Psychometric equivalence of electronic and telephone completion of the ICIQ modules

Abstract

Aims: To assess the equivalence of touch-screen (hand-held iPad) and telephone completion of patient-completed International Consultation on Incontinence Questionnaire (ICIQ) modules by comparison with corresponding data collected using conventional paper-and-pencil methods.

Methods: Men and women, attending urology outpatients complaining of LUTS, were randomised to one of three groups which determined the order in which they completed three administrations of the same questionnaire: paper, iPad and telephone. Four ICIQ questionnaires were evaluated: ICIQ-MLUTS, ICIQ-LUTSqol, ICIQ-OABqol and ICIQ-UI SF.

Results: From August 2012 to October 2014 a total of 448 out of 491 (91%) recruits completed the first two administrations and were included in the analysis. 348 out of 491 (71%) completed the phone administration. The intra-class correlation coefficient (ICC) and Kappa statistic were calculated where appropriate between completed pairs of administrations. Mean ICC correlations were high (>0.8) between paper and iPad administrations. Paired paper and phone administrations were less well correlated, although still high (mean ICC > 0.75). This may be partly due to the practical limitation that the phone interview was completed up to a week later than the initial two administrations. There was no evidence that potential moderator effects (gender, age, and experience with computers or touch screen devices) significantly affected overall reliability of scores between administrations.

Conclusions: We can recommend the interchangeable use of ICIQ electronic or paper based questionnaires in a clinical or research setting. Self-report is preferable to telephone delivery where possible.

Key words: electronic; equivalence; ICIQ; paper-and-pencil; patient reported outcomes; questionnaire; validation.

Introduction

The International Consultation on Incontinence Questionnaire (ICIQ) modules offer a range of patient completed questionnaires for lower urinary tract, bowel and vaginal symptoms (www.ICIQ.net) (1). Each has been given the highest grade of recommendation based on their degree of validation by the International Consultation on Incontinence (ICI) and are recommended for use in all clinical trials in order to standardise outcome assessment (2). The questionnaires are in widespread international clinical use and have been translated into more than 40 different languages.

The availability of psychometrically robust PRO questionnaires is important for the initial assessment, follow-up and monitoring of treatment strategies. The ICIQ questionnaires were originally developed and validated for pencil-and-paper administration. However, a questionnaire which can be shown to be reliable in different administrative formats has further potential for utility in different contexts (3). There are a number of potential advantages to electronic data capture when collecting PRO data. These include the reduction of administrative data entry workload, greater accuracy and completeness of data (4–6). Patient acceptance of using tablet computers to complete questionnaires is generally high which also increases compliance (7–9). A questionnaire which has been validated for equivalence over the telephone can provide additional flexibility for delivery (10–12).

The conversion to electronic patient reported outcomes (ePROs) requires the demonstration of equivalence to the paper-and-pencil version (13,14). Scores cannot be assumed to be equivalent (15), and must not differ simply due to the method of data collection that is used. Current evidence recommends that full psychometric validation may be unnecessary for minor modifications to questionnaire format (13,16). However, in light of the international use and extensive number of ICIQ questionnaires available it was deemed necessary to evaluate equivalence. This study aimed to evaluate whether scores obtained from patient-completed entry of four different ICIQ PRO questionnaires on a touch screen device (iPad) or when administered by proxy over the telephone are sufficiently well correlated with corresponding data collected using conventional pencil-and-paper methods.

Material and Methods

Participants

Men and women attending the urological outpatients department of the Bristol Urological Institute (UK) were recruited between August 2012 and October 2014. Patients were included if they complained of lower urinary tract symptoms (LUTS), including overactive bladder or stress incontinence symptoms. Written consent was taken and participants were subsequently asked to complete the most appropriate questionnaire according to their gender and the nature of their symptoms.

