**EQUIVALENCE TESTING OF A NEWLY DEVELOPED INTERVIEWER-LED TELEPHONE SCRIPT FOR THE EORTC QLQ-C30**

Claire Piccinin1, Madeline Pe1, Dagmara Kuliś1, James W. Shaw2, Sally J. Wheelwright3, & Andrew Bottomley1

1. Quality of Life Department, EORTC, Brussels, Belgium
2. Patient-Reported Outcomes Assessment, Bristol-Meyers Squibb, Lawrenceville, New Jersey
3. Macmillan Survivorship Research Group, University of Southampton, Southampton, United Kingdom

**Corresponding author:** [claire.piccinin@eortc.org](mailto:claire.piccinin@eortc.org)

**Word count (excluding abstract, tables, and references):** 3941

**PURPOSE**

The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life-Core Questionnaire (QLQ-C30) [1] is currently one of the most widely used self-report measure of health-related quality of life (HRQOL) for cancer patients [2]. Patient reports using the EORTC QLQ-C30 are carried out using paper-and-pencil administration or through electronic methods. However, to increase the accessibility of the questionnaire across different research settings and populations (e.g., in patients that could be at risk of exclusion because of illiteracy), and to minimize the need for otherwise unnecessary clinic trips, the EORTC Quality of Life Group (QLG) set out to develop and validate a voice script for phone administration of the questionnaire. The voice script would incorporate all relevant elements of the QLQ-C30, along with additional instructions and text, where necessary, to facilitate its interviewer-led administration,

The EORTC QLQ-C30 was first released in 1987 by Aaronson and colleagues [3], and underwent further revisions leading to the development of its third version [1], which is still in use today. Comprised of 30 different items (questions), it is made up of eight multi-item functional (physical, role, emotional, cognitive, and social) and symptom (fatigue, pain, and nausea) scales, one global health status and quality of life (QOL) scale, and six single items (dyspnoea, insomnia, appetite loss, constipation, diarrhoea, and financial difficulties). Covering the majority of the core symptoms recommended for patient-reported outcome (PRO) measurement in cancer clinical trials [4], it is available for use in over 110 language versions, having undergone extensive testing to demonstrate its psychometric [5] and cultural [6] validity.

The majority of QLQ-C30 items (n=28) are measured by a four-option Likert response scale that ranges from 1, indicating “not at all”, to 4, indicating “very much”, capturing the presence and/or severity of a symptom or issue and its impact on QOL. The final two items which make up the global health status and QOL scale are rated on a scale from 1 to 7, with 1 indicating “very poor” and 7 “excellent”. The time scale for all items is “during the past week” with the exception of the first 5 items (the physical functioning scale), for which no specific timeframe is used, given the intent to capture a more global impact on physical functioning, not limited to a one-week recall period. All single items and multi-item scales in the questionnaire are scored and transformed onto a 0-100 scale, with higher scores for the functional and global health status/QOL scales indicating higher levels of functioning and QOL, and higher scores for symptom scales and single items indicating a higher degree of symptomatology and problems.

In addition to its frequent use in cancer research and clinical trials [2, 7], the QLQ-C30 is being increasingly used for monitoring purposes in clinical practice [8]. By providing a direct means of measuring core symptoms and issues from the patient’s perspective, the QLQ-C30 provides clinically meaningful information, distinct from that offered by clinical markers and clinicians’ ratings [2, 7, 9]. In 2018, the EORTC QLG published guidelines to help facilitate the use and migration of EORTC questionnaires into electronic PRO (ePRO) formats (e.g., computer, tablet) [10]. A computerised adaptive testing (CAT) version of the QLQ-C30, the EORTC CAT Core [11], is also available, and consists of dynamic item banks which correspond with the QLQ-C30’s 14 functional and symptom domains.

The purpose of this study was to pilot test the provisional QLQ-C30 phone script through cognitive debriefing interviews to ensure its acceptability and relevance, amending it if needed, and to subsequently validate the QLQ-C30 phone-administered version by carrying out equivalence testing between the paper and phone administration modes in a population of patients actively undergoing cancer treatment. An intraclass correlation coefficient (ICC) of >0.70, the recommended threshold to demonstrate equivalence between various modes of administration, was employed in this study for the purpose of equivalence testing [12]. Previous research supports the use of ICC >0.70, as demonstrated in studies by Lundy and colleagues [13, 14], in which an interactive voice response (IVR) version of the QLQ-C30 was developed. Similarly, in an equivalence study aimed at comparing tablet computer, IVR, and paper-based administration of the PRO-CTCAE [15], the degree of mode equivalence was assessed using ICC >0.70

Although previous work conducted by Lundy and colleagues demonstrated the equivalence of an IVR version of the QLQ-C30 to its paper administration [13, 14], this is the first project aimed at validating a voice script for phone administration of the QLQ-C30 by an interviewer. A considerable body of research comparing paper to screen-based (e.g., tablet, computer) administration of PROs has demonstrated high levels of reliability between both modes [16–18] but less work has compared paper administration to auditory modes (e.g., IVR, phone interview). Still, the existing research suggests that equivalence can be established between paper and phone PRO administration [13, 15].

**METHODS**

Patient recruitment and data collection, management, and analysis were subcontracted to contract research organization (CRO) Mapi/ICON plc who provided a final report to the EORTC detailing the methodology and findings. Study approval was obtained in the United Kingdom (UK) by the Quorum Review independent review board.

**Sample**

Recruitment was carried out through a UK-based recruitment agency (Global Perspectives Limited) and patients were eligible to participate if they were 18 years or older, currently receiving cancer treatment as confirmed by a clinician, able to read and understand English, voluntarily agreed to participate in the study, and provided written informed consent.

***Pilot testing***

Five patients were interviewed to test the acceptability, understanding, and relevance of the instructions for the QLQ-C30 voice script. Given that the QLQ-C30 has already been extensively tested and validated, and pilot-testing of the script had the aim of ensuring that the inclusion of additional text (e.g., instructions) was understandable, 5 patients was determined to be an acceptable number.

***Equivalence testing***

In addition to the previously described eligibility criteria, patients in the equivalency testing were required to have no changes in treatment planned between the paper and phone version completion. To support equivalence between paper-and-pen and phone administration modes using an ICC >0.70 and a minimally acceptable level of 0.50, a sample size of 63 patients was required [12]. Two waves of recruitment were conducted. In the first wave, 50 patients were recruited, the appropriate number for an equivalence threshold of ICC >0.90. Since protocol deviations were observed in which only 26 patients completed the paper and phone versions of the QLQ-C30 within the pre-specified 2-day timeframe, a second wave of recruitment was therefore conducted to address these limitations. The deviation relative to the 2-day time frame was due to issues with mailing of the questionnaires, and missing time stamps. The initial decision to use an equivalence threshold of ICC >0.90 garnered significant discussion among the research team members and CRO. Following the protocol deviation, it was deemed an appropriate moment to adapt the criteria to the widely used ICC >0.70. As such, thirty-seven additional patients were recruited based on the same eligibility criteria, bringing the total sample size to 63.

**Study Design**

***Pilot testing***

Patients’ interviews were conducted by trained qualitative researchers and audio-recorded for the purpose of analysis. Interviews lasted approximately 60 minutes and were based on a study-specific interview guide, which contained a summary of the methods to conduct the interview, along with semi-structured questions. The guide also contained questions regarding demographic and clinical variables to capture during the interview. Patients’ responses were recorded anonymously on a grid. The QLQ-C30 phone script was subsequently revised accordingly, and the updated version was used for equivalence testing. The interview recordings were destroyed after completion of the analysis, with an anonymised copy of the recordings retained for the study files.

