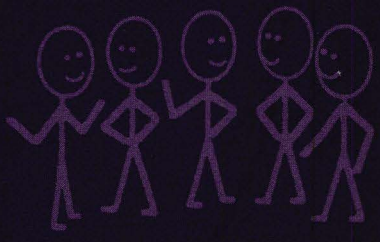
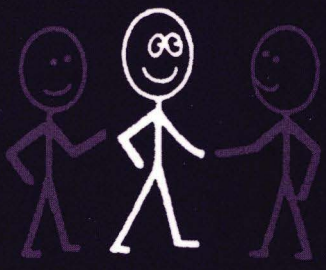
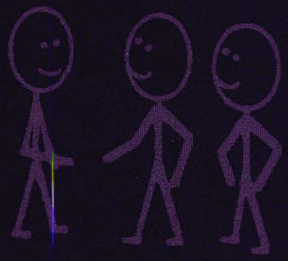


Quality of life assessment by proxy









Kommer C.A. Sneeuw

Quality of life assessment by proxy

Kommer C.A. Sneeuw

K.C.A. Sneeuw
Quality of life assessment by proxy
Thesis: Vrije Universiteit, Amsterdam, 2002
ISBN 90-9015724-7

The research described in this thesis was financially supported by the Dutch Cancer Society (Nederlandse Kankerbestrijding; NKI 93-139 and NKI 90-A).
Final preparations of the thesis were facilitated by TNO Prevention and Health, Leiden.
The printing of this thesis was financially supported by the Dutch Cancer Society.

Cover design: Jaap van der Plas

Printed by: Ridderprint offsetdrukkerij, Ridderkerk, en drukkerij De Bink, Leiden

VRIJE UNIVERSITEIT

Quality of life assessment by proxy

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op woensdag 15 mei 2002 om 15.45 uur
in het auditorium van de universiteit,
De Boelelaan 1105

door

Kommer Cornelis Arie Sneeuw

geboren te Noordoostpolder

promotor: prof.dr. N.K. Aaronson
copromotor: dr. M.A.G. Sprangers

To my proxies

CONTENTS

Chapter 1	Introduction	1
Chapter 2	Value of caregiver ratings in evaluating the quality of life of patients with cancer *Journal of Clinical Oncology 1997; 15: 1206-1217*	7
Chapter 3	Evaluating the quality of life of cancer patients: Assessments by patients, significant others, physicians and nurses *British Journal of Cancer 1999; 81: 87-94*	21
Chapter 4	Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients *Journal of Clinical Epidemiology 1998; 51: 617-631*	31
Chapter 5	The use of significant others as proxy raters of the quality of life of patients with brain cancer *Medical Care 1997; 35: 490-506*	49
Chapter 6	Comparison of patient and spouse assessments of health related quality of life in men with metastatic prostate cancer *Journal of Urology 2001; 165: 478-482*	69
Chapter 7	Assessing quality of life after stroke: The value and limitations of proxy ratings *Stroke 1997; 28: 1541-1549*	77
Chapter 8	Literature review and discussion The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: An update *Submitted for publication*	89
Summary		121
Samenvatting		129
Dankwoord		139
Curriculum Vitae		141

Chapter 1

Introduction

GENERAL INTRODUCTION

Health-related quality of life

Health-related quality of life (HRQL) is increasingly recognized in clinical research as an important outcome of disease and treatment, frequently as a supplement to more traditional endpoints. An on-going research effort over the last several decades has produced a range of generic and disease-specific HRQL questionnaires for use in clinical research.[1] Although there is no general definition of HRQL, most of these questionnaires are designed to assess how patients experience various aspects of their health, including physical and psychosocial functioning, physical symptoms, and overall health or well-being. Incorporating such assessments into clinical research results in a more comprehensive picture of the benefits and costs that accrue from a given therapy.

More recently, attention has been directed toward the possibility of employing individual HRQL assessments in daily clinical practice. Most typically, patients are asked to complete a HRQL questionnaire, either in the traditional paper-and-pencil form or via a computer, the responses are computer-scored and a (graphic) summary is provided to the physician. In this way, physicians receive structured feedback about their patients' HRQL, which can be used for identifying and prioritizing problems, facilitating communication, screening for hidden problems, facilitating shared clinical decision making and monitoring changes over time or responses to treatment.[2]

Role of proxy respondents

Given that the patient is the most appropriate source of information regarding his or her quality of life, HRQL assessments are derived primarily from patients themselves. However, there are several patient groups and situations in which the ability to complete a questionnaire may be compromised. Problems with self-report may arise when patients have insufficient cognitive or communication abilities, when they experience severe symptom distress, or when they find an interview to be physically or emotionally too burdensome. For those patients unable or unwilling to provide HRQL information themselves, their significant others (e.g., spouses, parents, relatives, friends) or health care providers (e.g., physicians, nurses) might be employed as alternative sources of such information, so-called proxy respondents.[3,4]

Thus, the use of proxy respondents may be an effective means of obtaining information that would otherwise be lost. The inability of highly relevant subgroups of patients to participate in clinical HRQL studies may generate findings

that cannot be generalized to the total patient population of interest. This is an important issue of concern in a range of populations, such as the elderly,[5] cancer patients,[6] stroke survivors,[7] patients with neurological deficits,[8] and pediatric patients.[9] When studying such patient populations, researchers frequently rely on information provided by proxy raters. In clinical practice, informed clinical decision-making may be hampered by the inability of patients to provide information about their HRQL. These may be precisely the patients for whom information on quality of life is most needed for delivering the most adequate patient care.

Both the problem of missing data for highly relevant patient subgroups in clinical studies and the factoring of HRQL considerations into the clinical decision-making process lead to the same basic question: to what extent are health care providers and other individuals involved in the care of patients able to assess accurately the patients' quality of life?

AIM AND OUTLINE OF THE THESIS

This thesis represents a systematic effort to investigate the value and limitations of proxy ratings of patients' HRQL. Evaluation of the quality of proxy-generated information typically involves a comparison of patient and proxy ratings. A decade ago, a literature review of 49 studies addressing this issue indicated that the concordance between patient and proxy HRQL ratings was far from optimal, irrespective of the type of proxy rater.[3] However, it was also noted that the literature in this field was characterized by a high degree of heterogeneity and weaknesses in research design. Most importantly, patients and proxy ratings were frequently found to have been derived from different or unstandardized instruments, and the studies were often based on very small sample sizes.

Chapters 2 to 4 of this thesis present the results of a study examining patient-proxy agreement in which an attempt was made to address several of these methodological shortcomings. The study sample was composed of a heterogeneous sample of 320 cancer patients under active treatment with chemotherapy, their significant others (most often spouses), their treating physicians, and nurses for those receiving inpatient chemotherapy. Patients and significant others completed two standardized multidimensional HRQL questionnaires, the Dartmouth COOP Funtional Health Assessment charts/WONCA (COOP/WONCA charts; 7 global health status questions),[10] and the European Organization for Research and Treatment of Cancer Quality of Life Core Questionnaire (EORTC QLQ-C30; 30

cancer-specific questions).[11] Physicians and nurses completed the COOP/WONCA charts only. Additional information was collected to identify variables affecting the level of patient-proxy agreement. The respondents completed the questionnaires at two points in time, during an early phase of treatment and three months later. *Chapter 2* presents a head-to-head comparison of COOP/WONCA chart ratings provided by all patients, significant others and physicians at two points in time. This included not only examination of the level of patient-proxy agreement, but also assessment of the relative validity (i.e., responsiveness to changes over time) of patient- versus proxy-generated information. *Chapter 3*, focusing on the subgroup of inpatients for whom nurse COOP/WONCA chart ratings were obtained as well, investigates the relative effects of the (three) types of proxy raters, the (seven) types of questions/HRQL domains, the patients' clinical status, and several background characteristics of all raters on the level of patient-proxy agreement. *Chapter 4* examines the level and pattern of agreement between patients' and significant others' EORTC QLQ-C30 ratings, the reliability and validity of both types of information, and the influence of several factors on the extent of agreement.

Chapters 5 to 7 describe the results of three clinical studies among specific patient populations, whereby proxy HRQL ratings were collected in addition to patients' self-report. Two of these studies were conducted among cancer patients. *Chapter 5* investigates the level of response agreement between 103 patients with brain cancer and their significant others on the EORTC QLQ-C30 and a brain cancer-specific questionnaire module. *Chapter 6* compares the responses of 72 men with metastatic prostate cancer and their spouses on the EORTC QLQ-C30 and a prostate cancer-specific questionnaire module. *Chapter 7* reports on a study among 437 patients who had suffered a stroke six months earlier. HRQL was assessed by means of the Sickness Impact Profile (SIP).[12] For one-quarter of the patients, who were not able to provide self-report ratings, SIP ratings were provided by their significant others. For 228 of the remaining patients, both patient and significant other SIP ratings were obtained. In addition to evaluating the level of patient-proxy agreement, this study estimated the impact of using proxy HRQL ratings for one-quarter of the patient sample on the results pertaining to one of the research question under investigation (i.e., the relationship between stroke type and HRQL).

Chapter 8 provides a quantitative analysis of the results of the 6 studies described in this thesis and 17 other recent studies examining patient-proxy agreement for well-known, multidimensional HRQL instruments. A number of methodological issues are discussed that require additional attention in determining the value and limitations of proxy data in HRQL studies.

REFERENCES

1. Spilker B: Quality of life and pharmacoeconomics in clinical trials. Philadelphia: Lippincott-Raven publishers, 1996.
2. Higginson IJ, Carr AJ: Measuring quality of life: using quality of life measures in the clinical setting. BMJ 2001; 322: 1297-1300.
3. Sprangers MAG, Aaronson NK: The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992; 45: 743-760.
4. Neumann PJ, Araki SS, Gutterman EM: The use of proxy respondents in studies of older adults: lessons, challenges, and opportunities. J Am Geriatr Soc 2000; 48: 1646-1654.
5. Zimmerman SI, Magaziner J: Methodological issues in measuring the functional status of cognitively impaired nursing home residents: the use of proxies and performance-based measures. Alzheimer Dis Assoc Disord 1994; 8 Suppl 1: S281-S290.
6. Aaronson NK: Methodologic issues in assessing the quality of life of cancer patients. Cancer 1991; 67: 844-850.
7. De Haan R, Aaronson NK, Limburg M, Hewer RL, van Crevel H: Measuring quality of life in stroke. Stroke 1993; 24: 320-327.
8. Murrell R: Quality of life and neurological illness: a review of the literature. Neuropsychol Rev 1999; 9: 209-229.
9. Wallander JL, Schmitt M, Koot HM: Quality of life measurement in children and adolescents: issues, instruments, and applications. J Clin Psychol 2001; 57: 571-585.
10. Van Weel C: Functional status in primary care: COOP/WONCA charts. Disabil Rehabil 1993; 15: 96-101.
11. 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 1993; 85: 365-376.
12. Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981; 19: 787-805.

Chapter 2

Value of caregiver ratings in evaluating the quality of life of patients with cancer

K.C.A. Sneeuw, N.K. Aaronson, M.A.G. Sprangers,
S.B. Detmar, L.D.V. Wever, J.H Schornagel

Journal of Clinical Oncology 1997, 15: 1206-1217

Value of Caregiver Ratings in Evaluating the Quality of Life of Patients With Cancer

By Kommer C.A. Sneeuw, Neil K. Aaronson, Mirjam A.G. Sprangers, Symone B. Detmar, Lidwina D.V. Wever, and Jan H. Schornagel

Purpose: **To evaluate the usefulness of caregiver ratings of cancer patients' quality of life (QL), we examined the following: (1) the comparability of responses to a brief standardized QL questionnaire provided by patients, physicians, and informal caregivers; and (2) the relative validity of these ratings.**

Methods: **The study sample included cancer patients receiving chemotherapy, their treating physicians, and significant others involved closely in the (informal) care of the patients. During an early phase of treatment and 3 months later, patients and caregivers completed independently the COOP/WONCA charts, covering seven QL domains. At baseline, all sources of information were available for 295 of 320 participating patients (92%). Complete follow-up data were obtained for 189 patient-caregiver triads.**

Results: **Comparison of mean scores on the COOP/WONCA charts revealed close agreement between patient and caregiver ratings. At the individual patient level, exact or global agreement was observed in the majority of cases (73% to 91%). Corrected for chance agreement, moderate intraclass correlations (ICC) were noted (0.32 to 0.72). Patient, physician, and informal caregiver COOP/WONCA scores were all responsive to changes over time in specific QL domains, but differed in their relative performance. Relative to the patients, the physicians were more efficient in detecting changes over time in physical fitness and overall health, but less so in relation to social function and pain.**

Conclusion: **For studies among patient populations at risk of deteriorating self-report capabilities, physicians and informal caregivers can be useful as alternative or complementary sources of information on cancer patients' QL.**

J Clin Oncol 15:1206-1217. © 1997 by American Society of Clinical Oncology.

MEASUREMENT of health-related quality of life (QL) is increasingly more common in clinical cancer research.[1] An on-going research effort over the last several decades has produced a range of generic and cancer-specific QL questionnaires for use in clinical trials, a comprehensive summary of which has recently been published.[2] One important starting point in QL research is that the assessment is essentially subjective, with the patient being the primary source of information on his or her QL. This should not, however, imply a wholesale rejection of alternative sources of such information.[3] There are several reasons why it is important to study the value of proxy QL ratings provided by the patients' caregivers at home (eg, family members or close companions) and in the clinic (eg, physicians or nurses).

First, the selective use of such proxy ratings of the patients' QL might contribute to resolving the problem of poor compliance rates in the collection of self-report QL data that has been encountered frequently in cancer clinical studies.[4-9] A significant proportion of patients in these studies failed to complete the QL questionnaires at the required follow-up intervals. Moreover, patient loss to follow-up in QL investigations does not appear to be a random event, but rather is often related directly to the patients' poor health.[6,9,10] Yet it is precisely at the point of disease progression or acute symptom experience that we may be most interested in assessing changes in the QL. Unacceptable levels of missing data, especially non-randomly missing data, may lead to substantial bias in the analysis of QL data.[11,12] This raises the question as to whether caregivers can provide accurate proxy ratings of the patients' QL.

Secondly, proxy judgements of patients' QL can and often do play a role, at least implicitly, in decisions regarding treatment and patient care.[13-15] Particularly in oncology, where many patients are treated with palliative rather than curative intent, QL considerations may weigh heavily in delivering the most adequate patient care.[16] For this reason, it is important to understand the extent to which caregivers can assess accurately the patients' level of functioning and well-being.

Most typically, the accuracy of proxy QL ratings is determined by examining the extent to which proxy ratings correspond to those provided by the patients themselves. The extant literature in this field is characterized by a high degree of heterogeneity in research methodologies, and a diversity of results.[3] Yet the prevailing opinion, at least in oncology research, is that the capacity of caregivers to rate accurately the patients' QL is limited.[1,17]

From the Division of Psychosocial Research and Epidemiology and the Department of Internal Medicine, the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Amsterdam; and the Department of Medical Psychology, University of Amsterdam, Amsterdam, the Netherlands.

Submitted June 24, 1996; accepted September 10, 1996.

Supported by grants no. NKI 93-139 and NKI 90-A from the Dutch Cancer Society, Amsterdam, the Netherlands.

Address reprint requests to Neil K. Aaronson, PhD, Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam; Email naaron@nki.nl.

© 1997 by American Society of Clinical Oncology.

0732-183X/97/1503-0051$3.00/0

An influential study in this area was performed by Slevin et al,[18] who compared patient and caregiver ratings of three items inquiring about QL, anxiety, and depression. They found moderate correlations (ranging from 0.31 to 0.50; interpreted as poor by the investigators themselves) between 100 cancer patients and their physicians, and slightly higher correlations (ranging from 0.41 to 0.54) between a subsample of 50 patients and their relatives. However, in two earlier studies, Spitzer et al[19] and Selby et al[20] reported, as part of the validation of their QL instruments, higher correlations (generally > 0.60) between cancer patients and their physicians on several general health- and disease-related items. More recently, two studies have examined patient-proxy agreement on the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30).[21] In a small study ($N = 52$), Blazeby et al[22] reported poor to moderate levels of agreement (weighted kappa ranging from 0.14 to 0.61) between patients with esophageal cancer and their treating physician, but generally better agreement between a subsample of 39 patients and their family caregivers. Finally, Sneeuw et al[23] found moderate to good correlations (between 0.40 and 0.75 for most QL dimensions) between ratings provided by patients with brain cancer and their informal caregivers on the EORTC QLQ-C30 and a brain cancer–specific questionnaire module.

The assumption underlying the comparison of patient and proxy ratings is that the patient is the primary source of information, and should, consequently, be taken as a gold standard to which the proxy rating should conform. However, patients' ratings themselves are not perfectly reliable. For most questionnaires, reliability estimates fail to meet the 0.90 criterion recommended for interpretation of scores at the individual level.[24,25] Moreover, patients' ratings are also subject to several biases.[26,27] Minimally, discrepancies between patient and proxy ratings should not necessarily be interpreted as evidence of the poor quality of proxy-derived information. Therefore, several investigators have suggested that new studies should extend beyond examination of patient-proxy agreement, by addressing the relative validity of ratings provided by patients and their caregivers.[23,26,28]

In the current study we used two different strategies to examine whether physicians and informal caregivers can provide useful information on the health-related QL of a heterogeneous group of cancer patients. First, we investigated the level of agreement between patient and caregiver responses to a brief standardized QL questionnaire. This included exploring a possible relationship between patient-proxy agreement and the patients' health status, which is highly relevant given the notion that problems with patient self-report are most likely to occur among more impaired patients. Secondly, we extended our analyses beyond the examination of patient-proxy agreement by determining the relative validity of patient- versus caregiver-generated information. More specifically, we compared the responsiveness to changes over time in QL of both patient- and proxy-derived scores. In both analytic strategies, a head-to-head comparison was made between physicians and informal caregivers as proxy raters of patients' QL.

METHODS

Study Sample

In examining the concordance between patient and proxy ratings of patients' QL, it is useful to use a heterogeneous patient sample in terms of disease severity, thereby optimizing the variability in QL ratings. In turn, this can increase the generalizability of the obtained results. Therefore, the patient sample was composed of patients with a range of cancer diagnoses who, during the period between November 1993 and September 1995, attended the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital for treatment involving chemotherapy. Patients were recruited from either the outpatient clinic or one of the clinical wards. To further increase the likelihood of optimal variability in QL ratings, we excluded patients receiving adjuvant chemotherapy, most of whom usually exhibit a relatively good performance status, and we planned the initial assessment at the second (for inpatients) or third (for outpatients) cycle of treatment. Further exclusion criteria included participation in a concurrent QL study, having a treating physician not participating in the study, being younger than 18 years, and a lack of basic proficiency in Dutch.

Eligible patients received a full, verbal and written explanation of the purpose and procedures of the study. Consenting patients were requested to identify an informal caregiver (ie, spouses or others in a close relationship to the patient) and to ask them to participate in the study. The informal caregivers were also provided with verbal and written information on the study. Given their central role in the treatment of patients receiving outpatient chemotherapy, medical oncologists working in the Internal Medicine Department were asked to participate in the study. For inpatients, all ward physicians (interns and residents) employed in the hospital over the entire study period were asked to take part in the study. The physicians also received a full explanation of the purposes and procedures of the study.

Measures and Procedures

Health-related QL was assessed by means of the Dartmouth COOP Functional Health Assessment charts/WONCA.[29-31] The COOP/WONCA charts are an adapted version of the Dartmouth COOP charts,[32] developed by a cooperative group of community medical practices to fill the need for a brief tool for assessing patients' overall functioning. The reliability and validity of the original COOP charts has been established in a number of studies.[32,33] While psychometric testing of the revised version is ongoing, there is ample evidence that the COOP/WONCA charts used here also yield reliable and valid data.[30,31] The COOP/WONCA charts assess QL at a generic level, covering a core set of domains, including physical fitness, feelings, daily and social activities, overall health, and pain. An additional chart assessing overall QL was also included. Each chart consists of a descriptive title, a question referring to a single aspect

of the patient's QL in the past 2 weeks, and five response categories illustrated by drawings. Scores range from 1 to 5, with 1 representing the best and 5 indicating the worst level of functioning or well-being (see Appendix).

Patients, physicians, and informal caregivers were asked to complete the COOP/WONCA charts independently of each other. The proxy questions were identical to those of the patients, but were slightly rephrased so that each question referred to the patient. Also, standard instructions were provided in which proxies were asked to try to view the situation from the perspective of the patient, and to complete the questionnaire as they thought the patient would. While patients and informal caregivers received each question on a separate sheet, the seven questions were concentrated on a single form for use by the physicians.

The COOP/WONCA charts were completed at a baseline assessment (during the second or third cycle of chemotherapy) and at a follow-up assessment 3 months later. Both administrations were planned to take place during patients' scheduled visits to the outpatient clinic or during clinical ward stays. Data were collected by self-administration, with a research assistant present to check for missing data. In most cases, the informal caregiver completed the questionnaire while accompanying the patient to the outpatient clinic or while visiting the patient at the clinical ward. In these cases, the informal caregiver was asked to fill out the questionnaire in a separate room in the presence of another research assistant. When the informal caregiver could not be approached at the hospital, the questionnaire was given to the patient for the proxy to complete at home. In these cases, the patients and proxies received explicit instructions not to discuss the questions with each other. Self-addressed, stamped envelopes were provided for return of the questionnaires, and telephone reminders were used occasionally to maximize response rate. The physicians at the outpatient clinic were asked to complete the COOP/WONCA charts immediately following the medical consultation with the patient. The physicians on the clinical ward completed the form in their office, on the same day as did their patients.

For purposes of establishing the test-retest reliability of the COOP/WONCA charts, subgroups of patients and informal caregivers completed the charts at a third point in time. Patients visiting the hospital at the follow-up assessment were randomly allocated to one of two groups. In one group, the patients themselves were given an additional set of charts, with the instruction to complete them 1 day later and return them by mail. In the second subgroup of patients, the same procedure was followed with their informal caregiver. For practical reasons, the test-retest reliability of the physicians' COOP/WONCA scores was not evaluated.

At the follow-up assessment, the patients also completed a seven-item questionnaire that inquired about the extent to which they had experienced changes over the study period in the seven analogous domains of the COOP/WONCA charts. These so-called transition items are designed to elicit information regarding such perceived changes over time. These questions, based on the Subjective Significance Questionnaire (SSQ),[34] asked the patients to indicate whether their condition (ie, physical condition, emotional state, etc) had changed since the last time they had completed the questionnaire. Seven response categories were available, ranging from "very much worse" to "very much better" ("no change" as middle category).

Patients and informal caregivers provided additional information, including sociodemographic data and information on the nature of their relationship. Clinical information was collected by trained research assistants through medical chart audit. The research assistants also rated the patients' performance status, using the Eastern Cooperative Oncology Group (ECOG) performance status scale, ranging from 0 (normal activity without restriction) to 4 (completely disabled).[35,36]

Statistical Methods

Mean scores, SDs, and the percentages of respondents with the maximum and minimum possible scores on the COOP/WONCA charts were calculated for each source of information. Test-retest reliability of patient and informal caregiver scores on the COOP/WONCA charts was determined by intraclass correlations (ICC) between the follow-up and retest assessments.[24,37,38] The ICC coefficient can vary from 0 to 1, with 1 indicating perfect reliability.

Patient-proxy agreement at the level of the individual patient was assessed by calculating the proportions of exact and global agreement, and the intraclass correlation between the patient and caregiver scores. Exact agreement refers to the proportion of identical patient and proxy responses. Global agreement was defined as the proportion agreement within one response category in either direction. Intraclass correlations were used to discount chance agreements.[39] Patient-proxy ICC may vary from 0 to 1, with 0 indicating no agreement beyond chance and 1 indicating perfect agreement. For ordinal data, as used here, the ICC coefficient has been demonstrated to be mathematically equivalent to the weighted kappa statistic.[40]

Patient-proxy agreement at the group level was evaluated by comparing patient and caregiver mean scores on the seven COOP/WONCA charts. Statistically significant differences between mean scores, using paired Student's t tests, were interpreted as evidence of systematic bias (ie, caregivers tending to report a lower or higher level of QL than the patients themselves).[39] To estimate the statistical magnitude of any observed systematic bias, the mean difference scores were standardized by relating these scores to their SDs. Given the similarity to effect size (d) calculations for paired observations,[41] a standardized difference of $d = 0.2$ was taken to indicate a small difference, $d = 0.5$ a moderate difference, and $d = 0.8$ a large difference.

The relationship between patient-proxy agreement and the patients' clinical status was explored by comparing, for each COOP/WONCA chart, the mean of the absolute differences between patient and proxy scores (possible range from 0 to 4) among patients with a relatively good performance status (ECOG 0 or 1) with that observed among more impaired patients (ECOG 2 or 3). Between-group differences were tested by Student's t test.

Responsiveness, frequently denoted in the literature as sensitivity to change,[38] refers to the ability of an instrument to detect relevant changes over time. For each source of information, changes on the COOP/WONCA charts assessed longitudinally (calculated by subtracting the baseline from the follow-up score) were compared with direct patient perceptions of changes over time in each underlying domain. For the latter direct reports, we used the above-mentioned transition questions, asking the patients directly whether their condition on the seven analogous domains of the COOP/WONCA charts had changed. For purpose of analysis, the seven response categories on these transition items were collapsed into three: patients who felt worse, the same, or better at follow-up. Analysis of variance (ANOVA) was used to test for statistically significant differences in mean change scores between the three subgroups of patients (ie, "worse", "same" or "better" groups). To evaluate the responsiveness of the physician and informal caregiver scores relative to the patients' own scores, relative validity (RV) estimates were calculated, defined as the ratio of the proxy ANOVA F-value to the patient ANOVA F-value.[42,43] The RV estimates indicate, for each domain, and in proportional terms, the ability of physicians and informal caregivers to detect changes over time in the patients' QL, relative

to the patients' self-report. An RV greater than 1 (or RV < 1) indicates that the proxies are more (or less) efficient in assessing changes over time than the patients themselves.

RESULTS

Level of Participation and Follow-Up

Of 378 eligible patients (263 outpatients and 115 inpatients), 320 agreed, after a written request, to participate in the study (85% response rate). Of the 58 nonrespondents, 29 patients chose not to participate in the study because of very poor physical or emotional condition, and the remaining 29 patients reported not being interested or having enough time. At the baseline assessment, physicians' COOP/WONCA ratings were available for 307 of the 320 patients (96%). Similarly, informal caregiver COOP/WONCA ratings were obtained for 308 patients (96%). Nine patients did not have or did not want to ask an informal caregiver to take part in the study, and three caregivers chose not to participate. The baseline analyses focus on the 295 patients (92%) for whom all three sources of QL ratings were available.

Follow-up patient ratings on the COOP/WONCA charts were available for 235 of the 295 (80%) patients with complete baseline data. The reasons for failure to participate at follow-up were death (36 patients), too great a physical or emotional burden (12 patients), lack of interest or time (seven patients), or logistical problems (five patients). At the follow-up assessment, physicians' COOP/WONCA ratings were available for 197 of these 235 patients (84%). For 34 patients, the latter ratings were not available because the patients had changed to a treating physician not participating in the study. Informal caregiver COOP/WONCA ratings were obtained for 227 of 235 patients (97%). All sources of information combined, complete follow-up COOP/WONCA data were available for 189 of the 295 (64%) baseline triads (ie, patients, physicians, and informal caregivers). For 52 of these 189 complete triads (28%), the baseline and follow-up physician scores were provided by different physicians, most often related to changed treatment location (ie, inpatient to outpatient clinic and vice versa). The average duration between baseline and follow-up was 3.4 months (SD, 0.9 months).

After the follow-up assessment, 177 retest questionnaires were randomly distributed to either patients or informal caregivers. Seventy-five patients and 75 informal caregivers returned the questionnaires within 1 week, most of which were completed, as intended, 1 day after the follow-up assessment (79% and 76% for patients and informal caregivers, respectively).

Sample Characteristics

The sociodemographic and medical characteristics of the patients are listed in Table 1. Sixty percent of the sample was female. The patients had a mean age of 52 years (range, 19 to 80). Patients had a range of cancer diagnoses, with advanced breast cancer being the most prevalent diagnosis (34%). All patients were treated with chemotherapy on either an outpatient (69%) or inpatient (31%) basis. While the majority of patients (72%) entered the study with a relatively good performance status (ECOG 0 or 1), a substantial minority of patients exhibited more impaired performance status levels (ECOG 2 or 3).

Table 1. Patients' Sociodemographic and Medical Characteristics

	No. of Patients (N = 295)	%
Sex		
Female	176	60
Male	119	40
Age (years)		
Mean ± SD	52.0 ± 13.6	
Range	19-80	
Marital status		
Married/cohabiting	232	79
Widowed/divorced	28	9
Unmarried	35	12
Education		
Primary	27	9
Secondary	154	52
Advanced secondary	84	28
University	30	10
Tumor site		
Breast	99	34
Gastrointestinal	43	15
Lymphoma	43	15
Melanoma	23	8
Lung	21	7
Genitourinary	21	7
Gynecological	15	5
Soft-tissue/osteosarcoma	13	4
Other	17	6
ECOG performance status*		
0	48	16
1	166	56
2	63	21
3	18	6
Treatment setting		
Outpatient	204	69
Inpatient	91	31

*ECOG performance status score at baseline.

Twenty-nine physicians participated in the study, 14 of whom were medical oncologists working at the outpatient clinic, and 15 were ward physicians working at the inpatient clinic. The mean age of the medical oncologists was 45 years (range, 34 to 56), with an average of 18 years work experience (range, 8 to 28). The ward physicians were younger (mean age, 30 years; range, 26 to 36), and had, on average, 25 months of work experience (range, 2 to 60).

The informal caregivers were most often the patients'

Table 2. Distribution of Patient and Proxy Scores on the COOP/WONCA Charts

	Patient		Physician		Informal Caregiver	
	Mean ± SD	%Max-Min*	Mean ± SD	%Max-Min	Mean ± SD	%Max-Min
Baseline (n = 295)						
Physical fitness	3.1 ± 1.2	12-15	3.3 ± 1.1	5-14	3.3 ± 1.2	12-16
Feelings	2.2 ± 1.0	27-2	2.4 ± 1.0	18-2	2.5 ± 1.1	18-5
Daily activities	2.7 ± 1.3	20-10	2.7 ± 1.2	18-8	2.9 ± 1.3	17-13
Social activities	2.2 ± 1.3	39-6	2.5 ± 1.2	25-6	2.4 ± 1.3	32-6
Overall health	3.3 ± 1.0	7-9	3.1 ± 1.0	6-5	3.5 ± 1.0	3-16
Pain	2.1 ± 1.1	36-3	1.7 ± 1.0	58-1	2.3 ± 1.2	32-3
QL	3.0 ± 1.1	13-5	3.0 ± 1.0	9-3	3.3 ± 0.9	4-9
Follow-up (n = 189)						
Physical fitness	3.0 ± 1.1	11-10	3.2 ± 1.0	7-9	3.3 ± 1.0	4-12
Feelings	2.1 ± 1.0	31-3	2.3 ± 1.0	24-3	2.4 ± 1.1	23-2
Daily activities	2.7 ± 1.2	21-8	2.6 ± 1.2	22-6	2.7 ± 1.2	21-9
Social activities	2.1 ± 1.3	42-5	2.3 ± 1.2	31-5	2.4 ± 1.3	34-8
Overall health	3.2 ± 1.0	7-8	2.9 ± 1.1	16-5	3.4 ± 1.0	5-12
Pain	2.1 ± 1.1	40-2	1.8 ± 1.1	57-2	2.3 ± 1.2	35-2
QL	3.0 ± 1.0	9-3	2.8 ± 1.1	17-3	3.2 ± 1.0	9-6

NOTE. Scores range from 1 to 5 with a higher score representing a more impaired level of functioning or well-being.
*Proportion of patients with the maximum (score 1) and minimum (score 5) level of functioning or well-being.

spouse or partner (74%). The remaining informal caregivers were children (8%), parents (4%), other relatives (8%), or friends (6%). Most caregivers were living in the same household as the patients (80%). The mean age of the informal caregivers was 51 years (range, 18 to 78), and 51% was male.

Score Variability and Reliability

Table 2 lists the distribution of the patient and caregiver scores on the COOP/WONCA charts. At both the baseline and follow-up assessment, all of the charts had substantial variation in scores. The full range of scores was observed on all charts. The scores for feelings, social activities, and pain were somewhat skewed towards the positive end of the scale. The most skewed distribution was found for the pain ratings provided by the physicians, with 58% and 57% (at baseline and follow-up, respectively) of the patients being rated as having no pain.

The test-retest reliability of both patient and informal caregiver scores on the COOP/WONCA charts was good to excellent (Table 3). ICC of patient scores ranged from 0.79 to 0.89. ICC of informal caregiver scores were comparable for physical fitness, feelings, overall health, and pain (ICC = 0.77 to 0.85), but were somewhat lower for daily and social activities, and QL (ICC = 0.63-0.76).

Patient-Proxy Agreement at the Individual Level

The results pertaining to the extent of agreement between patient and caregiver ratings at the level of the individual patient are listed in Table 4. At baseline, proportions of exact agreement varied between 36% and 48% for patient-physician pairs, and between 39% and 50% for patient-informal caregiver pairs. When allowing for one response category of difference in either direction, the proportions of global agreement were approximately 85% for six of the seven domains. For social activities, the proportion of global agreement was approximately 75%, indicating patient-proxy discrepancies of two or more response categories for one quarter of the patients. ICC for patient-physician pairs ranged from 0.32 for social activities to 0.63 for daily activities (mean ICC = 0.48). ICC for patient-informal caregiver pairs were slightly higher (mean ICC = 0.54), especially for feelings, social activities, and pain (ICC = 0.43 to 0.64 compared with ICC = 0.32 to 0.53 between patients and physicians). At the follow-up assessment, higher levels of both patient-physician and patient-informal caregiver agreement were observed for feelings, social activities, pain, and QL (Ta-

Table 3. Test-Retest Reliability of Patient and Informal Caregiver Scores on the COOP/WONCA Charts

	Patient ICC	Informal Caregiver ICC
Physical fitness	.89	.85
Feelings	.79	.77
Daily activities	.86	.76
Social activities	.85	.63
Overall health	.84	.82
Pain	.85	.84
QL	.86	.72

NOTE. Test-retest reliability of the physician scores was not evaluated.

Table 4. Agreement Between Patient and Proxy Scores on the COOP/WONCA Charts at the Individual Level

	Baseline (n = 295)			Follow-up (n = 189)		
	Exact Agreement (%)	Global* Agreement (%)	ICC	Exact Agreement (%)	Global* Agreement (%)	ICC
Patient-physician						
Physical fitness	38	85	.53	41	86	.50
Feelings	37	82	.37	43	86	.47
Daily activities	43	86	.63	39	85	.61
Social activities	36	73	.32	43	79	.50
Overall health	42	87	.51	38	85	.52
Pain	48	86	.53	59	90	.71
QL	38	85	.45	40	85	.53
Patient-informal caregiver						
Physical fitness	42	84	.56	48	87	.57
Feelings	40	85	.48	50	88	.58
Daily activities	43	88	.67	42	87	.65
Social activities	39	76	.43	43	80	.53
Overall health	50	88	.51	49	91	.60
Pain	47	88	.64	55	91	.72
QL	44	84	.48	48	91	.58

*Agreement within one response category in either direction.

ble 4). As noted at baseline, patient-physician correlations (mean ICC = 0.55) at follow-up were slightly lower than patient-informal caregiver correlations (mean ICC = 0.60).

Patient-Proxy Agreement at the Group Level

Agreement at the group level was evaluated by comparing patient and caregiver mean scores. At both baseline and follow-up, statistically significant differences were noted for the majority of patient-physician and patient-informal caregiver comparisons (Table 5). Mean differences between patient and informal caregiver scores were all in the same direction, with the caregivers reporting more impaired levels of functioning and lower levels of well-being than the patients themselves. This trend was not as consistent when examining mean differences between patient and physician scores. Physicians rated the patients as having more problems with physical, emotional, and social functioning, but better overall health and QL (at follow-up only) and less pain than the patients themselves. It is important to note that, when relating the observed mean differences to their SDs, all differences

Table 5. Agreement Between Patient and Proxy Scores on the COOP/WONCA Charts at the Group Level

	Baseline (n = 295)		Follow-up (n = 189)	
	Mean Difference* (Mean ± SD)	Standardized Difference† (*d*)	Mean Difference* (Mean ± SD)	Standardized Difference† (*d*)
Patient-physician				
Physical fitness	0.2 ± 1.1§	0.2	0.2 ± 1.1	0.2
Feelings	0.2 ± 1.1§	0.2	0.2 ± 1.0‡	0.2
Daily activities	0.0 ± 1.0	0.0	−0.1 ± 1.0	−0.1
Social activities	0.3 ± 1.4§	0.2	0.2 ± 1.2‡	0.2
Overall health	−0.2 ± 1.0‡	−0.2	−0.3 ± 1.0§	−0.3
Pain	−0.4 ± 1.0§	−0.4	−0.3 ± 0.8§	−0.4
QL	0.0 ± 1.1	0.0	−0.2 ± 1.0§	−0.2
Patient-informal caregiver				
Physical fitness	0.2 ± 1.1§	0.2	0.3 ± 1.0§	0.3
Feelings	0.3 ± 1.0§	0.3	0.3 ± 0.9§	0.3
Daily activities	0.2 ± 1.0‡	0.2	0.0 ± 1.0	0.0
Social activities	0.2 ± 1.4	0.1	0.3 ± 1.3‡	0.2
Overall health	0.2 ± 1.0§	0.2	0.2 ± 0.9‡	0.2
Pain	0.2 ± 1.0§	0.2	0.2 ± 0.8§	0.2
QL	0.3 ± 1.0§	0.3	0.2 ± 0.9‡	0.2

*Proxy minus patient score.

†*d* = mean difference/SD of difference (*d* = 0.2 small; *d* = 0.5 moderate; *d* = 0.8 large).

‡$P < .05$.

§$P < .01$.

between patient and proxy mean scores were found to be of a relatively small magnitude ($d = -0.4$ to 0.3).

Agreement as a Function of Patients' Performance Status

Analyses of the association between the level of patient-proxy agreement, as indicated by absolute differences between patient and proxy scores, and the patients' performance status yielded an inconsistent pattern of results across the seven COOP/WONCA charts (data not presented in tabular form). At baseline, patients' and physicians' ratings of feelings, social activities, and pain were significantly less often in agreement among patients with a more impaired performance status (ECOG 2 or 3) than among patients with a good performance status (ECOG 0 or 1). For physical fitness, overall health, and QL, the results were in the opposite direction, with more agreement noted among patients having a more impaired performance status. For patient-informal caregiver pairs, similar results were observed for feelings and social activities (ie, less agreement among more impaired patients), and for physical fitness and QL (ie, more agreement among more impaired patients). At the follow-up evaluation, the findings concerning the association between patient-proxy agreement and performance status were partly confirmed for patient-physician pairs (ie, for physical fitness, pain, and QL). For patient-informal caregiver pairs, no significant differences in levels of agreement at follow-up were noted between patients with a good performance status and more impaired patients.

