Automated Collection of Quality-of-Life Data: A Comparison of Paper and Computer Touch-Screen Questionnaires

By G. Velikova, E.P. Wright, A.B. Smith, A. Cull, A. Gould, D. Forman, T. Perren, M. Stead, J. Brown, and P.J. Selby

Purpose: To evaluate alternative automated methods of collecting data on quality of life (QOL) in cancer patients. After initial evaluation of a range of technologies, we compared computer touch-screen questionnaires with paper questionnaires scanned by optical reading systems in terms of patients' acceptance, data quality, and reliability.

Patients and Methods: In a randomized cross-over trial, 149 cancer patients completed the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire–Core 30, version 2.0 (EORTC QLQ-C30), and the Hospital Anxiety and Depression Scale (HADS) on paper and on a touch screen. In a further test-retest study, 81 patients completed the electronic version of the questionnaires twice, with a time interval of 3 hours between questionnaires.

Results: Fifty-two percent of the patients preferred the touch screen to paper; 24% had no preference. The quality of the data collected with the touch-screen system was good, with no missed responses. At the group level, the differences between scores obtained with the two modes of administration of the instruments were small, suggesting equivalence for most of the QOL scales, with the possible exception of the emotional, fatigue, and nausea/vomiting scales and the appetite item, where patients tended to give more positive responses on the touch screen. At the individual patient level, the agreement was good, with a kappa coefficient from 0.57 to 0.77 and percent global agreement from 61% to 97%. The electronic questionnaire had good test-retest reliability, with correlation coefficients between the two administrations from 0.78 to 0.95, kappa coefficients of agreement from 0.55 to 0.90, and percent global agreement from 56% to 100%.

Conclusion: Computer touch-screen QOL questionnaires were well accepted by cancer patients, with good data quality and reliability.

J Clin Oncol 17:998-1007. © 1999 by American Society of Clinical Oncology.

THE OUTCOMES OF health care interventions in cancer patients have traditionally been measured by objective tumor response and survival. One of the major changes in cancer medicine during the past decade has been the increase in attention to formal assessment of quality of life (QOL) and psychosocial issues in patient care and the recognition that patients' subjective well-being is an important outcome of anticancer treatment. Researchers have developed many cancer-specific QOL-assessment instruments with carefully tested and documented psychometric properties. These instruments are now frequently used in clinical research as outcome measures in clinical trials, as predictors of survival and response to treatment,[1-3] and as screening tools for psychosocial morbidity.[4] More recently, attention has focused on incorporating health-related QOL assessment across the course of care.[5] There are recent data on the clinical meaning and the significance to patients of changes in QOL scores and reference data in a general Norwegian population, and these data will facilitate the interpretation of QOL scores in individual patients.[6,7] If QOL measures are to become an intrinsic part of the monitoring and evaluation of cancer patients' care, automated methods for collection of QOL data are needed that are easy, quick, inexpensive, and reliable and that can be integrated into oncology practice routine.

Traditionally, QOL data have been collected through patient self-report questionnaires printed on paper forms. The responses are usually entered into a database manually. More recently, optical mark-recognition systems have been used to transfer data from paper forms to a database. Optical scanning allows processing of large amounts of data and is useful for mailed surveys. However, special forms are required, multiple answers may be recognized, data may be missing, and verification and examination of the database for errors are necessary. Electronic methods of data collection (eg, QOL recorders[8] and interactive computer pro-

From the Imperial Cancer Research Fund Cancer Medicine Research Unit, St James's University Hospital; Research School of Medicine, Centre for Cancer Research, University of Leeds; Northern and Yorkshire Cancer Registry and Information Service; and Northern and Yorkshire Clinical Trials and Research Unit, Leeds, UK; Department of Psychology and Imperial Cancer Research Fund Medical Oncology Unit, Western General Hospital, Edinburgh, UK; and Information and Statistics Division, National Health Service in Scotland, Edinburgh, UK.

Submitted March 27, 1998; accepted October 22, 1998.

Supported by a grant from the NHS National Research and Development Programme, the Imperial Cancer Research Fund (G.V., A.C., and P.J.S.), and the National Lottery Charities Board (G.V.).

The views expressed in this article are those of the authors and not necessarily those of the National Health Service Executive.

The software was designed by A.B.S., who can be contacted for further information at the correspondence address.

Address reprint requests to G. Velikova, MD, Imperial Cancer Research Fund Cancer Medicine Research Unit, St James's University Hospital, Beckett Street, Leeds LS97TF, UK.

0732-183X/99/1703-998

grams) look more promising for implementation in clinical practice. With the use of electronic questionnaires, some of the problems with the process of data entry may be overcome.[9] Results can be compiled automatically in a database and can be immediately available for use in clinical practice, health services outcome studies, and clinical trials.

Although a concern exists that patients, particularly elderly patients, may be resistant to using new technology, several studies outside oncology have suggested that electronic data–capture methods were preferred over traditional paper-and-pen methods by elderly volunteers and patients.[10-12] Surveys in patients with diabetes or gastrointestinal diseases and in psychiatry found that interactive computer programs were well accepted by the patients and provided reliable information.[13-15] The development of automated computer systems is therefore a promising approach to collection of QOL data in busy clinical oncology practices. Before a new method of administration of an existing questionnaire can be recommended for wider use, the method should be evaluated for its effect on the reliability and validity of the instrument and for possible effect on patient responses.

The aims of this project were (1) to develop automated methods for measuring QOL and screening for psychologic morbidity in oncology clinics and wards based on two widely used QOL instruments: the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire–Core 30, version 2.0 (EORTC QLQ-C30),[16] and the Hospital Anxiety and Depression Scale (HADS)[17]; and (2) to evaluate the feasibility and practicality of introducing these methods into everyday clinical practice.

