The Journal of Arthroplasty 29 (2014) 2136–2139

Contents lists available at ScienceDirect

The Journal of Arthroplasty

journal homepage: www.arthroplastyjournal.org

The Equivalence of Remote Electronic and Paper Patient Reported Outcome (PRO) Collection

W. Griffiths-Jones, MRCS, M.R. Norton, FRCS (Orth), E.D. Fern, FRCS (Orth), D.H. Williams, MSc, FRCS (Orth)

Royal Cornwall Hospital, Truro TR1 3LJ

ARTICLE INFO

Article history:
Received 10 June 2014
Accepted 1 July 2014

Keywords:
outcomes
hip impingement
electronic health
patient reported outcome measures
hip surgery

ABSTRACT

Individual patient level Patient Reported Outcomes (PROs) are increasingly important in clinical practice. Web-based collection enables clinicians to remotely collect scores at regular intervals, away from the clinic setting. In this randomized crossover study, 47 patients, having undergone hip surgery, were allocated to two groups. Group 1 completed the web-based scores followed by the paper equivalents one week later; Group 2 completed the scores the other way around. The Intraclass Correlation Coefficient (ICC) for the Oxford Hip Score was 0.99, 0.98 to 0.99 (ICC, 95% CI) and the ICCs for the other scores were between 0.95 and 0.97. We conclude that remote ePRO collection using this web-based system reveals excellent equivalence to paper PRO collection of the Oxford Hip, McCarthy, UCLA and howRu scores.

Patient Reported Outcomes (PROs) are increasingly being used to measure what is important to patients in terms of the severity of symptoms and the level of function. PROs were primarily developed as assessment tools for use in medical research but are increasingly used to enhance decision-making during the doctor-patient consultation [1–7]. While clinic based electronic systems enable easier collection of PRO data in some areas [2,3], long term PRO follow up at regular intervals will rely on web-based electronic PRO (ePRO) collection remote or away from the hospital setting. Remote ePRO collection has yet to be fully assessed.

The primary aim of this study is to assess whether a number of PROs – the Oxford Hip, McCarthy hip, University of California Los Angeles activity (UCLA) and the howRU general health score – collected remote or away from the hospital setting via an electronic web-based system are equivalent to the same scores collected via traditional pen and paper in a group of patients who have already undergone open hip surgery.

Materials and Methods

In 2009 the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) electronic Patient Reported Outcomes (ePRO) task force published recommendations on evidence needed to support measurement of the equivalence of electronic and paper based PROs [1]. The goal of these recommendations was to be explicit about how much additional validation is required when converting a paper version of a PRO score to an electronic version. The report describes examples of minor, moderate and substantial modifications together with the respective level of testing required but does not classify location i.e. whether or not it is completed at or remote from the clinic.

The web-based system in use at our institution collects PROs at regular intervals throughout the patient care pathway (Fig. 1). The system gives patients and their medical team the opportunity to monitor PRO scores over time and to compare response to treatment with that of other patients. The Oxford Hip, McCarthy hip, University of California Los Angeles activity (UCLA) and the howRU general health score are collected for patients registering a hip problem on the system [4–7]. As per system design, the content and format of the electronic versions of the scores were identical to the paper questionnaires. The Oxford Hip Score, a twelve question set scored from 0 to 48 (worst to best), was designed to assess the outcome of hip arthroplasty and is commonly used for the pre-operative assessment of hip pain and function in arthritis and other hip pathology [4]. The McCarthy hip score or 'nonarthritic hip score' is a twenty question set scored out of a possible 80 designed to assess the severity of hip symptoms in a younger more active group of patients [5]. The UCLA activity score is a single point scale from 1 to 10 designed to evaluate current activity level [6]. The howRU is a general health and quality of life score composed of 4 questions scored out of 12 [7].

Equivalence testing is designed to ensure that PROs collected electronically do not vary significantly from those scores collected via pen and paper and there are two recommended study designs: 1) the parallel group and 2) the crossover design. Any difference between the modes of administration should not exceed the minimally important differences (MIDs) of the measures being assessed. In studies of health-related quality of life where MIDs were estimated, most estimates clustered around half a standard deviation defined as "medium" by Cohen [8]. Psychology research supports half a standard

The Conflict of Interest statement associated with this article can be found at http://dx.doi.org/10.1016/j.arth.2014.07.003.