Responsiveness of Patient and Proxy Scores

Table 6 lists longitudinally assessed changes in patient and caregiver scores on the COOP/WONCA charts for three patient subgroups: patients who felt worse, the same, or better at follow-up on the underlying domains (as assessed with the transition questions). Across all raters and domains, patients' direct perceptions of change over time were accompanied by corresponding changes (ie, in the expected direction) in baseline to follow-up scores on the analogous COOP/WONCA charts. For patients reporting worsened functioning and well-being (first column of Table 6), mean change scores varied from 0.2 to 1.1. Similarly, for those patients reporting better health, mean change scores ranged from -0.1 to -0.9. For patients who did not change, mean change scores approached 0 (-0.2 to 0.2). Between-group differences in mean change scores were statistically significant ($P < .001$) for all three raters across each domain, except for physical fitness as measured by patients ($P = .02$) and informal caregivers ($P = .05$). While the patient, physician, and informal caregiver COOP/WONCA scores were all responsive to changes over time in the specific domains, the raters differed in relative performance. Relative to the patients, the physicians were more efficient in detecting changes in physical fitness (RV = 3.42) and overall health (RV = 2.64), but performed less well in detecting changes in social functioning (RV = 0.36) and pain (RV = 0.40). Similar physician RV estimates were noted when restricting these analyses to the 137 patients for whom the baseline and follow-up scores were provided by the same physician. Compared with the patients, informal caregivers were equally efficient in detecting changes in feelings, overall health, and QL, but were less efficient in relation to the remaining domains (RV = 0.61 to 0.69).

DISCUSSION

The purpose of the current study was to evaluate the accuracy of physicians and informal caregivers in rating the QL of cancer patients. The availability of such information could facilitate significant reductions of missing QL data in clinical investigations. Also, to the extent that QL considerations play a role in decisions regarding treatment and patient care, such information can contribute to our understanding of the place of subjective judgement in such decision-making processes. Toward that purpose, two different strategies were used to examine whether persons playing a central role in the care of cancer patients can provide reliable and valid information on various general aspects of the patients' QL.

First, we examined the level of agreement between patient and caregiver responses to the seven questions of the COOP/WONCA charts, each representing a specific QL domain. As expected on the basis of earlier studies among cancer patients,[18-20,22,23] the results indicated generally moderate correlations between patient ratings and those provided by physicians and informal caregivers. The level of agreement between patients and their physicians was slightly lower than that observed between patients and their spouses or close companions, being in line with two previous studies.[18,22] At the initial assessment point, lower levels of agreement were noted for more private domains, such as feelings, social function, and overall QL. However, agreement levels for these domains, as well as for pain, were somewhat higher at the follow-up evaluation. This finding suggests that monitoring of the patients' QL over time may increase caregivers' awareness of patients' psychosocial problems and pain intensity.

When interpreting the magnitude of the observed correlations between patient and caregiver scores, it is important to use the reliability of those scores as a frame of reference. That is, high levels of agreement between patient and caregiver responses cannot reasonably be expected when either one would provide ratings with compromised reliability.[44,45] The present study indicates satisfactory test-retest reliability of patient scores on the COOP/WONCA charts, being well within the range of that observed for QL instruments used frequently in cancer clinical trials.[46,47] With

Table 6. Measured Changes Between Baseline and Follow-up Scores* on the COOP/WONCA Charts as a Function of Patient-Reported Changes at Follow-up†

		Patient-Reported Change at Follow-up				
	No. of Patients‡	Worse Change (Mean ± SD)	Same Change (Mean ± SD)	Better Change (Mean ± SD)	Analysis of Variance¶ (F)	Relative Validity‖ (RV)
Physical fitness	181	(65)§	(55)	(61)		
Patient		0.3 ± 0.9	0.0 ± 0.7	−0.1 ± 1.0	4.19	1.00
Physician		0.5 ± 0.9	−0.1 ± 0.9	−0.4 ± 0.9	14.34	3.42
Informal caregiver		0.4 ± 1.2	0.0 ± 1.0	−0.1 ± 1.0	2.77	0.66
Feelings	179	(42)	(94)	(43)		
Patient		0.4 ± 1.0	−0.1 ± 0.9	−0.7 ± 1.1	12.18	1.00
Physician		0.2 ± 0.9	0.1 ± 0.9	−0.5 ± 1.0	10.13	0.83
Informal caregiver		0.4 ± 1.0	0.1 ± 0.9	−0.6 ± 1.0	12.33	1.01
Daily activities	177	(57)	(77)	(43)		
Patient		0.5 ± 1.0	0.1 ± 0.9	−0.7 ± 0.8	23.15	1.00
Physician		0.8 ± 1.1	−0.1 ± 1.0	−0.7 ± 1.1	23.95	1.03
Informal caregiver		0.6 ± 1.1	−0.1 ± 1.1	−0.6 ± 1.0	14.20	0.61
Social activities	176	(32)	(118)	(26)		
Patient		1.1 ± 1.5	−0.2 ± 1.1	−0.9 ± 1.2	21.23	1.00
Physician		0.6 ± 1.3	−0.1 ± 1.2	−0.7 ± 1.3	7.72	0.36
Informal caregiver		1.0 ± 1.3	0.1 ± 1.4	−0.7 ± 1.2	12.92	0.61
Overall health	176	(50)	(65)	(61)		
Patient		0.4 ± 0.9	0.1 ± 1.0	−0.4 ± 1.0	9.07	1.00
Physician		0.6 ± 1.0	−0.2 ± 1.0	−0.6 ± 0.8	23.92	2.64
Informal caregiver		0.5 ± 0.9	0.0 ± 0.9	−0.3 ± 1.0	10.50	1.16
Pain	178	(47)	(96)	(35)		
Patient		1.0 ± 1.0	−0.1 ± 0.9	−0.6 ± 1.2	31.59	1.00
Physician		0.8 ± 1.2	0.0 ± 1.0	−0.3 ± 1.0	12.67	0.40
Informal caregiver		0.9 ± 1.0	−0.1 ± 1.0	−0.7 ± 1.4	21.65	0.69
QL	163	(35)	(76)	(52)		
Patient		0.4 ± 0.8	0.2 ± 0.9	−0.5 ± 1.0	12.32	1.00
Physician		0.5 ± 0.9	−0.1 ± 1.0	−0.4 ± 0.9	10.51	0.85
Informal caregiver		0.5 ± 1.1	0.0 ± 0.8	−0.5 ± 1.0	15.13	1.23

*Follow-up minus baseline score.

†Patients' perception of change over time in each domain during the study period.

‡Varies because of missing data, but was held constant across raters for each domain.

§No. of patients reporting worse, same, or better health in each domain.

¶ANOVA F-statistics for between-group differences in measured changes between baseline and follow-up scores.

‖RV estimates represent the ratio of the proxy F-values to the patient F-values; for patients, RV is set to 1; an RV > 1 (or RV < 1) indicates that the proxies are more (or less) efficient in assessing changes over time than the patients themselves.

one exception (ie, social function), reliability estimates of informal caregiver scores on the COOP/WONCA charts also exceeded the 0.70 standard recommended for group comparisons in clinical studies.[24,25] However, the reliability estimates of both patient and proxy scores failed to meet the 0.90 criterion recommended for interpretation of scores at the individual level.

When taking the reliability of scores into account, the correlations between patient and caregiver QL ratings are in keeping with expectations. Moreover, the results indicate that, despite suboptimal patient-caregiver correlations, the physicians and informal caregivers provided identical or similar (ie, within one response category) ratings as the patients themselves in the vast majority of cases. Minimally, these findings do not support the currently held view that caregivers are poor judges of the patients' QL.[1,17,18] At the same time, the caregiver ratings are clearly not identical to those of the patients. Thus, it remains important to verify one's perception by eliciting feedback directly from the patients.

Examination of patient-caregiver agreement at the group level is of particular importance for use of proxy QL ratings in clinical trials, where groups of patients are compared rather than individual patients. When comparing group means of patient and caregiver responses, encouraging results were observed. Although systematic differences between patient and proxy mean scores were noted, similar in direction to those reported in previous studies,[3] these differences were small in magnitude. The most pronounced difference was the lower pain ratings of physicians compared with the patients themselves. This confirms the findings of several other studies,[48-50] in which the routine use

of pain assessment tools is recommended to improve caregivers' understanding of their patients' pain status. Nevertheless, overall, the small patient-proxy differences at the group level indicate that only a modest degree of response bias would be introduced when substituting patients' self-report of their QL by caregivers' ratings.

A second analytic strategy used to examine the usefulness of proxy-derived ratings of the patients' QL was to evaluate the validity of both patient and caregiver QL scores. The underlying rationale for performing such analyses is that discrepancies between patient-proxy ratings should not necessarily be interpreted as evidence of the inaccuracy or biased nature of proxy-derived information. An indicator of validity that is particularly important in clinical trials is responsiveness to changes over time.[24,38] The results of this study provide support for the responsiveness of the ratings provided by physicians and informal caregivers to changes in patients' QL. Relative to those of the patients, the physicians' COOP/WONCA ratings were even more efficient in reflecting direct patient reports of change over time in physical function and overall health. While proxy COOP/WONCA ratings, and especially those provided by the physicians, were less responsive to change than patient self-report for social function and pain, the overall findings lend support to the validity of proxy-derived information on the patients' QL.

These results should be interpreted in light of some possible generalizability and methods limitations. The current results were obtained in a study whose specific objective was to examine the usefulness of proxy ratings of cancer patients' QL. One might question the extent to which the results, especially those pertaining to physicians' ratings, are generalizable to actual clinical practice and research situations. However, the fact that the physicians, varying widely in years of experience, did not receive any training in completing the questionnaire (apart from a general introduction for purposes of obtaining consent) and completed it during busy clinical practice, argues for the generalizability of the results to other clinical settings. Interestingly, the fact that the baseline and follow-up QL ratings were obtained from different physicians in slightly more than one quarter of the cases did not appear to compromise their validity (ie, responsiveness). A second concern is that the observed results may not be generalizable to patients unable to provide QL information themselves as a consequence of more severe cognitive, physical or emotional disability.[3,51] The current study did not yield consistent evidence of a relationship between patient-proxy agreement and the patients' overall performance status. However, we would recommend that this issue be examined further in future studies.

For practical reasons (ie, limited time during busy clinical practice), we used a brief questionnaire comprised of seven global questions, each representing a specific QL domain. One might conjecture that the results would be different if a lengthier, more detailed questionnaire was used. Patient-proxy agreement might be poorer when more detailed information is requested; information that demands more precise knowledge of the patients' level of functioning and well-being. Conversely, one could argue that the level of agreement would be heightened by more detailed questions, in that the requested information would be more specific and concrete. Additionally, aggregation of several questions in multi-item scales might also lead to higher levels of patient-proxy agreement, given that multi-item scales are theoretically more reliable than single-item measures.[25] Possible effects of using lengthier, multi-item instruments rather than global questions on the relative validity results are also speculative. Multi-item scales have been shown to detect clinical differences between patients more precisely than single-item measures.[52] If this would be the case for all raters, however, one would expect very similar relative validity estimates to those observed in the current study. Clearly, additional studies are needed to investigate whether equally valid and reliable proxy ratings can be elicited when using lengthier QL questionnaires composed of multi-item scales.

In conclusion, this study provides encouraging findings on the validity of caregiver ratings of several general aspects of cancer patients' QL, and on the comparability of patient and caregiver QL scores at the group level. At the individual patient level, however, both physicians and informal caregivers may sometimes provide different information than patients themselves, especially in the area of psychosocial function. The findings provide support, albeit with the necessary caution, for the feasibility of using proxies for estimating patients' QL in clinical research settings. For studies among patient populations at risk of serious cognitive impairment (eg, patients with brain tumors or brain metastases) or expected deterioration of self-report capabilities, we recommend obtaining caregiver QL ratings parallel to patients' own reports. In case of substantial amounts of missing self-report QL data, consideration should be given to relying primarily on caregiver ratings, rather than on mixing patient- and proxy-derived information. In studies where the amount of missing QL data based on patients' self-reports is relatively modest, one might consider substituting the corresponding proxy ratings. In such situations, however, we would recommend replicating the statistical analyses with and without data substitution to evaluate the impact of such procedures on study results and conclusions.

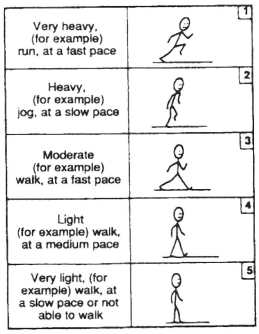
ACKNOWLEDGMENT

We are grateful to the medical and nursing staff of the Department of Internal Medicine of the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, the patients, and their significant others for their willingness to participate in this study.

APPENDIX. The COOP/WONCA Charts

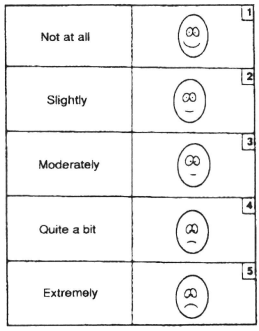
PHYSICAL FITNESS

During the past 2 weeks...
What was the hardest physical activity you could do for at least 2 minutes?



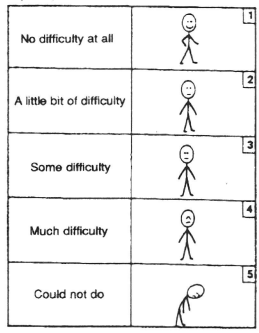
FEELINGS

During the past 2 weeks...
How much have you been bothered by emotional problems such as feeling anxious, depressed, irritable or downhearted and sad?



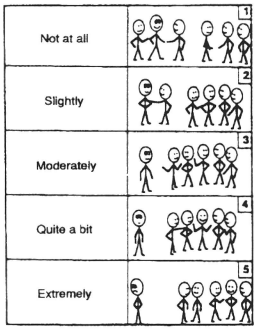
DAILY ACTIVITIES

During the past 2 weeks...
How much difficulty have you had doing your usual activities or task, both inside and outside the house because of your physical and emotional health?



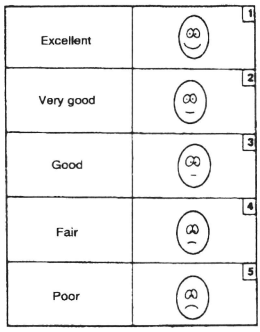
SOCIAL ACTIVITIES

During the past 2 weeks...
Has your physical or emotional health limited your social activities with family, friends, neighbors or groups?



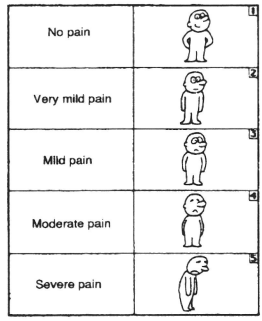
OVERALL HEALTH

During the past 2 weeks...
How would you rate your health in general?



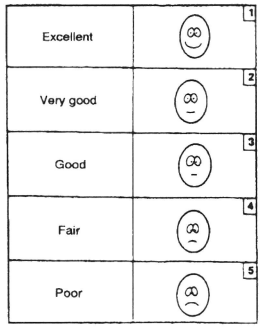
PAIN

During the past 2 weeks...
How much bodily pain have you generally had?



QUALITY OF LIFE

During the past 2 weeks...
How would you rate your overall quality of life?



REFERENCES

1. Osoba D: Lessons learned from measuring health-related quality of life in oncology. J Clin Oncol 12:608-616, 1994
2. Spilker B: Quality of Life and Pharmacoeconomics in Clinical Trials. Philadelphia, PA, Lippincott-Raven, 1996
3. Sprangers MA, Aaronson NK: The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. J Clin Epidemiol 45:743-760, 1992
4. Coates A, Gebski V, Bishop JF, et al: Improving the quality of life during chemotherapy for advanced breast cancer. A comparison of intermittent and continuous treatment strategies. N Engl J Med 317:1490-1495, 1987
5. Finkelstein DM, Cassileth BR, Bonomi PD, et al: A pilot study of the Functional Living Index-Cancer (FLIC) Scale for the assessment of quality of life for metastatic lung cancer patients. Am J Clin Oncol 11:630-633, 1988
6. Ganz PA, Haskell CM, Figlin RA, et al: Estimating the quality of life in a clinical trial of patients with metastatic lung cancer using the Karnofsky performance status and the Functional Living Index-Cancer. Cancer 61:849-856, 1988
7. Hurny C, Bernhard J, Joss R, et al: Feasibility of quality of life assessment in a randomized phase III trial of small cell lung cancer—A lesson from the real world. Ann Oncol 3:825-831, 1992
8. Loizou LA, Rampton D, Atkinson M, et al: A prospective assessment of quality of life after endoscopic intubation and laser therapy for malignant dysphagia. Cancer 70:386-391, 1992
9. Medical Research Council Lung Cancer Working Party: A randomised trial of three or six courses of etoposide cyclophosphamide methotrexate and vincristine or six courses of etoposide and ifosfamide in small cell lung cancer (SCLC) II: Quality of life. Br J Cancer 68:1157-1166, 1993
10. Hopwood P, Stephens RJ, Machin D: Approaches to the analysis of quality of life data: Experiences gained from a medical research council lung cancer working party palliative chemotherapy trial. Qual Life Res 3:339-352, 1994
11. Aaronson NK: Assessing the quality of life of patients in cancer clinical trials: Common problems and common sense solutions. Eur J Cancer 28A:1304-1307, 1992
12. Fairclough DL, Gelber RD: Quality of life: Statistical issues and analysis, in Spilker B (ed): Quality of Life and Pharmacoeconomics in Clinical Trials. Philadelphia, PA, Lippincott-Raven, 1996, pp 427-435
13. Nelson EC, Berwick DM: The measurement of health status in clinical practice. Med Care 27:S77-S90, 1989 (suppl)
14. Schor EL, Lerner DJ, Malspeis S: Physicians' assessment of functional health status and well-being: The patient's perspective. Arch Intern Med 155:309-314, 1995
15. Pearlman RA, Uhlmann RF, Jecker NS: Spousal understanding of patient quality of life: Implications for surrogate decisions. J Clin Ethics 3:114-121, 1992
16. Gough IR, Dalgleish LI: What value is given to quality of life assessment by health professionals considering response to palliative chemotherapy for advanced cancer? Cancer 68:220-225, 1991
17. Nayfield SG, Ganz PA, Moinpour CM, et al: Report from a National Cancer Institute (USA) workshop on quality of life assessment in cancer clinical trials. Qual Life Res 1:203-210, 1992
18. Slevin ML, Plant H, Lynch D, et al: Who should measure quality of life, the doctor or the patient? Br J Cancer 57:109-112, 1988
19. Spitzer WO, Dobson AJ, Hall J, et al: Measuring the quality of life of cancer patients: A concise QL-index for use by physicians. J Chronic Dis 34:585-597, 1981
20. Selby PJ, Chapman JA, Etazadi Amoli J, et al: The development of a method for assessing the quality of life of cancer patients. Br J Cancer 50:13-22, 1984
21. 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
22. Blazeby JM, Williams MH, Alderson D, et al: Observer variation in assessment of quality of life in patients with oesophageal cancer. Br J Surg 82:1200-1203, 1995
23. Sneeuw KCA, Aaronson NK, Osoba D, et al: The use of significant others as proxy raters of the quality of life of patients with brain cancer. Med Care (in press)
24. Hays RD, Anderson R, Revicki D: Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 2:441-449, 1993
25. Nunnally JC, Bernstein IH: Psychometric theory. New York, NY, McGraw-Hill, 1994
26. Gotay CC: Patient-reported assessment versus performance-based tests, in Spilker B (ed): Quality of Life and Pharmacoeconomics in Clinical Trials. Philadelphia, PA, Lippincott-Raven, 1996, pp 413-420
27. Rubenstein LZ, Schairer C, Wieland GD, et al: Systematic biases in functional status assessment of elderly adults: Effects of different data sources. J Gerontol 39:686-691, 1984
28. Farrow DC, Samet JM: Comparability of information provided by elderly cancer patients and surrogates regarding health and functional status, social network, and life events. Epidemiology 1:370-376, 1990
29. Van Weel C: Functional status in primary care: COOP/WONCA charts. Disabil Rehabil 15:96-101, 1993
30. Scholten JHG, Van Weel C: Functional Status Assessment in Family Practice: The Dartmouth COOP Functional Health Assessment Charts/WONCA. Lelystad, the Netherlands, Meditekst, 1992
31. Van Weel C, König-Zahn C, Touw-Otten FWMM, et al: Measuring Functional Health Status With the COOP/WONCA Charts: A Manual. Groningen, the Netherlands, Northern Centre of Health Care Research, 1995
32. Nelson E, Wasson J, Kirk J, et al: Assessment of function in routine clinical practice: Description of the COOP chart method and preliminary findings. J Chron Dis 40:55S-69S, 1987 (suppl 1)
33. Nelson EC, Landgraf JM, Hays RD, et al: The COOP function charts: A system to measure patient function in physicians' offices, in Lipkin M (ed): Functional Status Measurement in Primary Care. New York, NY, Springer-Verlag, 1990, pp 97-131
34. Osoba D, Rodrigues G, Myles J, et al: Significance of changes in health-related quality of life (QOL) scores in women receiving chemotherapy for recurrent or metastatic breast cancer. Qual Life Res 4:468-469, 1995
35. Zubrod CG, Schneiderman M, Frei E, et al: Appraisal of methods for the study of chemotherapy of cancer in man: Comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. J Chronic Dis 11:7-33, 1960
36. Conill C, Verger E, Salamero M: Performance status assessment in cancer patients. Cancer 65:1864-1866, 1990
37. Bartko JJ: The intraclass correlation coefficient as a measure of reliability. Psychol Rep 19:3-11, 1966
38. Deyo RA, Diehr P, Patrick DL: Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. Control Clin Trials 12:142S-158S, 1991 (suppl)
39. Lee J, Koh D, Ong CN: Statistical evaluation of agreement

between two methods for measuring a quantitative variable. Comput Biol Med 19:61-70, 1989

40. Fleiss JL, Cohen J: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33:613-619, 1973

41. Cohen J: The t-test for means, in Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ, Lawrence Erlbaum, 1988, pp 19-74

42. Ware JE, Kosinski M, Bayliss MS, et al: Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the Medical Outcomes Study. Med Care 33:AS264-AS279, 1995

43. Liang MH, Larson MG, Cullen KE, et al: Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 28:542-547, 1985

44. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307-310, 1986

45. Nelson LM, Longstreth WT, Koepsell TD, et al: Proxy respondents in epidemiologic research. Epidemiol Rev 12:71-86, 1990

46. 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

47. Cella DF, Tulsky DS, Gray G, et al: The Functional Assessment of Cancer Therapy scale: Development and validation of the general measure. J Clin Oncol 11:570-579, 1993

48. Hodgkins M, Albert D, Daltroy L: Comparing patients' and their physicians' assessments of pain. Pain 23:273-277, 1985

49. Grossman SA, Sheidler VR, Swedeen K, et al: Correlation of patient and caregiver ratings of cancer pain. J Pain Sympt Manage 6:53-57, 1991

50. Au E, Loprinzi CL, Dhodapkar M, et al: Regular use of a verbal pain scale improves the understanding of oncology inpatient pain intensity. J Clin Oncol 12:2751-2755, 1994

51. De Haan R, Aaronson NK, Limburg M, et al: Measuring quality of life in stroke. Stroke 24:320-327, 1993

52. McHorney CA, Ware JE, Rogers W, et al: The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: Results from the Medical Outcomes Study. Med Care 30:MS253-MS265, 1992

Chapter 3

Evaluating the quality of life of cancer patients: Assessments by patients, significant others, physicians and nurses

K.C.A. Sneeuw, N.K. Aaronson, M.A.G. Sprangers,
S.B. Detmar, L.D.V. Wever, J.H Schornagel

British Journal of Cancer 1999, 81: 87-94

Evaluating the quality of life of cancer patients: assessments by patients, significant others, physicians and nurses

KCA Sneeuw[1], NK Aaronson[1], MAG Sprangers[3], SB Detmar[1], LDV Wever[1] and JH Schornagel[2]

[1]Division of Psychosocial Research and Epidemiology, the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands; [2]Department of Internal Medicine, the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; [3]Department of Medical Psychology, Academic Medical Center, Amsterdam, The Netherlands

Summary This study examined the usefulness of caregiver ratings of cancer patients' quality of life (QOL), an issue of relevance to both adequate patient care and to the possible use of proxy QOL raters in clinical studies. We compared QOL ratings of 90 cancer patients receiving inpatient chemotherapy with those provided by their significant others (most often the spouse), physicians and nurses. During patients' scheduled appointment for receiving chemotherapy on a clinical ward, all raters completed independently the Dartmouth COOP Functional Health Assessment charts/WONCA, an instrument developed by a cooperative group of primary care physicians to briefly assess a core set of seven QOL domains (physical fitness, feelings, daily and social activities, overall health, pain and quality of life) by single items with five response options. With few exceptions, mean scores of the proxy raters were equivalent or similar to those of the patients. Most patient–proxy correlations varied between 0.40 and 0.60, indicating a moderate level of agreement at the individual level. Of all comparisons made, 41% were in exact agreement and 43% agreed within one response category, leaving 17% more profound patient–proxy discrepancies. Disagreement was not dependent on the type of proxy rater, or on raters' background characteristics, but was influenced by the QOL dimension under consideration and the clinical status of the patient. Better patient–proxy agreement was observed for more concrete questions (daily activities, pain) and for patients with either a very good (ECOG 0) or poor (ECOG 3) performance status. The results indicate that both significant others and health care providers can be useful sources of information about cancer patients' QOL.

Keywords: quality of life; proxy respondents; agreement; questionnaire

Quality of life (QOL) assessment is increasingly being used in clinical cancer research as an important outcome for assessing treatment effects (Osoba, 1994; Medical Research Council Lung Cancer Working Party, 1996). Additionally, recent attention has been directed toward the possibility of employing individual QOL assessments in daily clinical practice (Wasson et al, 1992; Detmar and Aaronson, 1998). Both efforts are aimed at the factoring of quality of life considerations explicitly into the medical decision-making process. Given that the patient is the most appropriate source of information on his or her QOL, such assessments are primarily derived from the patients themselves. Yet, there are several reasons for studying the value of proxy QOL ratings provided by the patients' caregivers at home (e.g. spouses, other family members or friends) and in the clinic (e.g. physicians and nurses).

First, it is important to know the extent to which caregivers can assess accurately a patient's level of functioning and well-being, in that such assessments can influence significantly decisions regarding treatment and patient care (Ford et al, 1994; Schor et al, 1995; Macquart-Moulin et al, 1997). Second, there are a number of research situations in which the patient may not be able or willing to provide QOL ratings. Problems with self-report may arise when patients have cognitive impairments or communication deficits, when they experience severe symptom distress, or when an interview is physically or emotionally too burdensome. For such patients, caregivers might be employed as complementary or alternative sources of information on their QOL (Magaziner, 1992; Sprangers and Aaronson, 1992).

Historically, physicians and nurses have played a central role in evaluating patients' QOL, albeit in the limited sense of providing ratings of performance status, treatment toxicity and pain intensity. In earlier work (Sprangers and Aaronson, 1992), we identified 35 published reports evaluating QOL in patients with chronic disease in which ratings from health care providers and patients were compared. These studies indicated that the concordance between patients' and caregivers' QOL ratings was far from optimal, but also suggested a clear need for more methodologically sound studies using larger sample sizes and standard QOL questionnaires. Recently, two studies among large samples of cancer patients have shown more promising results. Stephens et al (1997) reported high levels of agreement between patients' and physicians' ratings on a range of key physical symptoms of the Rotterdam Symptom Checklist as assessed in two randomized trials of palliative treatment for patients with lung cancer. Importantly, they found that the conclusions based on the between-treatment comparisons for these symptoms were essentially the same whether one used the physicians' or the patients' QOL ratings. Sneeuw et al (1997a) also provided encouraging findings

Received 24 August 1998
Revised 29 January 1999
Accepted 12 April 1999

Correspondence to: NK Aaronson

on the validity of physicians' ratings of several general aspects of cancer patients' QOL as measured by the COOP/WONCA charts, an instrument developed by a cooperative group of primary care physicians to briefly assess a core set of QOL domains. Relative to the patients, the physicians were more efficient in detecting changes over time in physical fitness and overall health, but less so in relation to social function and pain.

Increasingly, attention has been paid to the potential use of significant others, particularly spouses, other relatives or friends taking care of the patient in the home situation, as raters of cancer patients' quality of life (Grassi et al, 1996; Kurtz et al, 1996; Sneeuw et al, 1997*a*, 1997*b*, 1998). Theoretically, significant others would seem to be a better choice as proxy raters of patients' QOL than health care providers. They have the opportunity to observe the patient engaging in a wide range of activities over extended periods of time and may have better access to the patient's thoughts and feelings than do health care professionals (Aaronson, 1991). This position has been supported by a number of small studies among cancer patients (Slevin et al, 1988; Blazeby et al, 1995; Sigurdardottir et al, 1996), showing slightly elevated levels of patient–proxy agreement for significant others as compared to either physicians or nurses. On the other hand, as we suggested earlier (Sprangers and Aaronson, 1992), ratings provided by significant others can also be biased by the caregiving function of the rater. In line with this suggestion, other studies have reported that significant others and health care providers evaluate patients' QOL with a comparable degree of (in)accuracy (Grassi et al, 1996; Sneeuw et al, 1997*a*).

The purpose of the current study was to examine the usefulness of caregiver ratings of the QOL of a heterogeneous sample of cancer patients by assessing the level of agreement between patient and proxy responses to a brief standardized QOL instrument. This study contributes to the growing body of research on the value of proxy QOL ratings by providing a head-to-head comparison of the levels of patient–significant other, patient–physician and patient–nurse agreement. Secondly, we investigated the relative effects of the type of proxy rater, the type of question/QOL domain, the patients' clinical status, and several sociodemographic characteristics of all raters on the level of patient–proxy agreement. Finally, in addition to the usual pairwise comparisons, the availability of responses from four different sources also allowed for comparisons of four ratings simultaneously.

METHODS

Study sample

The current analysis was based on data obtained from participants in a larger study examining the value and limitations of proxy ratings of cancer patients' QOL (Sneeuw et al, 1997*a*, 1998). The total patient sample was composed of patients with a range of cancer diagnoses who attended the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital for treatment involving either inpatient or outpatient chemotherapy. For all patients, QOL ratings were obtained from their significant others and treating physicians. Nurse ratings were obtained for hospitalized patients only, given the more frequent and consistent nurse–patient contact in the inpatient clinic setting. Since the aim of the current analysis was to provide a head-to-head comparison of significant others, physicians and nurses, we focused on the patients who were recruited from one of the inpatient wards. Exclusion criteria included participation in a concurrent QOL study, age less than 18 years and a lack of basic proficiency in Dutch.

Eligible patients received a full, verbal and written explanation of the purpose and procedures of the study. Consenting patients were requested to identify a significant other (i.e. spouse or other close companion) and to ask them to participate in the study. The significant others were also provided with verbal and written information on the study. Given their central role in the treatment and care of patients receiving inpatient chemotherapy, all ward physicians (interns and residents) and nurses who worked at the specific inpatient ward over the entire study period were asked to take part in the study. The physicians and nurses also received a full explanation of the purposes and procedures of the study.

Measures and procedures

Quality of life was assessed by means of the Dartmouth COOP Functional Health Assessment charts/WONCA (Scholten and Van Weel, 1992; Van Weel, 1993; Van Weel et al, 1995). The COOP/WONCA charts are an adapted version of the Dartmouth COOP charts (Nelson et al, 1987), developed by a cooperative group of primary care physicians. The reliability and validity of the original COOP charts have been established in a number of studies (Nelson et al, 1987, 1990). While psychometric testing of the revised version is ongoing, there is ample evidence that the COOP/WONCA charts used here also yield reliable and valid data (Scholten and Van Weel, 1992; Van Weel et al, 1995). The COOP/WONCA charts assess QOL at a generic level, covering a core set of domains, including physical fitness, feelings, daily and social activities, overall health and pain. An additional chart assessing overall QOL was also included. Each chart consists of a descriptive title, a question referring to a single aspect of the patient's QOL in the past 2 weeks, and five response categories illustrated by drawings. Scores range from 1 to 5, with 1 representing the best and 5 indicating the worst level of functioning or well-being (Figure 1).

Patients, significant others, physicians and nurses were asked to complete the COOP/WONCA charts independently of each other. The proxy questions were identical to those of the patients, but were rephrased slightly so that each question referred to the patient. Also, standard instructions were provided in which proxies were asked to try to view the situation from the perspective of the patient, and to complete the questionnaire as they thought the patient would. While patients and significant others received each question on a separate sheet, for practical reasons, the seven questions were concentrated on a single form for use by the physicians and nurses.

Data were collected by self-administration during patients' scheduled clinical ward stay at which they received a second cycle of chemotherapy. A research assistant was present to check for missing data. In most cases, the significant other completed the questionnaire while visiting the patient at the clinical ward. In these cases, the significant other was asked to fill out the questionnaire in a separate room in the presence of another research assistant. When the significant other could not be approached at the hospital, an arrangement was made to have the questionnaire completed at home and returned in a self-addressed, stamped envelope. The physicians and nurses on the clinical ward were asked to complete the COOP/WONCA charts in their office, on the same day as did their patients.

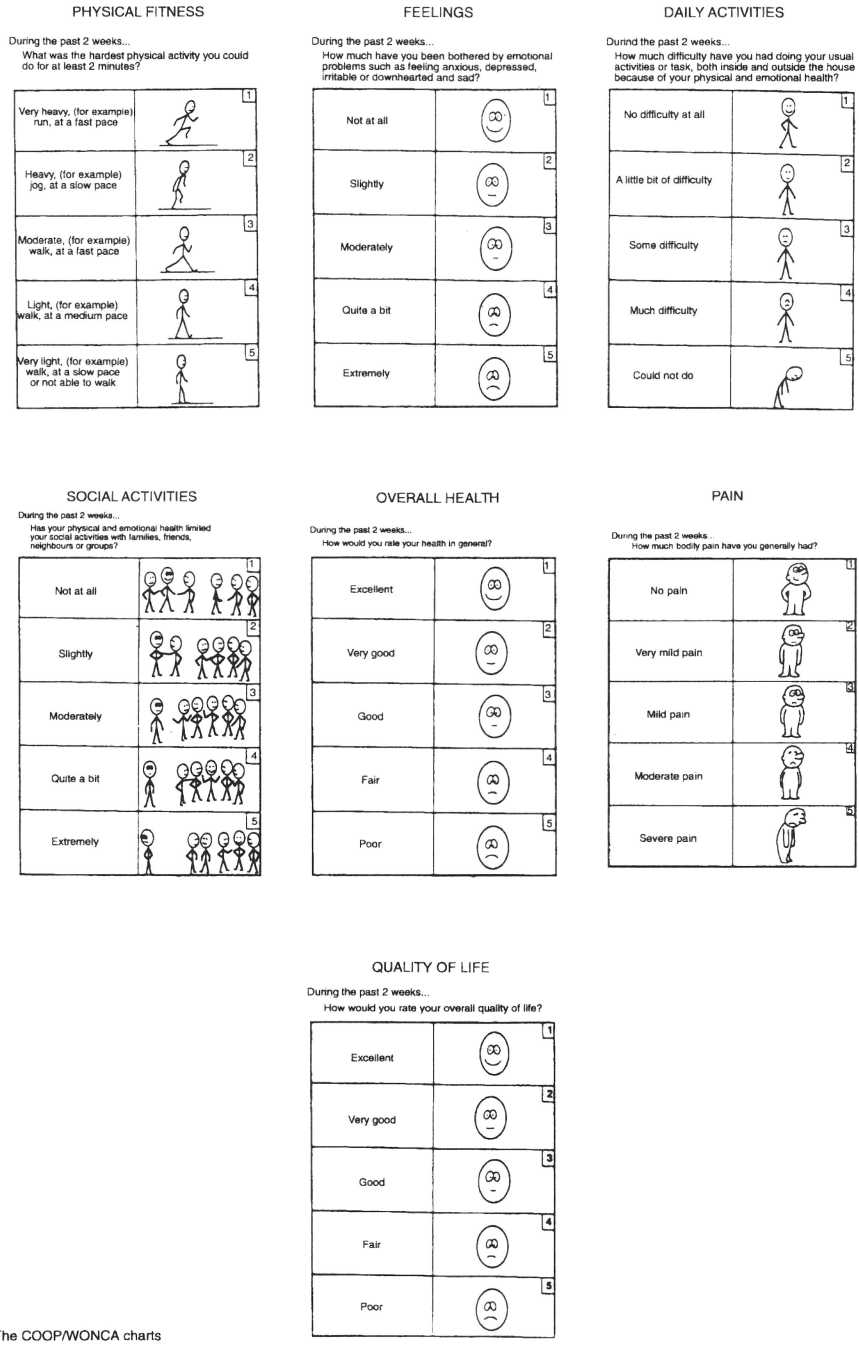


Figure 1 The COOP/WONCA charts

The research assistants also rated the performance status of the patients, using the Eastern Cooperative Oncology Group (ECOG) performance status scale (Zubrod et al, 1960; Sorensen et al, 1993). The ECOG scale describes the patients' level of functioning in terms of activity, ambulatory status and need for care. An ECOG score of 0 means normal activity, ECOG 1 means some restriction in activity but ambulatory, ECOG 2 means capable of self-care but some daytime spent in bed, ECOG 3 means more than 50% of daytime in bed, and ECOG 4 means completely bedridden. To establish the patients' performance status, the research assistants used a standard set of questions based on guidelines recommended by Schag et al (1984).

Data analysis

The level of agreement between patient and proxy ratings was examined in several ways. Mean scores of patients and proxies were compared to examine agreement at the group level. Statistically significant differences in mean scores, as indicated by paired Student's *t*-tests, were interpreted as providing evidence of systematic differences between raters (Lee et al, 1989; Marshall et al, 1994). To interpret the size of observed differences, the mean difference scores were standardized by relating these scores to their standard deviations. Given the similarity to effect size (d) calculations for paired observations (Cohen, 1988), a standardized difference of $d = 0.2$ was taken to indicate a small difference, $d = 0.5$ a moderate difference, and $d = 0.8$ a large difference.

The intraclass correlation (ICC) coefficient was used as an indicator of chance-corrected agreement between patient and proxy ratings at the individual level (Bartko, 1966; Lee et al, 1989). Guidelines for the ICC as a measure of the strength of agreement were labelled as follows: ≤0.40, poor to fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, good agreement; 0.81–1.00, excellent agreement (Landis and Koch, 1977). For ordinal data, as used here, the ICC coefficient has been demonstrated to be mathematically equivalent to the weighted kappa statistic (Fleiss and Cohen, 1973).

Additionally, response agreement between patient and proxy raters was assessed by calculating the proportions of exact agreement, adjacent-category differences, and differences of more than one response category. The latter, being interpreted as large patient–proxy discrepancies, were used to investigate further the relative effects of the type of question (the seven COOP/WONCA charts), the type of proxy rater (significant other, physician, nurse), the patients' performance status (ECOG score), and sociodemographic characteristics (age, sex, education) of all raters on the level of patient–proxy agreement.