***Equivalence testing***

A randomised, cross-over design was used to compare the self-administered paper version and the phone-administered version of the QLQ-C30 in patients currently receiving treatment for cancer, following recommendations as set out in the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) PRO Mixed Methods Task Force [19]. Patients were randomised (1:1) to complete either the paper or the phone-administered version first. Randomisation was conducted using a random number generator. After providing informed consent, each patient completed a brief sociodemographic and clinical form. Depending on randomisation, patients were then asked to either complete the paper version of the QLQ-C30 and return it to the recruitment agency in a prepaid envelope or respond by phone to the questionnaire following the phone script as presented by the interviewer, a trained qualitative researcher. The interviewers recorded patients’ responses on a paper version of the QLQ-C30. The paper version of the QLQ-C30 was estimated to take approximately 30 minutes to complete and administration time for the phone version was recorded for each patient. Any comments or observations made by the patient during the phone administration were recorded on a feedback form.

Two days after the first completion of the QLQ-C30, patients were asked to complete it again using the other mode of administration. The date of completion of the paper version was noted for each patient, to assess compliance with the pre-specified two-day time frame. For patients who completed the phone interview first, the recruitment agency waited for confirmation of interview completion from the study team before sending the paper version by post.

**Data Analysis**

Patients were described in terms of clinical and socio-demographic variables, as reported during the phone interview (pilot testing) or on the socio-demographic/clinical form (equivalence testing). Age, gender, educational status, and disease history were reported. All data processing and analyses were performed with SAS® software for Windows, Version 9.2 or later (SAS Institute, Inc., Cary, NC, USA).

***Pilot testing***

Feedback from patients was compiled in an analysis grid, and reported per patient based on a qualitative assessment of the questionnaire, its instructions and individual items, with any additional comments also recorded.

***Equivalence testing***

All patients who met the inclusion criteria and completed enough items in the QLQ-C30 questionnaire during each administration for each domain to be scored were included in the equivalence testing analysis. Responses to items from the QLQ-C30 were described based on completion and distribution of responses per administration mode. Missing data were described in terms of number and percent of missing responses per item along with number and percent of missing items per patient, including the number of patients with at least one missing item. Continuous variables were described based on their frequency, mean, standard deviation, median, first and third quartiles, and minimum and maximum values. Categorical variables were described based on the frequency and percentage of each response choice, with missing data included in the calculation of percentage.

Equivalence testing was performed at both the item and domain score levels, with the primary objective to evaluate equivalence at the score level between both modes of administration using ICC [20]. A two-way mixed effects, consistency, single rater/measurement approach, described by Fleiss and colleagues [20], was used to calculate ICCs and their confidence intervals. The widely used benchmark of ICC of >0.70 was used [21], with ICC values between 0.75 and 0.90 indicating good agreement and values greater than 0.90 indicating excellent agreement [22]. Weighted kappa coefficients [23] were used to assess the extent to which both administration modes produced the same responses by patients to the QLQ-C30 items (results are reported in Appendix A). Following Fleiss’ guidelines [24], a kappa value greater than 0.75 was characterized as excellent, 0.40-0.75 as fair to good, and less than 0.40 as poor. Mean differences in item-level scores were also calculated and are displayed in Appendix B.

To ensure robustness of results between the two waves of recruitment, a sensitivity analysis was conducted to compare the ICC values between patients included prior to the study amendment (first wave of recruitment: n=26) and those included after (second wave of recruitment: n=37) using scores from the paper and phone administration modes of the QLQ-C30. Additional sensitivity analyses were conducted on the full group of patients included in the equivalence testing (n=63) to compare ICC scores by age (<60 vs. >60) and gender.

**RESULTS**

**Sample**

***Pilot Testing***

Five patients (three males and two females) with a mean age of 51 years completed the pilot testing interviews. Patients had either liver, testicular, bowel cancer, or lymphoma, and one patient had breast, lung, and bowel cancer, as well as secondary liver cancer. More details regarding demographic and clinical characteristics are provided in Table 1.

Table 1. Demographic and clinical characteristics of the pilot testing population (n=5)

|  |  |
| --- | --- |
| Variable | Pilot testing sample  (n=5) |
| Age, years  Median  Min-Max | 51  (29-64) |
| Gender, n  Male/Female | 3 / 2 |
| Living status, n  Living as a couple  Living alone | 3  2 |
| Occupation, n  Full- or part-time employment  Retired | 4  1 |
| Education, n  Left high school with no qualifications  Completed high school with qualifications  Other | 1  3  2 |
| Type of cancer, n  Breast\*  Lung\*  Bowel\*  Other\* | 1  1  2  4 |
| Therapy  Surgery\*\*  Chemotherapy\*\*  Other | 1  3  2 |

\*One patient had breast, lung and bowel cancer, and secondary liver cancer

\*\*One patient had both surgery and chemotherapy

***Equivalence Testing***

Sixty-three patients (26 from the first wave and 37 from the second wave) made up the total sample included in the equivalence testing. Patients had a mean age of 55 years and 65% were female. Almost half of the sample (48%) was employed full- or part-time and 76% of patients were living as a couple. Education levels varied with 41% of patients having obtained a bachelor’s or postgraduate degree. Breast cancer was the most common disease type, reported in 29% of patients, followed by prostate (11%), lung (10%), and bowel (6%) cancers. A large proportion of patients (41%) reported “other” disease types. The majority of patients were undergoing chemotherapy (25%) or hormone therapy (16%) and other types of treatment included surgery (11%), radiotherapy (10%), biological therapy (13%), mixed therapy (8%) and “other” types of treatment (18%). Detailed demographic and clinical characteristics are provided in Table 2, presented to indicate patients who completed the paper (n=31) or phone (n=32) versions of the QLQ-C30 first.

Table 2. Demographic and clinical characteristics of the equivalence testing population (n=63)