Reprint requests: W Griffiths-Jones, MRCS, Royal Cornwall Hospital, Truro TR1 3LJ.

http://dx.doi.org/10.1016/j.arth.2014.07.003
0883-5403/

deviation as close to the limit of most people's ability to discriminate between changes over a wide range of tasks [9]. Therefore a difference between modes of administration of less than half a standard deviation should be considered.

So, to rule out a difference of 0.3 standard deviations using the first parallel group design, a two-sample t test based on 234 patients per group would provide 80% power with a two-tailed alternative and a 5% Type I error rate. However, as patients act as their own controls in the second, crossover study design, there is greater statistical power and a smaller sample size is required. The sample size calculated above can be multiplied by a factor of $(1 - p)/2$ where p is an estimate of the expected correlation between the two modes of administration [1]. Given a previous mean published correlation of 0.90 [3], a conservative estimate of $P = 0.80$ was chosen, giving a multiplier of 0.1 and requiring a total target sample size of 47 patients.

The United Kingdom National Research and Ethics Service provided approval for the study. Consecutive patients, consenting to participate in the study, had previously undergone open hip debridement surgery for the treatment of femeroactetabular impingement at least two years previously and were randomly allocated to one of two groups. Group 1 patients were asked to remotely register on the web-based system, complete the specified electronic PRO (ePRO) questionnaires and open a separate envelope one week later. Instructions inside that envelope asked patients to complete and return the paper PRO questionnaires via a stamped, addressed envelope. Group 2 patients were asked to complete the paper PROs first and to complete the ePROs on the web-based system one week later. A period of one week was specified to minimize memory or testing effects from the first administration (referred to as a carryover effect), but not so long that the underlying symptoms being reported might actually change. Of the sixty-seven patients recruited, 19 patients completed only a paper score and one patient completed only an electronic score resulting in 47 complete data sets. There were 27 patients in Group 1 comprising 5 male and 22 female patients with a mean age of 41.6 years who completed each score a mean of 7.3 days apart. There were 20 patients in Group 2 comprising 7 male and 13 female patients with a mean age of 43.2 years who completed each score a mean of 7.9 days apart.

The scores from Group 1 and 2 were combined to allow comparison of the two modes of administration, firstly by comparing the mean paper PRO and the mean ePRO scores and secondly by calculating the Intraclass Correlation Coefficient (ICC). Statistical analysis was performed using SPSS for Windows 2012 and Student's t-tests applied to determine if there were differences between the group means, with a P value of less than 0.05 deemed significant.

The Intraclass Correlation Coefficient (ICC) is the descriptive statistic used to measure agreement or equivalence between two quantitative groups of (approximately) continuous distributions. Cohen's kappa coefficient is not appropriate as it is a statistical measure of inter-rater agreement for categorical items and use of Pearson's or Spearman's correlation coefficients alone is not recommended because these tests are not sensitive to systematic mean differences between groups and can therefore can overestimate agreement. The ICC is able to assess both the covariance and degree of agreement between score distributions [10,11] and has been used to examine the equivalence of paper PROs and ePROs previously [3].

Fig. 1. The web-based system showing surgeon side aggregated data.

The recommended levels of acceptable reliability are at least 0.70 for group comparisons and 0.85 to 0.95 for applications at the individual level [12,13].

Results

The mean difference between the web-based ePRO and paper PRO Oxford score in Group 1 was 0.1, −1.0 to 1.1 (mean, 95% confidence interval (CI)) and in Group 2 was −0.7, −1.3 to −0.2 (mean, 95% CI). There was no significant difference between Group 1 and Group 2 (Table 1).

Combining the scores into paper PRO and an ePRO groups revealed a paper Oxford Hip Score of 32.8, 29.7 to 35.8 (mean, 95% CI) and an ePRO Oxford Hip Score of 33.0, 29.9 to 36.1 (mean, 95% CI) ($P = 0.99$). The Intraclass Correlation Coefficient (ICC) for the Oxford Hip Score was 0.99, 0.98–0.99 (ICC, 95% CI) (Table 2).