Finally, simultaneous comparisons of the four raters were made, establishing the proportions of complete agreement and deviant scores. Complete (or nearly complete) agreement was defined as those cases in which all four raters agreed within one response category (e.g. 2,2,3,2). A deviant score was defined as the situation in which only one of the raters differed more than one response category from the remaining three raters (e.g. 1,4,4,4).

RESULTS

Sample characteristics

Of 115 eligible patients, 100 agreed to participate in the study (87% response rate). Seven of the 15 non-respondents chose not to participate in the study due to very poor physical or emotional condition. The remaining eight non-respondents did not provide any specific reason for not participating other than a general lack of interest in the study. Significant other, physician and nurse COOP/WONCA ratings were obtained for 93, 97 and 99 of the 100 patients respectively. The current analyses focus on the 90 patients for whom all four sources of QOL ratings were available.

Table 1 Comparison of patient and proxy mean scores on the COOP/WONCA charts

	Patient Mean ± s.d.	Significant other Mean ± s.d.	Physician Mean ± s.d.	Nurse Mean ± s.d.
Physical fitness	3.3 ± 1.3	3.5 ± 1.2	3.5 ± 1.0	3.4 ± 1.0
Feelings	2.2 ± 1.0	2.7 ± 1.1[a]	2.6 ± 1.0[a]	2.3 ± 0.9
Daily activities	3.1 ± 1.3	3.3 ± 1.2[a]	3.0 ± 1.1	3.0 ± 1.1
Social activities	2.7 ± 1.4	2.7 ± 1.3	2.7 ± 1.2	2.6 ± 1.0
Overall health	3.4 ± 1.0	3.7 ± 0.9[a]	3.4 ± 0.9	3.5 ± 0.8
Pain	2.2 ± 1.2	2.5 ± 1.2[a]	1.8 ± 1.1[b]	2.1 ± 1.1
Quality of life	3.2 ± 1.1	3.5 ± 0.9[a]	3.2 ± 0.9	3.3 ± 0.8

Note. Scores range from 1 to 5 with a higher score representing a more impaired level of functioning or well-being. [a] Proxy scores significantly higher ($P < 0.05$) than patient scores. [b] Proxy scores significantly lower ($P < 0.05$) than patient scores.

Patients had a range of cancer diagnoses, with lung cancer (23%), advanced breast cancer (22%), testicular cancer (16%) and soft tissue sarcoma (14%) being the most prevalent. The patient sample was heterogeneous in terms of performance status (eight patients ECOG 0, 54 patients ECOG 1, 21 patients ECOG 2, and seven patients ECOG 3). The patient sample included 48 men (53%) and 42 women (47%). Patients' age ranged from 19 to 75 years, with a mean age of 49 years.

The significant others were most often the patients' spouse or partner (76%). The remaining significant others were parents (7%), adult children (6%), other relatives (7%), or friends (4%). Most of them were living in the same household as the patients (82%). Fifty-one significant others were women (57%) and 39 men (43%). Their mean age was 49 years, with a range of 20–78 years.

Fifteen ward physicians (ten female, five male) participated in the study. The physicians were 26–36 years of age (mean age 30 years) and had, on average, 25 months (range 2–60 months) of work experience as a physician. The nurse sample consisted of 35 nurses (29 female, six male). The nurses varied in age from 25 to 56 years (mean age 32 years). On average, they had 11 years (range 4–33 years) of work experience, of which 5 years (range 1–22 years) as an oncology nurse.

Patient–proxy agreement

Table 1 shows the mean scores of the patients and proxy respondents on the COOP/WONCA charts. For seven of the 21 patient–proxy comparisons statistically significant differences were noted, indicating systematic differences between patient and proxy ratings. The significant others rated the patients as having more impaired levels of feelings and daily activities, more pain, and poorer overall health and quality of life than did the patients themselves. The physicians' ratings indicated more impaired feelings, but less pain than those of the patients. No statistically significant differences were observed between the patients' and

Table 2 Intraclass correlations between patient and proxy scores on the COOP/WONCA charts

	Patient vs Significant other ICC	Patient vs Physician ICC	Patient vs Nurse ICC	Average ICC
Physical fitness	0.57	0.53	0.38	*0.49*
Feelings	0.48	0.37	0.43	*0.43*
Daily activities	0.66	0.56	0.58	*0.60*
Social activities	0.47	0.20	0.43	*0.37*
Overall health	0.44	0.45	0.41	*0.43*
Pain	0.64	0.50	0.66	*0.60*
Quality of life	0.37	0.51	0.36	*0.41*
Average	*0.52*	*0.45*	*0.46*	*0.48*

Table 3 Percentage of large discrepancies between patient and proxy scores on the COOP/WONCA charts

	Patient vs Significant other %	Patient vs Physician %	Patient vs Nurse %	Total[a] %
Physical fitness	18	13	21	*17*
Feelings	13	16	11	*13*
Daily activities	13	20	14	*16*
Social activities	27	38	25	*30*
Overall health	9	10	9	*9*
Pain	17	20	10	*16*
Quality of life	21	10	16	*16*
Total[b]	*17*	*18*	*15*	*17*[c]

[a]Across the three pairs of raters for 90 patients (3 × 90 = 270 comparisons).
[b]Across the seven questions for 90 patients (7 × 90 = 630 comparisons).
[c]Across the three pairs of raters and seven questions for 90 patients (3 × 7 × 90 = 1890 comparisons).

nurses' ratings. The standardized differences (effect size *d*) for the five systematic differences observed between patients and significant others ranged between 0.22 and 0.30 for daily activities, overall health and quality of life, and was 0.47 for feelings (not shown in tabular form). The standardized differences of the two systematic patient–physician differences were 0.40 for feelings and –0.40 for pain.

To examine patient–proxy agreement at the individual patient level, a 3 × 7 matrix was constructed of the intraclass correlations between the ratings of the three patient–proxy pairs on the seven COOP/WONCA charts. The average ICC over all 21 correlations was 0.48 (Table 2). Similar levels of agreement were noted between the three patient–proxy pairs, with the average ICCs ranging from 0.45 to 0.52. Average ICCs on the seven COOP/WONCA charts ranged from 0.37 for social activities to 0.60 for daily activities and pain. Relatively lower levels of agreement, as indicated by ICCs <0.40, were observed for five of the 21 patient–proxy comparisons: between patients' and significant others' ratings of overall quality of life, between patients' and physicians' ratings of feelings and social activities, and between patients' and nurses' ratings of physical fitness and overall quality of life.

Factors affecting response agreement

Given three pairs of ratings on the seven COOP/WONCA charts for each of the 90 patients, a potential total of 1890 patient–proxy comparisons could be made. Due to missing data, 11 comparisons were not possible, leaving 1879 comparisons between patient and proxy ratings. Of these, 764 (41%) were in exact agreement, and 801 (43%) agreed within one response category. Large patient–proxy discrepancies (i.e. more than one response category of difference) were noted on 314 (17%) occasions. As shown in Table 3, the percentages of large discrepancies varied across the

Table 4 Percentage of large discrepancies across all comparisons (n = 1890)[a] broken down by explanatory variables

	Proxy characteristics			Patient characteristics	
	No. of comparisons	% Large discrepancies		No. of comparisons	% Large discrepancies
Type of proxy rater			Performance status		
Significant other	628	17%	ECOG 0	165	10%
Physician	624	18%	ECOG 1	1128	16%
Nurse	627	15%	ECOG 2	440	24%
			ECOG 3	146	10%
Proxies' age			Patients' age		
≤ 40	1362	16%	≤ 40	609	16%
41–55	321	18%	41–55	623	19%
55+	189	15%	55+	647	15%
Proxies' sex			Patients' sex		
Male	520	16%	Male	1000	16%
Female	1359	17%	Female	879	17%
Proxies' education			Patients' education		
Low	203	19%	Low	524	22%
Intermediate	264	14%	Intermediate	793	14%
High	1405	17%	High	562	16%

[a]Across three pairs of raters and seven questions for 90 patients (3 × 7 × 90 = 1890 comparisons); number of comparisons varies due to missing data.

Table 5 Number of occasions with one of the raters having a deviant score[a]

	No. of Comparisons	No. of deviant scores				
		Patient	Significant other	Physician	Nurse	Total
Physical fitness	90	6	6	1	3	*16*
Feelings	89	3	1	4	3	*11*
Daily activities	90	4	2	1	3	*10*
Social activities	85	4	5	7	2	*18*
Overall health	89	3	0	3	0	*6*
Pain	88	1	0	4	1	*6*
Quality of life	90	5	1	0	0	*6*
Total	*621*	*26*	*15*	*20*	*12*	*73*

[a]Score of one rater being more than one response category different from those of the other three raters.

seven COOP/WONCA charts. For all patient–proxy pairs, relatively few large discrepancies (on average 9%) were noted for the overall health ratings, and relatively many large discrepancies (on average 30%) for ratings of social activities. The large discrepancies were evenly distributed across the three patient–proxy pairs, indicating that disagreement was not dependent on the type of proxy rater.

Table 4 displays the effects of the type of proxy rater, the patients' performance status and sociodemographic characteristics of the patients and proxy raters on the percentages of large patient–proxy discrepancies. As was observed for the type of proxy rater, the proportions of large patient–proxy discrepancies varied within narrow margins across the different age, sex and education subgroups. For patients' performance status, however, a more substantial effect was observed. For patients with either a very good (ECOG 0) or very poor (ECOG 3) performance status, the percentage of large discrepancies was 10%. For patients with a slightly impaired (ECOG 1) or moderately impaired (ECOG 2) performance status, the corresponding figures were 16% and 24% respectively.

Simultaneous comparison of four raters

Given the seven COOP/WONCA items and 90 patients, a potential total of 630 simultaneous comparisons of four responses could be made. Due to missing data for nine patients, 621 comparisons were possible. Complete or nearly complete agreement, defined as those cases in which the four raters agreed within one response category (e.g. 2,2,3,2), was noted on 340 occasions (55%). Deviant scores, defined as those cases in which only one of the raters differed more than one response category from the remaining three (e.g. 1,4,4,4), were found in 73 cases (12%). Table 5 shows that, of those 73 deviant scores, the patient was the deviant rater on 26 occasions, the significant other on 15 occasions, the physician on 20 occasions and the nurse on 12 occasions. Deviant scores were most often found for physical fitness and social activities (16 and 18 occasions respectively). For physical fitness, 13 of the 16 deviant ratings were in the positive direction (i.e. the deviant rater scoring 'very heavy', the other raters scoring 'moderate' to 'very light').

DISCUSSION

This report describes a head-to-head comparison of significant others, physicians and nurses as proxy raters of the QOL of patients with cancer receiving inpatient chemotherapy, as assessed by the seven questions of the COOP/WONCA charts. To examine patient–proxy agreement at the group level, mean scores of patients and proxy respondents were compared. This is of particular importance for using proxy QOL ratings in clinical trials, where groups of patients are compared rather than individual patients. Significant others systematically reported more problems than did the patients themselves for five of the seven QOL domains. Physicians also rated more emotional problems than did the patients, but underreported pain. The latter finding confirms earlier reports of physicians' tendency to underestimate patients' pain intensity (Hodgkins et al, 1985; Grossman et al, 1991; Au et al, 1994). Interestingly, no systematic differences in mean scores were noted between patient and nurse ratings, suggesting that nurses may be the most suitable source of proxy information in clinical trials of hospitalized cancer patients.

It is important to note that the statistical significance of observed differences is, in part, dependent on sample size. Unfortunately, there are no clear-cut ways of interpreting the importance of statistically significant differences. Ideally, one would like to know at which size a systematic difference is clinically meaningful. Although attempts have been made in this direction (King, 1996; Osoba et al, 1998), the QOL literature does not yet provide unequivocal recommendations. As a 'second best' alternative for interpreting the size of systematic differences, standardized differences (or effect sizes) were employed. In view of guidelines recommended by Cohen (Cohen, 1988), the systematic differences that were observed for seven of the 21 patient–proxy comparisons were small to moderate in magnitude. The larger differences observed between the patients' ratings and those of their significant others and physicians in the area of emotional functioning suggest that the use of proxy respondents may introduce a bias in this QOL domain. Overall, however, the results indicate that only a modest degree of response bias would be introduced when substituting patients' self-report of their QOL by proxy ratings.

To examine patient–proxy agreement at the individual patient level, we employed a combination of the ICC, being a suitable statistical measure for ordinal data (Nelson et al, 1990) and more appealing measures such as the proportions of exact agreement, adjacent-category differences (which can be described as global or approximate agreement) (Sneeuw et al, 1997*a*, 1997*b*), and differences of more than one response category. With few exceptions, the patient–proxy correlations varied between 0.40 and 0.60, usually interpreted as representing a moderate level of agreement (Landis and Koch, 1977). The proportions exact and global agreement indicated that significant others, physicians and nurses provided identical or very similar ratings to those of the patients in the vast majority of cases. Larger differences (of more than one response category on a 1–5 range) between patient and proxy ratings were found in approximately 15% of the cases.

Concordance between patient and proxy ratings appears to be dependent, in part, on the QOL dimension under consideration. As has been suggested earlier (Magaziner, 1992, 1997; Sprangers and Aaronson, 1992), both the type of question and the way in which questions are asked can affect the level of patient–proxy agreement. Specifically, the visibility of the functional problem or

symptom, the concreteness of the question, as well as the number, type and content of response categories can all influence levels of patient–proxy agreement. For the COOP/WONCA charts, higher rates of agreement were therefore expected for physical fitness, daily activities and pain. For the latter two domains, patient–proxy correlations were indeed relatively high. For the physical fitness question, as will be discussed later, responses of questionable validity may have been provided on several occasions. Since problems with the validity of this question were also noted in earlier work, (Sneeuw et al, 1997a; Siu et al, 1993) we would encourage efforts to improve the content of the physical function question and/or the response options. Relatively lower levels of patient–proxy agreement were expected and found for the more private domains of emotional and social function, and the broad concepts of overall health and quality of life.

As is the case for mean scores at the group level, there are no predefined ways to interpret the level of patient–proxy agreement at the individual level. Based on general statistical guidelines (Landis and Koch, 1977), poor to fair agreement was noted for five of the 21 patient–proxy comparisons. One might conclude that such low correlations make proxy ratings unacceptable for either clinical or research use. However, as for all other patient–proxy comparisons, the proportions of exact and global agreement for four of these five comparisons indicated that identical or similar ratings were provided for the large majority of patients. The only exception to this rule was the patient–physician comparison for social functioning, showing clear differences between patient and physician scores in almost 40% of the cases. Overall, we conclude that significant others and health care professionals can provide useful information about general aspects of cancer patients' QOL. At the same time, the proportion of exact agreement (about 40% across all comparisons) and the moderate correlations underscore the fact that patient and proxy ratings are frequently not identical.

Head-to-head comparison of the significant others, physicians and nurses as proxy raters of patients' QOL indicated that (dis)-agreement was not consistently associated with the type of proxy respondent. This finding is at odds with some earlier studies among cancer patients (Slevin et al, 1988; Blazeby et al, 1995; Sigurdardottir et al, 1996), in which it has been suggested that health care professionals are particularly poor judges of patients' QOL. Rather, the current findings support our earlier conclusion, based on a careful review of the literature (Sprangers and Aaronson, 1992), that significant others and health care professionals evaluate patients' QOL with a comparable degree of (in)accuracy. Also, given the modest effect of the age, gender and level of education of the proxy raters on the degree of patient–proxy agreement, there is insufficient evidence to prefer one type of proxy respondent over the other.

The level of patient–proxy agreement was dependent, in part, on the patients' performance status. Large discrepancies between patient and proxy ratings occurred most frequently among patients with a slightly or moderately impaired performance status, and less frequently among patients with either a very good or very poor performance status. This finding is in line with an earlier study (Sneeuw et al, 1998), suggesting a U-shaped relationship between patient–proxy concordance and the level of patients' functioning. This pattern can also be understood intuitively, given the smaller potential for score discrepancies in patients with either a very good or very poor functional status. While for such patients the answers to many questions will be evident (i.e. either at the top or bottom end of the scale), ratings are more likely to diverge for patients with an intermediate performance status. This finding is of particular relevance for the possible use of proxy respondents in clinical studies, because it implies that we can rely on proxy-derived QOL information when the need for proxy QOL ratings is most salient. That is, the use of proxy respondents is of particular relevance for those patients who cannot provide QOL ratings themselves due to their poor clinical status.

Interestingly, when comparing responses of the four raters simultaneously, deviant scores of more than one response category appeared to be caused most often by the patients themselves. This might be interpreted as indicating that all proxy raters were unaware of the patients' health experience. A more likely explanation, however, is that the patients' deviant scores reflect responses of questionable validity. For instance, six patients reported having a high level of physical fitness while their significant other, physician, and nurse reported that the patient could carry out only moderate to very light physical activity. On several occasions, such suspect responses by patients were noted by the research assistants. This finding supports the view that discrepancies between patient and proxy ratings should not necessarily be interpreted as evidence of the poor quality of proxy-derived information (Sneeuw et al, 1997*a*, 1997*b*).

The results of this study indicate that judgements made by significant others and professional caregivers about general aspects of cancer patients' QOL are reasonably accurate. The current study does not support an *a priori* preference for significant others over health care providers. One might conjecture that the results would be different if a lengthier, more detailed questionnaire was used. Patient–proxy agreement might be poorer when more detailed information is requested, demanding more precise knowledge of the patients' level of functioning and well-being. Conversely, one could argue that the level of agreement would be increased by more detailed questions, in that the requested information would be more specific and concrete. Additionally, aggregation of several questions in multi-item scales might also lead to higher levels of patient–proxy agreement, given that multi-item scales are theoretically more reliable than single-item measures (Nunnally and Bernstein, 1994). Results pertaining to the U-shaped relationship between patient–proxy concordance and the level of patients' functioning might also be different if a lengthier questionnaire was used. In a study of stroke survivors, which employed patient and proxy ratings on the Sickness Impact Profile, a linear relationship was found, with disagreement increasing with the level of impairment (Sneeuw et al, 1997*c*). This may be related to the fact that scales of this lengthier questionnaire yield many distinctions at the bottom end of the scale.

We conclude that for clinical studies among patient populations at risk of deteriorating self-report capabilities, both patients' significant others and their health care providers can be useful sources of proxy QOL information. At the same time, researchers need to continue to exercise the necessary caution in the analysis and interpretation of their data when using proxy ratings. Additionally, our findings suggest that, in a routine care situation, informal and professional caregivers of cancer patients are reasonably aware of their patients' level of functioning and well-being. Occasionally, however, substantial discrepancies can occur between patient and caregiver QOL judgements. Thus, for optimal patient care, it remains important to verify one's perception by eliciting feedback directly from the patients, whenever possible.

ACKNOWLEDGEMENTS

The authors are grateful to the medical and nursing staff of the Department of Internal Medicine of the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, the patients, and their significant others for their cooperation in this study. This study was supported by grants no. NKI 93-139 and NKI 90-A from the Dutch Cancer Society.

REFERENCES

Aaronson NK (1991) Methodologic issues in assessing the quality of life of cancer patients. *Cancer* **67**: 844–850

Au E, Loprinzi CL, Dhodapkar M, Nelson T, Novotny P, Hammack J and O'Fallon J (1994) Regular use of a verbal pain scale improves the understanding of oncology inpatient pain intensity. *J Clin Oncol* **12**: 2751–2755

Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* **19**: 3–11

Blazeby JM, Williams MH, Alderson D and Farndon JR (1995) Observer variation in assessment of quality of life in patients with oesophageal cancer. *Br J Surg* **82**: 1200–1203

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn, pp. 19–74. Hillsdale, New Yersey: Lawrence Erlbaum Associates

Detmar SB, Aaronson NK (1998) Quality of life assessment in daily clinical oncology practice: a feasibility study. *Eur J Cancer* **34**: 1181–1186

Fleiss JL and Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* **33**: 613–619

Ford S, Fallowfield L and Lewis S (1994) Can oncologists detect distress in their out-patients and how satisfied are they with their performance during bad news consultations? *Br J Cancer* **70**: 767–770

Grassi L, Indelli M, Maltoni M, Falcini F, Fabbri L and Indelli R (1996) Quality of life of homebound patients with advanced cancer: assessments by patients, family members, and oncologists. *J Psychosoc Oncol* **14**: 31–45

Grossman SA, Sheidler VR, Swedeen K, Mucenski J and Piantadosi S (1991) Correlation of patient and caregiver ratings of cancer pain. *J Pain Symptom Manage* **6**: 53–57

Hodgkins M, Albert D and Daltroy L (1985) Comparing patients' and their physicians' assessments of pain. *Pain* **23**: 273–277

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

Kurtz ME, Kurtz JC, Given CC and Given B (1996) Concordance of cancer patient and caregiver symptom reports. *Cancer Pract* **4**: 185–190

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–174

Lee J, Koh D and Ong CN (1989) Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput Biol Med* **19**: 61–70

Macquart-Moulin G, Veins P, Bouscary M-L, Genre D, Resbeut M, Gravis G, Camerlo J, Maraninchi D and Moatti J-P (1997) Discordance between physicians' estimations and breast cancer patients' self-assessments of side-effects of chemotherapy: an issue for quality of care. *Br J Cancer* **76**: 1640–1645

Magaziner J (1992) The use of proxy respondents in health studies of the aged. In: *The Epidemiologic Study of the Elderly*, Wallace RB and Woolson RF (eds), pp. 120–129. Oxford University Press: New York

Magaziner J (1997) Use of proxies to measure health and functional outcomes in effectiveness research in persons with Alzheimer disease and related disorders. *Alzheimer Dis Assoc Disord* **11**: 168–174

Marshall GN, Hays RD and Nicholas R (1994) Evaluating agreement between clinical assessment methods. *Int J Methods Psychiat Res* **4**: 249–257

Medical Research Council Lung Cancer Working Party (1996) Randomised trial of four-drug versus less intensive two-drug chemotherapy in the palliative treatment of patients with small cell lung cancer (SCLC) and poor prognosis. *Br J Cancer* 73: 406–413

Nelson EC, Wasson JH, Kirk JW, Keller A, Clark D, Dietrich A, Stewart A and Zubkoff M (1987) Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. *J Chronic Dis* **40**: 55S–69S

Nelson EC, Landgraf JM, Hays RD, Kirk JW, Wasson JH, Keller A and Zubkoff M (1990) The COOP function charts: a system to measure patient function in physicians' offices. In: *Functional Status Measurement in Primary Care*, Lipkin M (ed), pp. 97–131. Springer-Verlag: New York

Nelson LM, Longstreth WT Jr, Koepsell TD and Van Belle G (1990) Proxy respondents in epidemiologic research. *Epidemiol Rev* **12**: 71–86

Nunnally JC and Bernstein IH (1994) *Psychometric Theory* . McGraw-Hill: New York

Osoba D (1994) Lessons learned from measuring health-related quality of life in oncology. *J Clin Oncol* **12**: 608–616

Osoba D, Rodrigues G, Myles J, Zee B and Pater J (1998) Interpreting the significance of changes in health-related quality of life scores. *J Clin Oncol* **16**: 139–144

Schag CC, Heinrich RL and Ganz PA (1984) Karnofsky performance status revisited: reliability, validity, and guidelines. *J Clin Oncol* **2**: 187–193

Scholten JHG and Van Weel C (1992) *Functional Status Assessment in Family Practice: the Dartmouth-COOP Functional Health Assessment Charts/WONCA*. Meditekst: Lelystad

Schor EL, Lerner DJ and Malspeis S (1995) Physicians' assessment of functional health status and well-being: the patient's perspective. *Arch Intern Med* **155**: 309–314

Sigurdardottir V, Brandberg Y and Sullivan M (1996) Criterion-based validation of the EORTC QLQ-C36 in advanced melanoma: the CIPS questionnaire and proxy raters. *Qual Life Res* **5**: 375–386

Siu AL, Ouslander JG, Osterweil D, Reuben DB and Hays RD (1993) Change in self-reported functioning in older persons entering a residential care facility. *J Clin Epidemiol* **46**: 1093–1101

Slevin ML, Plant H, Lynch D, Drinkwater J and Gregory WM (1988) Who should measure quality of life, the doctor or the patient? *Br J Cancer* **57**: 109–112

Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV and Schornagel JH (1998) Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. *J Clin Epidemiol* **51**: 617–631

Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV and Schornagel JH (1997*a*) Value of caregiver ratings in evaluating the quality of life of patients with cancer. *J Clin Oncol* **15**: 1206–1217

Sneeuw KCA, Aaronson NK, Osoba D, Muller MJ, Hsu M-A, Yung WKA, Brada M and Newlands ES (1997*b*) The use of significant others as proxy raters of the quality of life of patients with brain cancer. *Med Care* **35**: 490–506

Sneeuw KCA, Aaronson NK, De Haan RJ and Limburg M (1997*c*) Assessing quality of life after stroke: the value and limitations of proxy ratings. *Stroke* **28**: 1541–1549

Sorensen JB, Klee M, Palshof T and Hansen HH (1993) Performance status assessment in cancer patients: an inter-observer variability study. *Br J Cancer* **67**: 773–775

Sprangers MAG and Aaronson NK (1992) The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *J Clin Epidemiol* **45**: 743–760

Stephens RJ, Hopwood P, Girling DJ and Machin D (1997) Randomized trials with quality of life endpoints: are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? *Qual Life Res* **6**: 225–236

Van Weel C (1993) Functional status in primary care: COOP/WONCA charts. *Disabil Rehabil* **15**: 96–101

Van Weel C, König-Zahn C, Touw-Otten FWMM, Van Duijn NP and Meyboom-de Jong B (1995) *Measuring Functional Health Status with the COOP/WONCA Charts: a Manual*. Northern Centre of Health Care Research (NCH): Groningen

Wasson J, Keller A, Rubenstein L, Hays R, Nelson E, Johnson D and The Dartmouth Primary Care COOP Project (1992) Benefits and obstacles of health status assessment in ambulatory settings. The clinician's point of view. *Med Care* 30: MS42–MS49

Zubrod CG, Schneiderman M, Frei E, et al (1960) Appraisal of methods for the study of chemotherapy of cancer in man: comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. *J Chronic Dis* **11**: 7–33

Chapter 4

Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients

K.C.A. Sneeuw, N.K. Aaronson, M.A.G. Sprangers,
S.B. Detmar, L.D.V. Wever, J.H Schornagel

Journal of Clinical Epidemiology 1998, 51: 617-631

Comparison of Patient and Proxy EORTC QLQ-C30 Ratings in Assessing the Quality of Life of Cancer Patients

Kommer C. A. Sneeuw,[1] Neil K. Aaronson,[1,] Mirjam A. G. Sprangers,[3] Symone B. Detmar,[1] Lidwina D. V. Wever,[1] and Jan H. Schornagel[2]*

[1]Division of Psychosocial Research and Epidemiology and [2]Department of Internal Medicine, The Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, NL-1066 CX Amsterdam, the Netherlands; and [3]Department of Medical Psychology, University of Amsterdam/Academic Hospital, NL-1105 AZ Amsterdam, the Netherlands

ABSTRACT. The aim of this study was to examine whether significant others can provide useful proxy information on the health-related quality of life (QL) of cancer patients. We examined the level and pattern of agreement between patient and proxy ratings of the EORTC QLQ-C30, the reliability and validity of both types of information, and the influence of several factors on the extent of agreement. QL ratings were obtained for 307 and 224 patient-proxy pairs (at baseline and follow-up, respectively). Agreement was moderate to good (ICC = 0.42 to 0.79). Multitrait-multimethod analysis showed good convergence and discrimination of specific QL domains. Comparison of mean scores revealed a small but systematic bias between patient and proxy ratings. The maximum level of disagreement was found at intermediate levels of QL, with smaller discrepancies noted for patients with either a relatively poor or good QL. Both patient and proxy QL ratings were reliable and responsive to changes over time. Several characteristics of the patients and their significant others were found to be associated with the level of agreement, but explained less than 15% of the variance in patient-proxy differences. In conclusion, the present findings lend support to the viability of employing significant others as proxy respondents of cancer patients' quality of life where this is necessary. J CLIN EPIDEMIOL 51;7:617–631, 1998. © 1998 Elsevier Science Inc.

KEY WORDS. Quality of life, proxy respondent, agreement, reliability, validity, questionnaire

INTRODUCTION

Quality of life (QL) assessment is increasingly incorporated in clinical research as an important outcome of disease and treatment. Given that the patient is the primary source of information regarding his or her QL, self-report questionnaires are most often used for such assessments. However, there are several patient groups and situations in which the ability to complete a questionnaire may be impaired. Problems with self-report may arise when patients have cognitive impairments or communication deficits, when they experience severe symptom distress, or when an interview is physically or emotionally too burdensome. The inability of such patients to participate in QL studies may yield results that are not representative of the total patient population of interest [1,2].

One possible approach to circumvent this methodological difficulty in assessing patients' QL is the use of proxy respondents [3–5]. For those patients unable or unwilling to provide QL information themselves, health care providers or significant others (e.g., spouses, relatives, or friends) might be employed as alternative sources of information. While the use of proxy raters may be an effective means of obtaining information that might otherwise be lost, it assumes that the proxy respondent can provide reliable and valid data on several aspects of patients' QL. A number of studies have addressed this issue, especially with concern to the value of proxy ratings provided by patients' significant others. Such studies have been carried out in a range of patient populations, including specific elderly patient groups [6–11], stroke survivors [12,13], cancer patients [14–19], and epilepsy patients [20].

Studies evaluating the use of significant others as proxy respondents of patients' QL have primarily examined the extent to which proxy ratings are in agreement with those provided by the patients themselves. Generally, this has included assessing patient-proxy agreement at the level of the

*Address for correspondence: Neil K. Aaronson, Ph.D., Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands.

Accepted for publication on 31 March 1997.

individual patient, most often by means of correlations between patient and proxy scores, as well as agreement at the group level, by comparing patient and proxy mean scores. The former method provides a direct indication of the extent to which the proxy respondents' QL ratings concur with those of the patients themselves. The latter method allows one to determine the direction and magnitude of any systematic bias that might be introduced in QL investigations when using proxy respondents. The accumulated results to date suggest that, at the individual patient level, patient-proxy agreement is generally moderate. Most typically, measures of agreement (i.e., correlation coefficients or weighted kappa), range between 0.40 and 0.60. Lower levels of patient-proxy agreement are often found in ratings of psychosocial versus physical functioning. At the group level, systematic differences between patient and proxy mean scores are frequently observed, with a tendency of proxies to rate patients as having a more impaired QL than the patients themselves. In general, however, the magnitude of this bias tends to be limited.

One methodological difficulty in studies examining patient-proxy agreement is that, by definition, both patients and proxies are required to complete the same questionnaire. Hence, patient-proxy agreement cannot be examined for those patients for whom the need for proxies is most salient (i.e., those who are not capable of self-report). Thus, it is important to examine trends in patient-proxy agreement as a function of patients' level of QL, so that findings can potentially be extrapolated to more impaired patient subgroups. Findings from studies addressing this issue are inconclusive. In three studies [8,13,18], lower levels of agreement and more biased ratings were observed for patients with more impaired levels of functioning. This trend was not observed in several other studies [6,9,19,20], in which the differences between patient and proxy ratings were either unrelated or not consistently related to the patients' level of functioning.

The assumption underlying the comparison of patient and proxy ratings is that the patient is the primary source of information and should, consequently, be taken as the gold standard to which the proxy rating should conform. However, patients' ratings themselves are not perfectly reliable and valid. Minimally, discrepancies between patient and proxy ratings should not necessarily be interpreted as evidence of the poor quality of proxy-derived information. Therefore, it is important to extend beyond examination of patient-proxy agreement by determining the reliability and validity of patient and proxy ratings separately. A number of studies have reported a head-to-head comparison of reliability estimates based on both patient and proxy ratings [7,13,18–20]. In general, the findings indicate that the reliability of proxy-generated data is similar to that of the patient. However, the reliability estimates of both patient and proxy scores usually fail to meet the 0.90 criterion recommended for interpretation of scores at the individual level [21].

The validity of proxy ratings has been examined in two studies using different methods [11,19]. Sneeuw *et al.* [19] examined the responsiveness to change over time in specific QL domains of both patient- and proxy-derived data. All patient and proxy scores were responsive to changes over time, but differed somewhat in their relative performance. As compared to the patients themselves, significant others (i.e., most often the spouse or partner) were equally efficient in detecting changes in emotional function, overall health, and quality of life, but slightly less so in relation to other QL domains. Rothman *et al.* [11] explored the validity of patient and proxy scores of patients' QL by comparing regression equations predicting these respective QL scores. That is, they assumed both patient and proxy QL ratings to be related primarily to other indicators of patients' health and functioning, and to be confounded to the extent that other characteristics (e.g., sociodemographic variables, proxies' own health and well-being) are associated with these scores. The patients' own QL ratings and proxy-generated physical function scores were found to be relatively free of the influence of other variables. The proxy-generated psychosocial function scores, however, appeared to be heavily influenced by the perceived burden of caregiving and the significant others' own psychological distress.

Several studies comparing patient and proxy ratings of patients' QL have investigated the influence of a range of proxy characteristics on the level of agreement [6–10,18,20]. These studies might help to identify the most appropriate proxy respondent (if there is a choice) and to interpret responses from varying types of proxy raters [9]. Characteristics such as age, sex, education, relationship to patient, living arrangement in relation to patient, frequency of contact, and caregiving function were examined. While most variables have been shown to exert an influence on the extent of agreement between patients and their significant others, the findings have not been consistent across the studies and types of information being requested [3,4]. Moreover, the outcomes from these studies were based, with few exceptions [6,20], on univariate rather than multivariate analytic methods. Thus, they did not provide an indication of the extent to which the level of patient-proxy agreement might be affected by several characteristics combined (i.e., which proportion of variance in patient-proxy differences can be explained by proxy characteristics).

The purpose of the current study was to examine whether significant others can provide useful information on the health-related quality of life of patients with cancer. We employed a comprehensive analytical framework, drawing on approaches employed in earlier studies, thereby enabling a meaningful comparison of results across studies. The following issues were addressed: (1) the agreement between patient and proxy responses to a QL instrument frequently used in cancer clinical research; (2) the association between patient-proxy agreement and the patients' level of QL; (3) the reliability and responsiveness of both patient- and proxy-derived information; (4) the influence of patient and

proxy characteristics on their respective ratings of patients' QL; and (5) the influence of patient and proxy characteristics on the absolute level and direction of differences between their ratings.

METHODS

Study Sample

In examining the concordance between patient and proxy ratings of patients' quality of life, it is useful to employ a heterogeneous patient sample in terms of disease severity, thereby optimizing the variability in QL ratings. In turn, this can increase the generalizability of the obtained results.

PATIENTS. The patient sample was composed of patients with a range of cancer diagnoses who, during the period between November 1993 and September 1995, attended the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital for treatment involving chemotherapy. Patients were recruited from either the outpatient clinic or one of the clinical wards. To further increase the likelihood of optimal variability in QL ratings, we excluded patients receiving adjuvant chemotherapy, most of whom usually have a relatively good performance status, and we planned the initial assessment at the second (for inpatients) or third (for outpatients) cycle of treatment. Further exclusion criteria included participation in a concurrent QL study, less than 18 years of age, and a lack of basic proficiency in Dutch. Eligible patients received a full, verbal and written explanation of the purpose and procedures of the study.

PROXY RESPONDENTS. Consenting patients were requested to identify a significant other (i.e., spouse or other person in close relationship to the patient) and to ask him/her to participate in the study. The significant others were also provided with verbal and written information on the study. The study was approved by the institutional review board of the hospital.

Quality of Life Measurement

Quality of life was assessed by means of the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ-C30), version 2.0 [22,23]. This questionnaire is composed of 30 items, organized into a number of multi-item scales and single items that reflect a range of physical, emotional, and social health issues relevant to a broad spectrum of cancer patients. It incorporates five functioning scales (physical, role, cognitive, emotional, and social functioning), three symptom scales (fatigue, nausea and vomiting, and pain), six single items (dyspnea, insomnia, anorexia, constipation, diarrhea, and financial impact), and one scale assessing global quality of life. The questionnaire employs a 1-week time frame and a mix of dichotomous response categories ("yes/no"), 4-point Likert-type response scales (ranging from "not at all" to "very much"), and 7-point response scales (numbered visual analogue scales). The QLQ-C30 has been shown to be reliable and valid in a wide range of patient populations and treatment settings. Across a number of studies, internal consistency estimates (Cronbach's coefficient alpha) of the multi-item scales met or approached the 0.70 reliability criterion recommended for group comparisons [24]. Test-retest reliability coefficients have been found to range between 0.80 and 0.90 for most multi-item scales and single items [25]. Tests of validity have shown the QLQ-C30 to be sensitive to meaningful between-group differences (e.g., local vs. metastatic disease, active treatment vs. follow-up) and changes in clinical status over time [22,24].

PROXY VERSION. Proxy respondents completed a slightly modified version of the QLQ-C30. Standard instructions were provided in which proxies were asked to try to view the situation from the patients' perspective, and to complete the questionnaires accordingly. Additionaly, all item statements were made from the third-person perspective (e.g., "Would the patient say that he/she. . . .").

SCORING METHOD. Following the scoring procedures recommended by the EORTC [22,24], all scale and single-item scores of the QLQ-C30 were linearly transformed to a 0 to 100 scale. For the functioning scales higher scores represent a better level of functioning; for the symptom measures a higher score corresponds to a higher level of symptomatology. While not part of the usual scoring method, we also calculated, for analytic purposes, a total QL score by aggregating the 15 functioning and symptom measures. For calculation of this total QL score, the nine symptom measures were reversed so that a higher total QL score represents a better quality of life.

CHANGE IN QL. At the follow-up assessment, the patients also completed six questions which inquired about the extent to which they had experienced changes over the study period in specific domains of their life. These so-called "transition items" are designed to elicit information regarding such perceived changes over time. These questions, based on the Subjective Significance Questionnaire (SSQ) [26], asked the patients to indicate whether their condition (i.e., physical, role, emotional, and social functioning, pain, and global quality of life) had changed since the last time they had completed the questionnaire. Seven response categories were available, ranging from "very much worse" to "very much better" ("no change" as middle category).

Patient and Proxy Characteristics

A range of variables was selected to explore the influence of patient and proxy characteristics on their respective ratings of the patients' QL, and on the differences between their ratings.

PATIENT CHARACTERISTICS. Characteristics of the patients included: (1) indicators of health (i.e., performance status, weight loss, and mental health); (2) sociodemo-

graphic data (i.e., age, sex, and education); and (3) indicators of reporting styles (i.e., social desirability, positive appraisal, and social expressiveness). The latter three variables were included to explore the potential effects of patients' habitual reporting styles on the way in which they report on their symptoms and functional problems. It may well be that patients' self-reports of their QL do not always correspond to their actual experiences. In turn, this may influence the level of patient-proxy agreement.