We reviewed commercially available automated methods for collection of questionnaire data, and we conducted a pilot study involving 213 patients to assess the performance of three data-capture methods: (1) paper questionnaires, with data transferred into a computer database using an optical mark-recognition system (Cardiff Software Co, San Marcos, CA); (2) two handheld computers—Psion Workabout (Psion plc, London, UK) and Apple Newton Message Pad (Apple Computer Inc, Cupertino, CA); and (3) a desktop computer with a touch-screen monitor (Geosoft UK Ltd, Leeds, UK). The handheld computers proved difficult for patients to use because of the small screens as well as the software design, which prevented patients from changing their responses. There were technical problems resulting in unreliable downloaded data and administrative problems in ensuring the security of the portable computers and the data they contained. As a result of this preliminary work, the touch-screen monitor was selected for more detailed evaluation and comparison with paper questionnaires with optical scanning for automated transfer of data. A computer program was designed for recording QOL data from the touch-screen monitor, and an assessment of the reliability of the electronic method and its feasibility for use with oncology patients was commenced.

In this article, we present the results from two studies. The first study compared the two methods of data capture—paper forms (with optical scanning for data entry) and computer touch-screen questionnaires—in terms of patients' acceptance and preference, data quality, feasibility, and reliability. The feasibility was assessed in terms of the logistics of administration of the questionnaires and the time taken for their completion. To assess the reliability of the computer questionnaires, we investigated the extent of agreement between the QOL scores obtained with the paper questionnaires and those obtained with the touch-screen questionnaires. The second study evaluated the test-retest reliability of the computer touch-screen questionnaires.

QOL INSTRUMENTS AND EQUIPMENT

EORTC QLQ-C30 is a 30-item cancer-specific questionnaire including five functional scales (physical, emotional, cognitive, social, and role), three symptom scales (fatigue, pain, and nausea/vomiting), a global health/QOL scale, and six single items assessing symptoms and perceived financial impact of disease and treatment. The validity of the instrument has been demonstrated in international clinical trials in cancer patients with heterogeneous diagnoses.[16,18] The raw scores were linearly transformed to give standard scores in the range of 0 to 100 for each of the scales and single items. Higher scores on the functioning and global health scales indicate better functioning, whereas higher scores on the symptom scales represent more severe symptoms.

HADS is a 14-item instrument that was designed to measure anxiety and depression in physically ill patients, with somatic symptoms not included in the assessment. There are two separate seven-item subscales for anxiety and depression. Scores range from 0 to 21 on each scale, with higher scores indicating more distress. The creators of HADS recommended using a threshold of 11 to identify probable cases of anxiety disorder and depressive illness and a threshold of 8 to 10 to include possible cases.[17]

In the case of the paper questionnaires, the original formats of EORTC QLQ-C30 and HADS were adapted for optical mark recognition; ie, patients were asked to fill in a circle rather than circle a number when making a response.

A computer program was developed for administering the computerized versions of the two questionnaires. Questions were presented individually on the screen, and respondents entered their answers by touching the corresponding buttons on the screen. Questions were presented with the same instructions and in the same response format as the originals. It was not possible to move on to the next question without completing the previous one, but it was possible to go back and change previous responses. Although results could be immediately printed out, we did not take advantage of this option in these studies.

A flatbed monochrome scanner with a bulk sheetfeeder and an IBM (Armonk, NY)-compatible PC were used to scan in responses to the paper questionnaires, using an optical mark-recognition system (Cardiff Software). A 15-inch touch-screen monitor with a privacy screen (ie, the computer screen could be seen only by facing the monitor) and an IBM-compatible PC with touch-screen software were used to present the questionnaires electronically to the patients.

STUDY COMPARING TWO METHODS OF ADMINISTRATION OF QOL QUESTIONNAIRES

Patients and Design

Consecutive cancer patients from the inpatient wards of St. James's University Hospital, Cookridge Hospital, and St. Gemma's Hospice, Leeds were considered for participation in the study between January 1997 and March 1997. To be enrolled, patients were required to have cancer (any type except primary or secondary brain tumors), be able to read English, and not be radioactive, isolated in a single room, or visually impaired. Patients who were too ill physically or who were cognitively impaired were excluded. Where possible, reasons for refusing to take part in the study were recorded with the patient's diagnosis, age, and sex. Written informed consent was obtained from all participating patients. Demographic and clinical details of participating patients were collected from case notes.

The study used an open randomized cross-over design, with the two modes of administration (paper followed by touch screen and touch screen followed by paper) allocated in a predetermined random order. The questionnaires were presented first or second an equal number of times within each combination. Each patient was asked to complete one version of the questionnaire in the morning and the second version in the afternoon, with a target interval of 3 hours between the two tests. The time taken to complete each questionnaire was recorded for each mode of administration. For the computer method, the patient entered identification data (ie, name, date of birth, and postal code) by pressing letters and numbers appearing on the touch-screen monitor. On completion of the study, the patient's preferred mode of administration was recorded (ie, touch screen, paper, or no preference).

Data collected using paper forms were transferred into a computer database using optical scanning. Data obtained using the touch screen were transferred electronically. The project was approved by the Ethical Committees at St James's University Hospital, Leeds and United Leeds Teaching Hospitals.