| Variable | Randomisation group | | Total  (n=63) |
| --- | --- | --- | --- |
| Paper version first (n=31) | Phone version first (n=32) |  |
| Age, years  n (Missing)  Mean (SD)  Median  Min-Max | 31 (0)  56.84 (13.83)  59.00  24.00 - 78.00 | 32 (0)  54.16 (11.36)  54.50  31.00 - 79.00 | 63 (0)  55.48 (12.61)  58.00  24.00 - 79.00 |
| Gender, n (%)  Male  Female | 12 (38.7%)  19 (61.3%) | 10 (31.3%)  22 (68.8%) | 22 (34.9%)  41 (65.1%) |
| Living status, n (%)  Living alone  Living as a couple  Other | 6 (19.4%)  23 (74.2%)  2 (6.5%) | 3 (9.4%)  25 (78.1%)  4 (12.5%) | 9 (14.3%)  48 (76.2%)  6 (9.5%) |
| Occupation, n (%)  Full/part time employment Homemaker/housewife  Student  Unemployed  Retired  Other  Full/part time employment and Homemaker | 11 (35.5%)  1 (3.2%)  1 (3.2%)  2 (6.5%)  9 (29.0%)  4 (12.9%)  3 (9.7%) | 19 (59.4%)  2 (6.3%)  0 (0.0%)  1 (3.1%)  6 (18.8%)  4 (12.5%)  0 (0.0%) | 30 (47.6%)  3 (4.8%)  1 (1.6%)  3 (4.8%)  15 (23.8%)  8 (12.7%)  3 (4.8%) |
| Education level, n (%)  Left high school with no qualifications  Completed high school with qualifications  Bachelor's degree  Post-graduate degree  Other | 2 (6.5%)  10 (32.3%)  6 (19.4%)  8 (25.8%)  5 (16.1%) | 4 (12.5%)  12 (37.5%)  4 (12.5%)  8 (25.0%)  4 (12.5%) | 6 (9.5%)  22 (34.9%)  10 (15.9%)  16 (25.4%)  9 (14.3%) |
| Type of cancer, n (%)  Breast  Prostate  Lung  Bowel  Other  Kidney (metastatic)  Lung (Non-small cell) | 10 (32.3%)  5 (16.1%)  5 (16.1%)  1 (3.2%)  9 (29.0%)  0 (0.0%)  1 (3.2%) | 8 (25.0%)  2 (6.3%)  1 (3.1%)  3 (9.4%)  17 (53.1%)  1 (3.1%)  0 (0.0%) | 18 (28.6%)  7 (11.1%)  6 (9.5%)  4 (6.3%)  26 (41.3%)  1 (1.6%)  1 (1.6%) |
| Treatment\*, n (%)  Surgery  Chemotherapy  Radiotherapy  Hormone therapy  Biological therapy  Other  Mixed | 4 (12.9%)  9 (29.0%)  2 (6.5%)  6 (19.4%)  2 (6.5%)  6 (19.4%)  2 (6.5%) | 3 (9.4%)  7 (21.9%)  4 (12.5%)  4 (12.5%)  6 (18.8%)  5 (15.6%)  3 (9.4%) | 7 (11.1%)  16 (25.4%)  6 (9.5%)  10 (15.9%)  8 (12.7%)  11 (17.5%)  5 (7.9%) |

\*Patients could select more than one treatment

Participants in both waves of testing were largely similar, with more considerable differences observed based on treatment type. Whereas 16.2% of patients reported undergoing surgery in the second wave of testing, only 3.8% reported it in the first wave. Moreover, no patients reported use of biological therapy in the second wave of testing, while 30.8% of patients reported it in the first wave. The full comparison of sociodemographic differences is presented in Table 3.

Table 3. Sociodemographic differences between patients from the two study waves

| Variable | Patients | | Total  (n=63) |
| --- | --- | --- | --- |
| Second wave of patients (n=37) | First wave of patients (n=26) |  |
| Age, years  n (Missing)  Mean (SD)  Median  Min-Max | 37 (0)  52.51 (12.79)  56.00  24.00 - 76.00 | 26 (0)  59.69 (11.27)  59.50  36.00 - 79.00 | 63 (0)  55.48 (12.61)  58.00  24.00 - 79.00 |
| Gender, n (%)  Male  Female | 8 (21.6%)  29 (78.4%) | 14 (53.8%)  12 (46.2%) | 22 (34.9%)  41 (65.1%) |
| Living status, n (%)  Living alone  Living as a couple  Other | 7 (18.9%)  25 (67.6%)  5 (13.5%) | 2 (7.7%)  23 (88.5%)  1 (3.8%) | 9 (14.3%)  48 (76.2%)  6 (9.5%) |
| Occupation, n (%)  Full/part time employment  Homemaker/housewife  Student  Unemployed  Retired  Other  Full/part time employment and Homemaker | 17 (45.9%)  3 (8.1%)  1 (2.7%)  3 (8.1%)  7 (18.9%)  3 (8.1%)  3 (8.1%) | 13 (50.0%)  0 (0.0%)  0 (0.0%)  0 (0.0%)  8 (30.8%)  5 (19.2%)  0 (0.0%) | 30 (47.6%)  3 (4.8%)  1 (1.6%)  3 (4.8%)  15 (23.8%)  8 (12.7%)  3 (4.8%) |
| Education level, n (%)  Left high school with no qualifications  Completed high school with qualifications  Bachelor's degree  Post-graduate degree  Other | 3 (8.1%)  12 (32.4%)  7 (18.9%)  9 (24.3%)  6 (16.2%) | 3 (11.5%)  10 (38.5%)  3 (11.5%)  7 (26.9%)  3 (11.5%) | 6 (9.5%)  22 (34.9%)  10 (15.9%)  16 (25.4%)  9 (14.3%) |
| Type of cancer, n (%)  Breast  Prostate  Lung  Bowel  Other  Kidney (metastatic)  Lung (Non-small cell) | 11 (29.7%)  2 (5.4%)  3 (8.1%)  2 (5.4%)  17 (45.9%)  1 (2.7%)  1 (2.7%) | 7 (26.9%)  5 (19.2%)  3 (11.5%)  2 (7.7%)  9 (34.6%)  0 (0.0%)  0 (0.0%) | 18 (28.6%)  7 (11.1%)  6 (9.5%)  4 (6.3%)  26 (41.3%)  1 (1.6%)  1 (1.6%) |
| Treatment\*  Surgery  Chemotherapy  Radiotherapy  Hormone therapy  Biological therapy  Other  Mixed | 6 (16.2%)  11 (29.7%)  2 (5.4%)  4 (10.8%)  0 (0.0%)  11 (29.7%)  3 (8.1%) | 1 (3.8%)  5 (19.2%)  4 (15.4%)  6 (23.1%)  8 (30.8%)  0 (0.0%)  2 (7.7%) | 7 (11.1%)  16 (25.4%)  6 (9.5%)  10 (15.9%)  8 (12.7%)  11 (17.5%)  5 (7.9%) |

\*Patients could select more than one treatment

**Pilot Testing**

All patients considered the instructions in the phone script to be clear and straightforward. Three comments were raised concerning the time and response scales of the questionnaire. Two patients made comments regarding the time scales, specifying that “during the past week” was too short of a time frame. However, these comments deviated from the source questionnaire and were thus not integrated into the script. One patient suggested numbering the response options from 1 to 4, for clarity. After discussion with the study team, numbers 1 to 4 were added to the response options in the phone script, thereby creating the final version of the phone script in UK English.

**Equivalence Testing**

All patients from both testing waves (n=63) completed all items in both the paper and phone versions of the QLQ-C30 and there were no missing data. Table 4 displays the results of the equivalence testing for QLQ-C30 scores between the paper and phone administration modes. Total ICC values were above the 0.70 threshold and ranged from 0.72 for nausea and vomiting (95% CI 0.48-0.86) to 0.90 for global health status/QoL (95% CI 0.80-0.95), cognitive functioning (95% CI 0.80-0.95), and pain (95% CI 0.79-0.95). ICCs were also calculated to compare paper first versus phone first administration. For paper first administration, the ICCs ranged from 0.55 for nausea and vomiting (95% CI 0.24-0.76) to 0.90 for appetite loss (95% CI 0.80-0.95) and constipation (95% CI 0.81-0.95), with only two scores, nausea and vomiting and financial difficulties (ICC 0.60; 95% CI 0.31-.079), falling below the 0.70 threshold. For phone first administration, all ICC values were above the 0.70 threshold and ranged from 0.79 for dyspnoea (95% CI 0.60-0.89) to 0.94 for physical functioning (95% CI 0.87-0.97). Results for equivalence testing at the single item level are displayed in Appendix A.