Performance status was rated by the research assistants, using the Eastern Cooperative Oncology Group (ECOG) performance status scale, ranging from 0 (normal activity without restriction) to 4 (completely disabled) [27,28]. The degree of weight loss (none, $\leq$10%, or >10%) was determined on the basis of patients' reports of their current weight and weight loss in the past 2 months. Mental health was measured by the 5-item version of the Mental Health Inventory (MHI-5), which has proven to be a robust measure of psychological functioning [29]. Social desirability, assessing patients' tendency to answer questions in such a way as to represent themselves favorably, was measured by a 5-item instrument of socially desirable response set (SDRS-5) [30]. This short-form measure contains selected items from the Marlowe-Crowne scale (e.g., "I am always courteous even to people who are disagreeable") [31]. Positive appraisal, defined as patients' habitual tendency to employ a positive style of coping with problems and unfavorable events, was measured by six items derived from the Utrecht Coping List (i.e., five items of the coping style characterized as "reassuring thoughts," and one item measuring "being optimistic about the future") [32]. Social expressiveness, defined as patients' habitual tendency to cope with problems and unfavorable events by expressing their feelings to others, was measured by six items derived from the "seeking support" subscale of the Utrecht Coping List (e.g., "sharing concerns," "discussing problems with family or friends").

PROXY CHARACTERISTICS. Characteristics of significant others pertained to: (1) sociodemographic data; (2) indicators of health (i.e., mental health and global health/QL); and (3) the perceived intensity and burden of their caregiving function.

As was the case for the patients, the significant others' mental health was measured by the 5-item version of the Mental Health Inventory (MHI-5). The significant others' own global health/QL was assessed by means of the global quality of life scale of the EORTC QLQ-C30. The intensity of the caregiving function was measured by six items assessing the extent to which they needed to assist the patient with activities (e.g., shopping, housekeeping or jobs about the house, administrative activities) that were usually done by the patient him/herself. The burden of the caregiving function was assessed by a 5-item measure, which was adapted from a questionnaire described by Wijker *et al.* [33]. Respondents were asked how often they felt burdened by the patients' disease and treatment (e.g., patients' need for assistance, limited time for own usual activities).

PATIENT-PROXY RELATIONSHIP. Variables indicative of the nature of the patient-proxy relationship were the type of relationship (spouse/partner, family member, or friend), the living situation (i.e., in the same household as patient or not), and the frequency of contact with the patient. To assess the quality of the patient-proxy relationship, both the patients and their significant others completed a 4-item measure adapted from the Norbeck Social Support Questionnaire (e.g., "how much does this person make you feel liked or loved") [34]. Responses from the patients and significant others were aggregated, yielding a single 8-item measure of the quality of their relationship. To measure the quality of the patient-proxy communication, the patients completed six statements about the ability to communicate with their significant other about their disease and treatment (e.g., "I sometimes feel that he/she avoids talking about my disease and treatment"), which were adapted from a subscale of the Cancer Rehabilitation Evaluation System (CARES) [35] and from a measure developed by Van den Borne and Pruyn [36].

Sum scores were calculated for all multi-item measures of patient and proxy characteristics (for summary, see Table 2). For ease of presentation, all multi-item scales were linearly transformed to a 0 to 100 scale.

Procedures

The questionnaires were completed at a baseline assessment (during the second or third cycle of chemotherapy) and at a follow-up assessment 3 months later. Both administrations were planned to take place during patients' scheduled visits to the outpatient clinic or during clinical ward stays. Data were collected by self-administration, with a research assistant being present to check for missing data. In most cases, the significant other completed the questionnaire while accompanying the patient to the outpatient clinic or while visiting the patient at the clinical ward. In these cases, the significant other was asked to fill out the questionnaire in a separate room in the presence of another research assistant. When the significant other could not be approached at the hospital, the questionnaire was given to the patient for the proxy respondent to complete at home. In these cases, the patients and significant others received explicit instructions not to discuss the questions with each other. Self-addressed, stamped envelopes were provided for return of the questionnaires, and telephone reminders were used occasionally to maximize response rate.

Statistical Methods

Mean scores and standard deviations for the patient- and proxy-generated scores on the QLQ-C30 measures were cal-

culated, as well as for the characteristics of the patients and their significant others. The internal consistency of the multi-item scales of the QLQ-C30 and the patient/proxy characteristics was assessed by Cronbach's coefficient alpha [37].

PATIENT-PROXY AGREEMENT. A range of analyses were carried out to examine the response agreement between patients and proxy respondents on the QLQ-C30 measures. First, Pearson correlations (r) and intraclass correlations (ICC) were calculated between the patient and proxy ratings on the corresponding measures. In previous proxy studies [6–8,11,14,17], r has often been used as an indicator of agreement. However, r does not necessarily provide an indication of actual agreement, because it disregards any systematic bias (i.e., when proxy ratings are consistently lower than patient ratings, r can be excellent but agreement can be poor) [38,39]. The intraclass correlation [40] accounts for systematic mean differences, and can be interpreted as a chance-corrected index of agreement. Guidelines for the ICC as a measure of the strength of agreement were labeled as follows: ≤0.40, poor to fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, good agreement; 0.81–1.00, excellent agreement [41].

Second, in line with Hays *et al.* [20], multitrait-multimethod analysis [42] of the patient-proxy correlations (r) between all 15 QLQ-C30 measures was used to evaluate the degree of convergence and discrimination between patient and proxy scores. The average correlation (r) between patient and proxy scores on corresponding measures was calculated, as well as the average off-diagonal correlation (i.e., correlations between patient and proxy scores on different QL domains).

Third, the mean of the absolute patient-proxy differences for each QLQ-C30 measure (i.e., irrespective of the direction of differences) was calculated. This provides an additional indicator of agreement between patient and proxy responses. The mean of directional differences (i.e., accounting for the direction of differences) was also calculated, being indicative of bias in proxy scores relative to those of the patient. When the mean of the directional differences was significantly different from zero, as determined by paired Student's t-tests, this was interpreted as evidence of systematic bias [38,43]. To examine the statistical magnitude of any observed systematic bias, the mean difference score was standardized by relating this score to its standard deviation. Given the similarity to effect size (d) calculations for paired observations [44], a standardized difference of $d = 0.2$ was taken to indicate a small bias, $d = 0.5$ a moderate bias, and $d = 0.8$ a large bias.

ASSOCIATION WITH LEVEL OF QL. To determine whether patient-proxy agreement varied as a function of the patients' level of QL, the pattern of agreement between patient and proxy total QL scores was visually examined by means of a scatter plot. That is, for each patient, the difference between the patient and proxy total QL scores (proxy minus patient score) was plotted against the average for each pair of scores (patient plus proxy score divided by 2) [38,45]. When depicted graphically, using the y axis to show difference scores and the x axis to show average scores, perfect correspondence would be represented by a horizontal line through an ordinate of zero. Any observed differences between the patient and proxy scores as a function of the range of their average scores (e.g., comparable scores at high levels of QL, but diverging scores at low levels of QL) can be taken as evidence of so-called scatter bias [43].

RELATIVE VALIDITY. An important indicator of validity is responsiveness [46,47], referring to the ability of an instrument to detect relevant changes over time. For each source of information, changes on six specific QLQ-C30 measures (i.e., physical, role, emotional, and social functioning, pain, and global quality of life) assessed longitudinally (calculated by subtracting the baseline from the follow-up score) were compared to direct patient perceptions of changes over time in each underlying domain. For the latter direct reports, we used the above mentioned transition questions, asking the patients directly whether their condition on the six specific domains had changed. For purpose of analysis, the seven response categories on these transition items were trichotomized: feeling worse, the same, or better at follow-up. Analysis of variance (ANOVA) was employed to test for statistically significant differences in mean change scores between the three subgroups of patients (i.e., "worse," "same," or "better" groups). To evaluate the responsiveness of the significant others' scores relative to the patients' own scores, relative validity (RV) estimates were calculated, defined as the ratio of the proxy ANOVA F-value to the patient ANOVA F-value [48,49]. The RV estimates indicate, for each domain, and in proportional terms, the ability of significant others to detect changes over time in the patients' quality of life, relative to the patients' self-report. An RV >1 (or RV <1) indicates that the proxy respondents are more (or less) efficient in assessing changes over time than the patients themselves.

FACTORS AFFECTING PATIENT AND PROXY QL SCORES. To determine the influence of patient and proxy characteristics on their respective ratings of patients' QL, we examined the contribution of these characteristics to the patient- and proxy-derived total QL score. Following the method described by Rothman *et al.* [11], hierarchical regression analysis was used to compare the factors predicting the patient and proxy total QL score. Predictor variables were entered into the regression equation in three steps: (step 1) measures of patients' health; (step 2) other patient characteristics; and (step 3) proxy characteristics and measures of the patient-proxy relation. For parsimony, only those predictors were entered into the regression model which exhibited statistically significant correlations ($P < 0.05$) with the total QL score at the bivariate level. The order of entry was

based on the hypothesis that measures of patients' health rather than the other characteristics are the best predictors of patients' QL. Validity of the patient- and proxy-derived QL score is supported if the data confirm this hypothesis. Validity is not supported if other characteristics account for a large proportion of the variance in their respective total QL score after variance due to patients' health has been accounted for.

FACTORS AFFECTING PATIENT-PROXY AGREEMENT. To determine the influence of patient and proxy variables on the differences between their respective QL scores, we regressed the absolute difference (as an indicator of agreement) and the directional difference (as an indicator of bias) between the patient- and proxy-derived total QL score on the range of patient and proxy characteristics. All variables exhibiting statistically significant correlations ($P < 0.05$) with the criterion measure at the bivariate level were entered simultaneously into the regression equation. Overall R-square was used to determine the proportion of variance in patient-proxy differences which could be explained by the patient and proxy variables combined.

RESULTS

Sample Characteristics

Of 378 eligible patients (263 outpatients and 115 inpatients), 320 agreed to participate in the study (85% response rate). Of the 58 nonrespondents, 29 patients chose not to participate in the study due to very poor physical or emotional condition, and the remaining 29 patients reported not being interested or having enough time. At the baseline assessment, proxy-derived QLQ-C30 ratings were obtained for 307 patients (96%). Nine patients did not have or did not want to ask a significant other to take part in the study, and four significant others chose not to participate. The baseline analyses focus on these 307 patients.

Follow-up patient ratings on the QLQ-C30 were available for 232 of the 307 (76%) patients with complete baseline data. The reasons for loss to follow-up were death (37 patients), too great a physical or emotional burden (24 patients), lack of interest or time (9 patients) or logistical problems (5 patients). At the follow-up assessment, significant other QLQ-C30 ratings were obtained for 224 of 232 patients (97%). The average duration between baseline and follow-up was 3.4 months (SD = 0.8 months).

Patients had a range of cancer diagnoses, with advanced breast cancer (33%), gastrointestinal tumors (15%), and lymphomas (15%) being the most prevalent. All patients were treated with chemotherapy on either an outpatient (70%) or inpatient (30%) basis. Additional characteristics of the patients, the proxy respondents, and their relationship are reported in Tables 1 and 2. Those patient and proxy characteristics assessed by means of multi-item scales are summarized in Table 2. All the scales had substantial variation in scores, although the observed scores did not cover the full range of possible scores (except for social desirability). The internal consistency reliability of the patient and proxy multi-item scales ranged between 0.67 and 0.85. This suggests that both the score variability and reliability of the multi-item measures were adequate for meaningful analysis of the impact of these characteristics on the (difference between) self-report and proxy ratings of patients' QL.

Variability and Reliability of QL Scores

Table 3 shows the means and standard deviations of the patient- and proxy-generated QLQ-C30 scores at the baseline assessment. The score distributions of the multi-item scales and the total QL score were roughly symmetrical, except for cognitive functioning and nausea/vomiting. The latter scales, as well as the single-item measures, exhibited a negative skew, with 39% to 84% of the scores observed in the best possible category (i.e., symptom or functional limitation not present). Nonetheless, with the exception of cognitive functioning, the full range of scores was observed for all functioning scales and symptom measures. For the total QL score, both the patient and proxy scores spanned a relatively large segment of the possible range of scores as well (37.3 to 100 and 30.4 to 100 for patients and proxy respondents, respectively).

The internal consistency reliability for the multi-item scales of the QLQ-C30 and the total QL score all surpassed the 0.70 criterion for group-level comparison [21], except for cognitive functioning. At the follow-up assessment (not presented in tabular form), the results were very similar. That is, apart from cognitive functioning, substantial variability and good internal consistency (range, $\alpha = 0.72$–0.88) was observed for both patient- and proxy-generated scores.

Patient-Proxy Agreement

The results pertaining to agreement between patient and proxy scores on the QLQ-C30 are summarized in Table 4. At baseline, Pearson correlations between patient and proxy scores on corresponding QLQ-C30 measures ranged from 0.46 to 0.74. Intraclass correlations were similar or slightly lower (ICC = 0.46–0.73), indicating a moderate to good level of agreement between patient and proxy ratings. Good agreement was noted for the patient and proxy total QL score ($r = 0.73$; ICC = 0.71). The small differences between Pearson and intraclass correlations also suggest that the amount of systematic bias between patient and proxy ratings is limited.

Additionally, to evaluate the degree of convergence and discrimination between patient and proxy scores on specific dimensions, Pearson correlations between patient and proxy scores on all 15 QLQ-C30 measures were calculated (not presented in tabular form). Multitrait-multimethod analysis

TABLE 1. Characteristics of patients and proxy respondents (n = 307)

Characteristic	Patient		Proxy	
	Number	%	Number	%
Sex				
Female	183	60	152	50
Male	124	40	155	50
Age (years)				
Mean ± SD	51.8 ± 13.5		50.7 ± 13.4	
Range	19–80		18–78	
Education[a]				
Primary school	27	9	20	7
Secondary school	159	52	157	52
Advanced training	89	29	99	32
University level	31	10	28	9
Performance status[a,b]				
ECOG 0	50	16		
ECOG 1	172	56		
ECOG 2	65	21		
ECOG 3	18	6		
Weight loss				
None	179	58		
≤10%	95	31		
>10%	33	11		
Relationship to patient				
Spouse/partner			229	75
Child			24	8
Parent			13	4
Other relative			23	7
Friend			18	6
Living situation[a]				
Same household as patient			244	80
Not in same household			61	20
Frequency of contact[a]				
Daily			269	88
<7 days/week			36	12

[a]n varies due to missing data.
[b]Eastern Cooperative Oncology Group performance status score at baseline.

TABLE 2. Descriptive statistics for multi-item scales employed to measure patient and proxy characteristics (n = 307)

	Number of items	Number of response categories	Score distribution		Reliability (α)
			Mean ± SD	Range	
Patient characteristics					
Mental health	5	6	63.6 ± 28.1	16–100	0.82
Social desirability	5	2	33.7 ± 31.6	0–100	0.69
Positive appraisal	6	4	57.9 ± 16.5	17–100	0.67
Social expressiveness	6	4	40.7 ± 19.1	0–94	0.85
Proxy characteristics					
Global health/QL	2	7	75.8 ± 17.7	17–100	0.74
Mental health	5	6	65.9 ± 17.4	16–100	0.81
Intensity of caregiving function	6	5	26.5 ± 22.7	0–96	0.80
Caregiver burden	5	5	24.2 ± 16.3	0–70	0.68
Patient-proxy relationship					
Quality of relationship	8	5	81.2 ± 14.5	34–100	0.83
Quality of communication	6	5	74.9 ± 20.2	4–100	0.79

Note: Higher scores represent better mental health, more social desirability, more positive appraisal, more social expressiveness, better global health/QL, higher intensity of caregiving function, more caregiver burden, better quality of relationship, and better quality of communication.

TABLE 3. Distribution and reliability of patient and proxy baseline EORTC QLQ-C30 scores (n = 307)

	Number of items	Number of response categories	Patient score (Mean ± SD)	Proxy score (Mean ± SD)	Patient score (α)	Proxy score (α)
Functioning scales						
Physical	5	2	63.6 ± 28.1	58.4 ± 28.2	0.72	0.73
Role	2	4	59.8 ± 31.1	55.3 ± 32.4	0.79	0.82
Cognitive	2	4	82.2 ± 21.6	82.5 ± 19.1	0.51	0.48
Emotional	4	4	75.7 ± 20.6	66.0 ± 23.1	0.83	0.86
Social	2	4	75.7 ± 26.5	72.4 ± 28.1	0.74	0.80
Global QL	2	7	62.9 ± 22.1	55.8 ± 23.8	0.85	0.88
Symptom scales/items						
Fatigue	3	4	42.5 ± 25.1	50.0 ± 27.1	0.86	0.86
Nausea/vomiting	2	4	14.1 ± 21.4	17.8 ± 23.1	0.74	0.76
Pain	2	4	22.9 ± 25.3	28.2 ± 27.7	0.82	0.86
Dyspnea	1	4	20.0 ± 25.2	20.1 ± 27.0	—	—
Insomnia	1	4	25.2 ± 29.1	32.0 ± 31.2	—	—
Anorexia	1	4	23.7 ± 32.3	24.7 ± 30.3	—	—
Constipation	1	4	14.5 ± 25.8	15.7 ± 26.8	—	—
Diarrhea	1	4	8.0 ± 18.5	8.7 ± 18.2	—	—
Financial impact	1	4	9.1 ± 21.0	7.8 ± 19.4	—	—
Total QL score[a]	—	—	76.0 ± 14.2	72.3 ± 15.6	0.85	0.87

[a]The total QL score is an aggregated score of the 15 functioning and symptom measures. For calculation of the total QL score, the nine symptom measures were reversed so that a higher total QL score represents a better quality of life.

of all patient-proxy correlations indicated that the average correlation between patient and proxy scores for corresponding domains (average r = 0.58) was substantially higher than that for diverging domains (average r = 0.22).

The means of absolute differences between patient and proxy scores on the QLQ-C30 measures ranged from 8.2 to 19.9 points on the 0–100 scales (Table 4). Accounting for the direction of the difference, the mean difference ranged from −9.7 to 7.5 points. Statistically significant mean differences between patient and proxy scores were noted for nine of the 15 QLQ-C30 measures, and for the total QL score. As compared to the patients themselves, the signifi-

TABLE 4. Patient-proxy agreement on the EORTC QLQ-C30: baseline (n = 307)

	Patient-proxy correlation		Absolute difference[a]	Directional difference[b]	
	r	ICC	(Mean ± SD)	Mean ± SD	d[c]
Functioning scales					
Physical	0.74	0.73	12.9 ± 16.4	−5.2 ± 20.2[d]	−0.26
Role	0.63	0.63	19.7 ± 19.1	−4.5 ± 27.1[d]	−0.17
Cognitive	0.50	0.49	13.6 ± 15.3	0.3 ± 20.5	0.01
Emotional	0.52	0.47	18.0 ± 15.1	−9.7 ± 21.5[d]	−0.45
Social	0.46	0.46	19.8 ± 20.6	−3.3 ± 28.4[d]	−0.12
Global QL	0.56	0.54	17.5 ± 14.4	−7.1 ± 21.5[d]	−0.33
Symptom scales/items					
Fatigue	0.65	0.62	17.3 ± 15.3	7.5 ± 21.9[d]	0.34
Nausea/vomiting	0.63	0.62	11.5 ± 15.8	3.7 ± 19.2[d]	0.19
Pain	0.63	0.61	16.0 ± 17.4	5.3 ± 23.0[d]	0.23
Dyspnea	0.62	0.62	13.1 ± 18.8	0.1 ± 22.9	0.00
Insomnia	0.50	0.48	19.9 ± 23.8	6.8 ± 30.3[d]	0.22
Anorexia	0.67	0.67	14.5 ± 21.0	1.0 ± 25.5	0.04
Constipation	0.59	0.59	12.0 ± 20.5	1.2 ± 23.7	0.05
Diarrhea	0.50	0.50	8.2 ± 16.5	0.7 ± 18.4	0.04
Financial impact	0.47	0.47	9.2 ± 18.6	−1.3 ± 20.7	−0.06
Total QL score	0.73	0.71	8.9 ± 7.4	−3.7 ± 11.0[d]	−0.34

[a]Absolute difference between patient and proxy score (indicator of agreement)
[b]Difference between patient and proxy score (indicator of bias)
[c]Standardized difference d = mean difference/standard deviation of difference (d = 0.2 small, d = 0.5 moderate, d = 0.8 large bias)
[d]Statistically significant difference between patient and proxy mean score ($P < 0.05$).

cant others rated the patients as having lower levels of physical, role, emotional, and social functioning, a more impaired global QL, and a greater degree of fatigue, nausea, and vomiting, pain, and insomnia. The statistical magnitude of this bias, as defined by standardized mean differences, was small to moderate ($d = -0.45$ to 0.34 for the specific QLQ-C30 measures; $d = -0.34$ for the total QL score). The most pronounced systematic bias was observed for patient and proxy ratings of emotional functioning (-9.7 ± 21.5; $d = -0.45$).

At follow-up, results pertaining to patient-proxy agreement were very similar (not presented in tabular form). Again, moderate to good levels of agreement were noted, with ICC ranging from 0.42 to 0.79 ($r = 0.44$–0.79) for the 15 QLQ-C30 measures, and ICC = 0.73 ($r = 0.75$) for the total QL score. Good to excellent convergence and discrimination between the patient and proxy scores was observed, with average $r = 0.61$ between patient and proxy scores for corresponding domains compared to average $r = 0.23$ for diverging domains. As noted at baseline, a small to moderate amount of systematic bias between patient and proxy ratings (in the same direction) was found for several QLQ-C30 measures ($d = -0.46$ to 0.32) and the total QL score ($d = -0.29$), being most substantial for emotional function (-8.8 ± 19.3; $d = -0.46$).

Association with Level of QL

Figures 1A and 1B depict scatter plots of the difference between the patient and proxy total QL score against the average for each pair of scores. These plots show a curvilinear relationship between the patient-proxy differences in total QL scores and the patients' level of QL. Both at baseline and follow-up, the maximum level of disagreement was found at intermediate levels of QL (average total QL scores of approximately 60 to 80), with smaller patient-proxy differences noted for patients with a relatively poor or good QL.

Relative Validity

Table 5 shows longitudinally assessed changes in patient and proxy scores on specific QLQ-C30 measures for three patient subgroups: patients who felt worse, the same, or better at follow-up on the underlying domains (as assessed with the transition questions). Across all domains, patients' direct perceptions of change over time were accompanied by corresponding changes (i.e., in the expected direction) in baseline to follow-up scores on the analogous QLQ-C30 measures. Between-group differences in mean change scores were statistically significant ($P < 0.001$) for both raters across each domain. While both patient and proxy scores were responsive to changes over time in the specific domains, the raters differed slightly in relative performance. Relative to the patients, the proxy respondents were more efficient in detecting changes in physical functioning (RV = 1.58) and role functioning (RV = 1.25), but were less efficient in relation to the remaining domains (RV = 0.55 to 0.91).

Factors Affecting Patient and Proxy QL Scores

Table 6 shows the bivariate correlations, beta weights, and change in R-square due to variables in each step of the hierarchical regression for the patient- and proxy-derived total QL scores separately. For both sources of information, a substantial proportion of variance in total QL scores was explained by variables of patients' health ($R^2 = 0.62$ and 0.52 for patient and proxy total QL scores, respectively), and only a small proportion by the range of other patient and proxy characteristics ($R^2 = 0.03$ and 0.05 for patient and proxy total QL scores, respectively). The only additional patient characteristic contributing to explaining variance in patients' self-reported total QL scores was social expressiveness. Those patients who usually express their feelings and discuss their problems with others, tended to report a more impaired QL. In addition to patients' health and social expressiveness, proxy-derived total QL scores were also associated significantly with the intensity of the caregiving function as reported by the significant others.

At follow-up (not presented in tabular form), the results of the regression analysis were very similar, with a substantial proportion of variance in total QL scores being explained by indicators of patients' health ($R^2 = 0.62$ and 0.54 for patient and proxy total QL scores, respectively), and only a small proportion by the range of other patient and proxy characteristics ($R^2 = 0.04$ and 0.07 for patient and proxy total QL scores, respectively).

Factors Affecting Patient-Proxy Agreement

At the bivariate level, several factors were found to be associated significantly with the absolute difference between patient and proxy total QL scores (being an indicator of agreement). Larger differences between patient and proxy scores were noted for patients with a poorer performance status, more weight loss, and poorer mental health; patients who were older, female, and lower educated; patients with a tendency toward a socially desirable response set; proxy respondents who were older and male; and proxy respondents reporting poorer health/QL for themselves, and a higher intensity of their caregiving function (Table 7). However, in the multivariate analysis, these variables together explained only 14% of the variance in absolute patient-proxy differences, with patients' socially desirable response set and significant others' own health/QL being the only statistically significant predictors of the absolute difference between patient and proxy total QL scores.

The direction of differences between patient and proxy total QL scores (being an indicator of systematic bias) was

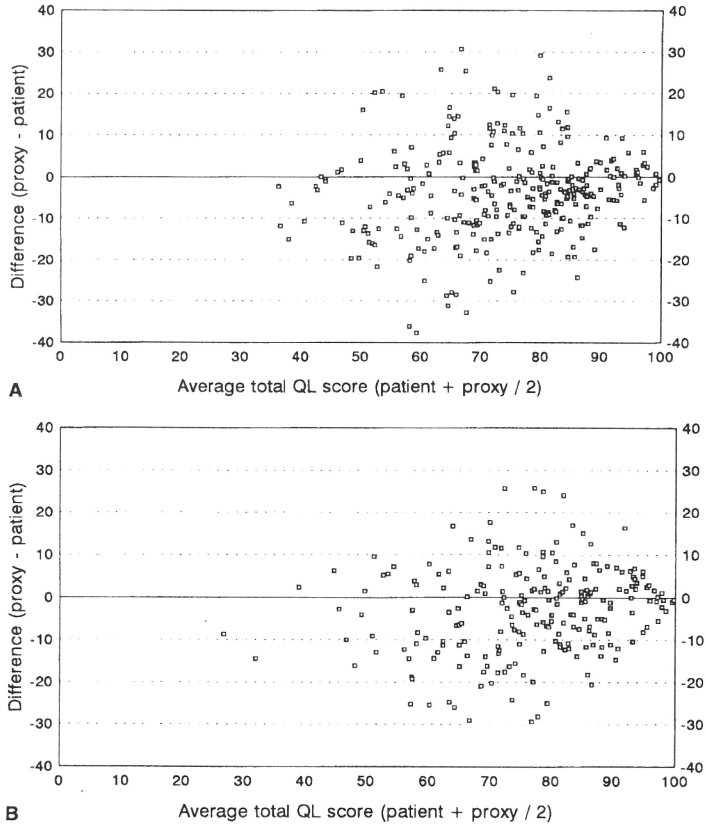


FIGURE 1. Difference between patient and proxy total QL scores plotted against their average at the baseline (A) (n = 307) and follow-up (B) (n = 224) assessment.

associated significantly with a limited number of variables. The observed bias between patient and proxy scores (i.e., significant others rating patients' QL as more impaired than the patients themselves) appeared to be greater for female patients; patients who exhibit a positive coping style and a socially desirable response set; and proxy respondents reporting poorer mental health, and a higher intensity of their caregiving function. In multivariate analysis, these variables explained only 9% of the variance in directional patient-proxy differences, with patients' positive coping style, and proxy respondents' mental health and caregiving experience being the most important predictors of the direction of patient-proxy differences. Similar results were obtained when employing the follow-up data (not presented in tabular form).

DISCUSSION

The aim of this study was to examine the viability of employing patients' significant others as either alternative or complementary sources of information on the health-related quality of life of patients with cancer. First, the quality of the proxy ratings was examined by comparing patient and proxy scores on the various subscales and a composite score of the EORTC QLQ-C30. In keeping with earlier studies of agreement between QL ratings of cancer patients and their significant others on the EORTC questionnaire [16–18], patient-proxy correlations were generally moderate at the individual patient level. Although agreement varied across the subscales, all patient-proxy correlations surpassed the 0.40 criterion below which agreement is interpreted as poor to fair [41]. Relatively high levels of agreement were found for the physical functioning scale and the total QL score. The former is presumably related to the concreteness of the five questions making up the scale, the latter to the psychometric benefits of a composite summary score.

Average absolute differences between patient and proxy scores on the QLQ-C30 scales ranged from approximately 10 to 20 points (on 0–100 scales). These observed differences were smaller than the range observed in a similar

TABLE 5. Changes between baseline and follow-up scores[a] on specific EORTC QLQ-C30 scales as a function of patient-reported changes at follow-up[b]

		Patient-reported change at follow-up				
	n[c]	Worse change (Mean ± SD)	Same no change (Mean ± SD)	Better change (Mean ± SD)	Analysis of variance[e] (F)	Relative validity[f] (RV)
Physical functioning	224	(83)[d]	(60)	(81)		
Patient		−9.4 ± 22.8	−0.7 ± 17.6	5.2 ± 22.8	9.51	1.00
Proxy		−11.7 ± 22.6	0.3 ± 14.5	6.9 ± 25.7	15.04	1.58
Role functioning	223	(78)	(85)	(60)		
Patient		−15.2 ± 27.1	− 2.5 ± 21.3	15.0 ± 30.3	22.78	1.00
Proxy		−13.0 ± 27.6	2.4 ± 25.0	20.0 ± 23.5	28.38	1.25
Emotional functioning	224	(51)	(115)	(58)		
Patient		−8.7 ± 20.1	1.4 ± 14.4	10.3 ± 17.8	17.46	1.00
Proxy		−12.5 ± 21.9	0.0 ± 18.7	7.6 ± 22.3	13.35	0.76
Social functioning	223	(46)	(139)	(38)		
Patient		−18.1 ± 22.5	4.4 ± 21.9	17.5 ± 32.4	24.50	1.00
Proxy		−15.9 ± 25.8	1.0 ± 27.4	15.8 ± 32.6	13.53	0.55
Pain	223	(61)	(120)	(42)		
Patient		20.8 ± 26.5	−0.4 ± 19.4	−9.9 ± 28.5	25.27	1.00
Proxy		22.7 ± 28.9	−1.1 ± 20.8	− 6.7 ± 30.8	22.90	0.91
Quality of life	221	(52)	(102)	(67)		
Patient		−15.7 ± 20.1	−2.1 ± 15.5	11.6 ± 18.4	35.28	1.00
Proxy		−16.8 ± 22.1	−0.7 ± 17.3	10.3 ± 20.9	28.09	0.80

[a]Follow-up minus baseline score.
[b]Patients' perception of change over time in each domain during the study period.
[c]*n* varies due to missing data, but was held constant across raters for each domain.
[d]Number of patients reporting worse, same, or better health in each domain.
[e]ANOVA F-statistics for between-group differences in measured changes between baseline and follow-up scores.
[f]RV estimates represent the ratio of the proxy F-values to the patient F-values; for patients, RV is set to 1; an RV >1 (or RV <1) indicates that the proxies are more (or less) efficient in assessing changes over time than the patients themselves.

analysis of an epilepsy-specific QL questionnaire [20], which was based in part on the MOS 36-item Short Form Health Survey (SF-36) [50]. Nevertheless, the magnitude of these average absolute differences, as well as of the patient-proxy correlations, indicate that significant others, sometimes provide different information on the QLQ-C30 than the patients themselves. In part, this may be due to less than perfect reliability of both the patient and proxy scores, which should be taken as a frame of reference when interpreting the magnitude of the observed patient-proxy differences and correlations [39,45]. Given that the reliability estimates of both patient and proxy scores fail to meet the 0.90 criterion recommended for interpretation of scores at the individual level [21], it is not realistic to expect very high levels of patient-proxy agreement.

Importantly, irrespective of the specific QLQ-C30 scale, the correlations between patient and proxy scores for corresponding domains were generally higher than that for diverging domains. This suggests that, although they sometimes provide different information than the patients themselves, proxy respondents are capable of making clear distinctions between various aspects of cancer patients' quality of life. Again, these results were more encouraging than those observed by Hays *et al.* [20] for their epilepsy-specific QL instrument.

At the group level, in keeping with expectations on the basis of previous studies [7–9,11,13,14,18,19], the significant others rated the patients as having a more impaired QL than the patients themselves. This systematic bias, however, was generally small in magnitude. The most pronounced bias was noted for patient and proxy ratings of emotional functioning (i.e., mean difference of −9.7 and −8.8 points at baseline and follow-up, respectively). In the present study, such a difference of almost 10 points is equivalent to the mean change in emotional functioning for those patients who reported a better or worse emotional condition at follow-up as compared to the baseline assessment. Thus, for this scale, the bias between patient and proxy ratings reflects a difference that is meaningful to the patients. Overall, however, the results suggest that there is only a modest degree of systematic bias in QL ratings provided by significant others versus cancer patients themselves.

It is important to note that the observed levels of patient-proxy agreement pertain, by definition, only to those patients who are able to complete a questionnaire themselves. In the current study, about 15% of eligible patients chose not to participate or were lost to follow-up due to very poor physical or emotional condition. To facilitate extrapolation of our findings to this more impaired group of patients (i.e., those for whom proxy respondents are most needed), we examined trends in patient-proxy agreement as a function

TABLE 6. Bivariate and multivariate correlates of patient and proxy baseline total QL scores (n = 307)

	Patient-derived			Proxy-derived		
	Bivariate correlation (r)	**Beta coefficient (β)**	**R-square increment (R^2)**	**Bivariate correlation (r)**	**Beta coefficient (β)**	**R-square increment (R^2)**
Step 1						
Patient characteristics			0.62			0.52
Performance status	−0.71[a]	−0.56[a]		−0.68[a]	−0.47[a]	
Weight loss	−0.31[a]	−0.09[c]		−0.32[a]	−0.11[b]	
Mental health	0.53[a]	0.31[a]		0.42[a]	0.19[a]	
Step 2						
Patient characteristics			0.02			0.01
Age	−0.06			−0.12[c]	−0.04	
Sex	−0.18[b]	0.07		−0.26[a]	−0.05	
Education	−.03			0.03		
Social desirability	0.12[c]	0.06		0.00		
Positive appraisal	0.15[b]	−0.01		0.02		
Social expressiveness	−0.16[b]	−0.11[b]		−0.16[b]	−0.11[b]	
Step 3						
Proxy characteristics			0.01			0.04
Age	−0.07			−0.10		
Sex	0.17[b]	−0.01		0.20[a]	−0.04	
Education	−0.09			−0.11		
Global health/QL	0.12[c]	0.01		0.15[b]	−0.01	
Mental health	0.13[c]	−0.07		0.22[a]	0.07	
Intensity of caregiving	−0.40[a]	−0.06		−0.48[a]	−0.16[c]	
Caregiver burden	−0.30[a]	−0.06		−0.28[a]	−0.01	
Patient-proxy relationship						
Relationship to partner	−0.04			0.02		
Living situation	−0.01			0.05		
Frequency of contact	−0.03			0.02		
Quality of relationship	0.03			0.03		
Quality of communication	0.23[a]	0.07		0.21[a]	0.08	
Overall R^2			0.65			0.57

Note: Hierarchical regression analysis was used, with predictors entered only when being statistically significant at the bivariate level.
[a] $P < 0.001$.
[b] $P < 0.01$.
[c] $P < 0.05$.

of patients' level of QL. Both at baseline and follow-up, the magnitude of the differences between patient and proxy scores was found to increase with the level of QL impairment. This is in line with the results of a study of stroke survivors which employed patient and proxy ratings on the Sickness Impact Profile [13]. However, our current results suggest that the magnitude of patient-proxy differences may decrease again at very low levels of QL. This finding is in line with our intuitive impression that the patients and their significant others are most likely to concur when the patients' QL is either very good or very poor. For the questionnaire employed in the present study, the answers to many questions for such patients will be evident (i.e., either "not at all" or "very much" for the symptom or functional limitation). Patient and proxy ratings will most likely diverge for patients with intermediate levels of QL. Given these considerations, we would suggest that there may be a U-shaped relationship between the degree of patient-proxy agreement and the patients' level of QL. This implies that the accuracy of proxy ratings may be better for patients with a very low level of QL (i.e., those for whom the need for proxy respondents is most salient) than for patients with intermediate levels of QL.

Differences between patient and proxy ratings should not necessarily be interpreted as evidence of the inaccuracy or biased nature of proxy-derived information. As in earlier studies [7,13,18–20], we found similar reliability estimates for patient and proxy ratings. The validity of patient- and proxy-derived information was examined by means of two types of analysis. First, we determined the relative responsiveness to change over time in patients' QL. The results provide support for the responsiveness of the ratings provided by the significant others to changes in several quality of life domains. This compares well with the results of an earlier study, in which the same analysis was employed for patient and proxy scores on the COOP/WONCA charts [19]. For emotional and social functioning, two domains for which moderate levels of patient-proxy agreement were

TABLE 7. Bivariate and multivariate correlates of differences between patient and proxy baseline total QL scores (n = 307)

	Absolute difference		Directional difference	
	Bivariate correlation (r)	Beta coefficient (β)	Bivariate correlation (r)	Beta coefficient (β)
Patient characteristics				
Performance status	0.23[a]	0.12	−0.04	
Weight loss	0.15[b]	0.09	−0.05	
Mental health	−0.13[c]	−0.07	−0.10	
Age	0.16[b]	0.05	−0.09	
Sex	0.16[b]	−0.02	−0.13[c]	−0.07
Education	−0.17[b]	−0.09	0.08	
Social desirability	0.16[b]	0.12[c]	−0.15[c]	−0.11
Positive appraisal	0.04		−0.16[b]	−0.16[b]
Social expressiveness	−0.01		−0.02	
Proxy characteristics				
Age	0.17[b]	0.03	−0.05	
Sex	−0.19[b]	−0.13	0.07	
Education	0.02		−0.05	
Global health/QL	−0.18[b]	−0.12[c]	0.07	
Mental health	−0.10		0.15[c]	0.12[c]
Intensity of caregiving	0.14[c]	0.00	−0.16[b]	−0.13[c]
Caregiver burden	0.06		−0.02	
Patient-proxy relationship				
Relationship to partner	0.03		0.08	
Living situation	−0.03		0.09	
Frequency of contact	0.00		0.07	
Quality of relationship	−0.05		0.01	
Quality of communication	−0.05		0.00	
Overall R^2	0.14		0.09	

Note: Simultaneous regression analysis was used, with predictors entered only when being statistically significant at the bivariate level.

[a] $P < 0.001$.

[b] $P < 0.01$.

[c] $P < 0.05$.

found, the proxy respondents were slightly less efficient than the patients in detecting changes over time. Nevertheless, they were clearly able to distinguish between those patients who reported an improvement, status quo, or deterioration in their emotional and social condition. Thus, it is not appropriate to interpret moderate patient-proxy agreement as evidence of a lack of validity of the proxy ratings.

Second, the validity of patient and proxy QL ratings was examined by determining the extent to which their respective scores could be predicted by patients' health rather than by specific characteristics of the patients and their significant others. As hypothesized, both at baseline and follow-up, patients' health accounted for a substantial proportion of variance in both the patient and proxy QL scores. Both the patient and proxy QL scores were relatively free of the influence of other variables. After variance due to patients' health had been taken into account, both patient and proxy scores were lower (i.e., reflecting a more impaired QL) for those patients who usually express their feelings and discuss their problems with others. Moreover, in line with Rothman *et al.* [11], QL scores provided by the proxy respondents were lower when they experienced a higher intensity of their caregiving function. However, given the marginal influence of this variable, our results suggest that significant others' own distress does not exert a strong influence on their ratings of patients' QL.

