5. PMID: 14754627.
75. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Danish version of the Oswestry disability index for patients with low back pain. Part 2: Sensitivity, specificity and clinically significant improvement in two low back pain populations. Eur Spine J. 2006 Nov;15(11):1717-28. PMID: 16736202.

76. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. European journal of pain (London, England). 2004 Aug;8(4):283-91. PMID: 15207508. doi: 10.1016/j.ejpain.2003.09.004.
77. Janwantanakul P, Sihawong R, Sitthipornvorakul E, Paksaichol A. A screening tool for non-specific low back pain with disability in office workers: a 1-year prospective cohort study. BMC Musculoskelet Disord. 2015 Oct 14;16:298. PMID: 26467434. doi: 10.1186/s12891-015-0768-y.
78. Scrimshaw SV, Maher C. Responsiveness of visual analogue and McGill pain scale measures. J Manipulative Physiol Ther. 2001 Oct;24(8):501-4. PMID: 11677548. doi: 10.1067/mmt.2001.118208.
79. Parker SL, Adogwa O, Paul AR, Anderson WN, Aaronson O, Cheng JS, McGirt MJ. Utility of minimum clinically important difference in assessing pain, disability, and health state after transforaminal lumbar interbody fusion for degenerative lumbar spondylolisthesis. J Neurosurg Spine. 2011 May;14(5):598-604. PMID: 21332281. doi: 10.3171/2010.12.SPINE10472.
80. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC musculoskeletal disorders. 2006;7:82. PMID: 17064410.
81. Kovacs FM, Abraira V, Royuela A, Corcoll J, Alegre L, Cano A, Muriel A, Zamora J, del Real MT, Gestoso M, Mufraggi N. Minimal clinically important change for pain intensity and disability in patients with nonspecific low back pain. Spine. 2007 Dec 1;32(25):2915-20. PMID: 18246018.
82. Stratford PW. Sensitivity to Change of the Roland-Morris Back Pain Questionnaire: Part 2. Phys Ther. 1998;78(11).
83. Jordan K, Dunn KM, Lewis M, Croft P. A minimal clinically important difference was derived for the Roland-Morris Disability Questionnaire for low back pain. Journal of clinical epidemiology. 2006 Jan;59(1):45-52. PMID: 16360560.
84. van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine. 2006 Mar 1;31(5):578-82. PMID: 16508555.
85. Rose G. Individuals and populations. The strategy of preventive medicine. Oxford, United Kingdom: Oxford University Press; 1992 p. 12, 53-63, 74.
86. De Vet HC BH, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. J Rheumatol. 2006;33(2):434.
87. Costa L, Meyer C, Latimer J. Self-report outcome measures for low back pain. Spine. 2007;32(9).
88. Mannion AF, Elfering A, Staerkle R, Junge A, Grob D, Semmer NK, Jacobshagen N, Dvorak J, Boos N. Outcome assessment in low back pain: how low can you go? Eur Spine J. 2005 Dec;14(10):1014-26. PMID: 15937673.
89. Parker SL, Mendenhall SK, Shau DN, Adogwa O, Anderson WN, Devin CJ, McGirt MJ. Minimum clinically important difference in pain, disability, and quality of life after neural decompression and fusion for same-level recurrent lumbar stenosis: understanding clinical versus statistical significance. J Neurosurg Spine. 2012 May;16(5):471-8. PMID: 22324801. doi: 10.3171/2012.1.SPINE11842.
90. Bird ML, Callisaya ML, Cannell J, Gibbons T, Smith ST, Ahuja KD. Accuracy, Validity, and Reliability of an Electronic Visual Analog Scale for Pain on a Touch Screen

Tablet in Healthy Older Adults: A Clinical Trial. Interact J Med Res. 2016 Jan 14;5(1):e3. PMID: 26769149. doi: 10.2196/ijmr.4910.
91. Underwood M, Parsons M, Eldridge S, Spencer A, Feder G. Asking older people about fear of falling did not have a negative effect. Journal of Clinical Epidemiology. 2006;59:629-34.
92. Bijur PE, Silver W, Gallagher EJ. Reliability of the visual analog scale for measurement of acute pain. Acad Emerg Med. 2001 Dec;8(12):1153-7. PMID: 11733293.
93. Gallagher EJ, Bijur PE, Latimer C, Silver W. Reliability and validity of a visual analog scale for acute abdominal pain in the ED. Am J Emerg Med. 2002 Jul;20(4):287-90. PMID: 12098173.
94. Bishop FL, Lewis G, Harris S, McKay N, Prentice P, Thiel H, Lewith GT. A within-subjects trial to test the equivalence of online and paper outcome measures: the Roland Morris disability questionnaire. BMC Musculoskelet Disord. 2010 Jun 08;11:113. PMID: 20529332. doi: 10.1186/1471-2474-11-113.
95. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who Uses Mobile Phone Health Apps and Does Use Matter? A Secondary Data Analytics Approach. J Med Internet Res. 2017 Apr 19;19(4):e125. PMID: 28428170. doi: 10.2196/jmir.5604.
96. Mokkink LB, Prinsen CA, Bouter LM, Vet HC, Terwee CB. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther. 2016 Jan 19;20(2):105-13. PMID: 26786084. doi: 10.1590/bjpt-rbf.2014.0143.
97. Froud R, Ellard D, Patel S, Eldridge S, Underwood M. Primary outcome measure use in back pain trials may need radical reassessment. BMC Musculoskelet Disord. 2015 Apr 14;16:88. PMID: 25887581. doi: 10.1186/s12891-015-0534-1.
98. Fawkes C, Carnes D, Froud R. Introducing electronic PROM data collection into clinical practice–learning the lessons from a pilot study. Physiotherapy. 2017;103:e111.
99. Fawkes CA. The development, evaluation, and initial implementation of a national programme for the use and collation of patient reported outcome measures (PROMs) in osteopathic back pain services in the UK.: Queen Mary University of London; 2017.
100. Deeny S, Steventon A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. BMJ Qual Saf. 2015 (0):1-11. doi: doi:10.1136/bmjqs-2015-004278.
101. Celi LA, Davidzon G, Johnson AE, Komorowski M, Marshall DC, Nair SS, Phillips CT, Pollard TJ, Raffa JD, Salciccioli JD, Salgueiro FM, Stone DJ. Bridging the Health Data Divide. J Med Internet Res. 2016 Dec 20;18(12):e325. PMID: 27998877. doi: 10.2196/jmir.6400.