Table 4. Equivalence testing for QLQ-C30 scores

| QLQ-C30 scores | All patients | |  |
| --- | --- | --- | --- |
|  | Paper first  (n=31) | Phone first  (n=32) | Total  (n=63) |
| Global health status score  ICC  95% CI | 0.89  0.79-0.95 | 0.91  0.81-0.95 | 0.90  0.80-0.95 |
| Physical functioning score  ICC  95% CI | 0.81  0.64-0.90 | 0.94  0.89-0.97 | 0.89  0.79-0.95 |
| Role functioning score  ICC  95% CI | 0.76  0.56-0.88 | 0.89  0.78-0.94 | 0.84  0.68-0.92 |
| Emotional functioning score  ICC  95% CI | 0.79  0.60-0.89 | 0.92  0.84-0.96 | 0.86  0.73-0.93 |
| Cognitive functioning score  ICC  95% CI | 0.85  0.71-0.92 | 0.94  0.87-0.97 | 0.90  0.80-0.95 |
| Social functioning score  ICC  95% CI | 0.76  0.55-0.87 | 0.85  0.70-0.92 | 0.79  0.59-0.89 |
| Fatigue score  ICC  95% CI | 0.76  0.56-0.88 | 0.82  0.66-0.91 | 0.79  0.59-0.89 |
| Nausea and Vomiting  ICC  95% CI | 0.55  0.24-0.76 | 0.91  0.83-0.96 | 0.72  0.48-0.86 |
| Pain  ICC  95% CI | 0.89  0.78-0.94 | 0.90  0.80-0.95 | 0.90  0.79-0.95 |
| Dyspnoea  ICC  95% CI | 0.78  0.59-0.89 | 0.79  0.60-0.89 | 0.76  0.55-0.88 |
| Insomnia  ICC  95% CI | 0.87  0.74-0.93 | 0.83  0.68-0.92 | 0.84  0.69-0.92 |
| Appetite loss  ICC  95% CI | 0.90  0.80-0.95 | 0.82  0.65-0.91 | 0.85  0.71-0.93 |
| Constipation  ICC  95% CI | 0.90  0.81-0.95 | 0.90  0.80-0.95 | 0.90  0.80-0.95 |
| Diarrhoea  ICC  95% CI | 0.77  0.58-0.88 | 0.85  0.71-0.92 | 0.80  0.62-0.90 |
| Financial difficulties  ICC  95% CI | 0.60  0.31-0.79 | 0.87  0.75-0.93 | 0.73  0.49-0.86 |

Mean differences in domain-level scores were assessed between administration modes and are shown in Table 5. Results for mean differences at the item level are displayed in Appendix B. At the domain level, differences between modes were minimal in absolute magnitude, ranging from 0.00 to 11.00 points.

Table 5. Mean differences in multi-item scores between paper and pen versions.

| QLQ-C30  scores | All patients | |
| --- | --- | --- |
| Paper first  (N=31) | Phone first  (N=32) |
| Global health status |  |  |
| Mean (SD) | -0.67 (9.01) | -0.52 (8.71) |
| Min - Max | -16.67 - 33.33 | -25.00 - 25.00 |
| Physical functioning |  |  |
| Mean (SD) | -2.37 (8.35) | 1.04 (6.36) |
| Min - Max | -20.00 - 20.00 | -6.67 - 26.67 |
| Role functioning |  |  |
| Mean (SD) | -0.54 (15.80) | 5.73 (14.42) |
| Min - Max | -33.33 - 33.33 | -16.67 - 33.33 |
| Emotional functioning |  |  |
| Mean (SD) | 0.54 (14.26) | 4.43 (10.37) |
| Min - Max | -25.00 - 33.33 | -25.00 - 25.00 |
| Cognitive functioning |  |  |
| Mean (SD) | 1.43 (10.92) | 4.17 (7.98) |
| Min - Max | -16.67 - 33.33 | -16.67 - 22.22 |
| Social functioning |  |  |
| Mean (SD) | -9.68 (19.61) | 1.04 (15.23) |
| Min - Max | -50.00 - 33.33 | -33.33 - 33.33 |
| Fatigue |  |  |
| Mean (SD) | 0.72 (16.71) | -6.25 (14.09) |
| Min - Max | -44.44 - 33.33 | -44.44 - 22.22 |
| Nausea and Vomiting |  |  |
| Mean (SD) | 4.30 (12.89) | -2.08 (5.60) |
| Min - Max | 0.00 - 66.67 | -16.67 - 0.00 |
| Pain |  |  |
| Mean (SD) | 1.08 (14.23) | 1.56 (13.63) |
| Min - Max | -16.67 - 50.00 | -16.67 - 33.33 |
| Dyspnoea |  |  |
| Mean (SD) | 2.15 (17.07) | -8.33 (14.66) |
| Min - Max | -33.33 - 66.67 | -33.33 - 0.00 |
| Insomnia |  |  |
| Mean (SD) | 1.08 (16.06) | -7.29 (20.27) |
| Min - Max | -33.33 - 33.33 | -66.67 - 33.33 |
| Appetite loss |  |  |
| Mean (SD) | 3.23 (10.02) | -0.00 (14.66) |
| Min - Max | 0.00 - 33.33 | -33.33 - 33.33 |
| Constipation |  |  |
| Mean (SD) | -1.08 (10.48) | -2.08 (14.51) |
| Min - Max | -33.33 - 33.33 | -33.33 - 33.33 |
| Diarrhoea |  |  |
| Mean (SD) | -1.08 (20.15) | -4.17 (14.04) |
| Min - Max | -100.00 - 33.33 | -33.33 - 33.33 |
| Financial difficulties |  |  |
| Mean (SD) | 10.75 (30.29) | 5.21 (12.30) |
| Min - Max | -66.67 - 66.67 | 0.00 - 33.33 |

The mean time for completion of the phone version of the QLQ-C30 was 8.6 ± 1.9 minutes and 39 participants (62%) made comments or asked questions during the interview.

Sensitivity analyses comparing patients included before the study amendment (n=26) with those included after (n=37) revealed significant differences (i.e., 95% CI overlapping) only for the nausea and vomiting ICC, which was lower in the first wave of patients, and the constipation ICC, which was lower in the second wave. The full results are displayed in Table 6.

Table 6. Sensitivity analysis of ICC scores by testing waves

| QLQ-C30 scores | All patients | |  |
| --- | --- | --- | --- |
|  | 1st wave  (n=26) | 2nd wave  (n=37) | Total  (n=63) |
| Global health status score  ICC  95% CI | 0.92  0.83-0.96 | 0.89  0.79-0.95 | 0.90  0.80-0.95 |
| Physical functioning score  ICC  95% CI | 0.86  0.72-0.93 | 0.93  0.58-0.96 | 0.89  0.79-0.95 |
| Role functioning score  ICC  95% CI | 0.76  0.55-0.88 | 0.77  0.57-0.88 | 0.84  0.68-0.92 |
| Emotional functioning score  ICC  95% CI | 0.84  0.69-0.92 | 0.87  0.74-0.94 | 0.86  0.73-0.93 |
| Cognitive functioning score  ICC  95% CI | 0.90  0.81-0.95 | 0.90  0.79-0.95 | 0.90  0.80-0.95 |
| Social functioning score  ICC  95% CI | 0.82  0.67-0.91 | 0.75  0.53-0.87 | 0.79  0.59-0.89 |
| Fatigue score  ICC  95% CI | 0.77  0.57-0.88 | 0.82  0.64-0.91 | 0.79  0.59-0.89 |
| Nausea and Vomiting  ICC  95% CI | 0.47  0.14-0.71 | 0.96  0.90-0.98 | 0.72  0.48-0.86 |
| Pain  ICC  95% CI | 0.85  0.71-0.93 | 0.94  0.87-0.97 | 0.90  0.79-0.95 |
| Dyspnoea  ICC  95% CI | 0.73  0.50-0.86 | 0.82  0.65-0.91 | 0.76  0.55-0.88 |
| Insomnia  ICC  95% CI | 0.78  0.59-0.89 | 0.89  0.80-0.95 | 0.84  0.69-0.92 |
| Appetite loss  ICC  95% CI | 0.85  0.70-0.92 | 0.85  0.71-0.93 | 0.85  0.71-0.93 |
| Constipation  ICC  95% CI | 0.98  0.95-0.99 | 0.84  0.70-0.92 | 0.90  0.80-0.95 |
| Diarrhoea  ICC  95% CI | 0.72  0.49-0.86 | 0.90  0.79-0.95 | 0.80  0.62-0.90 |
| Financial difficulties  ICC  95% CI | 0.52  0.20-0.74 | 0.79  0.60-0.90 | 0.73  0.49-0.86 |

The results of additional sensitivity analyses to assess possible differences in scores based on age (<60 versus >60) and gender are displayed in Tables 7 and 8.