Relatedly, none of the patient and proxy characteristics assessed in this study accounted for a substantial amount of variance in the differences between patient and proxy QL scores. At the bivariate level, several statistically significant but weak associations were observed. However, together, these variables explained less than 15% of the variance in both absolute patient-proxy differences (as a measure of agreement) and directional differences (as a measure of bias). This may be the reason why, in earlier studies, several proxy characteristics have been shown to exert an influence on the extent of patient-proxy agreement, but not in a particularly consistent way [3,4]. Minimally, there is insufficient evidence to prefer one type of proxy respondent over the other, as has been suggested by others [6,8]. The relatively low proportion of explained variance might be taken to mean that we have overlooked specific variables that

could predict the differences between patient and proxy scores, or that specific factors were not adequately measured. However, it is important to note that most patient-proxy discrepancies represent differences of only one response category (e.g., "quite a bit" versus "very much"). In our opinion, it is more reasonable to assume that discrepancies between patient and proxy ratings are primarily due to random error in both sources of information, which can result from lack of reliability or concreteness of the QL scales/items, differences in interpretations of the response categories, lack of precision or attention when completing particular questions, and situational factors (e.g., transient ideas, moodiness).

In conclusion, the present findings lend support to the use of significant others as proxy respondents of the quality of life of patients with cancer. While the ratings obtained from significant others will not always be identical to those provided by the patients themselves, the bias introduced by the use of proxy respondents is generally of a modest magnitude. One might question the extent to which the current findings are generalizable to other patient populations. Given the findings from earlier studies [6–20], proxy ratings of patients' QL seem to be useful for several other patient groups with chronic diseases as well. However, since patient-proxy differences were most pronounced for ratings of emotional function, the use of proxy respondents may be less suitable for conditions with mainly emotional symptoms.

Proxy respondents can be useful in both cross-sectional and longitudinal QL studies. In cross-sectional studies (e.g., [13]), proxy ratings can facilitate the inclusion of a more representative group of patients in the QL evaluation. In this situation, it is advisable to obtain both patient and proxy ratings for at least a substantial proportion of the patients, so that the potential impact of using proxy respondents can be critically assessed. In longitudinal studies, proxy ratings can be used to prevent patients' loss to follow-up because of disease progression or severe symptomatology. For studies among patient populations at risk of deteriorating self-report capabilities, it is advisable to obtain proxy ratings of patients' QL throughout the entire course of the study. Assuming that proxy ratings are obtained from the beginning of a study, we would commend against converting from self-report to proxy-report during the course of the study, but rather to rely entirely on the proxy ratings when the patient drops out. When proxy ratings are not obtained from the beginning, there is no other choice but to intermingle QL ratings of the patients and proxy respondents. In this situation, possible changes in patients' QL might be blurred or exaggerated, which should clearly be acknowledged in the interpretation of study results and conclusions.

The authors are grateful to the medical and nursing staff of the Department of Internal Medicine of the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, the patients, and their significant others for their cooperation in this study. The study was supported by grants no. NKI 93-139 and NKI 90-A from the Dutch Cancer Society.

References

1. Aaronson NK. Methodologic issues in assessing the quality of life of cancer patients. **Cancer** 1991; 67: 844–850.
2. De Haan R, Aaronson NK, Limburg M, *et al.* Measuring quality of life in stroke. **Stroke** 1993; 24: 320–327.
3. Magaziner J. The use of proxy respondents in health studies of the aged. In: Wallace RB, Woolson RF, Eds. **The Epidemiologic Study of the Elderly**. New York: Oxford University Press; 1992: 120–129.
4. Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. **J Clin Epidemiol** 1992; 45: 743–760.
5. Zimmerman SI, Magaziner J. Methodological issues in measuring the functional status of cognitively impaired nursing home residents: The use of proxies and performance-based measures. **Alzheimer Dis Assoc Disord** 1994; 8(Suppl. 1): S281–S290.
6. Bassett SS, Magaziner J, Hebel JR. Reliability of proxy response on mental health indices for aged, community-dwelling women. **Psychol Aging** 1990; 5: 127–132.
7. Epstein AM, Hall JA, Tognetti J, *et al.* Using proxies to evaluate quality of life. Can they provide valid information about patients' health status and satisfaction with medical care? **Med Care** 1989; 27: S91–S98.
8. Magaziner J, Simonsick EM, Kashner TM, Hebel JR. Patient-proxy response comparability on measures of patient health and functional status. **J Clin Epidemiol** 1988; 41: 1065–1074.
9. Magaziner J, Bassett SS, Hebel JR, Gruber-Baldini A. Use of proxies to measure health and functional status in epidemiologic studies of community-dwelling women aged 65 years and older. **Am J Epidemiol** 1996; 143: 283–292.
10. McCusker J, Stoddard AM. Use of a surrogate for the Sickness Impact Profile. **Med Care** 1984; 22: 789–795.
11. Rothman ML, Hedrick SC, Bulcroft KA, *et al.* The validity of proxy-generated scores as measures of patient health status. **Med Care** 1991; 29: 115–124.
12. Segal ME, Schall RR. Determining functional/health status and its relation to disability in stroke survivors. **Stroke** 1994; 25: 2391–2397.
13. Sneeuw KCA, Aaronson NK, De Haan RJ, Limburg M. Assessing quality of life after stroke: The value and limitations of proxy ratings. **Stroke** 1997; 28: 1541–1549.
14. Clipp EC, George LK. Patients with cancer and their spouse caregivers. **Cancer** 1992; 69: 1074–1079.
15. Farrow DC, Samet JM. Comparability of information provided by elderly cancer patients and surrogates regarding health and functional status, social network, and life events. **Epidemiology** 1990; 1: 370–376.
16. Blazeby JM, Williams MH, Alderson D, Farndon JR. Observer variation in assessment of quality of life in patients with oesophageal cancer. **Br J Surg** 1995; 82: 1200–1203.
17. Sigurdardottir V, Brandberg Y, Sullivan M. Criterion-based validation of the EORTC QLQ-C36 in advanced melanoma: The CIPS questionnaire and proxy raters. **Qual Life Res** 1996; 5: 375–386.
18. Sneeuw KCA, Aaronson NK, Osoba D, *et al.* The use of significant others as proxy raters of the quality of life of patients with brain cancer. **Med Care** 1997; 35: 490–506.
19. Sneeuw KCA, Aaronson NK, Sprangers MAG, *et al.* Value of caregiver ratings in evaluating the quality of life of patients with cancer. **J Clin Oncol** 1997; 15: 1206–1217.

20. Hays RD, Vickrey BG, Hermann BP, *et al.* Agreement between self reports and proxy reports of quality of life in epilepsy patients. **Qual Life Res** 1995; 4: 159–168.
21. Nunnally JC, Bernstein IH. **Psychometric Theory, 3rd Edition**. New York: McGraw-Hill; 1994.
22. 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** 1993; 85: 365–376.
23. Osoba D, Aaronson N, Zee B, *et al.* Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. **Qual Life Res** 1997; 6: 103–108.
24. Aaronson NK, Cull A, Kaasa S, Sprangers MAG. The European Organization for Research and Treatment of Cancer (EORTC) modular approach to quality of life assessment in oncology: An update. In: Spilker B, Ed. **Quality of Life and Pharmacoeconomics in Clinical Trials**. Philadelphia: Lippincott-Raven Publishers; 1996: 179–189.
25. Hjermstad MJ, Fossa SD, Bjordal K, Kaasa S. Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire. **J Clin Oncol** 1995; 13: 1249–1254.
26. Osoba D, Rodrigues G, Myles J, Pater J. Significance of changes in health-related quality of life (QOL) scores in women receiving chemotherapy for recurrent or metastatic breast cancer. **Qual Life Res** 1995; 4: 468–469.
27. Zubrod CG, Schneiderman M, Frei E, *et al.* Appraisal of methods for the study of chemotherapy of cancer in man: Comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. **J Chronic Dis** 1960; 11: 7–33.
28. Conill C, Verger E, Salamero M. Performance status assessment in cancer patients. **Cancer** 1990; 65: 1864–1866.
29. Berwick DM, Murphy JM, Goldman PA, *et al.* Performance of a five-item mental health screening test. **Med Care** 1991; 29: 169–176.
30. Hays R, Stewart A, Hayashi T. A five-item measure of socially desirable response set. **Educ Psychol Meas** 1989; 49: 629–636.
31. Crowne DP, Marlowe D. A new scale of social desirability independent of psychopathology. **J Consult Psychol** 1960; 24: 349–354.
32. Schreurs PJG, Van de Willige G, Brosschot JF, *et al.* **De Utrechtse Coping Lijst (UCL): handleiding** [The Utrecht Coping List (UCL): manual]. Lisse: Swets-Zeitlinger; 1993.
33. Wijker W, Monincx W, Birnie E, et al. Comparison of two forms of gynaecological care: At home and in the hospital. **Qual Life Newsletter** 1993; 7: 4.
34. Norbeck JS, Lindsey AM, Carrieri VL. The development of an instrument to measure social support. **Nurs Res** 1981; 30: 264–269.
35. Ganz PA, Schag CAC, Lee JJ, Sim MS. The CARES: A generic measure of health-related quality of life for patients with cancer. **Qual Life Res** 1992; 1: 19–29.
36. Van den Borne HW, Pruyn JFA. **Achtergronden en betekenis van lotgenotencontact bij kankerpatienten** [Background and meaning of support groups for cancer patients]. Tilburg: IVA; 1983.
37. Cronbach LJ. Coefficient alpha and the internal structure of tests. **Psychometrika** 1951; 16: 297–334.
38. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. **Comput Biol Med** 1989; 19: 61–70.
39. Nelson LM, Longstreth WT, Koepsell TD, Van Belle G. Proxy respondents in epidemiologic research. **Epidemiol Rev** 1990; 12: 71–86.
40. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. **Psychol Rep** 1966; 19: 3–11.
41. Landis JR, Koch GG. The measurement of observer agreement for categorical data. **Biometrics** 1977; 33: 159–174.
42. Campbell DT, Fiske DW. Convergent and discriminant validity by the multitrait-multimethod matrix. **Psychol Bull** 1959; 56: 81–105.
43. Marshall GN, Hays RD, Nicholas R. Evaluating agreement between clinical assessment methods. **Int J Methods Psychiat Res** 1994; 4: 249–257.
44. Cohen J. **Statistical Power Analysis for the Behavioral Sciences**. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
45. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. **Lancet** 1986; 8476: 307–310.
46. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. **Control Clin Trials** 1991; 12: 142S–158S.
47. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. **Qual Life Res** 1993; 2: 441–449.
48. Ware JE, Kosinski M, Bayliss MS, *et al.* Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the Medical Outcomes Study. **Med Care** 1995; 33: AS264–AS279.
49. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. **Arthritis Rheum** 1985; 28: 542–547.
50. Ware JE, Sherbourne CD. A 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. **Med Care** 1992; 30: 473–483.

Chapter 5

The use of significant others as proxy raters of the quality of life of patients with brain cancer

K.C.A. Sneeuw, N.K. Aaronson, D. Osoba, M.J. Muller,
M.A. Hsu, W.K.A. Yung, M. Brada, E.S. Newlands

Medical Care 1997, 35: 490-506

The Use of Significant Others as Proxy Raters of the Quality of Life of Patients with Brain Cancer

KOMMER C.A. SNEEUW, MSC,* NEIL K. AARONSON, PHD,* DAVID OSOBA, MD,† MARTIN J. MULLER, MSC,* MING-ANN HSU, MPH,‡ W.K. ALFRED YUNG, MD,§ MICHAEL BRADA, MRCP FRCR,¶ AND EDWARD S. NEWLANDS, PHD FRCP||

OBJECTIVES. The use of self-report questionnaires for the assessment of health-related quality of life (HRQOL) is increasingly common in clinical research. This method of data collection may be less suitable for patient groups who suffer from cognitive impairment, however, such as patients with brain cancer. In such cases, one can consider employing the patients' significant others as proxy raters of the patients' health-related quality of life. The authors examined the response agreement between patients with brain cancer and their significant others on a health-related quality of life instrument commonly used in cancer clinical trials, the EORTC QLQ-C30, and on a brain cancer-specific questionnaire module, the QLQ-BCM.

METHODS. The study sample consisted of 103 pairs of patients, with either recently diagnosed or recurrent brain cancer, and their significant others (75% spouses, 22% relatives, and 3% friends). Patients and proxies independently completed the EORTC QLQ-C30 and the QLQ-BCM at three different times.

RESULTS. Approximately 60% of the patient and proxy scores were in exact agreement, with more than 90% of scores being within one response category of each other. For most HRQOL dimensions assessed, moderate to good agreement was found. Statistically significant differences in mean scores were noted for several dimensions, with proxies tending to rate the patients as having a lower quality of life than the patients themselves. With the exception of fatigue ratings, this response bias was of a limited magnitude. Less agreement and a more pronounced response bias was observed for the more impaired patients, and particularly for patients exhibiting mental confusion. This finding was confirmed by longitudinal analyses, which indicated lower levels of patient–proxy agreement at follow-up for those patients whose physical or neurologic condition had deteriorated over time.

CONCLUSIONS. In general, patients and their significant others provide similar ratings of the patients' quality of life. Lower levels of agreement and more biased ratings can be expected among those patients for whom the need for proxies is most salient. It is argued, however, that discrepancies between patient–proxy ratings should not be interpreted, a priori, as evidence of the inaccuracy or biased nature of proxy-generated data. Future studies are needed to examine the relative validity and reliability of patient- versus proxy-generated health-related quality of life scores.

Key words: quality of life; proxy ratings; brain cancer. (Med Care 1997;35:490–506)

Although many conceptual and methodological issues surround the measurement of health-related quality of life (HRQOL), it is widely accepted that the patient should be the primary source of information regarding his or her quality of life.[1–3] Therefore, most HRQOL questionnaires today depend primarily on patient self-report. This method of data collection, however, may be less suitable for patients with severe physical disability or for patients who suffer from cognitive dysfunction (eg, stroke survivors, patients with primary brain tumors or brain metastases). In these populations, substantial numbers of patients may experience difficulty in completing a questionnaire or may be completely unable or unwilling to do so, thereby creating the problem of excluding a subgroup of patients from HRQOL assessment for whom it is highly relevant. For instance, in several stroke outcome studies, more than one quarter of the patients were excluded from HRQOL assessments because of serious cognitive, speech, and language disorders.[4–6] This problem is compounded in longitudinal studies, where patients may be lost to follow-up because of rapid disease progression and increased symptom levels. Study results can be seriously compromised and misleading if patients who are suffering from serious physical or cognitive deficits are excluded from the analyses.[1]

One possible approach to circumventing this methodological difficulty may be the use of secondary, or proxy, sources of information.[7,8] For those patients unable or unwilling to provide HRQOL information themselves, health care providers or significant others (eg, spouses, relatives, or friends) might be employed as alternative sources of information. To date, significant others have most frequently been used as proxy raters of the patients' HRQOL in health studies of the aged, and, more recently, in HRQOL studies of stroke survivors.[9,10] In these studies, the use of proxies as substitutes for incapacitated patients significantly increased the sample size and improved the representativeness of the patient population studied.

Reliance on significant others as alternative sources of information on the patients' HRQOL can only be justified, however, if one can demonstrate that the quality of such information is high. Evaluation of the quality of proxy-generated data typically involves a comparison of patient and proxy responses. Such studies have been performed in several research areas examining proxies' accuracy in rating factual information about the subjects, such as health care utilization and the presence of disease risk factors, and in rating patients' preferences in hypothetical life-sustaining treatment decisions.[11–13] When focusing on research in the area of functional status and quality of life, most studies on response comparability have been performed among the elderly.[14–19] Few studies comparing patient and significant other HRQOL ratings have been conducted among cancer patients.[20–22] To date, the literature yields few unequivocal findings. Despite the diversity of methodologies employed, two relatively consistent findings across studies are that the accuracy of proxy ratings is higher when the information sought is concrete and observable, and that

*From the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands.

†From the British Columbia Cancer Agency and University of British Columbia, Vancouver, British Columbia, Canada.

‡From Integrated Therapeutics, Inc. (subsidiary of Schering-Plough), Kenilworth, New Jersey.

§From the MD Anderson Cancer Center, Houston, Texas.

¶From the Royal Marsden Hospital, London, United Kingdom.

‖From Charing Cross Hospital, London, United Kingdom.

This study was sponsored by Integrated Therapeutics, Inc., a subsidiary of Schering-Plough.

Address correspondence to: Neil K. Aaronson, PhD, Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam.

proxies tend to underestimate the patients' quality of life.[8]

By definition, studies comparing patient and proxy ratings require that both patients and proxies complete the same questionnaires. Ironically, patient–proxy comparability cannot always be examined for those patients for whom the need for proxy ratings is most salient (ie, the cognitively and physically impaired who are incapable of completing questionnaires, either in written or verbal form). The resulting problem of generalizability may be addressed, in part, by employing a patient sample with a wide range of disease severities.[8] Use of a heterogeneous patient sample facilitates the examination of trends in the rates of patient–proxy agreement as a function of the patients' health status. At present, however, very little is known about the relationship between patient–proxy agreement and the physical and neurologic condition of the patients.

The purpose of the present study was to examine the level of agreement between HRQOL ratings provided by patients with brain cancer and their significant others. In this study, we attempted to address various methodological shortcomings observed in some of the earlier research in this area.[8] This included employing a sufficiently large study sample to facilitate necessary statistical analyses, inclusion of knowledgeable informants identified by the patients themselves, and the use of an existing HRQOL instrument with known psychometric properties. The reliability of the instrument, based on both patient-and proxy-generated data, was examined to provide a frame of reference for evaluating the patient–proxy agreement. That is, high levels of agreement between patient and proxy responses cannot reasonably be expected when either patients or proxies provide ratings with insufficient reliability.[12,23] We also examined whether patient–proxy agreement was associated with the patients' level of physical and neurologic impairment, as well as with changes in the degree of impairment over time. Finally, we examined the influence of several background characteristics of the patients and significant others on the degree of patient–proxy agreement.

Methods

Subjects

A consecutive series of 109 eligible patients was recruited between July 1993 and July 1994 from three participating centers: the MD Anderson Cancer Center in the USA (n = 61), and the Royal Marsden and Charing Cross Hospitals in the UK (n = 48). Patients who had either newly diagnosed, histologically documented, high grade glioma at least 2 weeks after surgery, or recurrent glioblastoma, documented radiologically, were eligible for recruitment. Patients had to be on a stable dose of steroid maintenance for at least 1 week and were allowed to enter the study any time during their chemotherapy or radiotherapy. Further inclusion criteria included an age of at least 18 years, life expectancy greater than 3 months, and the ability to provide informed consent and to complete an HRQOL questionnaire. Patients were excluded who lacked basic proficiency in English, were being treated for a psychiatric illness, or were receiving focal radiosurgery or brachytherapy.

Of the 109 eligible patients, 105 (96%) agreed to participate in the study. Participating patients were asked to identify a significant other (subsequently referred to as the proxy) with whom they had a close relationship. The proxy had to have known the patient for at least 1 year and had to live with or see the patient at least once a week. Proxies were excluded who were less than 18 years of age, lacked basic proficiency in English, or were not willing to complete the questionnaire. Two patients were not able to identify a significant other, leaving 103 patient–proxy pairs for the current analysis.

Measures and Procedures

Quality of life was assessed by means of the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30, a questionnaire that has been developed specifically for use in cancer clinical research.[24] The QLQ-C30 is composed of 30 items, organized into a number of multi-item scales and single items that reflect a range of physical, emotional, and social health issues relevant to a broad spectrum of cancer patients. It incorporates five functioning scales (physical, role, cognitive, emotional, and social functioning), three symptom scales (fatigue, nausea and vomiting, and pain), six single items (dyspnea, insomnia, anorexia, constipation, diarrhea, and financial impact), and a scale assessing global quality of life. Patients are asked to indicate the extent to which they have experienced specific functional limitations and symptoms over the past week. The questionnaire employs a mix of dichotomous response categories ("yes/no"), four-point Likert-type response scales (ranging from "not at all" to "very much"), and seven-point response scales (numbered visual analogue scales). The QLQ-C30 has been shown to have adequate reliability and validity in a wide range of patient populations and treatment settings. Across a number of studies, internal consistency estimates (Cronbach's coefficient α) of the multi-item scales met or approached the 0.70 reliability criterion recommended for group comparisons.[25] In a recent study, test–retest reliability coefficients ranged between 0.80 and 0.90 for most multi-item scales and single items.[26] Tests of validity have shown the QLQ-C30 to be sensitive to meaningful between-group differences (eg, local versus metastatic disease, active treatment versus follow-up) and changes in clinical status over time.[24,25]

The QLQ-C30 is designed to be supplemented by additional questionnaire modules for use among patients with specific problems related to a given tumor site or medical treatment.[27] In the current study, we included a recently developed brain cancer module (subsequently referred to as the QLQ-BCM).[28] This questionnaire module consists of 20 items, organized into four multi-item scales (future uncertainty, visual disorder, motor dysfunction, and communication deficit), and seven single items (headaches, seizures, drowsiness, hair loss, itching skin, weakness of both legs, and incontinence). All items employ a four-category response option (ranging from "not at all" to "very much"), and a 1-week time frame. As described in detail elsewhere, the QLQ-BCM has exhibited good internal consistency (Cronbach's α, 0.70 to 0.87), known-groups validity (eg, recently diagnosed versus recurrent brain cancer), and responsiveness to changes in functional and neurologic status over time.[28]

Following the scoring procedures recommended by the EORTC, all scale and single item scores of the QLQ-C30 and QLQ-BCM were linearly transformed to a 0 to 100 scale.[24,28] For the functioning scales, higher scores represent a better level of functioning; for the symptom measures, a higher score corresponds to a higher level of symptomatology.

Patients and proxies were asked to complete the QLQ-C30 and QLQ-BCM on three occasions: at the time of an initial clinic visit (baseline), 1 week later (retest), and at the time of a clinic visit at least 4 weeks after the baseline assessment (follow-up). The first and third questionnaire administrations took place while the patient waited to see the physician. The patient and proxy filled out the questionnaires separately and without discussion. If the proxy did not accompany the patient to the clinic, copies of the questionnaires were given to the patient for the proxy to complete at home. At baseline, patients and proxies were given an additional set of questionnaires, with the instruction to complete it at home 1 week later (retest). Both the patients and

proxies received explicit instructions to complete the questionnaires without conferring with each other. The proxy questionnaires were identical to those of the patients, except for minor rephrasing so that each item referred to the patient. The proxies were asked to answer the questions as they thought the patient would. Proxies were also asked to provide information on their age, gender, and the nature of their relationship with the patient. Patients' sociodemographic and clinical information was obtained from the medical records.

At both the baseline and follow-up clinic visits, the patients' physical and neurologic condition was rated by experienced neurooncologists, with the same physician doing all of the examinations at each study site. Performance status was assessed using the Karnofsky performance status rating, which is a widely used method of quantifying the functional status of cancer patients.[29,30] It is an 11-point rating scale that ranges from normal functioning (100) to dead (0). The neurooncologists routinely performed neurologic examinations and recorded the results in a standardized format. At the baseline clinic visit, a range of specific neurologic parameters was recorded, including motor deficit, mental confusion, dysphasia or aphasia, and dysarthria. At the follow-up clinic visit, the physicians indicated whether there had been a change in the patient's overall neurologic condition in relation to the baseline neurologic exam. Moreover, both at baseline and follow-up, the physicians completed a newly developed 10-item instrument assessing cognitive impairment (eg, difficulty remembering recent events, inappropriate responses to questions and to directions during examination, inappropriate affect), using four-point response options (none/mild/moderate/severe). The development of this short and practical instrument was deemed necessary in that it is often not feasible in clinical practice and in research settings to administer even relatively simple mental function tests, such as the Mini Mental State Exam (MMSE), let alone a lengthy battery of neuropsychological tests. The new instrument can be easily and quickly completed by an experienced neurologist or neurooncologist. In preliminary analyses (unpublished data), described in detail elsewhere, this physician-rated instrument exhibited good reliability (Cronbach's α 0.90) and known-groups validity (eg, dysphasic versus nondysphasic patients).

Statistical Analysis

A range of analyses was carried out to establish the reliability of the responses obtained from both patients and proxies, the level of agreement between patient and proxy responses, the presence and magnitude of systematic differences between patient and proxy ratings, and factors affecting the level of patient–proxy agreement.

Reliability. Test–retest reliability was assessed by calculating correlations between replicate scores (ie, baseline and 1 week later) on the QLQ-C30 and QLQ-BCM scales and single items. For this purpose, the intraclass correlation coefficient was used.[31] This is the preferred statistic for estimating the equivalence of repeated measurements for the same subjects.[32,33] The internal consistency of the multi-item scales, representing the average correlation among items within a scale, was assessed by Cronbach's coefficient alpha.[34,35] Both reliability coefficients vary from 0 to 1, with 1 indicating perfect reliability.

Patient–Proxy Agreement. Response agreement was assessed by calculating the proportion of exact and approximate agreement. Exact agreement, as the term suggests, was defined as those cases where the response category chosen by the patient and the proxy for a given item was identical. Approximate agreement was defined as the proportion agreement within one category in either direction. Proportion agreement for multi-item measures was calculated as the average percentage agreement across the number of items in each scale. To discount

chance agreements between patient and proxy responses, we also calculated the intraclass correlation coefficient between patient and proxy ratings.[36]

Systematic Response Bias. To evaluate the presence of a systematic tendency of proxies to rate the patients as having a higher or lower level of quality of life than did the patients themselves, mean difference scores between patient and proxy ratings were calculated (proxy minus patient score). A mean difference score significantly different from zero, using a paired Student's t test, provided evidence of systematic bias.[36] To examine the statistical magnitude of any observed systematic bias, the mean difference score was standardized by calculating the effect size for paired observations, which relates the mean difference score to the standard deviation of that mean score. Following guidelines provided by Cohen,[37] $d = 0.2$ was taken to indicate a small effect size, $d = 0.5$ was taken to indicate a moderate effect size, and $d = 0.8$ was taken to indicate a large effect size.

Factors Affecting Patient–Proxy Agreement. The number of discrepancies for each patient–proxy pair (ie, nonidentical responses) over all 30 questions of the QLQ-C30 and 20 questions of the QLQ-BCM was tabulated, yielding two variables with a theoretical range of 0 to 30 and 0 to 20, respectively. The association between these two dependent variables and a range of factors indicative of the patients' physical and neurologic condition, as well as the patients' and proxies' background, was examined. Five parameters were chosen to reflect the patients' physical and neurologic condition: (1) disease stage, (2) two parameters of physical function (performance status according to the physicians' Karnofsky rating and motor deficit as recorded by the physician in the standardized neurologic exam), and (3) two parameters of cognitive function (mental confusion as recorded by the physician in the neurologic exam and cognitive impairment according to the physician-assessed 10-item instrument). For the latter variable, the patient sample was divided into three subgroups of increasing cognitive impairment: none, minor (one or two mild impairments), and moderate to severe (at least three mild impairments and/or any moderate/severe impairment). Background characteristics of the patients and significant others included their age, gender, and culture (ie, American or English), and proxies' relationship to the patients and living arrangement in relation to the patients. Between-group differences were tested by means of analysis of variance, with linear trends for variables with more than two levels determined by a test of linearity.[38]

Change in Patient–Proxy Agreement Over Time. Two additional analyses were performed to evaluate the level of patient–proxy agreement as a function of change in physical and neurologic condition over time. First, repeated measures analysis of variance was used to compare the change in number of discrepancies on the QLQ-C30 and QLQ-BCM between patients whose physical or neurologic condition had deteriorated during the study period versus those whose condition remained stable or improved. Deterioration in physical and neurologic condition was defined as a shift of at least 20 points (ie, two levels of the 11-point scale) on the Karnofsky rating or an overall change in neurologic condition, as recorded by the physician at the follow-up neurologic exam. Secondly, for those patients whose condition had deteriorated over time, baseline intraclass correlations between patient and proxy ratings were compared with intraclass correlations at the follow-up examination.

Results

Patient Characteristics

The patient sample consisted of 65 men (63%) and 38 women (37%), ranging in age from 18 to 75 years, with a mean of 44.1 ± 12.8 years. The majority of the patients were

white (94%) and married (82%). The patients had an average of 14.4 ± 3.4 years (range 8–25 years) of education. Forty patients (39%) had recently diagnosed brain cancer and were recruited, on average, 5.6 ± 7.1 months (range 1–42 months) after their diagnosis of cancer. The remaining 63 patients (61%) had recurrent disease and were recruited, on average, 9.8 ± 13.0 months (range 1–84 months) after evidence of their first recurrence. At the time of enrollment into the study, 46% of the patients were under active treatment with chemotherapy and 10% with radiotherapy. The mean Karnofsky rating for the patients was 83.8 ± 14.3, with 27% scoring in the 50 to 70 range. For 72% of the patients, the baseline neurologic exam revealed signs of impairment, with 19% having a combination of four or more neurologic signs of variable severity. The most frequently observed neurologic signs were (any degree of) motor deficit (38%), minor mental confusion (23%), dysphasia or aphasia (24%), and intermittent slurred speech (13%).

Proxy Characteristics

The proxy sample consisted of 31 men (30%) and 72 women (70%). The age of the proxies ranged from 21 to 75 years, with a mean age of 46.7 ± 12.7 years. The majority of the proxies were the patients' spouses (75%). The remaining proxies were parents (13%), children (3%), siblings (6%), or friends (3%). Most proxies (87%) were living in the same household as the patients and had known the patients for, on average, 25.4 ± 12.9 years (range 1–60 years).

Descriptive Statistics and Reliability

Table 1 displays the means and standard deviations for the multi-item and single-item measures of the QLQ-C30 and QLQ-BCM for both patient- and proxy-generated scores at the baseline assessment. The score distributions of the multi-item measures were roughly symmetrical, except for nausea/vomiting, pain, and visual disorder. The latter symptom scales exhibited a negative skew (ie, more patients scoring toward minimal level of symptomatology). With the exception of financial difficulties and drowsiness, the single-item measures were negatively skewed as well, with more than 50% of the scores observed in the lowest category (ie, symptom or problem not present).

The evaluation of the test–retest reliability of the measures was restricted to 87 cases. Six patients and proxies did not return the retest questionnaire, and 10 patients and proxies completed them more than 10 days following the baseline administration. For the remaining 87 cases, the retest administration was completed within 2 to 10 days, with a mean of 7.2 ± 1.2 days. With the exception of the nausea/vomiting and pain scales, the test–retest reliability of the multi-item scales based on patient-generated scores was moderate to good, with intraclass correlations ranging from 0.54 to 0.81 (Table 2). The intraclass correlations for the multi-item measures based on proxy-generated scores were slightly higher (except for nausea/vomiting), ranging from 0.64 to 0.85. In general, lower test–retest reliability, based on either patient or proxy scores, was found for the single-item measures.

The internal consistency reliability for the multi-item scales of the QLQ-C30 ranged from 0.47 to 0.83, and from 0.51 to 0.86 for the patient-generated and proxy-generated scores, respectively. The internal consistency of the four multi-item scales of the QLQ-BCM were all above 0.70 for both versions of the questionnaire (Table 2).

Patient–Proxy Agreement

Exact agreement was greater than 50% for most QLQ-C30 and QLQ-BCM measures (Table 3). The average proportion of exact agreement was 60.8% for the QLQ-C30 and 61.6% for the QLQ-BCM. The level of exact agreement appeared to be primarily a function of the score distribution and the type of re-

TABLE 1. EORTC QLQ-C30 and QLQ-BCM Measures

	Number of Items	Number of Response Categories	Patient (n = 103)[a] (mean ± SD)	Proxy (n = 103)[a] (mean ± SD)
C30 Functioning scales[b]				
Physical	5	2	72.2 ± 30.3	65.8 ± 30.7
Role	2	2	65.7 ± 37.2	53.4 ± 36.7
Cognitive	2	4	71.8 ± 23.8	63.3 ± 27.1
Emotional	4	4	74.8 ± 20.9	71.2 ± 21.4
Social	2	4	66.3 ± 29.0	59.2 ± 29.1
Global quality of life	2	7	64.6 ± 21.2	61.4 ± 24.5
C30 Symptom scales/items[c]				
Fatigue	3	4	35.1 ± 20.5	46.8 ± 25.0
Nausea and vomiting	2	4	7.7 ± 16.4	9.0 ± 16.2
Pain	2	4	14.9 ± 19.9	18.3 ± 22.0
Dyspnea	1	4	12.2 ± 17.5	9.2 ± 15.7
Insomnia	1	4	20.6 ± 25.7	23.5 ± 27.6
Anorexia	1	4	16.0 ± 25.1	17.0 ± 23.3
Constipation	1	4	13.9 ± 24.2	15.5 ± 26.9
Diarrhea	1	4	8.3 ± 18.6	8.3 ± 17.3
Financial impact	1	4	31.0 ± 34.9	36.6 ± 34.3
BCM Symptoms scales/items[c]				
Future uncertainty	4	4	27.0 ± 20.8	29.8 ± 22.6
Visual disorder	3	4	13.1 ± 19.2	11.8 ± 21.6
Motor dysfunction	3	4	21.4 ± 22.7	25.4 ± 26.9
Communication deficit	3	4	22.8 ± 24.8	26.3 ± 26.7
Headaches	1	4	18.2 ± 22.0	18.5 ± 23.4
Seizures	1	4	6.6 ± 18.9	7.3 ± 17.4
Drowsiness	1	4	35.9 ± 26.8	39.2 ± 30.5
Bothered by hair loss	1	4	17.5 ± 30.2	20.2 ± 29.3
Bothered by itching skin	1	4	13.1 ± 23.5	13.7 ± 23.1
Weakness of both legs	1	4	9.6 ± 19.6	11.2 ± 20.2
Trouble controlling bladder	1	4	12.4 ± 21.5	9.5 ± 21.2

SD, standard deviation.
[a]Due to missing data, n varies from 97 to 103.
[b]Scores range from 0 to 100, with a higher score representing a higher level of functioning.
[c]Scores range from 0 to 100, with a higher score representing a greater degree of symptoms.

sponse format employed. Relatively high levels of exact agreement (as indicated by >70% exact agreement) were noted for infrequently endorsed symptoms (nausea/vomiting, dyspnea, constipation, diarrhea, visual disorder, weakness of both legs, and trouble controlling the bladder), and for the two measures containing items with a dichotomous response format (physical and role functioning). Conversely, relatively low levels of exact agreement (as indicated by <50% exact agreement) were noted for the more frequently reported symptoms and functional problems (social functioning, fatigue,

TABLE 2. Reliability of Patient and Proxy Ratings on the EORTC QLQ-C30 and QLQ-BCM

	Reproducibility[a] (n = 87)[c]		Internal Consistency[b] (n = 103)[d]	
	Patient ICC	Proxy ICC	Patient α	Proxy α
C30 Functioning scales				
Physical	0.79	0.83	0.76	0.76
Role	0.66	0.75	0.47	0.51
Cognitive	0.74	0.72	0.58	0.64
Emotional	0.63	0.70	0.83	0.83
Social	0.54	0.66	0.77	0.83
Global quality of life	0.66	0.82	0.75	0.85
C30 Symptom scales/items				
Fatigue	0.69	0.70	0.71	0.86
Nausea and vomiting	0.33	0.35	0.66	0.62
Pain	0.44	0.64	0.59	0.76
Dyspnea	0.28	0.58	—	—
Insomnia	0.70	0.65	—	—
Anorexia	0.62	0.74	—	—
Constipation	0.67	0.57	—	—
Diarrhea	0.24	0.48	—	—
Financial impact	0.71	0.66	—	—
BCM Symptom scales/items				
Future uncertainty	0.62	0.71	0.71	0.81
Visual disorder	0.81	0.79	0.72	0.87
Motor dysfunction	0.80	0.85	0.74	0.86
Communication deficit	0.66	0.74	0.87	0.90
Headaches	0.55	0.67	—	—
Seizures	0.66	0.28	—	—
Drowsiness	0.69	0.56	—	—
Bothered by hair loss	0.23	0.61	—	—
Bothered by itching skin	0.44	0.49	—	—
Weakness of both legs	0.43	0.51	—	—
Trouble controlling bladder	0.76	0.57	—	—

ICC, intraclass correlation coefficient.
[a]One-week test–retest intraclass correlation coefficient.
[b]Cronbach's coefficient α; not applicable for single item measures.
[c]Due to missing data, n varies from 80 to 86.
[d]Due to missing data, n varies from 97 to 103.

financial difficulties, future uncertainty, and drowsiness), and for the global quality-of-life measure that includes two items, each with a seven-point scale.

Approximate patient–proxy agreement, allowing one category of difference in either direction, was generally in excess of 90%. The average proportion of approximate agreement was 92.5% for the QLQ-C30 and 93.7% for the QLQ-BCM. Lower levels of approximate agreement (ie, below 90%) were noted for social functioning, pain, financial

TABLE 3. Patient–Proxy Agreement on EORTC QLQ-C30 and QLQ-BCM Measures

	Proportion Agreement[a]		Interclass Correlation Coefficient
	Exact (%)	Approximate (%)	
C30 Functioning scales			
Physical	82.4	—[b]	0.67
Role	77.1	—[b]	0.58
Cognitive	50.0	91.7	0.58
Emotional	53.2	94.1	0.62
Social	44.9	87.3	0.48
Global quality of life	30.5	91.5[c]	0.64
C30 Symptom scales/items			
Fatigue	44.3	93.2	0.52
Nausea and vomiting	77.6	94.6	0.37
Pain	62.6	89.3	0.23
Dyspnea	71.3	98.0	0.31
Insomnia	52.0	94.1	0.49
Anorexia	63.7	98.0	0.60
Constipation	70.3	94.1	0.55
Diarrhea	75.0	97.0	0.41
Financial impact	47.1	82.4	0.44
Total EORTC-QLQ-C30	60.8	92.5[d]	
BCM Symptom scales/items			
Future uncertainty	48.2	90.4	0.54
Visual disorder	72.4	93.7	0.57
Motor dysfunction	63.4	96.0	0.74
Communication deficit	51.5	94.8	0.64
Headaches	69.7	97.0	0.57
Seizures	87.1	99.0	0.73
Drowsiness	48.0	89.2	0.30
Bothered by hair loss	57.6	84.8	0.29
Bothered by itching skin	63.7	96.1	0.46
Weakness of both legs	71.3	95.0	0.32
Trouble controlling bladder	79.4	97.1	0.58
Total EORTC-QLQ-BCM	61.6	93.7	

[a]Proportion agreement for multi-item measures was calculated as the average percentage agreement across the number of items in each scale.
[b]Not applicable due to dichotomous response format.
[c]Proportion agreement within two categories in either direction.
[d]Seven items of physical/role functioning not included.

impact, drowsiness, and being bothered by hair loss.