Given that the correlation between these scores was very high, it was considered appropriate to compare the means and calculate the ICC between the ePRO and the paper PRO scores of the remaining PROs even though there were fewer complete data sets [1]. Thus, there were a paper McCarthy hip score of 48.7, 40.8 to 56.6 (mean, 95% CI) and an ePRO McCarthy hip score of 51.0, 43.2 to 58.8 (mean, 95% CI) ($P = 0.99$), a paper PRO UCLA activity score of 5.4, 4.5 to 6.4 (mean, 95% CI) and an ePRO UCLA activity score of 5.1, 4.2 to 6.1 (mean, 95% CI) ($P = 0.99$) and a paper howRU of 8.7, 7.7 to 9.7 (mean, 95% CI) and an ePRO howRU of 8.7, 7.6 to 9.8 (mean, 95% CI) ($P = 0.99$) (Table 2). The ICC for the McCarthy hip, UCLA activity and howRU scores was 0.97, 0.96 and 0.95 respectively (Table 2).

Table 2
Mean (95% CI) Overall ePRO and Paper PRO Score With Pearson Coefficient and Intraclass Correlation Coefficient (95% CI).

	Oxford Hip Score (n = 47)	McCarthy Hip Score (n = 24)	UCLA Activity Score (n = 23)	howRU (n = 23)
Mean Paper PRO score (95% CI)	32.8 (29.6–36.0)	48.7 (40.8–56.6)	5.4 (4.5–6.4)	8.7 (7.7–9.7)
Mean ePRO score (95% CI)	33.0 (29.8–36.3)	51.0 (43.2–58.8)	5.1 (4.2–6.1)	8.7 (7.6–9.8)
Pearson Coefficient	0.98	0.95	0.92	0.89
Intraclass Correlation Coefficient (95% CI)	0.99 (0.98-0.99)	0.97 (0.93–0.99)	0.96 (0.90–0.98)	0.95 (0.87–0.98)

Discussion

Patient Reported Outcomes (PROs) are standard questionnaires, developed with input from patients, that use the scores from individual questions (items) together, usually summed, to produce an overall score (or a number of scores) to represent a particular underlying construct or domain. This is justified and shown to be meaningful by testing and presenting evidence for the scores' measurement properties of reliability, validity and responsiveness. The collection of pre-operative and post-operative PROs has traditionally been used to assess the efficacy of surgical interventions [14] and can be applied to a wide range of chronic medical and surgical conditions into the long term.

The opportunity for the patient and their surgeon to compare the severity of their current symptom score – with their previous scores, with the scores of other similar patients and with the scores of patients who have already undergone treatment – is often lost because of the logistical delay in entering the individual paper score into a database. Some of the most useful information to be gained from PROs data is the variation in scores over time to enhance patient follow up – both in the early post-operative period [15] and into the long-term – thus enabling the real-time surveillance of most chronic clinical conditions. A robust system can potentially pick up a failing implant before traditional follow up time points [16], so long as collection does not rely upon a face-to-face clinic visit [17].

Such completion of pen and paper PRO questionnaires in routine practice, at the same time as the clinic appointment, results in data generally only useful for assessing individuals. As there is often no standardized period of assessment relative to the intervention, paper PROs can be less than useful – i.e. biased or misleading – for future analysis when aggregated with other peoples' scores. Centrally organized mailing of questionnaires to patients at particular time points is recommended to regularize follow-up appropriately but logistical constraints prevent the rollout of such programs to all patients who might benefit [18]. ePROs, deployed remote from the clinic setting, could potentially enhance the decision making process, promote patient centered care at the individual level and help to standardize assessment time periods of aggregated data, enabling reliable future analysis.

The use of technology in the collection of patient reported outcome measures is not new [3,19]. A meta-analysis of 46 unique studies provided extensive evidence that paper and computer administered PROs are equivalent [3]. Computer use in the clinical setting can improve the collection and collation of patient reported outcome data in many groups of patients, including the elderly demographic [20], and having the results of electronic scores directly available has been found to improve the quality of care [19]. The scores obtained from paper, touch screen and web-based clinic collection within the clinic setting have been shown to be equivalent in patients undergoing total hip arthroplasty at a single timepoint [21]. While web-based equivalence to paper-based collection has been determined in clinic for patient with psoriatic arthirits [22], it was previously not known if scores on a web-based system, collected remotely, away from the clinic setting are equivalent to scores collected by traditional pen and paper.