Statistical Methods

Statistical analysis was performed using the Statistical Package for Social Sciences (SPSS; SPSS, Inc, Chicago, IL). The sample size for the study was based on two a priori agreed-on criteria for acceptable levels of patients' preference and small systematic differences between the QOL scores obtained with the two modes of administration. We considered the touch-screen questionnaires to be acceptable if more than half of the patients either preferred them to paper questionnaires or had no preference. Therefore, we wanted to be able to detect a proportion of at least 61% with a 95% confidence interval (CI) of 10%, for which the required sample size was 100 patients. To detect a small systematic difference of 5 percentage points in the EORTC QLQ-C30 scores between the two modes of administration in a cross-over study with a power of 80% and level of significance of 5%, a sample of 150 patients was necessary. For the latter calculations, we used the average standard deviations and definitions for a small difference published by King[19] and based on results from 14 studies.

Patients' acceptance and preference. We assessed overall patient acceptability and preference by calculating the proportion of patients stating either a preference for the paper version or no preference, along with a 95% CI. In addition, we calculated the proportions in each of the three preference categories. We also compared groups of patients stating different preferences according to the order of presentation, time for completion, patient age, and patient sex. We used the χ^2 test for categoric variables (sex and order of presentation) and one-way analysis of variance for continuous variables (age and time for completion).

Data quality. The quality of the data collected using the two methods of administration was assessed by counting the number of errors in the final computer database and by discussing problems experienced by the researchers.

Feasibility. To examine the effect of the mode of administration on the time for completion of the two versions, we used the analysis recommended by Pocock[20] for two-period cross-over design trials. We checked for normality of the distribution of the times and transformed the data appropriately. We first tested for order effect and for mode-order interaction by analysis of covariance, adjusting for age and sex. Then the effect of mode of administration on time for completion was examined by the two-sample Student's *t* test.

Reliability. We calculated mean scores for all QOL dimensions on the computer and paper versions, as well as the mean of the individual paired differences in scores between the two versions (touch-screen questionnaire scores minus paper questionnaire scores). In addition, the scores were checked for normality. To examine the effect of the mode of administration on the QOL scores for the two versions, we used the analysis recommended by Pocock[20] for two-period cross-over design trials. Statistically significant differences between the mean differences in QOL scores were interpreted as evidence of systematic bias at the group level (ie, patients reporting systematically more or fewer problems on one of the versions). The range of the variation of scores between the two modes of administration was examined graphically using bar charts of paired differences.

To assess the agreement between the computer and paper questionnaire scores at the level of the individual patient, we calculated proportions of exact and global agreement and weighted kappa coefficients. ''Exact agreement'' referred to the percentage of patients who gave the same responses to individual questions on both occasions. ''Global agreement'' was defined as the proportion agreement within one response category in either direction. The percent agreement depends on the number of response categories; ie, it is expected to be higher for the single items with only four response categories and lower for the scales with higher numbers of possible response categories. Kappa is a coefficient of agreement that is corrected for chance agreement.[21,22] For ordinal data (as in QOL scores), weighted kappa is calculated by giving different weights to disagreements according to the magnitude of the discrepancy. Values of kappa range from 0 to 1, with 0 indicating no agreement beyond chance and 1 indicating perfect agreement. For interpretation of values between 0 and 1, we followed the guidelines of Landis and Koch[23]: kappa $\leq$ 0.2, poor agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, good agreement; and 0.81 to 1.00, very good agreement. The kappa coefficient was calculated for the QOL scores on the QOL scales and single items.

Results

Patient characteristics. There were 232 inpatients during the study period. Twenty (8.6%) were ineligible and 160 (75%) of the remainder agreed to take part. One hundred forty-nine patients completed both parts of the study. Eleven patients completed only the first version of the questionnaires (nine completed the paper version only and two completed the touch-screen version only); four were discharged, one was interrupted by a ward round, four felt too unwell to repeat the task, one needed a lot of help from researchers, and in one case there were problems with the computer program. These 11 cases were excluded from the

Table 1. Characteristics of Patients in the Cross-Over Study

	Participants	Nonparticipants
No. of patients (%)	160 (75%)	52 (25%)
Age, years		
Mean	57	65
SD	15.3	14.2*
Sex		
Male	71	30
Female	89	22
Diagnosis		
Gastrointestinal cancer	54	14
Lung cancer	14	4
Breast cancer	21	5
Female genital cancer	29	10
Male genital cancer	16	7
Head and neck cancer	7	4
Lymphoma	7	2
Other	12	6

*$P < .001$ using the independent sample t test.

analysis. Patient characteristics for participants and those who refused to participate are listed in Table 1. The patients who refused to participate in the study were significantly older than those who did take part ($P < .001$). The reasons given for refusal were as follows: ‘‘feel too ill’’ (16 patients), ‘‘too distressed’’ (three patients), ‘‘too busy’’ (four patients), ‘‘not interested’’ (four patients), ‘‘taking part in another QOL study’’ (one patient), and ‘‘no glasses’’ (one patient). Twenty-three patients gave no reason for refusal.

Patients’ acceptance and preference. Seventy-eight patients (52%; 95% CI, 44% to 60%) preferred the touch-screen version, 36 (24%; 95% CI, 17% to 31%) preferred the paper version, and 36 (24%; 95% CI, 17% to 31%) had no preference. Overall, 76% of the patients (95% CI, 69% to 83%) found the computer questionnaires acceptable (either preferred or as acceptable as the paper version). Patients’ preference for mode of administration was not related to age, sex, order of presentation, or time for completion.

Data quality. With both systems, entering the demographic data caused significant problems. We observed 285 errors due to misreading of the demographic details written on the paper questionnaires. Entering the identification data using the touch screen was difficult and time consuming. If patients made a mistake, they had to delete their responses and start again. Two patients used letters instead of numbers for their dates of birth and had to restart the program. Three patients had difficulty getting the touch screen to respond because of long nails, and one patient quit the program by mistake.