Intraclass correlations varied from 0.23 to 0.67 for the QLQ-C30 measures and from 0.29 to 0.74 for the QLQ-BCM measures. Lower levels of patient–proxy agreement (as indicated by intraclass correlations below 0.50) were found for three of the 13 multi-

TABLE 4. Differences[a] and Standardized Mean Differences (Effect Size *d*)[b] Between Ratings Provided by Patients and Proxies

	Total Sample (n = 103)		No Mental Confusion (n = 79)		Mental Confusion (n = 24)	
	Difference (mean ± SD)	*d*	Difference (mean ± SD)	*d*	Difference (mean ± SD)	*d*
C30 Functioning scales						
Physical	-6.4 ± 24.4[d]	0.26	-3.4 ± 20.9	0.16	-16.5 ± 32.3[e]	0.51
Role	-12.3 ± 32.6[c]	0.38	-13.9 ± 33.0[c]	-0.42	-6.5 ± 31.3	0.21
Cognitive	-8.5 ± 22.5[c]	0.38	-6.5 ± 19.9[d]	0.33	-15.3 ± 29.0[e]	0.53
Emotional	-3.6 ± 18.4	0.20	-1.7 ± 17.3	0.10	-9.7 ± 20.7[e]	0.47
Social	-7.1 ± 29.3[e]	0.24	-5.5 ± 28.3	0.19	-13.0 ± 32.6	0.40
Global quality of life	-3.2 ± 19.2	0.17	-1.1 ± 18.2	0.06	-9.8 ± 21.4[e]	0.46
C30 Symptom scales/items						
Fatigue	11.7 ± 20.8[c]	0.56	10.0 ± 19.4[c]	0.51	17.1 ± 24.3[d]	0.71
Nausea and vomiting	1.3 ± 18.3	0.07	-0.4 ± 19.0	0.02	6.9 ± 14.7[e]	0.47
Pain	3.4 ± 26.0	0.13	3.6 ± 22.8	0.16	2.8 ± 35.3	0.08
Dyspnea	-3.0 ± 19.5	0.15	-4.7 ± 16.9[e]	0.28	2.8 ± 25.9	0.11
Insomnia	2.9 ± 27.0	0.11	2.6 ± 25.6	0.10	4.2 ± 31.6	0.13
Anorexia	1.0 ± 21.7	0.05	3.0 ± 21.5	0.14	-5.8 ± 21.7	0.27
Constipation	1.6 ± 24.2	0.07	-0.4 ± 20.5	0.02	8.7 ± 33.7	0.26
Diarrhea	0.0 ± 19.5	0.00	-1.7 ± 19.4	0.09	5.8 ± 19.2	0.30
Financial impact	5.6 ± 36.4	0.15	6.0 ± 37.1	0.16	4.2 ± 34.5	0.12
BCM Symptom scales/items						
Future uncertainty	2.8 ± 20.8	0.13	0.3 ± 19.7	0.02	10.5 ± 22.8[e]	0.46
Visual disorder	-1.3 ± 19.0	0.07	0.1 ± 14.5	0.01	-6.3 ± 30.1	0.21
Motor dysfunction	4.0 ± 17.6[e]	0.23	4.3 ± 15.8[e]	0.27	2.6 ± 23.5	0.11
Communication deficit	3.5 ± 21.8	0.16	3.3 ± 21.0	0.16	4.3 ± 24.8	0.18
Headaches	0.3 ± 21.0	0.01	-0.9 ± 19.6	0.04	4.3 ± 25.2	0.17
Seizures	0.7 ± 13.3	0.05	-0.4 ± 11.4	0.04	4.3 ± 18.3	0.24
Drowsiness	3.3 ± 34.0	0.10	0.4 ± 33.8	0.01	12.5 ± 33.8	0.37
Bothered by hair loss	2.7 ± 35.5	0.08	2.6 ± 32.6	0.08	2.9 ± 44.8	0.06
Bothered by itching skin	0.6 ± 24.4	0.02	-0.4 ± 24.9	0.02	4.2 ± 22.7	0.18
Weakness of both legs	1.6 ± 23.3	0.07	1.7 ± 20.0	0.09	1.4 ± 32.5	0.04
Trouble controlling bladder	-2.9 ± 19.4	0.15	-3.4 ± 19.1	0.18	-1.4 ± 20.8	0.07

SD, standard deviation.

[a] Proxy minus patient score.

[b] *d* = mean difference/standard deviation of difference (*d* = 0.2, small effect; *d* = 0.5, moderate effect; *d* = 0.8, large effect.)

[c] $P < 0.001$.

[d] $P < 0.01$.

[e] $P < 0.05$.

item measures (social functioning, nausea/vomiting, and pain), and for eight of the 13 single-item measures (dyspnea, insomnia, diarrhea, financial impact, drowsiness, being bothered by hair loss and itching skin, and weakness of both legs). The remaining measures exhibited intraclass correlations ranging from 0.52 to 0.74.

Systematic Response Bias

Statistically significant differences in mean scores obtained from patients and proxies were noted for five of the 15 QLQ-C30 measures and one of the 11 QLQ-BCM measures (first column of Table 4). As compared to the patients themselves, the proxies rated the patients as having lower levels of physical, role, cognitive, and social functioning, and a greater degree of fatigue and motor dysfunction. The statistical magnitude of this bias, as defined by standardized mean differences (effect size d), was relatively small for physical, role, cognitive, and social functioning, and motor dysfunction (d = 0.23–0.38). Only fatigue showed a moderate degree of bias (d = 0.56).

Factors Affecting Patient–Proxy Agreement

The number of discrepancies for each patient–proxy pair ranged from three to 22 for

TABLE 5. Number of Discrepancies Over the 30 Questions of the EORTC QLQ-C30 and the 20 Questions of the QLQ-BCM as a Function of the Patients' Physical and Neurological Status

		EORTC QLQ-C30			EORTC QLQ-BCM		
	n	Mean ± SD	P^a	P^b	Mean ± SD	P^a	P^b
Total	103	11.7 ± 4.2			7.5 ± 3.2		
Disease stage							
Newly diagnosed	40	11.0 ± 4.3	0.24	—	7.0 ± 3.3	0.27	—
Recurrent	63	12.1 ± 4.2			7.8 ± 3.1		
Performance status							
Karnofsky 100	26	10.2 ± 4.8	0.07	0.01	6.5 ± 3.1	0.02	0.005
Karnofsky 90	33	11.7 ± 4.0			7.2 ± 2.7		
Karnofsky 70–80	30	11.9 ± 4.0			7.6 ± 3.6		
Karnofsky 50–60	14	13.9 ± 3.4			9.7 ± 2.9		
Motor deficit							
Normal function	64	11.1 ± 4.5	0.13	0.05	7.0 ± 3.2	0.10	0.19
Symptomatic weakness	20	11.8 ± 4.1			8.7 ± 2.9		
Decrease in function	19	13.4 ± 3.2			7.7 ± 3.4		
Mental confusion							
Normal function	79	11.0 ± 4.2	0.003	—	6.9 ± 3.0	0.0005	—
Minor mental confusion	24	13.9 ± 3.8			9.4 ± 3.0		
Cognitive impairment							
None	46	11.3 ± 4.6	0.21	0.14	6.5 ± 3.0	0.02	0.004
Minor	34	11.3 ± 4.0			7.9 ± 3.4		
Moderate	23	13.0 ± 3.7			8.7 ± 2.8		

SD, standard deviation.
[a] P value for test of between-groups differences (analysis of variance).
[b] P value for test of linearity (for variables with more than two subgroups).

the 30 items of the QLQ-C30, and from 0 to 14 for the 20 items of the QLQ-BCM. As shown in Table 5, the mean number of patient-proxy discrepancies was 11.7 ± 4.2 and 7.5 ± 3.2 for the QLQ-C30 and QLQ-BCM, respectively. The mean number of patient–proxy discrepancies was found to vary as a function of the patients' physical and neurologic condition. Specifically, significantly more patient–proxy discrepancies in QLQ-C30 scores were found for patients with minor mental confusion; based on tests of linearity, patients classified as having poor performance status and motor deficits also showed more patient–proxy discrepancies in these scores as well. Significantly more patient–proxy discrepancies in QLQ-BCM scores were observed for patients with poor performance status, mental confusion, and cognitive impairment. No statistically significant association was found between the number of patient–proxy discrepancies in QLQ-C30 and QLQ-BCM scores and the patients' and proxies' background characteristics, including their age, gender, and culture, the proxies' relationship with the patient, and the living arrangement in relation to the patient (data not presented in tabular form).

As the presence of minor mental confusion was found to be the most significant factor affecting the number of patient–proxy discrepancies, we examined in greater detail the agreement between patient and proxy responses on the QLQ-C30 and QLQ-BCM measures for patients without mental confusion (n = 79) and patients with mental confusion (n = 24). Expressed in terms of average proportion agreement, the levels of exact and approximate agreement were found to be lower for patients with mental confusion. For patients with mental confusion, the average proportion of exact agreement was 53.2% and 51.2%, and the average proportion of approximate agreement was 89.6% and 88.8% for the QLQ-C30 and QLQ-BCM, respectively. For patients without confusion, these rates were 63.0% and 64.7% exact agreement and 93.4% and 95.1% approximate agreement for the QLQ-C30 and QLQ-BCM, respectively.

The magnitude of the patient–proxy response bias, as indicated by the standardized mean differences between patient and proxy scores (effect size *d*), appeared to be greater for patients with mental confusion (last column of Table 4). Among mentally confused patients, response bias of a moderate magnitude was found for all functioning scales (except role functioning), fatigue, nausea and vomiting, future uncertainty, and drowsiness (d = 0.37–0.71). Interestingly, regarding the psychosocially oriented measures (emotional and social functioning, global quality of life, and future uncertainty), response bias was nearly absent for patients without mental confusion (d = 0.02–0.19), but moderate for patients with mental confusion (d = 0.40–0.47).

Change in Patient–Proxy Agreement Over Time

The availability of a follow-up assessment permitted an evaluation of possible changes in the level of patient–proxy agreement over time. The follow-up administration was completed by 89 patient–proxy pairs, on average 71.0 ± 43.2 days after the baseline administration. For the total study sample, the findings at follow-up were very similar to those at baseline. In line with the cross-sectional findings described above, we expected to find a lower level of patient–proxy agreement for patients whose neurologic or physical condition had worsened over time. For the stable/improved group (n = 67), the patient–proxy agreement was similar for both assessment points, whereas significant changes in agreement levels were noted for those patients whose physical or neurologic condition had deteriorated (n = 22). For the latter group of patients, the mean number of patient–proxy discrepancies on the QLQ-C30 and QLQ-BCM increased from 12.3 ± 3.2 to 13.7 ± 3.2 for the QLQ-C30 and from

8.3 ± 3.4 to 9.9 ± 3.8 for the QLQ-BCM. The between-group (stable/improved versus deteriorated) difference over time (baseline versus follow-up) approached statistical significance for the QLQ-C30 ($P = 0.07$) and was statistically significant for the QLQ-BCM ($P = 0.003$). Similarly, for those patients whose condition had deteriorated over time, the patient–proxy intraclass correlations of 14 of the 26 measures incorporated in the QLQ-C30 and QLQ-BCM were substantially lower at follow-up (intraclass correlations ranging from -0.13 to 0.65) as compared to baseline (ranging from 0.55 to 0.84).

Discussion

The aim of the present study was to examine the feasibility and appropriateness of using significant others as proxy raters of the health-related quality of life of patients with brain cancer. Toward that purpose, we evaluated the level of agreement between patient and proxy responses to the questions of the EORTC QLQ-C30 supplemented by a brain cancer specific module, the QLQ-BCM. For both questionnaires, the study provided generally encouraging results regarding the comparability of patient and proxy responses. The majority of the patient–proxy comparisons (approximately 60%) were in exact agreement. When disagreement was observed, the discrepancies most commonly involved adjacent-category differences. Substantial discrepancies between patients and proxies were noted in only a small minority of the comparisons made (5–10%).

The number of discrepancies per patient–proxy pair over the two questionnaires was found to be highly variable. While some pairs disagreed on only a few questions, others had discrepant scores on more than half of the items. Both cross-sectional and longitudinal analyses revealed that the number of discrepancies was related to the physical and neurologic condition of the patients. The more impaired the patients, the more discrepancies were noted on both the EORTC QLQ-C30 and the QLQ-BCM. The observed trend was most obvious for patients with some degree of mental confusion, which lends support to an earlier finding of lower patient–proxy agreement among elderly respondents with even mild mental impairment.[16]

When examining the patient–proxy agreement for the separate functional and symptom measures of the EORTC QLQ-C30 and QLQ-BCM, the proportion of exact agreement was found to be largely dependent on the score distributions. The percentage agreement, although informative in its own right, ignores the fact that a certain amount of agreement can be expected based on chance alone. Hence, the intraclass correlation coefficient was used as a chance corrected index of agreement. In earlier proxy studies, the Pearson correlation (r) was customarily used as an indicator of agreement.[14–16,18,20] r does not necessarily provide an indication of actual agreement, however, because it disregards any systematic bias (ie, when proxy ratings are consistently lower than patient ratings, r can be excellent, but agreement can be poor).[12,36] The intraclass correlation avoids the problem of a linear relationship being mistaken for agreement, but does not resolve other problems associated with correlation coefficients, such as the dependency on the range and variance of the measurements.[39] Low correlations are more likely to be found for measures with a low frequency of occurrence and for single-item measures, both of which can result in limited score variability.[34] Thus, low patient–proxy correlations can sometimes be attributed to a lack of score variability, rather than to a lack of agreement. For this reason, it is useful to interpret the findings separately for well-distributed multi-item measures versus those with a more restricted range.

Correlations between patient and proxy ratings on the roughly well-distributed multi-item measures (ie, all functioning

scales, fatigue, future uncertainty, motor dysfunction, and communication deficit) were moderate to good. Theoretically, the reliability of both patient and proxy scores limit the potential degree of patient–proxy agreement, because the correlation between the two scores can never exceed the square root of the product of the scores' reliability.[34] Given this frame of reference, the patient–proxy correlations of the well-distributed multi-item measures can be considered as reasonably high.

Correlations between patient and proxy ratings on the remaining measures (ie, nausea/vomiting, pain, visual disorder, and all symptom items) were highly variable. Low patient–proxy correlations were observed for a number of measures. As expected on the basis of the distributional considerations described above, the poor correlations were not always in accordance with the proportions of exact and approximate agreement. Except for pain, financial impact, drowsiness, and being bothered by hair loss, the low observed correlations might be explained by a lack of score variability, rather than by a lack of agreement. These findings support the position taken by Bland and Altman that it may not be possible to summarize agreement adequately using a single number, such as a correlation coefficient.[23,39] The proportion of approximate agreement (or its complement, the proportion of substantial discrepancies) may be the most useful indicator of agreement for measures with more than two response categories.

When comparing mean scores of the patients and their proxies, statistically significant differences were noted for a number of the QLQ-C30 and QLQ-BCM measures. We would caution that, given the relatively large number of comparisons made, some statistically significant differences could be found by chance alone. It should be noted, however, that the differences in patient and proxy mean scores were constantly in the same direction. As expected on the basis of previous studies, proxies rated the patients as having more disability than the patients themselves.[15,16,18–20] Importantly, except for fatigue, this systematic response bias was of a rather modest magnitude. As was the case with the level of response agreement, the magnitude of the response bias appeared to be dependent on the health status of the patient. Among patients exhibiting mental confusion, response bias of a moderate size was found for several HRQOL domains. These findings support the currently held view that proxies tend to rate patients as having more disability and a lower quality of life than do the patients themselves.[8,9] According to our findings, however, this tendency is greater for the more disabled patients.

The results of this study indicate that, in general, patients and their significant others will agree reasonably well on the patients' quality of life. Less agreement and more biased ratings, however, can be expected among those patients for whom the need for proxies is most salient (ie, the cognitively and physically impaired). It should be noted that our results are based on a study of a specific patient population and on the use of significant others as proxies. Whether these findings are generalizable to other conditions, other types of proxy raters (eg, physicians or nurses), and other HRQOL instruments requires further investigation.

Based on our findings, one might conclude that the use of significant others as an alternative source of quality-of-life ratings for those patients unable to provide such ratings themselves is inappropriate. An additional issue should be considered, however, when evaluating the potential role of significant others as proxy raters of the patients' quality of life. Given the central position of the patient in quality-of-life evaluations, the patient's rating is generally taken as a gold standard to which the proxy rating should conform. When adopting this assumption, the patient's rating is, by definition, the more valid one. Consequently, lack

of patient–proxy agreement has been interpreted as a lack of validity of the proxy rating, and the finding of more disability as reported by the proxy has been interpreted as overestimation of the patient's level of impairment.[15,16,18] Indeed, several characteristics of the proxy, such as his/her subjective well-being in terms of perceived caregiver burden and emotional distress, may contribute to the proxy providing an exaggerated rating of the patient's disability.[18] In the current study, however, the most discrepant patient–proxy pairs were those in which the patients had some degree of mental confusion or other cognitive deficits. Moreover, according to the treating physician, many of these patients showed a reduced ability to answer questions on a self-report questionnaire or an inappropriate affect during the clinical examination. Questions can be raised about the validity and reliability of the quality-of-life ratings provided by such patients. To date, no proxy studies have examined the relative validity and reliability of patient- and proxy-generated data. The findings of the current study suggest that the reliability (both test–retest reliability and internal consistency) of the proxy-generated data is slightly higher than that of the patients. New studies should extend beyond examination of patient–proxy agreement, by addressing the relative validity and responsiveness to clinical changes of patient- versus proxy-generated scores. It might well be that, for some patients, the significant other provides more reliable and valid data and should consequently be regarded as the primary source of information. Minimally, it should not be assumed, a priori, that discrepancies between patient–proxy ratings are evidence of the inaccuracy or biased nature of proxy-generated data.

Clearly, additional work is required to assess the potential impact of proxy-generated data on study outcomes in HRQOL research. For the present, researchers should think ahead when planning to include an HRQOL component in their studies. If the patient population of interest suffers from serious cognitive impairment (eg, stroke survivors, patients with brain tumors or brain metastases), or if it is expected that such impairment could develop over time, it is advisable to obtain proxy ratings of the patients' quality of life throughout the entire course of the study. This will enable the researcher to include a more representative group of patients in the HRQOL evaluation and offers a possibility to critically assess the potential impact of the use of proxy respondents on the study results.

Acknowledgments

The authors thank Cynthia Gelke and Christine Myers from the MD Anderson Cancer Center, Douglas Guerrero and Sue Sardell from the Royal Marsden Hospital, and Helen Evans from the Charing Cross Hospital for administration of the HRQOL questionnaires and collection of the medical chart information; David Sugano, DrPH, and Margaret Dugan, MD, from Schering-Plough for their contribution to the study; and Mirjam Sprangers, PhD, from the Netherlands Cancer Institute for comments on earlier versions of the manuscript. The authors are grateful to the patients and their significant others without whose cooperation this study would not have been possible.

References

1. Aaronson NK. Methodological issues in assessing the quality of life of cancer patients. Cancer 1991;67:844.

2. Cella DF, Tulsky DS. Quality of life in cancer: Definition, purpose, and method of measurement. Cancer Invest 1993;11:327.

3. Donovan K, Sanson-Fisher RW, Redman S. Measuring quality of life in cancer patients. J Clin Oncol 1989;7:959.

4. Viitanen M, Fugl-Meyer KS, Bernspang B, et al. Life satisfaction in long-term survivors after stroke. Scand J Rehab Med 1988;20:17.

5. Robinson RG, Bolduc PL, Kubos KL, et al. Social functioning assessment in stroke patients. Arch Phys Med Rehabil 1985;66:496.

6. Smith DS. Outcome studies in stroke rehabilitation: The South Australian stroke study. Stroke 1990;21:1156.

7. Aaronson NK. Assessing the quality of life of patients in cancer clinical trials: Common problems and common sense solutions. Eur J Cancer 1992;28A:1304.

8. Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. J Clin Epidemiol 1992;45:743.

9. Magaziner J. The use of proxy respondents in health studies of the aged. In: Wallace RB, Woolson RF, eds. The epidemiologic study of the elderly. New York, NY: Oxford University Press, 1992.

10. De Haan RJ, Limburg M, van der Meulen JH, et al. Quality of life after stroke: Impact of stroke type and lesion location. Stroke 1995;26:402.

11. Mosely RR, Wolinsky FD. The use of proxies in health surveys: Substantive and policy implications. Med Care 1986;24:496.

12. Nelson LM, Longstreth Jr WT, Koepsell TD, et al. Proxy respondents in epidemiologic research. Epidemiol Rev 1990;12:71.

13. Seckler AB, Meier DE, Mulvihill M, et al. Substituted judgment: How accurate are proxy predictions? Ann Intern Med 1991;115:92.

14. Bassett SS, Magaziner J, Hebel JR. Reliability of proxy response on mental health indices for aged, community-dwelling women. Psychol Aging 1990;5:127.

15. Epstein AM, Hall JA, Tognetti J, et al. Using proxies to evaluate quality of life: Can they provide valid information about patients' health status and satisfaction with medical care? Med Care 1989;27:S91.

16. Magaziner J, Simonsick EM, Kashner TM, et al. Patient-proxy response comparability on measures of patient health and functional status. J Clin Epidemiol 1988;41:1065.

17. McCusker J, Stoddard AM. Use of a surrogate for the Sickness Impact Profile. Med Care 1984;22:789.

18. Rothman ML, Hedrick SC, Bulcroft KA, et al. The validity of proxy-generated scores as measures of patient health status. Med Care 1991;29:115.

19. Rubenstein LZ, Schairer C, Wieland GD, et al. Systematic biases in functional status assessment of elderly adults: Effects of different data sources. J Gerontol 1984;39:686.

20. Clipp EC, George LK. Patients with cancer and their spouse caregivers. Cancer 1992;69:1074.

21. Curtis AE, Fernsler JI. Quality of life of oncology hospice patients: A comparison of patient and primary caregiver reports. Oncol Nurs Forum 1989;16:49.

22. Farrow DC, Samet JM. Comparability of information provided by elderly cancer patients and surrogates regarding health and functional status, social network, and life events. Epidemiology 1990;1:370.

23. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;8476:307.

24. 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 1993;85:365.

25. Aaronson NK, Cull A, Kaasa S, et al. The European Organization for Research and Treatment of Cancer (EORTC) modular approach to quality of life assessment in oncology: An update. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. New York, NY: Raven Press, 1996.

26. 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 1995;13:1249.

27. Sprangers MA, Cull A, Bjordal K, et al. The European Organization for Research and Treatment of Cancer approach to quality of life assessment: Guidelines for developing questionnaire modules. Qual Life Res 1993;2:287.

28. Osoba D, Aaronson NK, Muller MJ, et al. The development and psychometric validation of a brain cancer quality-of-life questionnaire for use in combination with general cancer-specific questionnaires. Qual Life Res 1996;5:139.

29. Karnofsky DA, Burchenal JH. The clinical evaluation of chemotherapeutic agents in cancer. In: MacLeod CM, ed. Evaluation of chemotherapeutic agents. New York, NY: Columbia University Press, 1949:191.

30. Mor V, Laliberte L, Morris JN, et al. The Karnofsky Performance Status Scale: An examination of its reliability and validity in a research setting. Cancer 1984;53:2002.

31. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3.

32 Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. J Clin Epidemiol 1991;44:381.

33. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. Control Clin Trials 1991;12:142S.

34. Nunnally JC, Bernstein IH. Psychometric theory. New York, NY: McGraw–Hill, 1994.

35. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297.

36. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med 1989;19:61.

37. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ:Lawrence Erlbaum Associates, 1988:19.

38. Norusis MJ/SPSS Inc. SPSS statistical data analysis: SPSS/PC+4.0 Base Manual. Chicago, IL: SPSS Inc., 1990:B137.

39. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990;20:337.

Chapter 6

Comparison of patient and spouse assessments of health related quality of life in men with metastatic prostate cancer

K.C.A. Sneeuw, P.C. Albertsen, N.K. Aaronson

Journal of Urology 2001, 165: 478-482

COMPARISON OF PATIENT AND SPOUSE ASSESSMENTS OF HEALTH RELATED QUALITY OF LIFE IN MEN WITH METASTATIC PROSTATE CANCER

KOMMER C. A. SNEEUW, PETER C. ALBERTSEN AND NEIL K. AARONSON

From the Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, and University of Connecticut Health Center, Farmington, Connecticut

ABSTRACT

Purpose: We examined the extent of agreement in health related quality of life ratings provided by patients with metastatic prostate cancer and their spouses. This agreement is important for determining the feasibility of using spouses as potential proxy raters in quality of life studies in this patient population.

Materials and Methods: The study sample consisted of 72 pairs of patients with metastatic prostate cancer in remission or progression and their spouses. Patients and spouses independently completed the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-C30 and a prostate cancer specific questionnaire module. Together the 2 questionnaires assess a wide range of symptoms and functional limitations for a total of 21 quality of life outcomes.

Results: For 5 of the 21 patient-proxy comparisons we noted systematic differences in the mean score with spouses rating more impairment in patients than patients indicated. Most patient-proxy correlations were 0.40 to 0.75, indicating moderate to good agreement in patient and spouse ratings. A low patient-proxy correlation of less than 0.40 was noted only for the 2 measures of sexual function.

Conclusions: Our findings suggest that the spouses of men with metastatic prostate cancer evaluate with a fair degree of accuracy how patients experience physical and psychosocial functioning, symptoms and overall quality of life. However, caution should be exercised when relying on proxy raters for assessing sexual functioning and satisfaction.

KEY WORDS: prostate, prostatic neoplasms, neoplasm metastasis, quality of life, questionnaires

Quality of life outcomes have become important end points of clinical research in oncology, especially in advanced disease trials. An important factor contributing to the increasingly frequent use of quality of life end points in clinical trials is growing evidence that despite their subjective nature quality of life data may be obtained in a reliable and valid manner. Recently several research groups have developed and validated quality of life questionnaires for use in prostate cancer that address general and condition specific quality of life issues.[1–5]

Given the subjective nature of quality of life data patients are generally considered to be the primary source of such information. However, self-reporting quality of life instruments are less suitable when patients have cognitive impairment or severe symptom distress, or when completing a questionnaire is physically or emotionally too burdensome. Therefore, in oncology trials the compliance rate in quality of life assessment is sometimes far from optimal, especially in studies of advanced disease.[6] To circumvent this methodological problem others have examined whether health care providers or significant others may provide quality of life information similar to that provided by patients,[7] including patients with advanced cancer.[8–11] These studies have had conflicting results.

We examined whether the spouses of patients with advanced prostate cancer may be used as complementary or alternative sources of information on patient quality of life. Spouses may be particularly sensitive raters of patient quality of life because they have access to patient thoughts, feelings and symptom experience associated with the disease and treatment. The extent of agreement in the quality of life ratings provided by prostate cancer patients and their spouses (so-called proxy ratings) was determined and compared with the results of earlier studies of patients with other types of advanced cancer.

Accepted for publication September 15, 2000.

Supported in part by Integrated Therapeutics Group, Inc., a subsidiary of Schering-Plough.

METHODS

Subjects. Our analysis was based on data obtained from participants in a multicenter cross-sectional study of quality of life in stage D2 prostate cancer treated with combined luteinizing hormone releasing hormone agonist and flutamide. Men at 23 hospitals who were receiving treatment for advanced prostate cancer were asked to participate in this study. Patients were classified into those in remission and responding to androgen ablation, and those with clinical evidence of disease progression or androgen insensitive disease. More detailed information on the patient sample has been previously reported.[6]

For patients with a spouse or partner the spouse, termed the proxy respondent, was also asked to participate in the study. Spouses included married and unmarried partners, that is those with a long standing and intimate relationship with patients. Patients and proxy respondents were asked to complete quality of life questionnaires at the initial office visit and again by mail 1 week later. Physicians of these patients were asked to rate the extent of patient disease as minimal or extensive depending on the number and location of bone metastases. Physicians also rated patient perfor-

mance status using the Eastern Cooperative Oncology Group scale.[12]

Measures. Quality of life was assessed by the EORTC Quality of Life Questionnaire (QLQ)-C30, version 2.0,[13,14] and an additional prostate cancer specific module in the same format.[5] The EORTC QLQ-C30 is a 30-item questionnaire that was developed to assess a range of physical, emotional and social health issues relevant to a broad spectrum of patients with cancer. Patients are asked to indicate the extent to which they have experienced 28 specific functional limitations and symptoms in the last week using a mixture of dichotomous yes-no response categories and a 4-point response scale ranging from not at all to very much. Questions are organized into a number of multi-item scales and single item symptom measures, including 5 functioning scales on physical, role, cognitive, emotional and social functioning, 3 symptom scales on fatigue, nausea and vomiting, and pain, and 6 single item measures on dyspnea, sleep disturbance, appetite loss, constipation, diarrhea and financial impact. The remaining 2 questions ask patients to rate overall health and quality of life on a 7-point scale ranging from very poor to excellent. These 2 items represent a global quality of life scale. All scales and single item scores were linearly transformed to a scale of 0 to 100.

The prostate cancer specific module was developed specifically for this project and it was partially based on previous studies of the EORTC Genitourinary Group.[15,16] This module contains 11 questions, including 4 assessing urinary symptoms that form a multi-item scale, single item measures assessing hot flashes, weight loss, weight gain and sexual satisfaction, and 3 items assessing sexual functioning aggregated into a multi-item scale. For all questions a 4-point response format ranging from not at all to very much is used. Scales and single item scores were linearly transformed to a scale of 0 to 100.

Proxy respondents completed a slightly modified version of these questionnaires. Standard instructions were provided in which proxies were asked to view the situation from the patient perspective and complete the questionnaires accordingly. In addition, all item statements were made from the third person perspective, for example "Would he say that . . . . "

Statistical analysis. Descriptive statistics were used to characterize the study population. The mean score plus or minus standard deviation (SD) was calculated for all QLQ-C30 and prostate cancer module scales and single item measures. Test-retest reliability of each measure was assessed by calculating the intraclass correlation of scores at baseline and at the 1-week assessment.

Patient-proxy agreement was assessed in 2 ways using previously reported approaches,[8,17,18] enabling a meaningful comparison of results among studies. In accordance with Sneeuw et al,[17] for all QLQ-C30 and prostate cancer module scales as well as single item measures we calculated the difference in patient and proxy mean scores, and the intraclass correlation of patient and proxy ratings. As indicated by paired t tests, statistically significant differences in mean scores were interpreted to be evidence of systematic differences in raters at the group level.[19] To interpret the size of observed differences mean difference scores were standardized by relating them to their SD with effect sizes d = 0.2, 0.05 and 0.08 indicating a slight, moderate and great difference, respectively.[20] Intraclass correlation was used as an indicator of chance corrected agreement in patient and proxy ratings at the individual patient level.[19] For ordinal data in our study the intraclass correlation was mathematically equivalent to the weighted κ statistic.[21]

In addition, in accordance with Stephens et al[8] we calculated the percent agreement of all individual items of the QLQ-C30 and prostate cancer module using a 4-point response format ranging from not at all to very much. Thus, the 7 questions in the QLQ-C30 using a dichotomous or 7-point response scale were excluded from analysis. To enable unambiguous interpretation of the ratings we also excluded from analysis 2 questions in the prostate cancer module on weight gain and sexual satisfaction that were posed in opposite directions, leaving 32 items assessing the degree of specific symptoms and functional limitations on a 4-point scale. Furthermore, using the approach of Litwin et al[18] we compared the percent of patients and proxy respondents reporting prostate cancer specific symptoms of any degree.

RESULTS

A total of 113 patients were enrolled in the study, including 60 (53%) in disease remission and 53 (47%) with progression. Table 1 lists additional patient clinical and sociodemographic characteristics. Overall 97 patients were living with a spouse or partner. Data were available from patients and spouses for 72 of these 97 patients (74%). Our analysis focused on these 72 patients. We observed no significant differences in the sociodemographic or clinical variables of the sample of 72 patients and the 25 with a nonparticipating spouse. Table 2 shows mean patient and proxy generated ratings of the scales and single item measures of the QLQ-C30 and prostate cancer module. Except for the sexual functioning and satisfaction measures scores were somewhat skewed toward the positive end of the scale with more patients reporting minimal functional limitations or symptomatology.

Test-retest reliability of patient and proxy scores was good for most scales and single item measures of the QLQ-C30 and prostate cancer module with an intraclass correlation of 0.68 to 0.96. The only exceptions were patient ratings of diarrhea (intraclass correlation 0.53), and patient and proxy ratings of sexual satisfaction (intraclass correlation 0.43 and 0.46, respectively). For 5 of the 21 patient-proxy comparisons we noted a statistically significant difference in the mean score, indicating systematic differences in patient and proxy ratings ($p < 0.05$, table 2). Proxy respondents rated patients with more impaired physical and role functioning, more sleep disturbance and weight loss, and a lower level of sexual satisfaction than the patients indicated. The standardized difference or effect size d for the 5 systematic differences observed was between 0.26 and 0.34.

Intraclass correlation in patient and proxy ratings was 0.47 to 0.73 for the QLQ-C30 and 0.24 to 0.75 for the prostate cancer module. The average patient-proxy correlation of all 21 comparisons was 0.58. As indicated by an intraclass correlations of less than 0.40, we noted a low level of agreement in 2 of the 21 quality of life domains (sexual functioning and sexual satisfaction). The low intraclass correlation of 0.31 for sexual functioning may largely be explained by 3 patients for

TABLE 1. *Clinical and sociodemographic characteristics*

	Total Sample No. (%)	No. Pts. Proxies (%)
Disease stage:		
Remission	60 (53)	43 (60)
Progression	53 (47)	29 (40)
Physician disease rating:		
Extensive	73 (65)	46 (64)
Minimal	40 (35)	26 (36)
Eastern Cooperative Oncology Group performance status:		
0	58 (51)	38 (53)
1	38 (34)	25 (35)
2	14 (12)	8 (11)
3	3 (3)	1 (1)
Mean age ± SD	72.5 ± 8.4	71.6 ± 7.8
Mean education ± SD (yrs.)	13.6 ±3.6	13.9 ±3.1
Living situation:		
Spouse, Partner	97 (86)	72 (100)
Relative, friend	3 (3)	0
Group home, residential center	2 (2)	0
Alone	11 (10)	0

TABLE 2. *Comparison of patient and proxy scores for the EORTC QLQ-C30, prostate cancer module scales and single items*

	Mean Score ± SD		Mean Difference ± SD	p Value	Intraclass Correlation
	Pt.	Proxy			
QLQ-C30 functioning scales:*					
Physical	78.6 ± 24.7	71.9 ± 28.0	−6.7 ± 19.6	0.01	0.71
Role	76.2 ± 29.7	68.8 ± 33.3	−7.4 ± 23.7	0.01	0.70
Cognitive	83.8 ± 17.2	82.9 ± 18.5	−0.9 ± 18.5	0.67	0.47
Emotional	78.9 ± 18.6	75.1 ± 25.0	−3.8 ± 22.4	0.15	0.48
Social	78.7 ± 27.6	77.5 ± 30.3	−1.2 ± 27.5	0.72	0.55
Global life quality	66.9 ± 24.3	65.3 ± 29.0	−1.6 ± 23.8	0.57	0.61
QLQ-C30 symptom scales, items:†					
Fatigue	32.6 ± 23.2	34.1 ± 27.6	1.5 ± 19.1	0.50	0.73
Nausea, vomiting	6.5 ± 12.0	6.3 ± 13.3	−0.2 ± 12.7	0.88	0.50
Pain	27.3 ± 28.6	26.9 ± 29.0	−0.4 ± 21.1	0.85	0.73
Dyspnea	18.1 ± 23.0	17.6 ± 24.4	−0.5 ± 22.7	0.86	0.54
Sleep disturbance	27.8 ± 28.0	35.2 ± 34.9	7.4 ± 28.1	0.03	0.59
Appetite loss	12.0 ± 24.6	16.2 ± 26.8	4.2 ± 19.3	0.07	0.71
Constipation	19.4 ± 27.3	22.2 ± 28.0	2.8 ± 23.6	0.32	0.64
Diarrhea	12.5 ± 22.0	13.4 ± 22.1	0.9 ± 22.4	0.73	0.49
Financial impact	19.9 ± 28.3	17.6 ± 26.2	−2.3 ± 23.3	0.40	0.64
Prostate Ca module scales, items:‡					
Urinary symptoms	27.1 ± 19.8	24.9 ± 21.5	−2.2 ± 16.7	0.27	0.67
Hot flashes	36.6 ± 29.2	42.1 ± 35.8	5.5 ± 30.1	0.12	0.57
Wt. loss	10.5 ± 22.4	15.7 ± 30.4	5.2 ± 18.5	0.02	0.75
Wt. gain	20.2 ± 27.9	23.9 ± 30.4	3.7 ± 26.8	0.24	0.58
Sexual functioning	14.7 ± 22.1	13.5 ± 27.4	−1.2 ± 29.3	0.73	0.31
Sexual satisfaction	39.8 ± 43.5	26.4 ± 38.7	−13.4 ± 50.3	0.02	0.24

Due to missing data there were 67 to 72 responses.
* Higher score represents better functioning and quality of life.
† Higher score represents a greater degree of symptoms.
‡ Higher score represents a greater degree of symptoms for urinary symptoms and hot flashes, more weight loss or gain for weight loss and weight gain, and better functioning and more satisfaction for sexual functioning and satisfaction.

whom the maximal discrepancy in patient and proxy respondents was indicated in all 3 items, namely severe limitation according to the patient but none according to the proxy. When these 3 patients were excluded from analysis, intraclass correlation was 0.55.

From the 32 items assessing specific symptoms and limitations on a 4-point response scale for each of the 72 patient-proxy dyads a potential of 2,304 patient-proxy comparisons were possible. Due to missing data 18 comparisons were not available, leaving 2,286 comparisons of patient and proxy ratings (table 3). Of the 2,286 comparisons we identified agreement in 1,433 (63%), disagreement by 1 response category in 717 (31%) and disagreement by 2 or 3 response categories in 136 (6%).

The figure shows the proportion of patients in whom prostate cancer specific symptoms were reported by patients and proxy respondents. A similar percent of impairment was reported by patients and spouses. We observed a statistically significant difference according to patients and proxies only for difficult urination control (62% versus 46%, $p = 0.02$).

DISCUSSION

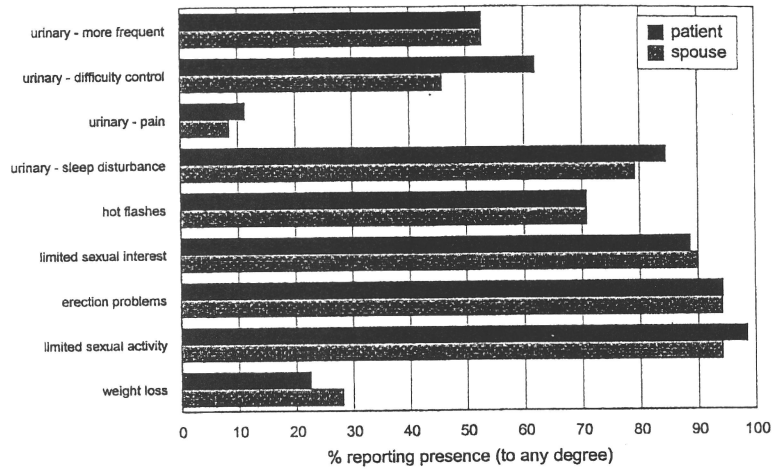
For the majority of quality of life dimensions measured by the QLQ-C30 and prostate cancer module we identified no evidence of systematic differences in patient and proxy mean scores. This finding is important in regard to the possible use of proxy quality of life ratings in clinical trials, in which the mean scores of patient groups are compared rather than individual patient scores. The observed systematic differences in 5 of the 21 quality of life dimensions were in the expected direction[7] with spouses rating more functional problems and symptoms in patients than patients indicated. The magnitude of these differences in terms of standard differences or effect size was slight to moderate.