Study Design

The study was based on a randomised crossover design (16). Each participant completed the assigned questionnaire a total of three times, in an order determined by a pre-generated randomisation list. The first two administrations were completed in the clinic with at least 20 minutes between administrations, in which time participants also completed other tasks related to their clinical appointment. Participants were telephoned approximately one week

later to complete the third (telephone) administration. Participants were randomised to one of three groups which determined the order in which they completed the paper and iPad questionnaires.

1) Paper: iPad: Phone
2) iPad: Paper: Phone
3) Paper: Paper: Phone

The crossover design was selected to avoid possible bias and ordering effects. The paper versus paper test-retest group (3) was the control group, from which any differences between paper and iPad could be compared. The phone administration was completed last by all participants for practical reasons. Data was uploaded in real-time during entry onto the iPad to a web-based database. The data from the other administrations were subsequently entered onto this database using unique patient identification numbers. Ethics committee approval was provided by NRES Committee South West – Central Bristol.

Questionnaires

A total of four ICIQ questionnaires were tested as part of this study: ICIQ-LUTSqol – Lower Urinary Tract Symptoms quality of life, ICIQ-MLUTS – Male Lower Urinary Tract Symptoms, ICIQ-OABqol – Overactive Bladder quality of life, and ICIQ-UI SF – Urinary Incontinence short form.

Electronic questionnaire development

The only modification from the original paper-based questionnaires was to present one-question-per-screen as opposed to the multiple-questions-per-page in the paper version. These were considered 'minor modifications' (16), as no changes were made to either content or meaning of the questionnaires. The interface was informed by qualitative cognitive interviewing in order to be as user-friendly as possible. This is a method which helps determine whether the respondent understands and uses the questionnaire as intended by the developer (17). On the basis of this evidence, modifications were made to the appearance and functionality of the electronic versions and were implemented for the quantitative testing phase. For example, some participants had difficulty recognising when a response option had been selected. In the final design, the selection clearly changes to a darker colour when touched, before moving on to the next question.

Sample Size

The sample size required for 80% power was estimated using the following assumptions. Assuming an underlying population intra-class correlation (ICC) coefficient of 0.85, 43 patients with complete paired observations would be required in order to declare that true population reliability is above an ICC of 0.75 (at 95% confidence). Based upon this calculation, a sample of 50 for each group of questionnaire administrations was considered appropriate for this study. This gave a required sample size of 150 recruits per questionnaire.

Statistical analysis

The statistical package R (18) was used for analysis alongside SPSS (version 21.0). The primary aim of this study was to determine whether answers differed between the electronically administered questionnaires and the original paper version. Equivalence was further assessed by comparison with any differences found between the test-retest

administrations of the paper versions. In addition, the phone administered version of the questionnaire was compared with responses to the paper version. The intra-class correlation coefficient (ICC) is appropriate to assess the test-retest reliability of PRO questionnaires and a coefficient value of greater than 0.7 is considered to be a good indicator of reliability (13). For each questionnaire, the ICC was calculated for paired questionnaire items between completed pairs of administrations. The mean of these item-level ICC values were then compared for each questionnaire and administration pairing. Due to the nominal response items of the ICIQ-UI SF, the kappa coefficient was used as a more appropriate statistical test. The kappa statistic provides a chance-corrected measure of agreement between ratings on an either nominal or ordinal scale (19).

Results

Recruitment

A total of 491 recruits with 448 out of 491 randomised patients (91%) completed the first two administrations and 348 out of 491 (71%) completed the phone administration. The minimum requirement of 43 patients with complete paired observations was achieved for all questionnaire sub-groups. Table I shows the number of complete paired administrations for each questionnaire. If a patient did not complete the first two administrations, another patient was recruited in their place. More than fifty data pairs were achieved for the ICIQ-UI Short Form as given its brevity it was completed alongside one of the other questionnaires being evaluated. Fifty data pairs were not achieved for the analysis in three of the administration groups as the second administration was not completed by some recruits. Reasons for participants not completing both initial administrations included changing their mind, worries about car-parking and being called away for other medical reasons. 30 individuals were approached and recorded as not willing to enter the study. Common reasons proffered were that they did not have their reading glasses or they did not have time. There was no indication that refusal or non-completion was related to questionnaire mode of administration. The phone administration responses were paired with the first occurring paper administration responses from each randomisation group.