The quality of the numerical QOL data collected with the electronic touch-screen version was excellent, with no missing or problematic responses, because patients could not progress through the questionnaire without answering each question. With the paper forms, we experienced problems that were either patient related (missing responses, multiple answers, or changed answers: 357 errors) or scanner related (no recognition of response: 725 errors). Only 12 of 158 optically read questionnaires had no scanning, verification, or database errors. Thus, there was significant additional work for the research staff, who had to check and verify the quality of the database after each scanning of paper forms. Each questionnaire took approximately 5 minutes to transfer into the database (1 minute of scanning time and 4 minutes of debugging). The development of the touch-screen version of the QOL questionnaires involved approximately 50 hours of programming and debugging, but once the program was functional and in use, the transfer of responses into the database was direct. The scanning software was a commercially available product, specifically designed to capture this type of data, and the initial work on installation took less than an hour.

Feasibility. The time for completion of the questionnaires had a positively skewed distribution, and a logarithmic transformation of the data was used. The times taken to complete the questionnaires are listed in Table 2. Mean values presented are geometric means (as a result of the transformation). Overall, patients were marginally quicker on the touch screen ($P < .0001$), but there was a significant order effect. Response to the second presentation was quicker, especially when the paper version was administered first ($P < .0001$). There was also a significant age effect, with older patients being slower both on the touch screen and on paper ($P < .001$).

Reliability. Table 3 lists the mean QOL scores for the two methods of administration as well as the mean differences between them and the results of the two-period cross-over analysis. There was a significant order effect on the role function scale and the diarrhea item. For the role function scale, the mean difference between the two administrations was 6.6 when the paper questionnaire was completed first and -2.9 when the touch-screen version was completed first. For the diarrhea item, the observed mean difference was -2.8 when the paper questionnaire was completed first and 3.2 when the touch-screen version was completed first. Patients seemed to report more problems (ie, lower role function and higher level of symptoms) during the first assessment, which made interpretation of the real effect of the methods of administration on scores impossible.

The mean differences between the QOL scores obtained with the two versions of the questionnaires were small (all $<$ 5 percentage points), suggesting equivalence for most of the scales and items. However, on the emotional, fatigue, and nausea/vomiting scales and on the appetite item, the mean difference was statistically significant (P values for Student’s t test for mode effect $< .05$). Patients reported better emotional function and less fatigue, nausea, and loss of

Table 2. Cross-Over Study: Mean Time for Completion of Touch-Screen and Paper Questionnaires and Mean Differences in Time

	No. of Patients	Touch-Screen Version		Paper Version		Mean Difference†	95% CL
		Mean Time*	95% CI	Mean Time*	95% CI		
All patients	146	8.3	7.7-8.9	9.6	8.9-10.4	−1.8‡	−2.7, −0.9
Patients by order of administration							
Touch screen, paper	75	9.0	8.2-9.9	8.7	7.9-9.8	0.17	−1.0, 1.3
Paper, touch screen	71	7.5	6.8-8.2	10.8	9.7-11.9	−3.9§	−5.1, −2.6
Patients by age, years							
< 40	23	7.5	6.5-8.7	8.6	7.1-10.3		
41-60	58	7.1	6.5-7.8	8.6	7.6-9.7		
61-74	46	9.4	8.6-10.3	10.7	9.6-12.0		
> 75	19	11.0	8.1-14.9	12.5	10.1-15.4		

Abbreviation: CL, confidence limits.

*Geometric mean.

†Time on touch screen minus time on paper.

‡$P < .0001$, indicating significant overall difference in time for completion.

§$P < .0001$, indicating significant order (carry-over) effect.

appetite on the touch screen, regardless of the order of presentation. A similar trend was observed on the anxiety subscale of HADS. These apparently significant differences must be interpreted in light of the fact that we performed multiple tests, each of which had a 5% chance of a false-positive result. It was therefore possible that these significant *P* values had occurred purely by chance. However, a general inspection of the values of the mean differences (Table 3) showed predominantly positive values for the functional scales (ie, better function on the touch-screen version) and predominantly negative values for the symptom scales (ie, fewer symptoms reported on the touch-screen version).

The small mean differences between the QOL scores obtained with the two versions of the questionnaires suggested equivalence at group level but did not address equivalence on an individual patient basis. When we examined the range of the magnitude of the differences graphically, we found that although the majority of patients gave either the same answer on both occasions or answers within

Table 3. Cross-Over Study: Mean QOL Scores and Mean Differences in QOL Scores on Touch-Screen and Paper Versions and Results of Two-Period Cross-Over Analysis

	Touch-Screen Version		Paper Version		Mean of Paired Differences*		*P*		
	Mean Score	SD	Mean Score	SD	TS-P	SD	Mode Effect	Order Effect†	Mode-Order Interaction†
EORTC QLQ-C30									
Function scales									
Overall QOL	52.3	25.2	50.5	26.8	1.7	14.6	.16	.59	.20
Physical function	59.7	31.0	59.7	32.5	0.0	17.8	.98	.81	.25
Role function	49.6	32.4	48.0	33.8	1.7	25.5	.38	.02	.20
Emotional function	70.7	23.6	67.0	24.5	3.6	18.1	.02	.51	.38
Social function	55.1	31.8	54.5	34.1	0.6	24.9	.78	.64	.10
Cognitive function	73.7	24.2	72.9	24.7	0.8	16.5	.55	.51	.66
Symptom scales and single items									
Fatigue	48.2	27.1	50.6	26.6	−2.3	14.5	.05	.08	.32
Pain	37.1	33.6	36.8	33.1	0.2	19.5	.88	.93	.91
Nausea/vomiting	17.8	24.5	20.1	27.2	−2.2	13.5	.05	.69	.72
Appetite	34.5	35.3	37.7	34.9	−3.4	20.3	.04	.80	.66
Dyspnea	32.6	35.3	33.1	35.4	−0.5	20.1	.77	.44	.10
Sleep	38.5	32.2	39.6	32.3	−1.1	22.1	.54	.85	.51
Constipation	24.4	33.5	26.7	33.5	−2.3	22.7	.23	.77	.93
Diarrhea	14.1	24.8	13.9	24.7	0.2	16.9	.89	.03	.91
Financial	23.1	30.2	25.4	31.6	−2.3	21.1	.20	.68	.41
HADS									
Anxiety	6.6	4.3	6.9	4.1	−0.4	2.5	.06	.78	.47
Depression	5.8	4.1	5.7	4.1	0.1	2.3	.90	.24	.40