Except for the measures of sexual functioning and satisfaction patient-proxy correlation was between 0.47 and 0.75, which may be interpreted as representing a moderate to high level of agreement.[22] The observed patient-proxy correlations are in accordance with those of earlier studies using the EORTC questionnaire to compare quality of life ratings of cancer patients and their significant others.[10,17,23,24] The low patient-proxy correlation for sexual functioning and sexual satisfaction may have been due to low reliability of the score provided by 1 or each rater, low score variance or outliers.[25] For sexual functioning test-retest reliability was high but score variance was low, that is most patients self-reported limitations in sexual interest, sexual activity and achieving or maintaining erection. In addition, the magnitude of the correlation was significantly influenced by a few outliers. Maximal discrepancy was observed in patient and proxy scores in 3 patients with patients reporting severe limitation and spouses reporting none. It is not clear whether

TABLE 3. *Patient-proxy agreement on 32 items assessing symptoms and limitations of the EORTC QLQ-C30 and prostate cancer module on a 4-point response scale*

Proxy Rating	Pt. Rating				
	No. Not at All	No. Little	No. Quite Bit	No. Very Much	Total No.
Not at all	874*	217	20	14	1,125
Little	186	313*	92	13	604
Quite bit	26	114	77*	36	253
Very much	23	40	72	169*	304
Totals	1,109	684	261	232	2,286†

All symptoms and functional limitations have a 4-point response format in the same direction, excluding 7 QLQ-C30 questions using a dichotomous or 7-point response format, and 2 module questions on weight gain and sexual satisfaction posed in opposite directions.
* Agreement.
† Across 32 items for 72 patients ($32 \times 72 = 2,304$ comparisons) with 18 comparisons missing ($2,304 - 18 = 2,286$).



Percent of patients and spouses reporting prostate cancer specific symptoms

this discrepancy was due to disagreement or to errors in completing the questionnaire by 1 of the 2 raters. For sexual satisfaction the low patient-proxy correlation was probably due to the low reliability of this item.

An additional and perhaps more straightforward approach to examining the extent of agreement of patient and proxy ratings was provided by the percent of agreement and disagreement on individual questions. Patients and proxy respondents did not respond to multi-item scales but to individual questions on functional limitations and symptoms. More than 60% of all patient-proxy comparisons were in exact agreement. When there was disagreement, the ratings most often differed by only 1 response category. Substantial discrepancies of 2 or 3 categories were noted in only 6% of the comparisons made. These results are similar to those of a study of brain cancer that compared patient and proxy ratings on the QLQ-C30 and a brain cancer specific module.[23] An even higher rate of agreement was reported in a study of patients with lung cancer,[8] in which comparisons were made of ratings by patients and treating physicians on 11 key physical symptoms. There was almost 80% exact agreement with only 5% disagreement by 2 of 3 response categories. Notably in all 3 studies, particularly the latter, a substantial proportion of concordant ratings represented agreement on absent symptoms.

The percent of cases in which prostate cancer specific symptoms were reported by patients and spouses was similar. This finding is in contrast to that of Litwin et al.[18] In their series urologists substantially underestimated all patient symptoms and low correlations were observed in patient and urologist ratings of patient quality of life.[18] The rather disappointing results of this study may have been partially due to the fact that patients and urologists did not complete the same questionnaire.[26] Recent studies involving direct comparisons of quality of life ratings by cancer patients, significant others and physicians indicate that the level of patient-physician agreement may only be slightly lower than that of patients and spouses or other close companions.[9,11,24]

CONCLUSIONS

Our study provides encouraging findings on the usefulness of quality of life ratings provided by the spouses of patients with advanced prostate cancer. The findings suggest that spouses evaluate with a relatively high degree of accuracy how patients experience physical and psychosocial functioning, symptoms and overall quality of life. Since such evaluations may have a role in delivering adequate patient care in the home setting, these results are reassuring. Furthermore, our findings provide support for the feasibility of using spouses as proxy respondents of patient quality of life in clinical studies when it is deemed necessary.

REFERENCES

1. Litwin, M. S., Hays, R. D., Fink, A. et al: The UCLA Prostate Cancer Index: development, reliability, and validity of a health-related quality of life measure. Med Care, **36:** 1002, 1998
2. Esper, P., Mo, F., Chodak, G. et al: Measuring quality of life in men with prostate cancer using the functional assessment of cancer therapy-prostate instrument. Urology, **50:** 920, 1997
3. Borghede, G. and Sullivan, M.: Measurement of quality of life in localized prostatic cancer patients treated with radiotherapy: development of a prostate cancer-specific module supplementing the EORTC QLQ-C30. Qual Life Res, **5:** 212, 1996
4. Stockler, M. R., Osoba, D., Corey, P. et al: Convergent discriminative, and predictive validity of the prostate cancer specific quality of life instrument (PROSQOLI): assessment and comparison with analogous scales from the EORTC QLQ-C30 and a trial-specific module. J Clin Epidemiol, **52:** 653, 1999
5. Albertsen, P. C., Aaronson, N. K., Muller, M. J. et al: Health-related quality of life among patients with metastatic prostate cancer. Urology, **49:** 207, 1997
6. Hopwood, P., Stephens, R. J. and Machin, D.: Approaches to the analysis of quality of life data: experiences gained from a Medical Research Council Lung Cancer Working Party palliative chemotherapy trial. Qual Life Res, **3:** 339, 1994
7. Sprangers, M. A. G. and Aaronson, N. K.: The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol, **45:** 743, 1992
8. Stephens, R. J., Hopwood, P., Girling, D. J. et al: Randomized trials with quality of life endpoints: are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? Qual Life Res, **6:** 225, 1997
9. Grassi, L., Indelli, M., Maltoni, M. et al: Quality of life of homebound patients with advanced cancer: assessments by patients, family members, and oncologists. J Psychosoc Oncol, **14:** 31, 1996

10. Sigurdardottir, V., Brandberg, Y. and Sullivan, M.: Criterion-based validation of the EORTC QLQ-C36 in advanced melanoma: the CIPS questionnaire and proxy raters. Qual Life Res, **5:** 375, 1996
11. Sneeuw, K. C. A., Aaronson, N. K., Sprangers, M. A. G. et al:: Value of caregiver ratings in evaluating the quality of life of patients with cancer. J Clin Oncol, **15:** 1206, 1997
12. Zubrod, C. G., Schneiderman, M., Frei, E. et al: Appraisal of methods for the study of chemotherapy of cancer in man: comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. J Chron Dis, **11:** 7, 1960
13. Aaronson, N. K., 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, 1993
14. Osoba, D., Aaronson, N., Zee, B. et al: Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. Qual Life Res, **6:** 103, 1997
15. Fossa, S. D., Aaronson, N. K., Newling, D. et al:: Quality of life and treatment of hormone resistant metastatic prostatic cancer. Eur J Cancer, **26:** 1133, 1990
16. Da Silva, F. C., Fossa, S. D., Aaronson, N. K. et al: The quality of life of patients with newly diagnosed M1 prostate cancer: experience with EORTC clinical trial 30853. Eur J Cancer, **32A:** 72, 1996
17. Sneeuw, K. C. A., Aaronson, N. K., Sprangers, M. A. G. et al: Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. J Clin Epidemiol, **51:** 617, 1998
18. Litwin, M. S., Lubeck, D. P., Henning, J. M. et al: Differences in urologist and patient assessments of health related quality of life in men with prostate cancer: results of the CaPSURE database. J Urol, **159:** 1988, 1998
19. Lee, J., Koh, D. and Ong, C. N.: Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med, **19:** 61, 1989
20. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum, 1988
21. Fleiss, J. L. and Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas, **33:** 613, 1973
22. Landis, J. R. and Koch, G. G.: The measurement of observer agreement for categorical data. Biometrics, **33:** 159, 1977
23. Sneeuw, K. C. A., Aaronson, N. K., Osoba, D. et al: The use of significant others as proxy raters of the quality of life of patients with brain cancer. Med Care, **35:** 490, 1997
24. Blazeby, J. M., Williams, M. H., Alderson, D. et al: Observer variation in assessment of quality of life in patients with oesophageal cancer. Br J Surg, **82:** 1200, 1995
25. Nunnally, J. C. and Bernstein, I. H.: Psychometric Theory. New York: McGraw-Hill, 1994
26. Sprangers, M. A. G. and Sneeuw, K. C. A.: Are healthcare providers adequate raters of patients' quality of life: perhaps more than we think? Acta Oncol, **39:** 5, 2000

Chapter 7

Assessing quality of life after stroke: The value and limitations of proxy ratings

K.C.A. Sneeuw, N.K. Aaronson, R.J. de Haan, M. Limburg

Stroke 1997, 28: 1541-1549

Assessing Quality of Life After Stroke

The Value and Limitations of Proxy Ratings

K.C.A. Sneeuw, MS; N.K. Aaronson, PhD; R.J. de Haan, RN, PhD; M. Limburg, MD, PhD

Background and Purpose Because many stroke survivors have cognitive and communication disorders, self-reported information on a patient's quality of life (QL) cannot always be obtained. Proxy ratings may be used to prevent exclusion of this highly relevant subgroup of patients from QL studies. The purpose of this study was to evaluate both the value and possible limitations of such proxy ratings.

Methods The patient sample was composed of 437 patients who had suffered a stroke 6 months earlier. QL was assessed by means of the Sickness Impact Profile (SIP). For 108 patients who were not communicative because of cognitive or linguistic deficits, proxy ratings on the SIP were provided by the patients' significant others. For 228 of the 329 communicative patients, both self-reported and proxy SIP ratings were obtained.

Results When mean SIP scores for patients with both self-reported and proxy-derived data available were compared, the proxy mean scores were generally in close agreement with those of the patients. However, systematic differences were noted for several SIP scales, with proxies rating patients as having more QL impairments than the patients themselves. Intraclass correlations were moderate to high for most SIP subscales (average intraclass correlation coefficient [ICC]=.63), the physical (ICC=.85) and psychosocial dimensions (ICC=.61), and the total SIP score (ICC=.77). The proxy SIP scores were sensitive to differences in patients' functional health, which supports the validity of these ratings. For all patients combined, more QL impairments were found for patients with supratentorial cortical or subcortical infarctions and hemorrhages than for patients with lacunar infarctions and infratentorial strokes. Although proxy respondents were more frequently needed for patients with the first two types of stroke, we found no evidence of biased results as a consequence of an unbalanced use of proxy respondents across the different types of stroke.

Conclusions These results suggest that the benefits of using proxy ratings for noncommunicative patients outweigh their limitations. The findings stress the need for inclusion of this important subgroup of patients in QL studies. Their significant others are able to provide useful information on these patients' QL. (***Stroke.* 1997;28:1541-1549.**)

Key Words • cerebrovascular disorders • quality of life • stroke assessment • stroke outcome

Several studies have shown that many stroke survivors experience a decline in their QL in terms of impaired physical, functional, psychological, and social health.[1-7] QL is most often assessed by means of either structured interviews or written questionnaires. However, it has been recognized that these methods of data collection are not always suitable for studies of stroke survivors.[8] Given the frequency of serious cognitive, speech, and language disorders, many patients are not able to communicate effectively or to understand what is being asked. The inability of a highly relevant subgroup of patients to participate in such studies may yield results that cannot be generalized to the total patient population of interest.

The use of family members or caregivers as alternative sources of information (ie, proxy respondents) may help to resolve the problem of excluding patients with limited self-reporting capabilities. Although the use of proxy respondents is common in epidemiological research[9] and health studies among elderly populations,[10,11] surprisingly little is known about this method of data collection in stroke research. In a study of emotional and personality changes following stroke, Nelson et al[12] relied solely on information obtained from family members or close companions. In two recent QL studies among stroke survivors,[1,2] proxy respondents were used for patients who were not able to communicate because of severe language or cognitive disturbances. De Haan et al[1] reported that proxy respondents (most often the partner) were used for 25% of their patient sample, thereby increasing the number of available patients at 6 months after stroke from 329 to 441. However, there are limitations of such proxy reports of patients' QL, and their impact on study outcomes is not well documented.

Although the use of proxies may be an effective means of obtaining information that might otherwise be lost, it assumes that the proxy can report accurately on several aspects of the patient's health and QL. The accuracy of proxy reports is most typically determined by examining the extent to which proxy ratings are in agreement with those provided by the patients themselves. To our knowledge, agreement between stroke survivors and their proxies has been examined in only one small study (n=38).[13] This study showed good agreement for two instruments measuring frequency of and independence in performing several activities but lower levels of agreement for an instrument assessing perceived limitations in such activities. Evidence from several other studies performed in the elderly and populations of patients

Received February 25, 1997; final revision received May 12, 1997; accepted May 13, 1997.

From the Division of Psychosocial Research and Epidemiology (K.C.A.S., N.K.A.), The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam; and the Departments of Clinical Epidemiology and Biostatistics (R.J. de H.) and Neurology (M.L.), Academic Medical Center, Amsterdam, The Netherlands.

Reprint requests to Neil K. Aaronson, PhD, Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, Netherlands. E-mail naaron@nki.nl

© 1997 American Heart Association, Inc.

Selected Abbreviations and Acronyms

ADL = activities of daily living
CT = computed tomography
ICC = intraclass correlation coefficient
QL = quality of life
SIP = Sickness Impact Profile

with chronic disease suggests that the response agreement between patients and proxies is far from optimal and that the use of proxy respondents may introduce considerable bias.[14]

By definition, studies comparing patient and proxy ratings can be carried out only among communicative patients. Thus, it is necessary to extrapolate findings of patient-proxy agreement for communicative patients to the subgroup of noncommunicative patients. This can be facilitated by examining trends in the level of agreement as a function of patients' health status.[14] When properly analyzed, examination of patient-proxy agreement can provide important information on the extent and direction of any bias introduced by using proxy respondents. In the current study, we examined, in a subgroup of communicative patients, the level of response agreement between stroke survivors and their significant others (ie, family members and close companions) on a standardized QL questionnaire. To facilitate extrapolation of results to the subgroup of noncommunicative patients, this analysis included determining whether agreement varied across the range of questionnaire scores.

Moreover, the validity of proxy QL ratings was examined by exploring the relationship between QL scores and the level of patients' functioning as indicated by the well-known Rankin scale.[15,16] This scale is a frequently used handicap index in stroke outcome research that can be viewed as a global functional health index with a strong emphasis on physical disability.[15] The validity of proxy QL scores would be supported by a substantial association between these scores and the patients' Rankin grade. In contrast to the former type of analysis (ie, patient-proxy agreement), this analysis can give a direct indication of the clinical validity of proxy ratings for noncommunicative patients.

In addition to evaluating the value and limitations of proxy ratings, we estimated the impact of using proxy QL ratings for noncommunicative patients on the results of the original QL study from which the data for this article were derived.[1] Based on a combined analysis of self-reported and proxy-derived QL data, this earlier QL study reported a significant relationship between stroke type and QL. Specifically, patients with infratentorial strokes were found to have less QL impairment than patients with supratentorial strokes. However, among the patients with supratentorial strokes, those with lacunar infarctions exhibited significantly less QL impairment than patients with cortical or subcortical infarctions and hemorrhages. Thus, it was concluded that patients with lacunar infarctions and infratentorial strokes exhibit less QL impairment than survivors of larger supratentorial strokes (ie, cortical or subcortical infarctions and hemorrhages). In the current analysis, based on a strategy suggested by Semaan[17] and Nelson et al,[9] we examined the relationship between stroke type and QL with and without substitution of proxy ratings for noncommunicative patients. As in the original study, we first analyzed the combined data, substituting proxy ratings for noncommunicative patients. Subsequently, we performed separate stratified analyses of the same relationship among communicative and noncommunicative subgroups of patients.

Subjects and Methods

Subjects

The study sample consisted of patients who had suffered a stroke 6 months earlier. They were the survivors of an original cohort of 760 consecutively admitted stroke patients who had participated in a multicenter quality of care study in the Netherlands. Two hundred fifty-eight patients died after the stroke, yielding a 6-month mortality rate of 34%. Of the remaining 502 eligible patients, 17 declined to participate. Of the 485 consenting patients, 112 were not able to communicate because of severe speech, language, or cognitive disorders. For these patients, data were obtained from proxy respondents. In the current analyses, 4 noncommunicative patients with a healthcare provider as proxy respondent were excluded to limit the proxy sample to the patients' family members and close companions. For 44 of the 373 communicative patients, no QL data were available because the interview was unacceptably lengthy and burdensome to the patients. For 228 of the remaining 329 communicative patients, both self-reported and proxy ratings were obtained. Thus, the patient sample could be divided into three subgroups: communicative patients with self-reported data only (n=101), communicative patients with self-reported and proxy-derived data (n=228), and noncommunicative patients with proxy ratings only (n=108).

Measures

Patients' sociodemographic and clinical information was obtained from the medical and nursing charts by trained research assistants shortly after discharge from the hospital. This information included age, gender, history of stroke, stroke type, and lesion location. CT data were available for 430 patients (98%). The scans were made within 2 weeks after the stroke and were evaluated by local radiologists. Stroke types were divided into supratentorial strokes (subcortical, cortical, and lacunar infarctions and hemorrhages) and infratentorial strokes. A hemorrhage was considered to be present if the CT scan showed evidence of a recent intracerebral hemorrhage. The diagnosis of lacunar stroke was made if there was a clinical picture of a lacunar syndrome and the CT scan was compatible with that diagnosis.[18]

The patients were interviewed 6 months after their stroke. Patients' handicaps and disabilities in ADL were assessed by means of the modified Rankin scale[15,16] and the Barthel Index.[19] The Rankin scale is a six-point index that ranges from no symptoms (grade 0) to severe handicap (grade 5). The Barthel Index consists of 10 items: continence of bowels, continence of bladder, grooming, toilet use, feeding, transfer, mobility, dressing, climbing stairs, and bathing. Scores can range from 0 to 20, with a higher score indicating more ADL independence.

QL was assessed with the SIP.[20] The SIP is a widely used, reliable, and valid health status questionnaire that addresses a wide range of health-related QL domains, with a focus on behavior rather than subjective expressions.[8,21] The questionnaire, which consists of 136 yes/no statements describing limitations or recent changes in functioning, is organized into 12 subscales: sleep and rest, emotional behavior, body care and movement, household management, mobility, social interaction, ambulation, alertness behavior, communication, work, recreation and pastimes, and eating. Respondents are asked to endorse items that apply to them on a given day because of ill health. The SIP is scored by summing the weights attached to all endorsed statements and is expressed as a percentage of the

maximum possible score. Thus, scores can range from 0 to 100, with a higher score representing more impaired QL. An aggregated score can be obtained for the total SIP as well as for physical and psychosocial dimensions. The subscales that make up the physical dimension are body care and movement, mobility, and ambulation. The subscales included in the psychosocial dimension are emotional behavior, social interaction, alertness behavior, and communication. Because most of the patients were retired at the time of their stroke, the work subscale (10 items) was excluded from all analyses.

Proxies completed a slightly modified version of the SIP in which the wording of individual items was changed so that each item clearly referred to the patient. Proxies were instructed to endorse only those statements that they were sure applied to the patient on a given day because of ill health. The proxies were also asked to provide information on their own age and gender and the nature of their relationship with the patient.

Data Analysis

Differences were examined between the three subgroups of patients (communicative patients with self-reported data only, communicative patients with self-reported and proxy-derived data, and noncommunicative patients with proxy ratings only) in terms of stroke type, lesion location, history of stroke, Rankin and Barthel Index scores, age, and gender. To determine the extent to which these patient characteristics were associated with the need to use proxy respondents, differences between communicative (subgroups 1 and 2 combined) and noncommunicative (subgroup 3) patients were tested with χ^2 tests. To examine whether the patients used in the direct patient-proxy agreement analyses were representative of the total communicative patient sample, subgroups 1 and 2 were compared as well.

Patient-proxy agreement was evaluated in the 228 communicative patients with both self-reported and proxy-derived SIP data (subgroup 2). For both sources of SIP data, mean scores, standard deviations, and the range of scores on the specific SIP scales were calculated. Internal consistency reliability for both patient and proxy scores, as indicated by Cronbach's coefficient α,[22] was established as a frame of reference for interpreting patient-proxy agreement. This is relevant because the level of agreement between patient and proxy ratings is dependent, in part, on the reliability of the instruments used.[9,23]

Three analytic strategies were used to examine patient-proxy agreement. First, agreement at the group level was evaluated by comparison of group means. For each SIP scale, we calculated the mean difference score between patient and proxy ratings (proxy minus patient score). Mean difference scores significantly different from zero, using paired Student's t tests, were interpreted as providing evidence of systematic bias.[24,25] To examine the statistical magnitude of any observed systematic bias, the mean difference score was standardized by relating this score to its standard deviation. Given the similarity to effect size (d) calculations for paired observations,[26] a standardized difference of $d=0.2$ was taken to indicate a small bias; $d=0.5$, a moderate bias; and $d=0.8$, a large bias.

Second, to determine whether agreement varied across the range of SIP scores, the pattern of agreement was visually examined by means of a scatterplot. That is, for each patient, the difference between the patient and proxy scores (proxy minus patient score) was plotted against the average for each pair of scores (patient plus proxy score divided by 2).[23,24] When depicted graphically, using the y axis to show difference scores and the x axis to show average scores, perfect correspondence would be represented by a horizontal line through an ordinate of zero. Any observed differences between the patient and proxy scores as a function of the range of their average scores (eg, comparable scores at high levels of functioning but diverging scores at low levels of functioning) were taken as evidence of scatter bias.[25]

Third, the ICC was used as an indicator of chance-corrected agreement between patient and proxy ratings at the individual patient level.[24,27] Guidelines for the ICC as a measure of the strength of agreement were labeled as follows: ≤0.40, poor to fair agreement; 0.41 through 0.60, moderate agreement; 0.61 through 0.80, good agreement; and 0.81 through 1.00, excellent agreement.[28]

The clinical validity of proxy QL ratings was determined by examining differences in mean physical, psychosocial, and total SIP scores between patients with different Rankin grades. ANOVA was used for statistical testing of the differences. The relationship between proxy-derived SIP scores and patients' Rankin grades was examined in two patient subgroups, communicative patients with both self-reported and proxy-derived SIP data (subgroup 2) and noncommunicative patients with proxy ratings only (subgroup 3). The availability of both self-reported and proxy-derived SIP data in subgroup 2 also allowed for a head-to-head comparison of the between-group differences in SIP scores based on patient and proxy ratings in one sample.

The effect of using proxy data for noncommunicative patients on the observed association between stroke type and QL was examined by performing both combined and stratified analyses of this relationship. Specifically, we examined differences in mean total SIP scores between patients with cortical or subcortical infarctions, intracerebral hemorrhages, lacunar infarctions, and infratentorial strokes. Statistical testing for differences between the four stroke types was conducted as follows. ANOVA was applied to test whether the mean total SIP score of at least one group (ie, stroke type) differed from one other group. Additionally, the Newman-Keuls procedure was used to test the statistical significance of any observed differences between each pair of groups.[29] The between-group differences were tested for all patients combined. as well as for the communicative and noncommunicative patients separately. Within the subsample of communicative patients with both self-reported and proxy-derived SIP data (subgroup 2), we also made a head-to-head comparison of the between-group differences in total SIP scores based on patient and proxy ratings in one sample.

Results

Study Sample

The characteristics of the patient sample are summarized in Table 1, both for the total study sample and broken down by the source of information on the SIP. Three hundred thirty-one patients had suffered a supratentorial stroke (200 subcortical and cortical infarctions, 49 intracerebral hemorrhages, and 82 lacunar infarctions), and 61, an infratentorial stroke. For 45 patients the stroke type was unknown or incompletely described. In terms of lesion laterality, 172 patients had right-hemisphere and 191 patients had left-hemisphere lesions (for 74 patients, lesion laterality was undetermined, or they had infratentorial strokes). For 65% of the patients, it was their first stroke. Moderate to severe handicaps (Rankin grades 3 to 5) were noted in 59% of the patients, but 77% of all patients were nevertheless considerably ADL independent (Barthel Index scores of 15 to 20). Fifty-nine percent of the patients were older than 70 years of age (mean age, 70 years; range, 20 to 94 years), and 55% were male.

As expected, the characteristics of the noncommunicative patients (n=108) differed significantly from those of the communicative patients (n=329, subgroups 1 and 2 combined). Noncommunicative patients more often had supratentorial cortical or subcortical infarctions and hemorrhages ($P<.001$), left-hemisphere lesions ($P=.01$),

TABLE 1. Characteristics of Stroke Patients in Relation to Availability of Patient Self-Reported and/or Proxy-Derived SIP Data

		Communicative Patients		Noncommunicative Patients
	Total Sample (n=437)	Patient Data Only (n=101)	Patient and Proxy Data (n=228)	Proxy Data Only (n=108)
Stroke type (n=392)*				
Supratentorial stroke				
Cortical or subcortical infarction	200 (51%)	36 (41%)	98 (48%)	66 (66%)
Intracerebral hemorrhage	49 (12%)	8 (9%)	23 (11%)	18 (18%)
Lacunar infarction	82 (21%)	25 (29%)	48 (23%)	9 (9%)
Infratentorial stroke	61 (16%)	18 (21%)	36 (18%)	7 (7%)
Lesion laterality (n=363)†				
Right-hemisphere	172 (47%)	42 (52%)	94 (51%)	36 (37%)
Left-hemisphere	191 (53%)	38 (48%)	91 (49%)	62 (63%)
Previous stroke (n=420)‡				
No	274 (65%)	68 (70%)	140 (64%)	66 (63%)
Yes	146 (35%)	29 (30%)	78 (36%)	39 (37%)
Rankin grade (n=434)‡				
0-2	180 (41%)	46 (46%)	127 (56%)	7 (7%)
3-5	254 (59%)	55 (54%)	100 (44%)	99 (93%)
Barthel Index score (n=435)‡				
15-20	337 (77%)	89 (89%)	198 (87%)	50 (47%)
0-14	98 (23%)	11 (11%)	30 (13%)	57 (53%)
Age (n=437)				
20-70 y	180 (41%)	42 (42%)	109 (48%)	29 (27%)
>70 y	257 (59%)	59 (58%)	119 (52%)	79 (73%)
Sex (n=437)				
Male	239 (55%)	47 (47%)	135 (59%)	57 (53%)
Female	198 (45%)	54 (53%)	93 (41%)	51 (47%)

n indicates number of patients.
*For 45 patients stroke type was unknown or incompletely described.
†For 74 patients lesion laterality was undetermined or the strokes were infratentorial.
‡Because of missing data, n varies.

and moderate to severe handicaps ($P<.001$) were more frequently ADL dependent ($P<.001$), and were older ($P<.001$) than the communicative patients. Among the communicative patients, no significant differences were noted between the subgroup of patients for whom both patient and proxy SIP data were obtained (n=228) and those with self-reported SIP data only (n=101).

The proxy respondents most often were the patient's spouse or partner (65%), with the remainder being family members or friends. The mean age of the proxies was 61 years (range, 18 to 90 years), and 73% were female.

Patient-Proxy Agreement

Agreement between patient and proxy SIP scores was examined in the subgroup of 228 patients for whom both sources of information were available. Table 2 presents patient and proxy mean scores and reliability coefficients for 11 of the 12 SIP subscales (work was excluded as mentioned above), the physical and psychosocial dimensions, and the total SIP. For most subscales and dimensions, there was substantial variation in scores, with both the patient and proxy scores spanning a relatively large segment of the possible range of scores. However, for the total SIP, the observed scores in this subgroup of communicative patients covered only half the possible score range (0 to 50.3 and 0 to 58.3 for patients and proxies, respectively). Internal consistency reliabilities surpassed or approached the .70 criterion for group-level comparisons[30] for 9 of the 11 SIP subscales. Based on both patient and proxy ratings, the internal consistency of the sleep/rest and eating subscales was relatively poor. High-reliability estimates were noted for the physical and psychosocial dimensions as well as for the total SIP score.

Differences between patient and proxy mean scores were statistically significant for 7 of the 11 SIP subscales, the physical and psychosocial dimensions, and the total SIP score (Table 3). Except for the eating subscale, the mean differences were all in the same direction, with the proxies rating the patients as having more functional limitations than the patients themselves. The magnitude of the differences between patient and proxy mean scores was low to moderate ($d=-.04$ to $.45$).

Evidence of scatter bias was found for 6 SIP subscales, the physical and psychosocial dimensions, and the total SIP. Specifically, the tendency of proxies to rate patients as having more functional limitations than the patients themselves was most notable among patients with more impaired levels of functioning. Fig 1 illustrates this pattern of bias for the total SIP score. This plot shows that both the magnitude and direction of the differences between patient and proxy total SIP scores are dependent on the patients' level of functioning. Thus, although the overall bias for the total SIP score is moderate ($d=0.44$), this bias is much more pronounced at more impaired levels of functioning.

At the individual patient level, chance-corrected agreement between patient and proxy scores for the SIP subscales ranged from moderate for the eating scale

TABLE 2. Distribution and Reliability* of Patient and Proxy SIP Scores† in the Subgroup of Communicative Patients With Both Self-Reported and Proxy-Derived SIP Data Available (n=228)‡

	No. of Items	Patient		Proxy		Patient	Proxy
		Mean±SD	Range	Mean±SD	Range	∝	∝
SIP subscale§							
Sleep and rest	7	18.2±17.8	0-90.2	19.7±17.0	0-73.7	0.56	0.45
Emotional behavior	9	13.8±16.8	0-81.3	16.5±17.4	0-72.5	0.68	0.64
Body care and movement	23	16.1±16.1	0-67.8	19.8±19.5	0-72.8	0.85	0.88
Household management	10	32.5±24.7	0-100.0	39.1±29.2	0-100.0	0.74	0.81
Mobility	10	19.1±18.9	0-85.3	24.2±22.0	0-100.0	0.74	0.79
Social interaction	19	12.5±12.8	0-61.3	18.7±16.7	0-75.1	0.74	0.79
Ambulation	12	25.1±17.3	0-71.1	25.2±21.3	0-90.1	0.65	0.76
Alertness behavior	10	21.1±25.9	0-100.0	27.8±28.2	0-100.0	0.85	0.86
Communication	9	12.6±16.0	0-74.5	14.2±16.7	0-100.0	0.67	0.68
Recreation and pastimes	8	31.8±21.8	0-100.0	37.0±26.4	0-100.0	0.61	0.71
Eating	9	7.7±8.0	0-46.2	7.3±10.2	0-100.0	0.36	0.59
SIP dimension							
Physical	45	18.9±15.1	0-64.3	22.1±18.6	0-72.7	0.90	0.93
Psychosocial	47	14.7±12.9	0-55.9	19.2±14.7	0-73.1	0.88	0.89
Total SIP score§	126	17.9±11.8	0-50.3	21.5±14.2	0-58.3	0.94	0.95

*Internal consistency, Cronbach's coefficient ∝.
†SIP scores range from 0 to 100, with a higher score representing a more impaired QL.
‡Because of missing data, n varies from 220 to 228.
§Work subscale not included.

(ICC=.47) to excellent for the ambulation (ICC=.80) and body care and movement scales (ICC=.82). Good to excellent agreement was noted for the physical (ICC=.85) and psychosocial (ICC=.61) dimension scores and for the total SIP score (ICC=.77).

Clinical Validity of Proxy Ratings

The relationship between patient and proxy SIP scores and patients' Rankin grades is depicted in Table 4. The first two columns show a head-to-head comparison of the patient and proxy SIP scores in the same sample of communicative patients (n=228). The mean patient- and proxy-rated SIP scores on the physical and psychosocial dimensions, as well as the total SIP, were significantly associated with the patients' Rankin grade. As expected, based on the results described above, the mean proxy SIP scores were systematically higher than the mean patient SIP scores, with this difference being most pronounced for patients with Rankin grade 4 (grade 5 was not observed among the communicative patients). Importantly, the third column of Table 4 shows that the mean physical, psychosocial, and total SIP scores for noncommunicative patients, based on proxy ratings, were also significantly associated with the Rankin grade (ranging from grade 2 to 5 among these patients). The between-group differences in mean scores were larger for the physical dimension than for the psychosocial dimension, as expected given the emphasis of the Rankin scale on physical disability.

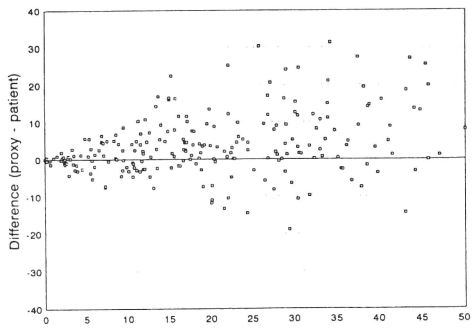
TABLE 3. Agreement Between Patient and Proxy SIP Scores in the Subgroup of Communicative Patients With Both Self-Reported and Proxy-Derived SIP Data Available (n=228)

	Difference (Proxy−Patient)		
	Mean±SD	*d*†	ICC
SIP subscale‡			
Sleep and rest	1.5±16.1	.09	.57
Emotional behavior	2.7±15.4*	.18	.59
Body care and movement	3.7±10.3*	.36	.82
Household management	6.6±20.7*	.32	.69
Mobility	5.1±16.5*	.31	.66
Social interaction	6.2±13.9*	.45	.52
Ambulation	0.1±12.2	.01	.80
Alertness behavior	6.7±24.2*	.28	.59
Communication	1.6±14.5	.11	.60
Recreation and pastimes	5.2±21.3*	.24	.60
Eating	−0.4±9.4	−.04	.47
SIP dimension			
Physical	3.2±8.7*	.37	.85
Psychosocial	4.5±11.6*	.39	.61
Total SIP score‡	3.6±8.2*	.44	.77

*Statistically significant difference between patient and proxy score ($P<.05$).
†Standardized difference d=mean difference/SD of difference (d=.2, small bias; d=.5, moderate bias; and d=.8, large bias).
‡Work subscale not included.

Impact of Proxy Data on Study Results

Table 5 shows the results of both the combined and stratified analyses of the association between stroke



Differences between patient and proxy total SIP scores against their averages.

TABLE 4. Relationship Between Patient and Proxy SIP Scores and Rankin Grade

	Communicative Patients		Noncommunicative Patients
	Patient Score Mean±SD (n)	Proxy Score Mean±SD (n)	Proxy Score Mean±SD (n)
Physical SIP dimension			
Rankin grade			
0 (no symptoms)	4.2±6.3 (13)	3.2±5.3 (13)	. . .
1 (minor symptoms)	6.9±6.8 (26)	9.4±14.4 (26)	. . .
2 (minor handicap)	10.8±8.4 (88)	12.5±10.0 (88)	7.1±4.9 (7)
3 (moderate handicap)	27.0±12.5 (67)	30.8±15.0 (67)	23.4±14.2 (31)
4 (moderately severe handicap)	39.5±9.5 (33)	47.9±12.2 (33)	44.8±13.4 (50)
5 (severe handicap)	. . .	. . .	58.6±11.2 (18)
P (ANOVA)	<.001	<.001	<.001
Psychosocial SIP dimension			
Rankin grade			
0 (no symptoms)	2.5±3.6 (13)	5.3±8.0 (13)	. . .
1 (minor symptoms)	6.0±6.5 (26)	10.5±13.3 (26)	. . .
2 (minor handicap)	13.4±12.0 (88)	18.5±13.9 (88)	13.0±10.8 (7)
3 (moderate handicap)	20.4±13.3 (67)	23.4±14.3 (67)	27.4±19.5 (31)
4 (moderately severe handicap)	19.2±13.5 (33)	26.0±13.6 (33)	34.3±15.1 (50)
5 (severe handicap)	. . .	. . .	37.3±12.9 (18)
P (ANOVA)	<.001	<.001	.003
Total SIP*			
Rankin grade			
0 (no symptoms)	4.1±4.1 (13)	5.0±5.4 (13)	. . .
1 (minor symptoms)	7.5±6.3 (26)	11.2±12.5 (26)	. . .
2 (minor handicap)	13.1±8.0 (88)	16.4±10.0 (88)	10.6±6.6 (7)
3 (moderate handicap)	25.2±9.4 (67)	28.4±11.3 (67)	26.3±10.5 (31)
4 (moderately severe handicap)	30.7±9.1 (33)	37.7±10.8 (33)	40.5±10.0 (50)
5 (severe handicap)	. . .	. . .	49.6±9.4 (18)
P (ANOVA)	<.001	<.001	<.001

n indicates number of patients; for 3 patients the Rankin grade was unknown.
*Work subscale not included.

type and the total SIP score. Overall, the stratified analyses demonstrated substantial differences between communicative and noncommunicative patients in their level of QL impairment, with the mean total SIP score of the noncommunicative patients (36.0±14.3) being almost twice as high as that of the communicative patients (18.6±11.8). Because the results described above indicated that for the communicative patients proxy-derived SIP scores were systematically higher than patient-derived SIP scores and that this bias was dependent on the patients' level of functioning, the observed difference may partly be due

TABLE 5. Combined and Stratified Analysis of Relationship Between Stroke Type and Total SIP Score

	Combined	Stratified	
	All Patients Patient or Proxy Score* Mean±SD (n)	Communicative Patients Patient Score Mean±SD (n)	Noncommunicative Patients Proxy Score Mean±SD (n)
All stroke types	22.9±14.5 (437)	18.6±11.8 (329)	36.0±14.3 (108)†
Cortical or subcortical infarction (A)	25.8±14.4 (200)	21.2±12.4 (134)	35.0±13.8 (66)
Intracerebral hemorrhage (B)	26.0±17.1 (49)	17.2±12.0 (31)	41.1±13.8 (18)
Lacunar infarction (C)	17.9±13.1 (82)	15.5±10.8 (73)	38.1±13.0 (9)
Infratentorial stroke (D)	19.3±12.7 (61)	18.2±10.8 (54)	29.2±23.3 (7)
P (ANOVA)	<.001	.007	.26
Between-group differences	A,B≠C,D	A≠C	

n indicates number of patients.
*Patient scores for communicative patients and proxy scores for noncommunicative patients combined.
†One might wish to adjust the raw proxy scores for systematic bias related to the source of information. In the current study, we calculated such adjusted proxy scores by means of a linear regression equation. This equation was based on regression of patient-derived total SIP scores on proxy-derived total SIP scores and Rankin grades in the subgroup of 228 patients with both sources of SIP data available: patient score=−0.94+(0.52×proxy score)+(3.28×Rankin grade); explained variance R^2=0.71. The resulting regression equation was used to adjust the raw proxy scores for the subgroup of patients when only proxy data were available. This strategy yielded a mean total SIP score of 30.0±9.6 for the 108 noncommunicative patients (as compared with 36.0±14.3 using raw proxy ratings). The observed relationship between stroke type and the total SIP score was not changed by the use of adjusted instead of raw proxy ratings.