Table 7. Sensitivity analysis of ICC scores by age above and below 60

| QLQ-C30 scores | All patients | |
| --- | --- | --- |
|  | Age <60  (n=40) | Age ≥60  (n=23) |
| Global health status score  ICC  95% CI | 0.89  0.78-0.94 | 0.90  0.81-0.95 |
| Physical functioning score  ICC  95% CI | 0.95  0.89-0.97 | 0.77  0.57-0.88 |
| Role functioning score  ICC  95% CI | 0.84  0.69-0.92 | 0.81  0.64-0.90 |
| Emotional functioning score  ICC  95% CI | 0.88  0.76-0.94 | 0.81  0.65-0.90 |
| Cognitive functioning score  ICC  95% CI | 0.90  0.80-0.95 | 0.88  0.77-0.94 |
| Social functioning score  ICC  95% CI | 0.81  0.65-0.91 | 0.69  0.45-0.84 |
| Fatigue score  ICC  95% CI | 0.80  0.61-0.90 | 0.75  0.55-0.87 |
| Nausea and Vomiting  ICC  95% CI | 0.63  0.35-0.81 | 0.92  0.84-0.96 |
| Pain  ICC  95% CI | 0.89  0.79-0.95 | 0.89  0.80-0.95 |
| Dyspnoea  ICC  95% CI | 0.79  0.61-0.90 | 0.72  0.49-0.85 |
| Insomnia  ICC  95% CI | 0.88  0.76-0.94 | 0.72  0.49-0.85 |
| Appetite loss  ICC  95% CI | 0.92  0.85-0.96 | 0.71  0.48-0.85 |
| Constipation  ICC  95% CI | 0.87  0.75-0.94 | 0.94  0.88-0.97 |
| Diarrhoea  ICC  95% CI | 0.68  0.42-0.83 | 0.95  0.90-0.97 |
| Financial difficulties  ICC  95% CI | 0.75  0.54-0.87 | 0.54  0.23-0.75 |

Table 8. Sensitivity analysis of ICC scores by gender

| QLQ-C30 scores | All patients | |
| --- | --- | --- |
|  | Male  (n=22) | Female  (n=41) |
| Global health status score  ICC  95% CI | 0.92  0.84-0.96 | 0.87  0.75-0.94 |
| Physical functioning score  ICC  95% CI | 0.77  0.57-0.88 | 0.94  0.89-0.97 |
| Role functioning score  ICC  95% CI | 0.81  0.65-0.90 | 0.84  0.69-0.92 |
| Emotional functioning score  ICC  95% CI | 0.80  0.62-0.90 | 0.88  0.76-0.94 |
| Cognitive functioning score  ICC  95% CI | 0.87  0.75-0.93 | 0.90  0.79-0.95 |
| Social functioning score  ICC  95% CI | 0.83  0.68-0.91 | 0.73  0.51-0.86 |
| Fatigue score  ICC  95% CI | 0.66  0.40-0.82 | 0.85  0.71-0.93 |
| Nausea and Vomiting  ICC  95% CI | 0.52  0.20-0.73 | 0.88  0.76-0.94 |
| Pain  ICC  95% CI | 0.88  0.76-0.94 | 0.91  0.82-0.95 |
| Dyspnoea  ICC  95% CI | 0.69  0.45-0.84 | 0.83  0.67-0.91 |
| Insomnia  ICC  95% CI | 0.75  0.55-0.87 | 0.88  0.76-0.94 |
| Appetite loss  ICC  95% CI | 0.87  0.74-0.93 | 0.84  0.69-0.92 |
| Constipation  ICC  95% CI | 0.87  0.75-0.93 | 0.90  0.81-0.95 |
| Diarrhoea  ICC  95% CI | 0.68  0.44-0.83 | 0.90  0.80-0.95 |
| Financial difficulties  ICC  95% CI | 0.62  0.34-0.79 | 0.74  0.52-0.87 |

**DISCUSSION**

This study aimed to develop and validate a voice script for phone administration of the QLQ-C30 and evaluate its equivalence to paper administration in a sample of patients actively undergoing cancer treatment. During pilot testing, the voice script was deemed understandable and relevant with minimal comments received from patients.

Results from the final sample of patients included in the equivalence testing indicated good equivalence between paper and phone administration modes, with all total ICC scores above the 0.70 threshold, ranging from 0.72 to 0.90. In the evaluation of paper administration first, two ICC scores were found to be below the 0.70 threshold, for nausea and vomiting (ICC 0.55; 95% CI 0.24-0.76) and financial difficulties (ICC 0.60; 95% CI 0.31-0.79). When comparing differences in means at the domain score level, the differences were still well below 10 points for the comparison of both administration modes, suggesting minimal clinically meaningful differences despite the ICCs [25]. Failure to reach the 0.70 ICC threshold for nausea and vomiting may also reflect the possibility of more ambiguity surrounding the rating of nausea. While vomiting is a more concrete occurrence, and it is unlikely that a patient’s recollection would change over a 2-day timeframe, nausea may be subject to broader interpretation. Moreover, medications are generally readily available to patients, which help to resolve these symptoms on a day-to-day basis, thus indicating that those symptoms can change within a two-day period. For financial difficulties, the potential source of discrepancy is less clear. There may have been an issue that was not detectable based on the scope of this study, or it may simply be due to random error or noise.

A more general limitation of using ICC to assess equivalence is that the absolute size of a given ICC is dependent on the variation observed within the sample. As such, minimal variation in nausea and vomiting and financial difficulties scores may have contributed to the lower ICCs. Still, the ICCs for nausea and vomiting and financial difficulties were well above 0.50 for paper administration first, indicating that they remain within the minimally acceptable range, especially since the total ICC scores for both scales were over 0.70. It is worth noting that the nausea and vomiting domain score has performed poorly in a previous test-retest study carried out by Hjermstad and colleagues [26], so there may be other factors influencing that scale, which were not identified in this study. Such factors could also account for the lower ICC score found for nausea and vomiting, when the paper version was administered first.

Differences in mean scores at the domain score level were uniformly minimal, suggesting that, overall, results from both administration modes were equivalent. The relatively short completion time of 8.6 ± 1.9 minutes for the voice script suggests that it can be integrated into a study protocol with relative ease and minimal patient burden.