Previous models have suggested that kappa coefficients, usually equivalent to the ICC, of less than 0.40 are poor, 0.40–0.59 are fair, 0.60–0.74 are good and greater than 0.74 are excellent [10]. Furthermore, the ISPOR ePRO taskforce suggests the use of the ICC in most cases with a level of at least 0.70 for group comparisons and 0.85 to 0.95 for individual applications [1]. Thus, using the Oxford Hip Score, McCarthy hip score, University of California Los Angeles activity (UCLA) score and howRU [4–7] score on this web-based system, Intraclass Correlation coefficients ranging from 0.95 to 0.99 here reveal excellent equivalence of remote ePRO collection with traditional pen and paper collection for both group and individual

Table 1
Mean Difference (95% CI) Between ePRO and Paper PRO Score in Each Group.

	Oxford Hip Score	McCarthy Hip Score	UCLA Activity Score	HowRU Score
Group 1	0.1 (−1.0 to 1.1) N = 27	2.0 (−1.3 to 5.3) N = 14	−0.3 (−0.7 to 0.1) N = 14	0.1 (−0.7 to 0.8) N = 14
Group 2	−0.7 (−1.3 to −0.2) N = 20	2.7 (−1.6 to 7.0) N = 10	−0.3 (−1.2 to 0.5) N = 9	−0.1 (−0.7 to 0.5) N = 9

applications. The web-based system in use at our institution is therefore suitable for collecting these ePROs away from the clinic setting. Further work to assess the remote collection of other ePRO scores in other subspecialties is required.

References

1. Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR epro good research practices task force report. Value Health 2009;12(4):419.
2. Williams D. The myClinicalOutcomes website: providing real-time, patient-level proms data. Bull R Coll Surg Engl 2012;94(1):20.
3. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value Health 2008;11(2):322.
4. Dawson J, Fitzpatrick R, Carr A, et al. Questionnaire on the perceptions of patients about total hip replacement. J Bone Joint Surg (Br) 1996;78(2):185.
5. Christensen CP, Althausen PL, Mittleman MA, et al. The nonarthritic hip score: reliable and validated. Clin Orthop Relat Res 2003;406:75.
6. Amstutz HC, Thomas BJ, Jinnah R, et al. Treatment of primary osteoarthritis of the hip. A comparison of total joint and surface replacement arthroplasty. J Bone Joint Surg Am 1984;66(2):228.
7. Benson T, Sizmur S, Whatling J, et al. Evaluation of a new short generic measure of health status: howRU. Inform Prim Care 2010;18(2):89.
8. Cohen J. A power primer. Psychol Bull 1992;112(1):155.
9. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41(5):582.
10. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420.
11. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613.
12. Fleiss JL. Statistical methods for rates and proportions. John Wiley & Sons; 1973.
13. Nunnally JC, Bernstein IH, Berge JMF. Psychometric theory. New York: McGraw-Hill; 1967.
14. Greenhalgh J. The applications of pros in clinical practice: what are they, do they work, and why? Qual Life Res 2009;18(1):115.
15. Rothwell AG, Hooper GJ, Hobbs A, et al. An analysis of the oxford hip and knee scores and their relationship to early joint revision in the new zealand joint registry. J Bone Joint Surg (Br) 2010;92(3):413.
16. Bannister GC. Total hip replacement: a guide to best practice. London: British Orthopaedic Association; 1999.
17. Valderas JM, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. Qual Life Res 2008;17(2):179.
18. Dawson J, Doll H, Fitzpatrick R, et al. The routine use of patient reported outcome measures in healthcare settings. BMJ 2010;340:c186.
19. Richter JG, Becker A, Koch T, et al. Self-assessments of patients via tablet PC in routine patient care: comparison with standardised paper questionnaires. Ann Rheum Dis 2008;67(12):1739.
20. Yarnold PR, Stewart MJ, Stille FC, et al. Assessing functional status of elderly adults via microcomputer. Percept Mot Skills 1996;82(2):689.
21. Shervin N, Dorrwachter J, Bragdon CR, et al. Comparison of paper and computer-based questionnaire modes for measuring health outcomes in patients undergoing total hip arthroplasty. J Bone Joint Surg Am 2011;93(3):285.
22. MacKenzie H, Thavaneswaran A, Chandran V, et al. Patient-reported outcome in psoriatic arthritis: a comparison of web-based versus paper-completed questionnaires. J Rheumatol 2011;38(12):2619.