Administration equivalence

Fig. 1 directly compares the mean item-level ICCs of the pairs of administrations of interest in this study. The first two columns (Paper versus iPad, iPad versus Paper) present the effect of reversing the order of the first two administrations and show a mean item-level ICC of >0.8. Thus, it was possible to combine the data for the first two administrations, presented as the ‘iPad/paper paper/iPad comb’ columns. For all pairs of administrations, 95% confidence intervals (CIs) are overlapping and of a narrow width of between 0.06 and 0.10, indicating consistent ICC variability at the questionnaire item-level. The paper versus paper test-retest is presented in the fourth column with overlapping CIs and a mean ICC of >0.8 for each of the three questionnaires. When the first paper administered responses are compared to responses to the phone delivered questionnaire responses (paper versus phone) the mean ICC is less at approximately 0.75 for each questionnaire.

Item-level reliability

Of the 182 question item-level ICCs calculated between paper and iPad administrations across each of these three questionnaires, 167 (92%) showed correlations of greater than 0.75. Item-level response inconsistencies were explored in more detail. Fig. 2 gives an example of response variation for the first question of ICIQ-LUTSqol. The majority of responses between pairs of administrations were identical but for this particular question, the iPad followed by paper test responses (B) were more variable than paper followed by iPad (A) responses. The paper followed by paper administrations (C) showed a similar level of variability which is also common to the other questions and questionnaires at the item-level.

ICIQ-UI SF

The responses for questions 1 to 3 on the ICIQ-UI SF are graded on an ordinal scale and Q4a-h are nominal (yes/no). A higher proportion of responses, between repeat administrations, would be expected to agree by chance for the nominal items. The kappa coefficient is therefore a more appropriate statistical test for this questionnaire. Kappa statistic (κ): 0 = poor, 0.01-0.20 = slight, 0.21-0.40 = fair, 0.41-0.6 = moderate, 0.61-0.8 = substantial, and 0.81-1 = almost perfect agreement. Table II shows for paper versus paper and paper versus iPad, 10 of the 11 questions have a kappa statistic which may be described as substantial (κ=0.61-0.8) to almost perfect agreement (κ=0.81-1) (20). For paper versus telephone, 7 of the 11 questions may be described at this level of agreement.

Moderator effects

Corresponding patient data on age, experience with touch screen devices, computer use and gender were analysed for any potential moderator effects on the equivalence of the iPad questionnaire responses. Table III details the demographic statistics for the sample analysed for moderator effects. The mean ICC when stratified by age (<65 years, ≥66 years), computer use (≤15 times a month, >15 times a month), gender (male, female) and experience with touch screen devices (yes, no) was consistently over the acceptable range (>0.7) for every category in the questionnaires tested. Fig. 3, graph A shows small reductions in mean ICC for the older sample (≥ 66 years). A reduction in mean ICC for the older adult sample was also evident in the paper versus paper sample. Wider CIs in the groups using computers ≤15 times a month (C) and having 'no' experience with touch screen devices (E), suggest some increased item-level ICC variability for these groups. Gender of participant had no statistically significant effect on mean ICC in the questionnaires tested (D).

Discussion

The results show that iPad and paper-and-pencil administrations of the ICIQ modules tested produce scores that are equivalent. This was demonstrated by very high overall correlations that were no different from those obtained by repeated paper administrations. As the modification made when migrating from paper to electronic versions was minor, it is reasonable to generalise that any of the fully validated ICIQ modules would have the same level of equivalence (16).