Following guidelines from ISPOR’s PRO Mixed Modes Good Research Practices Task Force, and drawing on methodology used in similar PRO equivalence studies [13, 15], this study had a number of strengths. The randomized cross-over design helped to minimize the potential for bias in either one of the administration modes, and the inclusion criteria ensured that the voice script was evaluated and tested by patients for whom it would be relevant and feasible (i.e., those actively undergoing treatment, with the appropriate language level). The final sample of patients was diverse, and sufficiently well-balanced in terms of demographic and clinical characteristics, helping to ensure representativeness across patients and disease types. Analyses were also strengthened by the fact that there were no missing data in either the paper or phone administered versions of the questionnaire, making the results more easily interpretable.

The decision to decrease the initial ICC threshold from >0.90 to >0.70 following the study amendment to include a second wave of testing, is well-supported by robust evidence in the literature, for which an ICC of >0.70 has also been used to evaluate equivalence in similar studies [13–15].

Moreover, the total ICCs for both waves were largely similar. Significant differences were only observed for nausea and vomiting, which was lower in the first wave, and constipation, which was lower in the second wave of testing. Although the same recruitment procedures and inclusion criteria were applied, and demographic and clinical characteristics were largely comparable between groups, differences in gender and age distribution were found between the two waves of testing. In light of these differences, sensitivity analyses were carried out across all participants by age and gender. While most scores were similar across groups, differences were found in ICC scores for the nausea and vomiting and diarrhoea domains scores by group, with younger patients scoring lower on these domains compared to older patients. Males also scored lower than females on both the nausea and vomiting and physical functioning scales.

Despite these findings, when examining all other ICCs and comparing them across subgroups, no consistent pattern was identified which would support a potential correlation between lower or higher ICCs and the age or gender of patients. Moreover, the limited sample size makes it difficult to draw robust conclusions at the subgroup level. Factors other than age and gender may be related to the experience of disease and treatment, and may help to account for the differences observed; however, such interpretations are beyond the scope of this study, which is limited to the available demographic and clinical data.

Overall, sensitivity analyses showed that differences observed between the two waves of testing are minimal, thereby further supporting the equivalence of paper-and-pen and phone administration modes.

**CONCLUSIONS**

Results from this study support the equivalence of paper and phone administration modes of the QLQ-C30. In addition to its initial source language (UK English) development, the QLQ-C30 voice script is now available in multiple other languages, with more translations anticipated in the future. By providing an alternative means of questionnaire completion, the QLQ-C30 voice script helps to ensure that the questionnaire remains accessible in multiple formats across a wide range of patients.

**DECLARATIONS**

**Acknowledgements:** The authors would like to acknowledge and thank the EORTC for their support of this work, Bristol-Meyers Squibb for their financial support, and Mapi/ICON plc for their scientific support, involving patient recruitment and data collection, management, and analysis.

**Funding:** This research was funded by Bristol-Meyers Squibb.

**Conflicts of interest/Competing interests:** James W. Shaw is an employee and shareholder of Bristol-Myers Squibb. All other authors have no financial interests/personal relationships, which may be considered as potential competing interests, to declare.

**Availability of data and material:** N/A

**Code availability:** N/A

**Authors’ contributions:**

* Claire Piccinin: Methodology; Writing - original draft
* Madeline Pe: Conceptualization; Methodology; Writing – review & editing
* Dagmara Kuliś: Conceptualization; Methodology; Project administration; Writing – review & editing
* James W. Shaw: Conceptualization; Methodology; Supervision; Writing – review & editing
* Sally J. Wheelwright: Conceptualization; Methodology; Supervision; Writing – review & editing
* Andrew Bottomley: Conceptualization; Methodology; Supervision; Writing – review & editing
* *\*Data curation and Formal analysis subcontracted to Mapi/ICON plc*

**Ethics approval:** Study approval was obtained in the United Kingdom (UK) by the Quorum Review independent review board.

**Consent to participate:** All patients provided written informed consent for their participation.

**Consent for publication:** N/A

**REFERENCES**

1. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, Haes JCJMD, Kaasa S, Klee M, Osoba D, Razavi D, Rofe PB, Schraub S, Sneeuw K, Sullivan M, Takeda F (1993) The European organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. https://doi.org/10.1093/jnci/85.5.365

2. Mierzynska J, Piccinin C, Pe M, Martinelli F, Gotay C, Coens C, Mauer M, Eggermont A, Groenvold M, Bjordal K, Reijneveld J, Velikova G, Bottomley A (2019) Prognostic value of patient-reported outcomes from international randomised clinical trials on cancer: a systematic review. Lancet Oncol 20:e685–e698 . https://doi.org/10.1016/S1470-2045(19)30656-4

3. Aaronson NK, Bullinger M, Ahmedzai S (1988) A modular approach to quality-of-life assessment in cancer clinical trials. Recent Results Cancer Res 111:231–249 . https://doi.org/10.1007/978-3-642-83419-6\_27

4. Reeve BB, Mitchell SA, Dueck AC, Basch E, Cella D, Reilly CM, Minasian LM, Denicoff AM, O’Mara AM, Fisch MJ, Chauhan C, Aaronson NK, Coens C, Bruner DW (2014) Recommended patient-reported core set of symptoms to measure in adult cancer treatment trials. J Natl Cancer Inst. https://doi.org/10.1093/jnci/dju129

5. Gundy CM, Fayers PM, Groenvold M, Petersen MA, Scott NW, Sprangers MAG, Velikova G, Aaronson NK (2012) Comparing higher order models for the EORTC QLQ-C30. Qual Life Res. https://doi.org/10.1007/s11136-011-0082-6

6. Scott NW, Fayers PM, Aaronson NK, Bottomley A, De Graeff A, Groenvold M, Koller M, Petersen MA, Sprangers MAG (2007) The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. Qual Life Res. https://doi.org/10.1007/s11136-006-9120-1

7. Gotay CC, Kawamoto CT, Bottomley A, Efficace F (2008) The prognostic significance of patient-reported outcomes in cancer clinical trials. J. Clin. Oncol.

8. Wintner LM, Sztankay M, Aaronson N, Bottomley A, Giesinger JM, Groenvold M, Petersen MA, van de Poll-Franse L, Velikova G, Verdonck-de Leeuw I, Holzner B (2016) The use of EORTC measures in daily clinical practice—A synopsis of a newly developed manual. Eur. J. Cancer

9. Quinten C, Coens C, Mauer M, Comte S, Sprangers MA, Cleeland C, Osoba D, Bjordal K, Bottomley A (2009) Baseline quality of life as a prognostic indicator of survival: a meta-analysis of individual patient data from EORTC clinical trials. Lancet Oncol. https://doi.org/10.1016/S1470-2045(09)70200-1

10. Kuliś D, Holzner B, Koller M, Ruyskart P, Itani A, Williams P, Bottomley A (2018) Guidance on the implementation and management of EORTC quality of life instruments in electronic applications. EORTC, Brussels, Belgium

11. Petersen MA, Aaronson NK, Arraras JI, Chie WC, Conroy T, Costantini A, Dirven L, Fayers P, Gamper EM, Giesinger JM, Habets EJJ, Hammerlid E, Helbostad J, Hjermstad MJ, Holzner B, Johnson C, Kemmler G, King MT, Kaasa S, Loge JH, Reijneveld JC, Singer S, Taphoorn MJB, Thamsborg LH, Tomaszewski KA, Velikova G, Verdonck-de Leeuw IM, Young T, Groenvold M (2018) The EORTC CAT Core—The computer adaptive version of the EORTC QLQ-C30 questionnaire. Eur J Cancer 100:8–16 . https://doi.org/10.1016/j.ejca.2018.04.016

12. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, Lenderking WR, Cella D, Basch E (2009) 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 Heal 12:419–429 . https://doi.org/10.1111/j.1524-4733.2008.00470.x

13. Lundy JJ, Coons SJ, Aaronson NK (2014) Testing the measurement equivalence of paper and interactive voice response system versions of the EORTC QLQ-C30. Qual Life Res. https://doi.org/10.1007/s11136-013-0454-1

14. Lundy JJ, Coons SJ, Aaronson NK (2015) Test–Retest Reliability of an Interactive Voice Response (IVR) Version of the EORTC QLQ-C30. Patient 8:165–170 . https://doi.org/10.1007/s40271-014-0071-2

15. Bennett A V., Dueck AC, Mitchell SA, Mendoza TR, Reeve BB, Atkinson TM, Castro KM, Denicoff A, Rogak LJ, Harness JK, Bearden JD, Bryant D, Siegel RD, Schrag D, Basch E, Heon N, Shaw M, Ryan S, Stark LP, Malveaux D, Pettus W, Gansauer L, Wind J, Thomassie A, Davila G, Alexander K (2016) Mode equivalence and acceptability of tablet computer-, interactive voice response system-, and paper-based administration of the U.S. National Cancer Institute’s Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO. Health Qual Life Outcomes 14:1–12 . https://doi.org/10.1186/s12955-016-0426-6

16. Gwaltney CJ, Shields AL, Shiffman S (2008) Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. Value Heal. https://doi.org/10.1111/j.1524-4733.2007.00231.x

17. Muehlhausen W, Doll H, Quadri N, Fordham B, O’Donohoe P, Dogar N, Wild DJ (2015) Equivalence of electronic and paper administration of patient-reported outcome measures: A systematic review and meta-analysis of studies conducted between 2007 and 2013. Health Qual Life Outcomes. https://doi.org/10.1186/s12955-015-0362-x

18. Ramachandran S, Lundy JJ, Coons SJ (2008) Testing the measurement equivalence of paper and touch-screen versions of the EQ-5D visual analog scale (EQ VAS). Qual Life Res. https://doi.org/10.1007/s11136-008-9384-8

19. Eremenco S, Coons SJ, Paty J, Coyne K, Bennett A V., McEntegart D (2014) PRO data collection in clinical trials using mixed modes: Report of the ISPOR PRO mixed modes good research practices task force. Value Heal 17:501–516 . https://doi.org/10.1016/j.jval.2014.06.005

20. Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420–428 . https://doi.org/10.1037/0033-2909.86.2.420

21. Nunnally J, Bernstein I (1994) Psychometric Methods, 3rd ed. McGraw-Hill, New York

22. Portney L, Watkins M (2000) Foundations of Clinical Research: Applications to Practice. Pearson Prentice Hall, Upper Saddle River, New Jersey

23. Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213–220. Psychol Bull

24. Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33:613–619 . https://doi.org/10.1177/001316447303300309

25. Osoba D, Rodrigues G, Myles J, Zee B, Pater J (1998) Interpreting the significance of changes in health-related quality-of- life scores. J Clin Oncol. https://doi.org/10.1200/JCO.1998.16.1.139

26. Hjermstad MJ, Fossa SD, Bjordal K, Kaasa S (1995) Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire. J Clin Oncol. https://doi.org/10.1200/JCO.1995.13.5.1249

**APPENDIX A**

A1. Equivalence testing for single items

| QLQ-C30 items | All patients | |  |
| --- | --- | --- | --- |
|  | Paper first  (n=31) | Phone first  (n=32) | Total  (n=63) |
| Item 1 (Strenuous activity)  Weighted Kappa  95% CI | 0.50  0.21-0.79 | 0.73  0.55-0.91 | 0.65  0.50-0.81 |
| Item 2 (Long walk)  Weighted Kappa  95% CI | 0.64  0.41-0.87 | 0.85  0.71-0.99 | 0.77  0.64-0.89 |
| Item 3 (Short walk)  Weighted Kappa  95% CI | 0.85  0.65-1.00 | 0.83  0.62-1.00 | 0.84  0.69-0.99 |
| Item 4 (Stay in bed or chair)  Weighted Kappa  95% CI | 0.51  0.26-0.77 | 0.73  0.55-0.91 | 0.65  0.50-0.81 |
| Item 5 (Assistance)  Weighted Kappa  95% CI | 0.83  0.53-1.00 | 0.72  0.44-1.00 | 0.77  0.56-0.98 |
| Item 6 (Work or daily activities)  Weighted Kappa  95% CI | 0.64  0.44-0.83 | 0.63  0.43-0.83 | 0.64  0.50-0.78 |
| Item 7 (Hobbies)  Weighted Kappa  95% CI | 0.65  0.47-0.82 | 0.68  0.52-0.84 | 0.67  0.55-0.79 |
| Item 8 (Short of breath)  Weighted Kappa  95% CI | 0.73  0.51-0.96 | 0.65  0.43-0.86 | 0.69  0.53-0.85 |
| Item 9 (Pain)  Weighted Kappa  95% CI | 0.78  0.61-0.96 | 0.82  0.68-0.95 | 0.80  0.70-0.91 |
| Item 10 (Rest)  Weighted Kappa  95% CI | 0.63  0.40-0.86 | 0.55  0.33-0.76 | 0.59  0.43-0.74 |
| Item 11 (Trouble sleeping)  Weighted Kappa  95% CI | 0.77  0.62-0.93 | 0.71  0.55-0.87 | 0.74  0.63-0.85 |
| Item 12 (Weak)  Weighted Kappa  95% CI | 0.60  0.38-0.81 | 0.41  0.19-0.63 | 0.52  0.36-0.68 |
| Item 13 (Lack appetite)  Weighted Kappa  95% CI | 0.85  0.68-1.00 | 0.74  0.58-0.89 | 0.79  0.67-0.91 |
| Item 14 (Nausea)  Weighted Kappa  95% CI | 0.68  0.43-0.92 | 0.87  0.73-1.00 | 0.78  0.64-0.93 |
| Item 15 (Vomiting)  Weighted Kappa  95% CI | 0.64  0.09-1.00 | 0.79  0.33-1.00 | 0.71  0.32-1.00 |
| Item 16 (Constipation)  Weighted Kappa  95% CI | 0.87  0.73-1.00 | 0.79  0.65-0.93 | 0.82  0.73-0.92 |
| Item 17 (Diarrhoea)  Weighted Kappa  95% CI | 0.79  0.53-1.00 | 0.75  0.60-0.90 | 0.77  0.63-0.92 |
| Item 18 (Tired)  Weighted Kappa  95% CI | 0.61  0.37-0.84 | 0.66  0.46-0.85 | 0.63  0.48-0.79 |
| Item 19 (Pain and activity)  Weighted Kappa  95% CI | 0.70  0.51-0.89 | 0.68  0.51-0.85 | 0.69  0.56-0.81 |
| Item 20 (Difficulty concentrating)  Weighted Kappa  95% CI | 0.75  0.58-0.92 | 0.65  0.45-0.86 | 0.70  0.57-0.84 |
| Item 21 (Tense)  Weighted Kappa  95% CI | 0.63  0.42-0.84 | 0.62  0.42-0.81 | 0.63  0.49-0.78 |
| Item 22 (Worry)  Weighted Kappa  95% CI | 0.41  0.17-0.65 | 0.58  0.38-0.79 | 0.52  0.37-0.68 |
| Item 23 (Irritable)  Weighted Kappa  95% CI | 0.63  0.43-0.82 | 0.67  0.51-0.83 | 0.65  0.52-0.77 |
| Item 24 (Depressed)  Weighted Kappa  95% CI | 0.65  0.48-0.82 | 0.90  0.79-1.00 | 0.79  0.69-0.89 |
| Item 25 (Difficulty remembering)  Weighted Kappa  95% CI | 0.73  0.53-0.93 | 0.56  0.39-0.74 | 0.64  0.51-0.78 |
| Item 26 (Family life)  Weighted Kappa  95% CI | 0.56  0.34-0.78 | 0.65  0.44-0.86 | 0.60  0.44-0.76 |
| Item 27 (Social)  Weighted Kappa  95% CI | 0.55  0.34-0.76 | 0.65  0.48-0.82 | 0.60  0.47-0.74 |
| Item 28 (Financial)  Weighted Kappa  95% CI | 0.49  0.25-0.72 | 0.78  0.60-0.96 | 0.63  0.48-0.78 |
| Item 29 (Overall health)  Weighted Kappa  95% CI | 0.73  0.58-0.89 | 0.70  0.54-0.86 | 0.72  0.61-0.83 |
| Item 30 (Overall QoL)  Weighted Kappa  95% CI | 0.69  0.56-0.83 | 0.77  0.61-0.94 | 0.74  0.63-0.85 |