*Touch-screen questionnaire score minus paper questionnaire score.

†Adjusted for age and sex.

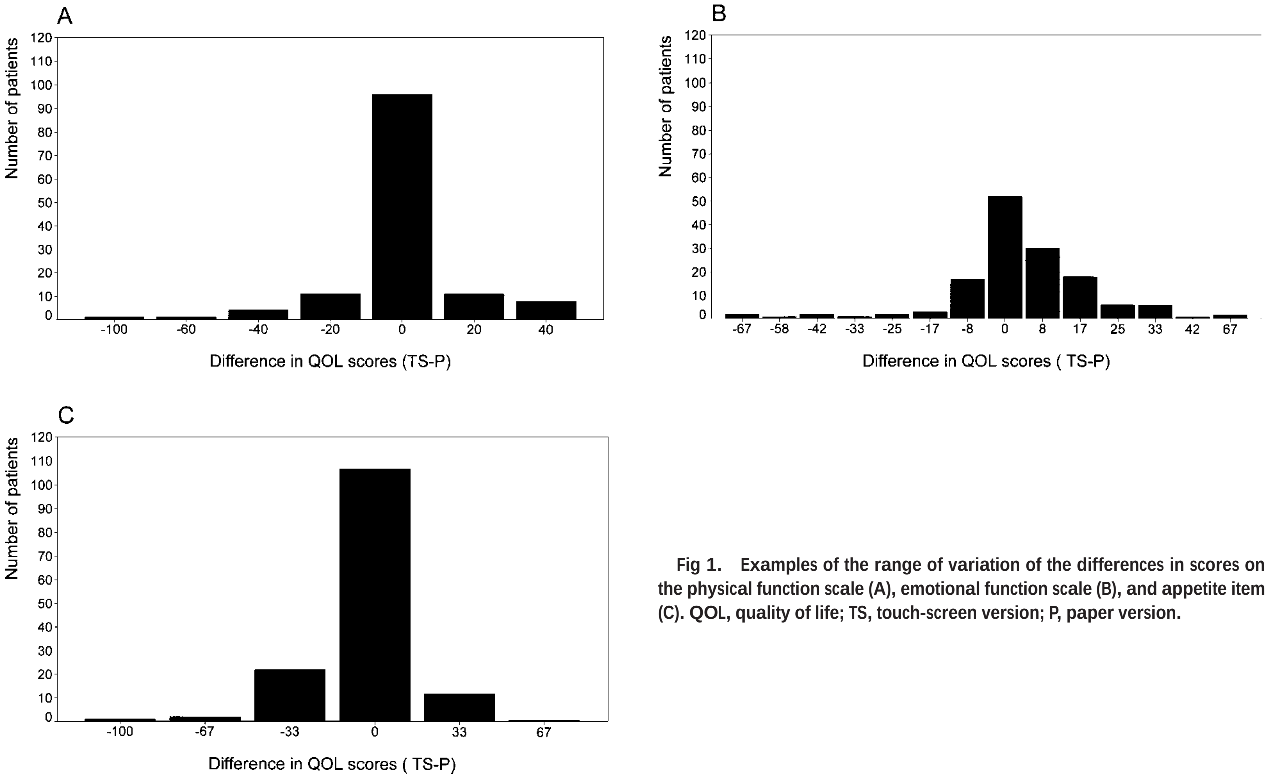

Fig 1. Examples of the range of variation of the differences in scores on the physical function scale (A), emotional function scale (B), and appetite item (C). QOL, quality of life; TS, touch-screen version; P, paper version.

the same response category, a small number of patients gave answers within more than one response category. Figure 1 presents examples of the range of variation of the differences in scores on the physical and emotional scales and the appetite item, and Table 4 lists the proportions of exact and global agreement on all scales and items between first and second administrations of the questionnaires (once using the touch-screen version and once using the paper version).

The agreement between computer and paper questionnaire scores at the individual patient level, calculated by kappa statistic, is also listed in Table 4. The weighted kappa coefficients showed good agreement for most of the scales (kappa $>$ 0.61). The role, emotional, and social function scales had moderate agreement but kappa coefficients were between 0.57 and 0.60.

STUDY EVALUATING TEST-RETEST RELIABILITY OF TOUCH-SCREEN COMPUTER VERSIONS OF EORTC QLQ-C30 AND HADS

Patients and Design

A second study of consecutive inpatients from Cookridge Hospital, Leeds was performed between May 1997 and June 1997. The same eligibility criteria were applied and 124 patients were invited to participate in the study. Patients completed the touch-screen versions of EORTC QLQ-C30 and HADS in the morning and again in the afternoon, with a target interval of 3 hours between the two administrations. The scores obtained with both questionnaires, the length of time taken to complete each questionnaire, and the time between the presentations were recorded. We also collected patients' demographic details and reasons for refusal.