Scores obtained over the telephone were slightly less reliable, although showed overall correlations that are still considered high. Previous studies comparing telephone interview questionnaires with paper versions have also found high correlations (3,12). The comparatively less reliable scores for telephone delivery in the current study may have been

due to the inherent nature of the proxy-delivery over the telephone. This makes several demands on the patient which are not comparable to the other modes of administration. The requirement of the patient to remember the question response categories before making an answer may introduce error as there is no opportunity for visual review of the possible responses before answering. There is also the perception of increased confidentiality or privacy when filling in a paper questionnaire in comparison to delivering answers to an investigator over the phone. In addition, there was the practical limitation that the phone administration was completed approximately one week following the initial two administrations allowing for some increased variability in responses. Telephone administration is therefore considered an acceptable method of delivery for ICIQ questionnaires, but self-completion using the electronic or paper formats should be used where possible due to their higher reliability.

The presence of equivalent questionnaire item-level variability in the paper versus paper test-retest responses indicates that any variability in the other administrative pairings were not due to mode of administration. Overall mean ICCs plus 95% CIs presented in figure 1 are a good demonstration of the similar overall variability of questionnaire answers between paired administrative formats. Individual item-level variation (fig. 2) may be expected due to participants simply making a mistake, changing their mind, or because their situation has changed since reading the question the first time. Correlations of repeated tests by different modes of administration are not expected to be 1.0 (16).

Age, gender, computer or touch screen experience was evaluated for any potential effect on equivalence. There was very little evidence of any effect as correlations were consistently high for all groups. The slight reduction in mean ICCs for patients aged 66 or over was also present in the paper versus paper retest group, suggesting this was unlikely to be due to any effect of the mode of administration. This is in agreement with other studies which conclude that the reduction of the test-retest reliability of questionnaires can be attributed to increasing age (13,21). There was some evidence for increased variation in item-level ICCs attributable to lack of experience with computers or touch screen devices. The validity of the conclusions remain unaffected as the overall mean ICCs were well above the acceptable level of reliability.

This study provides the validation required for the development and use of an application or ‘app’ to be made available for patients to be able to complete electronic versions of the questionnaires on mobile touch screen devices. In addition to the clear clinical advantages of electronic data capture, it is preferable to give patients the option of completing the questionnaires using the mode of administration most suitable to their needs or preferences (9). The results of this study provide justification for increasing the versatility of the ICIQ’s numerous modules, and the required evidence base that electronic versions of the questionnaires are as robust as their original versions.

Conclusions

We can recommend the use of electronic or paper based formats of ICIQ questionnaires in both clinical or research settings, as tested in this study. eICIQ modules will be available through the ICIQ website: www.iciq.net. Although correlations were high for the phone administered questionnaires, on the basis of the current evidence it is preferable to use self-

completion if at all possible, in accordance with the original intended mode of administration of the ICIQ modules.

Acknowledgements

The authors acknowledge the contribution of the staff and patients of the Bristol Urological Institute, Southmead Hospital, UK. The authors also acknowledge Liz Neagle for the provision of an educational grant by Astellas Pharma Ltd., UK and Ian Weir for the statistical programming for the project.

References

1. ICIQ | Home [Internet]. [cited 2015 Nov 18]. Available from: http://www.iciq.net/

2. Kelleher C. Patient Reported Outcome Assessment. In: Abrams P, Cardozo L, Khoury S, Wein A, editors. Incontinence, 5th International Consultation on Incontinence. 5th edition. Health Publications Ltd.; 2013. p. 398–429.

3. Limperopoulos C, Majnemer A, Steinbach CL, Shevell M. Equivalence Reliability of the Vineland Adaptive Behavior Scale Between In-Person and Telephone Administration. Phys Occup Ther Pediatr. 2006 Jun 27;26(1):115–27.

4. Bushnell DM, Reilly MC, Galani C, Martin ML, Ricci J-F, Patrick DL, et al. Validation of Electronic Data Capture of the Irritable Bowel Syndrome—Quality of Life Measure, the Work Productivity and Activity Impairment Questionnaire for Irritable Bowel Syndrome and the EuroQol. Value Health. 2006 Mar;9(2):98–105.

5. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. Bmj. 2002;324(7347):1193–4.

6. Velikova G, Wright EP, Smith AB, Cull A, Gould A, Forman D, et al. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. J Clin Oncol. 1999;17(3):998–998.

7. Aiello EJ, Taplin S, Reid R, Hobbs M, Seger D, Kamel H, et al. In a randomized controlled trial, patients preferred electronic data collection of breast cancer risk-factor information in a mammography setting. J Clin Epidemiol. 2006 Jan;59(1):77–81.

8. Bischoff-Ferrari HA. Validation and patient acceptance of a computer touch screen version of the WOMAC 3.1 osteoarthritis index. Ann Rheum Dis. 2005 Jan 1;64(1):80–4.

9. Hess R, Santucci A, McTigue K, Fischer G, Kapoor W. Patient Difficulty Using Tablet Computers to Screen in Primary Care. J Gen Intern Med. 2008 Apr;23(4):476–80.

10. Bellamy N, Campbell J, Hill J, Band P. A comparative study of telephone versus onsite completion of the WOMAC 3.0 osteoarthritis index. J Rheumatol. 2002;29(4):783–6.

11. Lungenhausen M, Lange S, Maier C, Schaub C, Trampisch HJ, Endres HG. Randomised controlled comparison of the Health Survey Short Form (SF-12) and the Graded Chronic Pain Scale (GCPS) in telephone interviews versus self-administered questionnaires. Are the results equivalent? BMC Med Res Methodol. 2007;7(1):50.

12. Resnik LJ, Clark MA, Borgia M. Telephone and face to face methods of assessment of veteran's community reintegration yield equivalent results. BMC Med Res Methodol. 2011;11(1):98.

13. 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 Mar;11(2):322–33.

14. Schulenberg SE, Yutrzenka BA. The equivalence of computerized and paper-and-pencil psychological instruments: Implications for measures of negative affect. Behav Res Methods Instrum Comput. 1999 Jun 1;31(2):315–21.

15. Juniper EF, Langlands JM, Juniper BA. Patients may respond differently to paper and electronic versions of the same questionnaires. Respir Med. 2009 Jun;103(6):932–4.

16. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, 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 Jun 1;12(4):419–29.

17. Beatty PC, Willis GB. Research Synthesis: The Practice of Cognitive Interviewing. Public Opin Q. 2007 Jun 20;71(2):287–311.

18. R Development Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2008 [cited 2014 Dec 3]. Available from: http://www.R-project.org

19. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Phys Ther. 2005 Mar 1;85(3):257–68.

20. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977 Mar;33(1):159.

21. Grafton KV, Foster NE, Wright CC. Test-retest reliability of the Short-Form McGill Pain Questionnaire: assessment of intraclass correlation coefficients and limits of agreement in patients with osteoarthritis. Clin J Pain. 2005;21(1):73–82.

Table I. Number of completed paired administrations included in the analysis.

Table II. Statistics of agreement and kappa statistic for ICIQ-UI SF.

Fig. 1. The mean of the question item-level intra class coefficients by pair of administration tested, with 95% confidence intervals. ICIQ-LUTSqol (A), ICIQ-MLUTS (B), and ICIQ-OABqol (C).

Fig. 2. An example of the variation in differences in scores between pairs of administrations for Q1from the ICIQ-LUTSqol. “To what extent does your urinary problem affect your household tasks (e.g. cleaning, shopping, etc.)” The options given are scored from 1 to 4: not at all, slightly, moderately, a lot.

Fig. 3. Questionnaires’ mean item-level intra class coefficients for: paper versus iPad (combined) stratified by age (A); for paper versus paper stratified by age (B); and for paper versus iPad (combined), stratified by computer use (C), gender (D), and experience with touch screen devices (E). Error bars represent 95% confidence intervals.