**APPENDIX B**

B1. Mean differences in single item scores between paper- and phone-administered versions.

| QLQ-C30 items | All patients | |
| --- | --- | --- |
| Paper first  (N=31) | Phone first  (N=32) |
| Item 1 (Strenuous activity)  Mean (SD)  Min - Max | -0.03 (0.66)  -2.00 - 1.00 | -0.13 (0.66)  -2.00 - 2.00 |
| Item 2 (Long walk)  Mean (SD)  Min - Max | 0.10 (0.60)  -1.00 - 2.00 | 0.09 (0.47)  -1.00 - 2.00 |
| Item 3 (Short walk)  Mean (SD)  Min - Max | 0.06 (0.25)  0.00 - 1.00 | -0.03 (0.40)  -2.00 - 1.00 |
| Item 4 (Stay in bed or chair)  Mean (SD)  Min - Max | 0.19 (0.48)  -1.00 - 1.00 | -0.16 (0.45)  -1.00 - 1.00 |
| Item 5 (Assistance)  Mean (SD)  Min - Max | 0.03 (0.18)  0.00 - 1.00 | 0.06 (0.25)  0.00 - 1.00 |
| Item 6 (Work or daily activities)  Mean (SD)  Min - Max | 0.00 (0.52)  -1.00 - 1.00 | -0.19 (0.59)  -1.00 - 1.00 |
| Item 7 (Hobbies)  Mean (SD)  Min - Max | 0.03 (0.55)  -1.00 - 1.00 | -0.16 (0.57)  -1.00 - 1.00 |
| Item 8 (Short of breath)  Mean (SD)  Min - Max | 0.06 (0.51)  -1.00 - 2.00 | -0.25 (0.44)  -1.00 - 0.00 |
| Item 9 (Pain)  Mean (SD)  Min - Max | 0.10 (0.54)  -1.00 - 2.00 | 0.00 (0.44)  -1.00 - 1.00 |
| Item 10 (Rest)  Mean (SD)  Min - Max | -0.03 (0.60)  -2.00 - 1.00 | -0.31 (0.69)  -2.00 - 1.00 |
| Item 11 (Trouble sleeping) |  |  |
| Mean (SD) | 0.03 (0.48) | -0.22 (0.61) |
| Min - Max | -1.00 - 1.00 | -2.00 - 1.00 |
| Item 12 (Weak) |  |  |
| Mean (SD) | 0.06 (0.73) | -0.09 (0.73) |
| Min - Max | -2.00 - 2.00 | -1.00 - 2.00 |
| Item 13 (Lack appetite) |  |  |
| Mean (SD) | 0.10 (0.30) | 0.00 (0.44) |
| Min - Max | 0.00 - 1.00 | -1.00 - 1.00 |
| Item 14 (Nausea) |  |  |
| Mean (SD) | 0.19 (0.48) | -0.09 (0.30) |
| Min - Max | 0.00 - 2.00 | -1.00 - 0.00 |
| Item 15 (Vomiting) |  |  |
| Mean (SD) | 0.06 (0.36) | -0.03 (0.18) |
| Min - Max | 0.00 - 2.00 | -1.00 - 0.00 |
| Item 16 (Constipation) |  |  |
| Mean (SD) | -0.03 (0.31) | -0.06 (0.44) |
| Min - Max | -1.00 - 1.00 | -1.00 - 1.00 |
| Item 17 (Diarrhoea) |  |  |
| Mean (SD) | -0.03 (0.60) | -0.13 (0.42) |
| Min - Max | -3.00 - 1.00 | -1.00 - 1.00 |
| Item 18 (Tired) |  |  |
| Mean (SD) | 0.03 (0.55) | -0.16 (0.51) |
| Min - Max | -1.00 - 1.00 | -1.00 - 1.00 |
| Item 19 (Pain and activity) |  |  |
| Mean (SD) | -0.03 (0.55) | 0.09 (0.64) |
| Min - Max | -1.00 - 1.00 | -1.00 - 2.00 |
| Item 20 (Difficulty concentrating) |  |  |
| Mean (SD) | -0.16 (0.45) | -0.19 (0.59) |
| Min - Max | -1.00 - 1.00 | -2.00 - 1.00 |
| Item 21 (Tense) |  |  |
| Mean (SD) | -0.06 (0.51) | -0.13 (0.75) |
| Min - Max | -1.00 - 1.00 | -2.00 - 2.00 |
| Item 22 (Worry) |  |  |
| Mean (SD) | 0.03 (0.75) | -0.22 (0.75) |
| Min - Max | -1.00 - 2.00 | -2.00 - 1.00 |
| Item 23 (Irritable) |  |  |
| Mean (SD) | -0.13 (0.56) | -0.16 (0.51) |
| Min - Max | -1.00 - 1.00 | -1.00 - 1.00 |
| Item 24 (Depressed) |  |  |
| Mean (SD) | 0.10 (0.54) | -0.03 (0.31) |
| Min - Max | -1.00 - 1.00 | -1.00 - 1.00 |
| Item 25 (Difficulty remembering) |  |  |
| Mean (SD) | -0.03 (0.55) | -0.03 (0.65) |
| Min - Max | -1.00 - 2.00 | -1.00 - 1.00 |
| Item 26 (Family life) |  |  |
| Mean (SD) | 0.39 (0.76) | 0.00 (0.62) |
| Min - Max | -1.00 - 3.00 | -2.00 - 1.00 |
| Item 27 (Social) |  |  |
| Mean (SD) | 0.19 (0.70) | -0.06 (0.67) |
| Min - Max | -2.00 - 1.00 | -2.00 - 1.00 |
| Item 28 (Financial) |  |  |
| Mean (SD) | 0.32 (0.91) | 0.16 (0.37) |
| Min - Max | -2.00 - 2.00 | 0.00 - 1.00 |
| Item 29 (Overall health) |  |  |
| Mean (SD) | -0.02 (0.64) | -0.03 (0.78) |
| Min - Max | -1.00 - 2.00 | -2.00 - 2.00 |
| Item 30 (Overall QoL) |  |  |
| Mean (SD) | -0.06 (0.73) | -0.03 (0.82) |
| Min - Max | -1.00 - 2.00 | -3.00 - 2.00 |