Statistical Methods

The distributions of the scores on EORTC QLQ-C30 and HADS were checked for normality. Test-retest reliability was measured using mean score differences between first and second presentations, weighted kappa coefficients, and percent exact and global agreement (defined as in the first study). We also calculated Pearson's correlation coefficients to compare our results with the only test-retest study of the paper version of EORTC QLQ-C30, which used correlation coefficients and percent exact agreement as statistical methods for assessment of reliability.[24]

RESULTS

Patient characteristics. One hundred two (82%) of 124 patients agreed to take part in the study (46 men and 56 women). Mean age was 57 years (SD, 14.7). Patients had the following diagnoses: gastrointestinal cancer (n = 40), breast cancer (n = 11), male genitourinary cancer (n = 9), cervical cancer (n = 7), lung cancer (n = 5), and other diagnoses, including sarcoma, myeloma, lymphoma, and cancer with an unknown primary site (n = 30). Twenty-two patients refused to participate (six men and 16 women). Mean age was 62 years (SD, 15.4). Five patients refused because they felt too ill or tired, three did not like questionnaires, one did not like computers, three were involved in other studies, and

Table 4. Cross-Over Study: Agreement Between Touch-Screen and Paper Questionnaire QOL Scores at the Individual Patient Level

	Weighted Kappa Coefficient	Percent Exact Agreement	Percent Global Agreement
EORTC QLQ-C30			
Function scales			
Overall QOL	0.73	50	83
Physical function	0.77	73	90
Role function	0.59	50	73
Emotional function	0.57	37	69
Social function	0.60	46	81
Cognitive function	0.64	60	88
Symptom scales			
Fatigue	0.67	40	80
Pain	0.72	61	86
Nausea/vomiting	0.76	74	92
Single items (symptoms)			
Appetite	0.74	74	97
Dyspnea	0.75	78	95
Sleep	0.70	74	96
Constipation	0.73	80	95
Diarrhea	0.75	89	97
Financial	0.72	79	96
HADS			
Anxiety	0.65	23	61
Depression	0.66	28	62

10 did not give a reason. There was no significant difference in age between those who agreed to take part and those who refused.

Eight patients completed only one version of the task: five withdrew from the study, one was discharged, and two were not available for the second assessment. The data from 13 patients had to be discarded because of a software problem (two patients) and building up of static electricity on the touch screen (11 patients). The latter made the screen oversensitive to touch, resulting in a number of questions being skipped by a single touch, especially if the patient's finger trembled. In the course of the study, the screen-sensitivity problem was overcome by placing an antistatic mat under the monitor.

Test-retest reliability. The times for completion of the questionnaires were similar to those for the touch screen in the cross-over study (mean time on first administration, 8.8 minutes, with 95% CI = 8.2 to 9.5 minutes). Patients were slightly quicker on the second presentation (mean time 6.8 minutes, with 95% CI = 6.2 to 7.3 minutes), and older patients took longer to complete the task (data not presented). The median time between the two administrations was 3 hours (range, 1.5 to 5 hours).

Table 5 lists the mean differences, Pearson's correlation coefficients, weighted kappa coefficients, percent exact agreement, and percent global agreement between the QOL scale and single-item scores on first and second administrations of the electronic questionnaires. The mean differences between the scores on the first and second administrations were small and the coefficients of agreement were very good. Eight of the scales and items showed very good agreement ($>$ 0.81), seven showed good agreement (range, 0.61 to 0.80), and two showed moderate agreement (role function, 0.60; depression [HADS], 0.55).

DISCUSSION

At the planning stage of the studies, it was hypothesized that the use of a computer for collection of QOL data might be a reason for some patients to refuse to take part in the studies. Therefore, we collected detailed information on all patients who were asked to participate. The proportion of patients who agreed to take part in the two studies (75% in the cross-over study and 82% in the test-retest study) was comparable with the rate of compliance with QOL studies in therapeutic clinical trials.[25] However, the aim of our studies was to recruit every cancer patient being admitted to our oncology or palliative care wards during the study period, to reflect the reality of oncology and palliative care practice. This attempt resulted in a heterogeneous sample of patients at different stages of disease and treatment, including terminally ill patients, with a wide age range. Therefore, compliance rates of 75% to 80% are likely to reflect the ‘‘true’’ acceptability of QOL studies in a broad oncology practice. The patients who refused to take part in the cross-over study were significantly older than the patients who accepted. The main reasons for refusal were related to the severity of disease. The use of a computer was mentioned as a reason for noncompliance by only one patient in the test-retest study. The computerized version of the QOL questionnaires was well accepted by the patients, with 76% either preferring it to the paper version or having no preference. Patients of all ages and both sexes showed similar preference patterns. The time required to complete the questionnaires on the touch screen was comparable to the time required for the paper versions (even slightly shorter), with an average of between 8 and 10 minutes. A significant learning effect was observed for all age groups, and most patients completed the survey more quickly on the second administration. These results are in keeping with the findings of Buxton et al[26] and demonstrate the feasibility of computerized administration of QOL questionnaires in a general clinical and medical oncology inpatient practice.

A major advantage of the computerized questionnaires is the ability to collect good-quality data without missing or problematic responses. Problems with missing data were detected in many of the paper questionnaires. In their review, Streiner and Norman[27] found that 5% to 10% of returned

Table 5. Test-Retest Study: Results of Comparisons Between QOL Scores From First and Those From Second Administration of Touch-Screen Questionnaires

	Mean Difference*	SD	Pearson's Correlation Coefficient†	Weighted Kappa Coefficient	Percent Exact Agreement	Percent Global Agreement
EORTC QLQ-C30						
Function scales						
Overall QOL	0.3	8.5	0.90	0.90	56	83
Physical function	−0.9	8.3	0.94	0.88	83	100
Role function	4.1	19.8	0.82	0.60	46	78
Emotional function	1.0	11.5	0.84	0.70	44	75
Social function	1.2	16.2	0.86	0.70	57	85
Cognitive function	−1.2	9.4	0.92	0.82	75	98
Symptom scales						
Fatigue	−3.3	14.4	0.83	0.65	38	80
Pain	−2.1	13.8	0.91	0.83	64	89
Nausea/vomiting	−2.3	11.1	0.95	0.78	80	96
Single items (symptoms)						
Appetite	−0.4	14.4	0.94	0.83	81	100
Dyspnea	0.4	13.4	0.91	0.84	84	100
Sleep	−2.1	16.9	0.86	0.75	74	100
Constipation	0.0	10.5	0.92	0.90	73	100
Diarrhea	−0.8	10.5	0.78	0.82	90	100
Financial	0.0	20.4	0.80	0.68	78	95
HADS						
Anxiety	−0.1	2.1	0.84	0.66	25	56
Depression	−0.1	2.1	0.84	0.55	25	56

*Scores obtained during first assessment minus scores obtained during second assessment.

†All correlations were significant ($P < .0001$).

paper questionnaires were reported as unusable because of omitted, illegible, or invalid responses. Our computer program was designed to allow only complete responses, thus overcoming the problem of missing data. Patients could alter a response on the touch screen by returning to the previous question, but they could not skip a question. This restriction was incorporated because we wanted the electronic version to be as close as possible to the original paper questionnaire. Although on the paper questionnaire patients could miss items for various reasons, they were not prompted to do so but rather were asked to answer all questions. Therefore, we thought that adding an option to the computer version to skip a question would prompt patients to consider missing questions. We chose to use standard questionnaires, the paper versions of which are widely used and of proven acceptability to patients. We did not receive any comments from patients regarding the issue of forced responses.

Scanning the paper questionnaires using our system was time consuming and prone to errors. There were substantial technical difficulties involving inaccurate recognition of responses. This required additional verification of the database, making the system difficult to use for routine collection of QOL data in clinical practice. With the electronic questionnaires, both data entry and editing were eliminated and data were transferred directly to the final computer database, allowing immediate printing out and use of the results. These benefits of computerized collection of questionnaire data were emphasized by other researchers.[9,11] One of the problems with this computer system was the buildup of static electricity, which was easily overcome with the use of an antistatic mat. A significant problem of both systems was the slow and inaccurate entry of names, dates of birth, and postal codes by patients themselves. To avoid this, a bar code system for identification can be used with both types of questionnaires, and we are currently using such a system for our subsequent studies with the touch-screen questionnaires. Despite the initial high cost of computer equipment and programming time, routine collection of data using electronic methods may prove to be less expensive in the long term than employing a staff member to scan and verify responses to paper questionnaires.

The equivalence of the touch-screen questionnaires to the paper questionnaires in terms of QOL results was assessed at group and individual levels. At the group level, relatively small mean differences between the QOL scores were found for most of the scales and items. The mean differences in scores on the emotional, fatigue, and nausea/vomiting scales and the appetite item were slightly larger and were statistically significant. These differences were observed in the context of multiple testing and were probably due to random variation. However, there was a consistent trend of patients' reporting fewer problems when using the touch-screen version. In a study comparing standard and computerized versions of QOL questionnaires, completed by diabetic patients, Pouwer et al[15] observed similar (more positive) responses on electronic questionnaires but attributed this finding to chance. In a Canadian study involving computer and paper versions of EORTC QLQ-C30, completed by 50 patients with breast cancer, Taenzer et al[28] did not find any difference in the QOL scores on all scales and items. Although the design of their study was similar to that of our study, there were important differences in patient characteristics. In the Canadian study, the population was relatively healthy and many subjects reported no troubles or negative experiences, which reduced the amount of variation. This floor effect was observed for most of the scales and raised the probability of type 2 error.

The small difference in scores between the electronic and the paper versions could be due to the different format of presentation. In the electronic questionnaire, each question was on a separate screen. This is generally viewed as a positive advantage because the patients cannot see the completed questions and each question must be answered on its own merit. The more positive reporting on the touch screen might be due to patients' perception that the questions and their answers can be more easily seen by other people nearby (including the research assistant), despite the use of monitors with privacy screens. Another possible explanation for the difference is that people respond differently if they are asked by a machine than if they are presented with questions on paper. The scales for which there were differences on the computer version consist of questions requiring judgment of individual subjective feelings rather than judgment of more objective physical limitations or well-defined symptoms. Thus, our findings could lead to a more general hypothesis of a human psychologic reaction to interactive computer systems changing the way people respond. From a practical point of view, researchers who administer electronic versions of paper questionnaires to groups of patients should be aware of a possible small effect of the mode of presentation on responses.

The agreement between the touch-screen questionnaire scores and the paper questionnaire scores on the individual patient level was generally good. It was just below the accepted level for good agreement for the role, emotional, and social function scales. Direct comparisons of agreement coefficients and percent exact and global agreement are not appropriate because different scales have different numbers of questions and different answer categories (eg, yes or no questions or one to four answers to one to seven questions). This point is particularly important for the interpretation of the results for the HADS scale, which showed very low exact and global agreement (because each of its subscales has seven questions with four possible answers) but good agreement coefficients.

The test-retest study showed good and excellent agreement between the scores for the two administrations, with the exception of the role function scale and the depression subscale of HADS. As expected, the proportions of exact and global agreement were slightly higher than in the cross-over study using two different modes of administration. Overall, the results from both studies suggest that the great majority of differences on the individual level are within one response category.

The correlation coefficients in the test-retest study are comparable to those in the test-retest study of the original EORTC QLQ-C30, even somewhat higher, likely because of the short interval between presentations and because of fewer patient response errors.[24] The performance of the role function scales in both studies may cause some concern in terms of reproducibility, given that the scores were consistently higher on the first assessment.

In our present studies, assessment of the new electronic touch-screen versions of EORTC QLQ-C30 and HADS was limited to obtaining test-retest reliability data and direct comparison to the paper versions, suitable for optical mark reading. However, electronic questionnaires also permit immediate calculation and printing out of summary information. This summary report may guide nurses and doctors in more focused inquiry, because it provides reliable assessment of QOL domains. The computer system also produces standardized documentation of the assessment for future reference.

Further research on the touch-screen version of the QOL questionnaires will focus on integration of automated computer-based QOL measurement in everyday oncology practice, assessment of the screening properties of the instruments for detection of psychologic morbidity, and the potential impact of the collected QOL information on patient care.

REFERENCES

1. Coates A, Porzsolt F, Osoba D: Quality of life in oncology practice: Prognostic value of EORTC QLQ-C30 scores in patients with advanced malignancy. Eur J Cancer 33:1025-1030, 1997

2. Dancey J, Zee B, Osoba D, et al: Quality of life scores: An independent prognostic variable in a general population of cancer patients receiving chemotherapy. Qual Life Res 6:151-158, 1997

3. Cella D, Fairclough DL, Bonomi PB, et al: Quality of life in advanced non-small cell lung cancer (NSCLC): Results from Eastern Cooperative Oncology Group (ECOG) study E5592. Proc Am Soc Clin Oncol 16:2a, 1997 (abstr 4)

4. Ibbotson T, Maguire P, Selby P, et al: Screening for anxiety and depression in cancer patients: The effects of disease and treatment. Eur J Cancer 30A:37-40, 1994

5. Ganz PA: Quality of life and the patient with cancer. Individual and policy implications. Cancer 74:1445-1452, 1994 (suppl 4)

6. Osoba D, Rodrigues G, Myles J, et al: Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 16:139-144, 1998

7. Hjermstad MJ, Fayers PM, Bjordal K, et al: Health-related quality of life in the general Norwegian population assessed by the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire: The QLQ-C30 (+ 3). J Clin Oncol 16:1188-1196, 1998

8. Sigle J, Porzsolt F: Practical aspects of quality-of-life measurement: Design and feasibility study of the quality-of-life recorder and the standardized measurement of quality of life in an outpatient clinic. Cancer Treat Rev 22:75-89, 1996 (suppl A)

9. Cella DF: Methods and problems in measuring quality of life. Support Care Cancer 3:11-22, 1995

10. Yarnold PR, Stewart MJ, Stille FC, et al: Assessing functional status of elderly adults via microcomputer. Percept Mot Skills 82:689-690, 1996

11. Drummond HE, Ghosh S, Ferguson A, et al: Electronic quality of life questionnaires: A comparison of pen-based electronic questionnaires with conventional paper in a gastrointestinal study. Qual Life Res 4:21-26, 1995

12. O'Connor KP, Hallam RS, Hinchcliffe R: Evaluation of a computer interview system for use with neuro-otology patients. Clin Otolaryngol 14:3-9, 1989

13. Skinner HA, Allen BA: Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. J Consult Clin Psychol 51:267-275, 1983

14. Lewis G, Sharp D, Bartholomew J, et al: Computerized assessment of common mental disorders in primary care: Effect on clinical outcome. Fam Pract 13:120-126, 1996

15. Pouwer F, Snoek FJ, van der Ploeg HM, et al: A comparison of the standard and the computerized versions of the Well-Being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). Qual Life Res 7:33-38, 1998

16. Aaronson NK, Ahmedzai S, Bergman B, et al: 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 85:365-376, 1993

17. Zigmond AS, Snaith RP: The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 67:361-370, 1983

18. Osoba D, Zee B, Pater J, et al: Psychometric properties and responsiveness of the EORTC Quality of Life Questionnaire (QLQ-C30) in patients with breast, ovarian and lung cancer. Qual Life Res 3:353-364, 1994

19. King MT: The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res 5:555-567, 1996

20. Pocock SJ: Crossover trials, in Pocock SJ (ed): Clinical Trials: A Practical Approach. New York, NY, Wiley, 1983, pp 100-123

21. Fleiss JL: Reliability of measurement, in Fleiss JL (ed): The Design and Analysis of Clinical Experiments. New York, NY, Wiley, 1986, pp 1-32

22. Armitage P, Berry G: Further analysis of categorical data: Kappa measure of agreement, in Armitage P, Berry G (eds): Statistical Methods in Medical Research (ed 3). Malden, MA, Blackwell Science, 1995, pp 443-447

23. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33:159-174, 1977

24. Hjermstad MJ, Fossa SD, Bjordal K, et al: Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire. J Clin Oncol 13:1249-1254, 1995

25. Osoba D: Lessons learned from measuring health-related quality of life in oncology. J Clin Oncol 12:608-616, 1994

26. Buxton J, White M, Osoba D: Patients' experiences using a computerized program with a touch-sensitive video monitor for the assessment of health-related quality of life. Qual Life Res 7:513-519, 1998

27. Streiner DL, Norman GR: Methods of administration, in Streiner DL, Norman GR (eds): Health measurement scales: A Practical Guide to Their Development and Use (ed 2). New York, NY, Oxford Medical, 1995, pp 189-205

28. Taenzer PA, Speca M, Atkinson MJ, et al: Computerized quality-of-life screening in an oncology clinic. Cancer Pract 5:168-175, 1997