TABLE 6. Relationship Between Stroke Type and Total SIP Score in the Subgroup of Communicative Patients With Both Self-Reported and Proxy-Derived SIP Data Available (n=228)

	Patient Score Mean±SD (n)	Proxy Score Mean±SD (n)
All stroke types	17.9±11.8 (228)	21.5±14.2 (228)
Cortical or subcortical infarction (A)	20.6±12.6 (98)	24.4±15.1 (98)
Intracerebral hemorrhage (B)	17.4±11.3 (23)	19.4±11.4 (23)
Lacunar infarction (C)	14.1±10.3 (48)	18.8±12.6 (48)
Infratentorial stroke (D)	18.0±10.9 (36)	21.2±13.8 (36)
P (ANOVA)	.02	.05
Between-group differences	A≠C	A≠C

n indicates number of patients.

to the source of SIP information (see footnote in Table 5).

Combined analysis of patient-derived scores for communicative patients and proxy-derived scores for noncommunicative patients yielded significantly higher total SIP scores (ie, more QL impairment) for patients with supratentorial cortical or subcortical infarctions and hemorrhages than for patients with lacunar infarctions and infratentorial strokes. When performing stratified analyses, this pattern of results could not be confirmed. Among the communicative patients, the highest mean total SIP score was observed for patients with cortical or subcortical infarctions, and the lowest for patients with lacunar infarctions, with intermediate scores for patients with intracerebral hemorrhages and infratentorial strokes. A statistically significant between-group difference was noted between patients with cortical or subcortical and lacunar infarctions only. For noncommunicative patients, the highest mean score was observed for patients with intracerebral hemorrhages, and the lowest for patients with infratentorial strokes, but no statistically significant differences were found between the four stroke types.

Table 6 shows the distribution of mean total SIP scores across the four stroke types within the subsample of communicative patients for whom both self-reported and proxy-derived SIP data were available (n=228). Although the magnitude of the patient and proxy scores differed, the pattern of mean scores across the stroke types was similar for self-reported and proxy-derived SIP scores. Regardless of the source of SIP data, statistically significant differences were noted between patients with cortical, subcortical, or lacunar infarctions, with intermediate levels of QL impairment for the remaining stroke types. This suggests that the use of proxies would not change the observed relationship between stroke type and the total SIP score.

Discussion

The primary aim of the current study was to examine the value and limitations of proxy ratings in evaluating the QL of stroke survivors. Proxy ratings of patients' QL may be needed for patients with cognitive or communication disorders who would otherwise be excluded from study participation. Exclusion of such patients can compromise the validity and generalizability of study outcomes. However, this type of selection bias must be weighed against the bias that may be introduced when significant others are used as proxy respondents for noncommunicative patients.

Comparison of the patient characteristics of the noncommunicative patients, representing one quarter of the total sample, with those of the communicative patients yielded several important differences. Not surprisingly, the noncommunicative patients more often had supratentorial cortical or subcortical infarctions and hemorrhages and left-hemisphere lesions and were more severely handicapped, more frequently ADL dependent, and older than the communicative patients. This finding supports empirically the argument that patients who are unable to provide self-reported data should not be excluded from QL assessments, as has been the case in several earlier studies (eg, Niemi et al[6] and Viitanen et al[7]).

The quality of the proxy ratings was evaluated by comparing patient and proxy responses to the SIP in a subsample of patients for whom both sources of information were available. This patient subgroup, including more than two thirds of the communicative patients, was representative of the total communicative patient sample with respect to a number of relevant clinical and sociodemographic characteristics. Importantly, this subsample of patients was highly heterogeneous in terms of stroke type, disease severity, and QL, thereby facilitating the examination of trends in the pattern of patient-proxy agreement across the range of SIP scores. In turn, this enabled us to estimate the potential degree of bias introduced by using proxy respondents for the noncommunicative patients.

The comparative analyses of the patient self-reported and proxy-derived SIP data yielded somewhat conflicting results. Correlations between the patient and proxy ratings on the SIP subscales ranged from moderate to excellent. For some subscales (ie, sleep and rest, eating), the lower correlations may be attributed, in part, to lack of scale reliability. The patient-proxy correlation was high for the aggregate physical dimension score and moderate for the aggregate psychosocial dimension score. This finding is at odds with the observations of an earlier study using the SIP,[31] in which the correlations between patients and their significant others were high for the physical dimension but poor for the psychosocial dimension. In this latter study, among more severely impaired patients (ie, as indicated by SIP scores), the psychosocial dimension score was found to be more closely associated with the proxy's own level of psychological distress and caregiver burden than with the patient's psychosocial health. Finally, for the total SIP score, we also observed a fairly strong correlation between patient and proxy ratings. Again, the observed correlation was higher than that reported in a study of more severely impaired patients (ie, chronically or terminally ill homebound patients) and their caregivers.[32]

Encouraging results were also noted when we evaluated the comparability of mean scores at the group level. Although the proxies systematically rated patients as having more functional impairments than the patients themselves (a finding in line with a consistent trend in the proxy literature),[10,14] these differences were relatively small in magnitude. This suggests that, at the group level, only a modest degree of bias would be introduced when substituting patients' self-report of their QL by ratings provided by significant others.

However, although overall agreement was generally quite high when viewed in terms of correlations and mean differences, the magnitude of patient-proxy agreement was found to be clearly associated with the patients' level of functioning. The tendency of proxies to rate patients as having more limitations than patients themselves was most pronounced among patients with more impaired levels of functioning. Because our patient sample was not as functionally impaired as those of earlier studies using the SIP,[31,32] this finding may explain the higher rates of agreement observed in our patient sample as compared with those reported in the latter investigations. In line with an earlier study among patients with brain cancer,[33] extrapolation of these findings suggests that lower levels of agreement and more biased ratings can be expected among noncommunicative patients.

Additionally, the validity of proxy QL ratings was determined by examining the association between the SIP scores and the level of patients' functional health as indicated by the Rankin scale. The results provided support for the clinical validity of the proxy QL ratings. Although the proxy SIP scores were consistently higher than the patients' own scores (being in line with the results described above), the proxy ratings were clearly sensitive to differences in patients' functional health. This was also true for the proxy SIP scores for noncommunicative patients, giving a direct indication of the validity of proxy QL ratings for those patients for whom proxies are really needed.

To illustrate the potential effects of combining patient and proxy ratings of patients' QL on relevant study outcomes, we performed a more detailed analysis of the relationship between stroke type and QL. The results of analyses that combined patient SIP scores for communicative patients with proxy SIP scores for noncommunicative patients indicated more QL impairment for patients with larger supratentorial strokes (ie, cortical or subcortical infarctions and hemorrhages) as compared with patients with lacunar infarctions and infratentorial strokes. When limiting the analysis to the communicative patient group, only the difference between cortical or subcortical and lacunar infarctions could be confirmed. This suggests that the other apparent differences in the combined analysis were due mainly to differences among the noncommunicative patients who were rated by proxy respondents. In turn, the observed differences in the combined analysis might be attributed, to some extent, to the more frequent use of proxy respondents for patients with larger supratentorial strokes. As suggested by our findings, the level of QL impairment of the latter patients may be overestimated because of the reliance on proxy ratings for a substantial percentage of these patients.

There are, however, several reasons to believe that the observed differences in levels of QL impairment across the four stroke types represent real differences in QL rather than an artifact of an unequal distribution of noncommunicative patients and, consequently, an unbalanced use of proxy respondents. First, patients with larger supratentorial strokes are likely to have more severe cerebral dysfunction. It can logically be expected that these patients also experience a more impaired QL than patients with lacunar and infratentorial strokes. Second, within the patient group for whom both self-reported and proxy ratings were available, the pattern of mean QL scores across the stroke types was not dependent on the source of information used. Although this subgroup of patients is not representative of the total patient sample, it suggests that the observed relationship between stroke type and QL would not be altered by the use of proxy-derived information. Together, these findings suggest that the use of proxy respondents for the noncommunicative patients did not affect the observed relationship between stroke type and QL.

In conclusion, although this study provides encouraging results on the validity of proxy ratings of patients' QL, researchers need to exercise some caution in interpreting their data when using proxy respondents for noncommunicative patients. Most likely, proxies will overrate the level of QL impairments for these patients. However, the impact on the study outcomes is not likely to be of clinical significance. Furthermore, the current findings stress the need for inclusion of this important subgroup of noncommunicative patients in QL investigations. If the analyses of the larger QL study[1] had been based on the communicative patients only, the level of QL impairment would have been underestimated, and, perhaps more importantly, questionable conclusions would have been drawn about the impact of several stroke types on the patients' QL. These findings indicate that noncommunicative patients should not be excluded from study participation and that their significant others can be used, with the necessary caution, as proxies to rate these patients' QL.

Acknowledgments

This study was supported by grants from The Netherlands Heart Foundation (NHS 40.004), Developmental Medicine (OG 1991-037), and the Dutch Cancer Society (NKI 93-139). Dr Limburg is a Clinical Investigator of The Netherlands Heart Foundation.

References

1. De Haan RJ, Limburg M, van der Meulen JHP, Jacobs HM, Aaronson NK. Quality of life after stroke: impact of stroke type and lesion location. *Stroke*. 1995;26:402-408.
2. Tuomilehto J, Nuottimäki T, Salmi K, et al. Psychosocial and health status in stroke survivors after 14 years. *Stroke*. 1995;26:971-975.
3. Angeleri F, Angeleri VA, Foschi N, Giaquinto S, Nolfe G. The influence of depression, social activity, and family stress on functional outcome after stroke. *Stroke*. 1993;24:1478-1483.
4. Johansson BB, Jadbäck G, Norrving B, Widner H, Wiklund I. Evaluation of long-term functional status in first-ever stroke patients in a defined population. *Scand J Rehab Med*. 1992; 26(suppl):105-114.
5. Åström M, Asplund K, Åström T. Psychosocial function and life satisfaction after stroke. *Stroke*. 1992;23:527-531.
6. Niemi ML, Laaksonen R, Kotila M, Waltimo O. Quality of life 4 years after stroke. *Stroke*. 1988;19:1101-1107.
7. Viitanen M, Fugl-Meyer KS, Bernspång B, Fugl-Meyer AR. Life satisfaction in long-term survivors after stroke. *Scand J Rehab Med*. 1988;20:17-24.
8. De Haan R, Aaronson NK, Limburg M, Hewer RL, Van Crevel H. Measuring quality of life in stroke. *Stroke*. 1993;24:320-327.
9. Nelson LM, Longstreth WT Jr, Koepsell TD, Van Belle G. Proxy respondents in epidemiologic research. *Epidemiol Rev*. 1990;12: 71-86.
10. Magaziner J. The use of proxy respondents in health studies of the aged. In: Wallace RB, Woolson RF, eds. *The Epidemiologic Study of the Elderly*. New York, NY: Oxford University Press; 1992:120-129.
11. Corder LS, Woodbury MA, Manton KG. Proxy response patterns among the aged: effects on estimates of health status and medical

care utilization from the 1982-1984 long-term care surveys. *J Clin Epidemiol*. 1996;49:173-182.
12. Nelson LD, Cicchetti D, Satz P, et al. Emotional sequelae of stroke. *Neuropsychology*. 1993;7:553-560.
13. Segal ME, Schall RR. Determining functional/health status and its relation to disability in stroke survivors. *Stroke*. 1994;25:2391-2397.
14. Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *J Clin Epidemiol*. 1992;45:743-760.
15. De Haan R, Limburg M, Bossuyt P, van der Meulen J, Aaronson N. The clinical meaning of Rankin "handicap' grades after stroke. *Stroke*. 1995;26:2027-2030.
16. Bamford JM, Sandercock PA, Warlow CP, Slattery J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1989;20:828. Letter.
17. Semaan S. Impact of proxy-reported data on the relationship between income and severity of functional impairment among impaired elderly. *J Appl Gerontol*. 1994;13:341-354.
18. Bamford J, Sandercock P, Jones L, Warlow C. The natural history of lacunar infarction: the Oxfordshire Community Stroke Project. *Stroke*. 1987;18:545-551.
19. Wade DT, Collin C. The Barthel ADL Index: a standard measure of physical disability? *Int Disabil Stud*. 1988;10:64-67.
20. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care*. 1981;19:787-805.
21. Anderson RT, Aaronson NK, Wilkin D. Critical review of the international assessments of health-related quality of life. *Qual Life Res*. 1993;2:369-395.
22. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334.
23. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;8476: 307-310.
24. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput Biol Med*. 1989;19:61-70.
25. Marshall GN, Hays RD, Nicholas R. Evaluating agreement between clinical assessment methods. *Int J Methods Psychiatr Res*. 1994;4:249-257.
26. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988:19-74.
27. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*. 1966;19:3-11.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
29. Snedecor GW, Cochran WG. *Statistical Methods*. 7th ed. Ames, Iowa: Iowa State University Press; 1980.
30. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd ed. New York, NY: McGraw-Hill; 1994.
31. Rothman ML, Hedrick SC, Bulcroft KA, Hickam DH, Rubenstein LZ. The validity of proxy-generated scores as measures of patient health status. *Med Care*. 1991;29:115-124.
32. McCusker J, Stoddard AM. Use of a surrogate for the Sickness Impact Profile. *Med Care*. 1984;22:789-795.
33. Sneeuw KCA, Aaronson NK, Osoba D, et al. The use of significant others as proxy raters of the quality of life of patients with brain cancer. *Med Care*. In press.

Chapter 8

Literature review and discussion

The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: An update

K.C.A. Sneeuw, M.A.G. Sprangers, N.K. Aaronson

Submitted for publication

ABSTRACT

Health-related quality of life (HRQL) studies sometimes rely, in part, on proxy information obtained from patients' significant others (spouse or close companion) or health care providers. This review: (1) provides a quantitative analysis of the results that have been reported in recent studies assessing the level of agreement between patient and proxy HRQL ratings, and (2) addresses a number of key methodological issues surrounding the use of proxy raters in HRQL research. This review concentrates on 23 studies, published between 1991-2000, that describe patient-proxy agreement for a number of well-known multidimensional HRQL instruments. In general, moderate to high levels of patient-proxy agreement were reported. Lower levels of agreement were found predominantly in studies employing a small sample size (approximately 50 patient-proxy pairs or less). In larger studies comparing patients and their significant others, median correlations were between 0.60-0.70 for physical HRQL domains and about 0.50 for psychosocial domains. Mixed results were reported in studies comparing patients and their health care providers, but most of these studies employed a relatively small sample size. Proxy raters tended to report more HRQL problems than patients themselves, but the magnitude of observed differences was modest (median standardized differences of about 0.20). Based on the current evidence, we conclude that judgements made by significant others and health care providers about several aspects of patients' HRQL are reasonably accurate. Substantial discrepancies between patient and proxy ratings occur in a minority of cases. We recommend that future studies focus on: (a) the reliability and validity of proxy ratings according to common psychometric methods, and (b) the balance between information bias due to proxy ratings and potential selection bias due to exclusion of important patient subgroups from HRQL studies.

INTRODUCTION

Health-related quality of life (HRQL) assessment is increasingly being used in clinical research as an important outcome of disease and treatment. Additionally, attention has been directed toward the possibility of employing individual HRQL assessments in daily clinical practice. Given that the patient is the most appropriate source of information regarding his or her HRQL, such assessments are derived primarily from the patients themselves. However, there are several patient groups and situations in which the ability to complete a questionnaire may be compromised. Problems with self-report may arise when patients have insufficient cognitive or communication abilities, when they experience severe symptom distress, or when they find an interview to be physically or emotionally too burdensome.

For those patients unable or unwilling to provide HRQL information themselves, their significant others (e.g., spouses, parents, relatives, friends) or health care providers (e.g., physicians, nurses) might be employed as alternative sources of such information.[1] The use of such proxy raters may be an effective means of obtaining information that might otherwise be lost. Study results can be substantially biased if highly relevant subgroups of patients are excluded from HRQL assessments. This is an important issue of concern in a range of populations, such as the elderly,[2,3] cancer patients,[4] stroke survivors,[5] patients with neurological deficits,[6] and pediatric patients.[7] When studying such patient populations, researchers frequently rely on information provided by proxy raters. In studies of the health of the aged, it is not uncommon for more than 20% of elderly and 50% of nursing home residents to be unable or unwilling to participate themselves.[8] For example, recent HRQL studies of older patients in emergency care settings required proxy raters in approximately 25% of the cases.[9,10] Similarly, HRQL studies among stroke survivors reported the use of proxy-based information in about 25% of the patients.[11,12] In other populations, such as persons with Alzheimer disease or young children, there may be no alternative to relying solely on data from proxy raters.

Reliance on significant others or health care providers as alternative sources of information on patients' HRQL can only be justified, however, if one can demonstrate that the quality of such proxy information is high. Given that HRQL is a multidimensional construct, proxy raters must be able to provide reliable and valid data on a range of HRQL domains, including patients' physical and psychosocial functioning, and a variety of physical symptoms. Evaluation of the quality of proxy-generated information typically involves a comparison of patient and proxy ratings. In earlier work,[1] we reviewed 49 studies addressing this issue. These studies indicated that the concordance between patient and proxy HRQL ratings was far

from optimal, irrespective of the type of proxy rater. However, we also found the literature in this field to be characterized by a high degree of heterogeneity and weaknesses in research design. Most importantly, patients and proxy ratings were frequently found to have been derived from different or unstandardized instruments, and the studies were often based on very small sample sizes.

In the past decade, many studies have been published examining the extent to which patients' HRQL ratings are in agreement with those provided by their significant others and/or their health care providers. Generally, this has included assessing patient-proxy agreement both at the level of the individual patient, most often by means of correlations, and at the group level, by comparing patient and proxy mean scores. The former method provides a direct indication of the extent to which the proxy ratings concur with those of the patient themselves. The latter method allows one to determine the direction and magnitude of any systematic bias that might be introduced in HRQL investigations when using proxy respondents.

The primary aim of this review is to provide a quantitative analysis of the results reported in recent studies examining the extent of agreement between patient and proxy HRQL ratings. At the time of our previous review,[1] patient-proxy studies employing standardized multidimensional HRQL instruments were scarce. Recently, a range of patient-proxy studies using such instruments have been reported. This enabled us to focus on studies using well-known multidimensional HRQL instruments completed by both raters. We raise a number of methodological issues that require additional attention in determining the value and limitations of proxy data in HRQL studies.

METHODS

Selection of studies

Studies included in this review were identified using Medline and Current Contents (Life and Social-Behavioral Sciences) databases. Snowball techniques (i.e., review of the references of articles thus obtained) were employed to identify additional relevant studies. The literature search covered the years 1991 to 2000. Criteria for inclusion in the present review were: (a) use of standardized multidimensional HRQL instruments, (b) patients and proxy raters completing the same (or proxy-adapted) instruments, and (c) published in English-language, peer-reviewed journals. Studies using domain-specific instruments which focus on one specific aspect of HRQL, such as activities of daily living (ADL), pain, other

physical symptoms, and psychological distress, were not included. The few studies in which a range of HRQL domains was assessed by a battery of instruments were also excluded. As in our previous review,[1] studies of pediatric subjects were excluded, since the conceptual and practical issues involved in pediatric health assessment may be different from those encountered with adult patients.

Statistical methods

For each study, the following information was extracted (Table 1): (a) the patient population of interest, (b) the HRQL instrument employed, (c) the number of HRQL domains assessed, (d) the type of proxy raters, (e) the number of questionnaire administrations (referred to as 'time'), (f) the number of patient-proxy paired assessments, (g) agreement at the individual patient level, and (h) agreement at the group level.

At the individual level, measures of association or agreement between patient and proxy ratings are presented. The following statistics were encountered in the studies reviewed: Pearson's correlation coefficient (r), intraclass correlation coefficient (ICC), Lin's concordance coefficient (LCC), and weighted or unweighted kappa (k). When available, the ICC was used as an indicator of chance-corrected agreement between patient and proxy ratings at the individual patient level.[13]

At the group level, differences in patient and proxy mean scores are presented (proxy minus patient score), both on a 0-100 scale and expressed in standard deviation units. When not already the usual scoring method, all scores were linearly transformed to a 0-100 scale with higher scores representing better quality of life. Negative differences indicate that proxies rate the patients as having worse functioning/health/quality of life and more symptoms than do the patients themselves. To examine the statistical magnitude of observed differences, the mean difference score was standardized by relating this score to its standard deviation. When available, the standard deviation of the difference score was employed as the denominator. Occasionally, the standard deviation of the difference score could be calculated from the standard error or the t statistic. Otherwise, the standard deviation of the patient score was used.

For each instrument, the range and median of results across the domains are presented (Table 1). Additionally, results are sorted for eight HRQL domains: physical, role, cognitive, emotional and social functioning, overall health, pain and fatigue (Tables 2 to 5). The strength of agreement can be interpreted as follows:[14] ≤ 0.40, poor to fair agreement; 0.41 - 0.60, moderate agreement; 0.61 - 0.80, good agreement, and 0.81 - 1.00, excellent agreement. Given the similarity to effect size (d)

Table 1. Studies assessing the level of agreement between patient and proxy reports of patients' health-related quality of life

							Agreement at individual level		Agreement at group level	
									Mean difference (proxy minus patient) [b]	
							Correlation [a]		On 0-100 scale	Standardized (*d*) [c]
Study	Subjects	Instrument	Domains	Proxy	Time	*n*	Type	Range (Median)	Range (Median)	Range (Median)
Blazeby *et al.*[16] 1995	Oesophageal cancer patients	EORTC QLQ-C30	15 Domains	S.O.		39	*k* (w)	0.39 to 0.59 (0.53)		
				Physician		52	*k* (w)	0.14 to 0.61 (0.35)		
Sigurdardottir *et al.*[17] 1996	Advanced melanoma patients	EORTC QLQ-C30 [d]	9 Domains (7 QLQ-C36 + 2 melanoma-specific domains)	S.O.		30	*r*	0.36 to 0.88 (0.57)	-6.7 to 4.6 (-1.0)	-0.22 to 0.19 (-0.09)
				Nurse		40	*r*	-0.25 to 0.51 (0.16)	-14.5 to 22.1 (6.1)	-0.43 to 0.85 (0.16)
Sneeuw *et al.*[18] 1997	Brain cancer patients	EORTC QLQ-C30	15 Domains	S.O.		103	ICC	0.23 to 0.67 (0.52)	-12.3 to 3.0 (-3.4)	-0.56 to 0.15 (-0.15)
		Brain cancer-specific module	11 Domains				ICC	0.29 to 0.74 (0.57)	-4.0 to 2.9 (-1.6)	-0.23 to 0.15 (-0.07)
Sneeuw *et al.*[19] 1998	Cancer patients	EORTC QLQ-C30	15 Domains	S.O.	T1	307	ICC	0.46 to 0.73 (0.59)	-9.7 to 1.3 (-3.7)	-0.45 to 0.06 (-0.17)
					T2	224	ICC	0.42 to 0.79 (0.61)	-8.8 to 2.9 (-3.7)	-0.46 to 0.13 (-0.17)
Sneeuw *et al.*[20; e] 2000	Prostate cancer patients	EORTC QLQ-C30	15 Domains	S.O.		72	ICC	0.47 to 0.73 (0.61)	-7.4 to 2.3 (-1.5)	-0.34 to 0.10 (-0.07)
		Prostate cancer-specific module	6 Domains				ICC	0.24 to 0.75 (0.58)	-13.4 to 3.7 (-3.2)	-0.28 to 0.14 (-0.11)
Segal *et al.*[21] 1994	Stroke patients	SF-36	8 Domains	S.O.		38	ICC	0.15 to 0.67 (0.33)		
		FIM	6 Domains				ICC	0.52 to 0.89 (0.79)		
			Physical score				ICC	0.91		
			Cognitive score				ICC	0.60		
			Total score				ICC	0.87		

- *continued overleaf*

Study	Subjects	Instrument	Domains	Proxy	Time	*n*	Agreement at individual level: Correlation [a] Type	Range (Median)	Agreement at group level: Mean difference (proxy minus patient) [b] On 0-100 scale Range (Median)	Standardized (*d*) [c] Range (Median)
Berlowitz *et al.*[22] 1995	Nursing home residents	SF-36	7 Domains (pain not included)	Physician		69	*r*	0.13 to 0.59 (0.24)	-21.0 to 3.0 (-12.0)	-0.88 to 0.07 (-0.44)
				Nurse		69	*r*	0.07 to 0.55 (0.24)	-32.0 to -4.0 (-17.0)	-1.33 to -0.09 (-0.53)
Meers *et al.*[23] 1995	End-stage renal disease patients	SF-36	8 Domains	Physician		30			-20.4 to 2.6 (-6.3)	-0.69 to 0.10 (-0.25)
				Nurse		30			-17.5 to 7.2 (-9.9)	-0.69 to 0.18 (-0.32)
Hays *et al.*[24] 1995	Epilepsy patients	SF-36 [f]/ QOLIE	8 Domains 9 Domains	S.O.		292	ICC ICC	0.29 to 0.56 (0.45) 0.32 to 0.56 (0.44)	-2.9 to 3.3 (0.4) -3.6 to 11.8 (0.5)	-0.14 to 0.09 (0.02) -0.13 to 0.50 (0.02)
			Total score				ICC	0.60	1.6	0.11
Wu *et al.*[25] 1997	Advanced HIV disease patients	SF-36 [g]/ MOS-HIV	7 Domains 3 Domains	S.O.		41			-11.9 to 1.9 (-6.7) -10.1 to -0.6 (-3.0)	-0.62 to 0.06 (-0.28) -0.45 to -0.03 (-0.15)
		EuroQol	Overall health						-5.4	-0.30
Rogers *et al.*[26] 1997	Critically ill patients	SF-36	8 Domains	S.O.	T1 T2	99 88	*r* *r*	0.22 to 0.73 (0.70) 0.49 to 0.92 (0.78)	-9.7 to 4.3 (-5.5) -4.3 to 2.7 (0.0)	
Pierre *et al.*[27] 1998	Elderly in rehabilitation	SF-36	8 Domains	S.O.		22	ICC	-0.11 to 0.58 (0.42)	-11.4 to 4.5 (-3.9)	-0.42 to 0.11 (-0.18)
				Nurse or therapist		41	ICC	0.01 to 0.60 (0.37)	-12.8 to 2.0 (-3.7)	-0.55 to 0.10 (-0.19)
	Elderly in day hospital			S.O.		19	ICC	-0.03 to 0.71 (0.27)	-10.5 to 12.5 (-5.2)	-0.80 to 0.59 (-0.24)
				Nurse or therapist		38	ICC	0.09 to 0.45 (0.31)	-11.3 to 15.8 (0.5)	-0.56 to 0.32 (0.02)

- *continued overleaf*

							Agreement at individual level		Agreement at group level	
									Mean difference (proxy minus patient) [b]	
							Correlation [a]		On 0-100 scale	Standardized (*d*) [c]
Study	Subjects	Instrument	Domains	Proxy	Time	*n*	Type	Range (Median)	Range (Median)	Range (Median)
Forjaz *et al.*[28] 1999	Hematological cancer patients	SF-36	8 Domains	S.O.[h]		33	*r*	0.17 to 0.59 (0.44)	-17.0 to -3.7 (-7.4)	-0.44 to -0.12 (-0.25)
						16	*r*	0.38 to 0.78 (0.69)	-9.2 to 13.3 (0.6)	-0.42 to 0.28 (0.01)
Rothman *et al.*[29] 1991	Chronically ill veterans	SIP	12 Domains	S.O.		275				
			Physical score				*r*	0.72	-2.5	-0.13
			Psychosocial score				*r*	0.33	0.0	0.00
			Total score							
Sneeuw *et al.*[30] 1997	Stroke patients	SIP	11 Domains (work not included)	S.O.		228	ICC	0.47 to 0.82 (0.60)	-6.7 to 0.4 (-3.7)	-0.45 to 0.04 (-0.24)
			Physical score					0.85	-3.2	-0.37
			Psychosocial score					0.61	-4.5	-0.39
			Total score					0.77	-3.6	-0.44
Brunelli *et al.*[31; i] 1998	Terminal cancer patients	TIQ	7 Domains	Physician or nurse		148				
			Physical symptoms							
			Therapy impact index							
Mathias *et al.*[32] 1997	Stroke patients	HUI-2	6 Domains (utility scores)	S.O.		33	ICC	0.39 to 0.81 (0.76)	-3.0 to 3.0 (0.5)	-0.38 to 0.23 (0.01)
			Global utility score				ICC	0.72	0.0	0.00
Grassi *et al.*[33] 1996	Advanced cancer patients	QL-Index	5 Domains	S.O.	T1	49	*r*	0.23 to 0.80 (0.50)		
					T2	49	*r*	0.64 to 0.92 (0.72)		
				Physician	T1	49	*r*	0.21 to 0.76 (0.47)		
					T2	49	*r*	0.27 to 0.81 (0.68)		

- continued overleaf

							Agreement at individual level		Agreement at group level	
									Mean difference (proxy minus patient) [b]	
							Correlation [a]		On 0-100 scale	Standardized (d) [c]
Study	Subjects	Instrument	Domains	Proxy	Time	n	Type	Range (Median)	Range (Median)	Range (Median)
Grassi *et al.* - *continued*			Total score	S.O.	T1	49	r	0.73	-2.0	-0.12
					T2	49	r	0.81	-2.1	-0.12
				Physician	T1	49	r	0.63	-4.5	-0.28
					T2	49	r	0.78	-6.4	-0.35
Sainfort *et al.*[34] 1996	Schizophrenic patients	QL-Index	5 Domains	Physician		37	k (w)	0.10 to 0.60 (0.43)		
			Total score				r	0.47	0.6	0.02
Moinpour *et al.*[35] 2000	Cancer patients with brain metastases	QL-Index	5 Domains	S.O.		40	k (w)	0.21 to 1.00 (0.29)	-14.0 to -1.0 (-6.0)	
			Total score				LCC	0.50	-6.0	-0.30
Sneeuw *et al.*[36,37; j] 1997	Cancer patients	COOP/WONCA Charts	6 Domains	S.O.	T1	295	ICC	0.43 to 0.67 (0.54)	-7.5 to -5.0 (-5.0)	-0.30 to -0.14 (-0.20)
					T2	189	ICC	0.53 to 0.72 (0.59)	-7.5 to 0.0 (-6.3)	-0.30 to 0.00 (-0.24)
				Physician	T1	295	ICC	0.32 to 0.63 (0.52)	-7.5 to 10.0 (-2.5)	-0.21 to 0.40 (-0.09)
					T2	189	ICC	0.47 to 0.71 (0.51)	-5.0 to 7.5 (-1.3)	-0.20 to 0.38 (-0.04)
1999				Nurse		90	ICC	0.38 to 0.66 (0.43)	-2.5 to 2.5 (-2.5)	-0.10 to 0.08 (-0.08)
Dorman *et al.*[38] 1997	Stroke patients	EuroQol	5 Domains	S.O.		130	k	0.30 to 0.64 (0.56)		
			Overall health				ICC	0.49	-2.2	-0.11

[a] When available the intraclass correlation coefficient (ICC), otherwise other measures of association or (weighted) kappa
[b] Negative differences mean that proxies report worse functioning/health/quality of life and more symptoms than patients themselves
[c] Standardized difference (effect size) d = mean difference/standard deviation of difference
[d] Based on previous version (EORTC QLQ-C36)
[e] This study was expected to be published in the year 2000, but was eventually published in February 2001
[f] Same or adapted domains as that of the SF-36 included in the QOLIE
[g] Same or adapted domains as that of the SF-36/SF-20 included in the MOS-HIV
[h] 33 spouses or partners and 16 other confidants
[i] This study fulfilled the criteria for inclusion in the present review, but employed other statistical methods for assessing agreement at both the individual and group level
[j] Two related studies based on the same dataset; nurse ratings were only obtained for the subgroup of patients receiving chemotherapy on a clinical ward at T1

calculations for paired observations,[15] the magnitude of the standardized difference can be interpreted as follows: d = 0.2, a small difference; d = 0.5, a moderate difference; d = 0.8, a large difference.

RESULTS

Description of studies

Twenty-three studies met the inclusion criteria.[16-38] Eleven studies were conducted among cancer patients,[16-20,28,31,33,35-37] four in stroke patients,[21,30,32,38] three in elderly populations,[22,27,29] and the remaining five in patients with end-stage renal disease,[23] epilepsy,[24] HIV disease,[25] schizophrenia,[34] and a critical illness requiring intensive care.[26] All studies are summarized in Table 1, ordered by the HRQL instrument employed.

Seventeen studies employed questionnaires in which the majority of HRQL domains are measured by two or more items.[16-32] Five studies in cancer patients[16-20] used the European Organization for Research and Treatment of Cancer Quality of Life Core Questionnaire (EORTC QLQ-C30)[39,40]. One of these studies[17] presented data on 7 domains of a previous version of this questionnaire (EORTC QLQ-C36)[41] and 2 domains of a study-specific malignant melanoma questionnaire, and two studies[18,20] provided additional data on a brain cancer-specific[42] and prostate cancer-specific[43] questionnaire module. Eight studies in different patient populations[21-28] employed the Medical Outcomes Study (MOS) 36-Item Short-Form Health Survey (SF-36)[44,45]. In one of these studies[21] the Functional Independence Measure (FIM)[46] was used as well, and in two studies[24,25] the same or adapted domains as that of the SF-36 were included in disease-specific questionnaires, the Quality of Life in Epilepsy Inventory (QOLIE)[47] and the MOS HIV Health Survey (MOS-HIV)[48]. Two studies[29,30] used the Sickness Impact Profile (SIP)[49], and one study[31] the Therapy Impact Questionnaire (TIQ)[50]. One further study[32] employed the Health Utilities Index Mark 2 (HUI-2)[51], an instrument in which utility values are assigned to the various domains.

Six studies used instruments in which the HRQL domains are assessed by single global items.[33-38] In three of these studies[33-35] the Spitzer Quality of Life Index (QL-Index)[52,53] was employed. Two related studies,[36,37] based on the same dataset, used the Dartmouth COOP Functional Health Assessment charts/WONCA (COOP/WONCA charts)[54], an adapted version of the Dartmouth COOP charts[55]. The first of these studies[36] employed a heterogeneous sample of cancer patients,

their significant others and physicians. The second study[37] focused on the subgroup of patients receiving chemotherapy on a clinical ward for whom nurses provided HRQL ratings as well. Furthermore, one study in stroke patients[38] employed the EuroQol instrument[56,57].

In eleven studies the number of patient-proxy pairs was approximately 50 or less.[16,17,21,23,25,27,28,32-35] In two of these studies, the sample size was further decreased by reporting results for two subgroups based on either the type of patients[27] or significant others[28]. Five studies employed samples of up to 100 patient-proxy pairs,[18,20,22,26,37] and seven studies used large samples ranging from 130 to 307 patient-proxy pairs.[19,24,29-31,36,38] In four studies patient-proxy pairs were assessed at two points in time.[19,26,33,36]

Nineteen studies assessed the level of agreement between HRQL ratings provided by the patients and their significant others,[16-21,24-30,32,33,35-38] of which six studies also included a comparison with ratings obtained from health care providers.[16,17,27,33,36,37] The remaining four studies were focused on a comparison of HRQL ratings provided by patients and their health care providers.[22,23,31,34]

Agreement between patients and significant others

With two exceptions,[25,29] all studies assessing agreement between patients and significant others reported correlation coefficients or kappa statistics. Most studies showed moderate to good agreement between patients and their significant others, as indicated by median correlations ranging from 0.42 to 0.78 (Table 1). Low levels of agreement, as indicated by median correlations below 0.40, were observed in only three studies,[21,27,35] all employing small patient-proxy samples (19 to 40 pairs).

Table 2 shows correlations between patient and significant other ratings for the eight specific HRQL domains. In total, 143 correlations were reported across the eight domains, with a median correlation of 0.56. Based on all studies combined, good agreement was observed for physical functioning (median correlation of 0.69) and moderate agreement for all other domains (median correlations ranging from 0.47 to 0.58). However, substantial differences were noted between smaller and larger studies. For larger studies ($n \geq 50$), correlations ranged from 0.22 to 0.92, with a median of 0.59. Only 3 of the 75 correlations (4%) were below 0.40. Moreover, a clear difference in the magnitude of correlations was observed between the physical domains (physical and role funtioning, overall health, pain and fatigue; median correlations 0.60 - 0.71) and the psychosocial domains (emotional, social and cognitive functioning; median correlations 0.48 - 0.50). For smaller studies ($n < 50$), a greater range in correlations was found (-0.11

to 1.00; median 0.52), and 21 of the 68 correlations (31%) were below 0.40. Median correlations ranged from 0.30 for social functioning to 0.67 for physical functioning. As compared to the larger studies, median correlations were substantially lower for role functioning, social functioning, pain and fatigue.

Table 2. Level of agreement between HRQL reports provided by patients and significant others for various domains

			PF[a]	RF[b]	CF[c]	EF[d]	SF[e]	OH[f]	PA[g]	FA[h]
Study	Instrument	*n*	ICC[i]	ICC	ICC	ICC	ICC	ICC	ICC	ICC
Blazeby *et al.*[16]	EORTC QLQ-C30	39	0.53	0.58	0.54	0.40	0.58	0.50	0.58	0.39
Sigurdardottir *et al.*[17]	EORTC QLQ-C30/C36	30	0.88	0.58	-	0.57	0.36	0.51	-	0.58
Sneeuw *et al.*[18]	EORTC QLQ-C30	103	0.67	0.58	0.58	0.62	0.48	0.64	0.23	0.52
Sneeuw *et al.*[19; j]	EORTC QLQ-C30	307	0.73	0.63	0.49	0.47	0.46	0.54	0.61	0.62
		224	0.75	0.67	0.46	0.55	0.42	0.56	0.71	0.66
Sneeuw *et al.*[20]	EORTC QLQ-C30	72	0.71	0.70	0.47	0.48	0.55	0.61	0.73	0.73
Segal *et al.*[21]	SF-36	38	0.67	0.20	-	0.29	0.21	0.47	0.36	0.40
Hays *et al.*[24]	SF-36/QOLIE	292	0.55	0.44	0.48	0.44	0.45	0.52	0.56	0.44
Rogers *et al.*[26; j]	SF-36	99	0.71	0.73	-	0.22	0.73	0.66	0.73	0.68
		88	0.89	0.86	-	0.49	0.76	0.71	0.92	0.79
Pierre *et al.*[27; k]	SF-36	22	0.55	0.40	-	-0.11	0.19	0.58	0.57	0.11
		19	0.71	-0.03	-	0.52	0.01	0.33	0.21	0.40
Forjaz *et al.*[28; l]	SF-36	33	0.59	0.55	-	0.42	0.17	0.53	0.33	0.46
		16	0.44	0.42	-	0.74	0.67	0.71	0.78	0.71
Sneeuw *et al.*[30]	SIP	228	0.80	0.69	0.59	0.59	0.52	-	-	-
Mathias *et al.*[32]	HUI-2	33	0.75	-	0.61	0.76	-	-	0.39	-
Grassi *et al.*[33; j]	QL-Index	49	0.80	0.50	-	0.77	0.23	0.46	-	-
		49	0.72	0.92	-	0.69	0.87	0.64	-	-
Moinpour *et al.*[35]	QL-Index	40	0.29	0.25	-	0.21	1.00	0.34	-	-
Sneeuw *et al.*[36; j]	COOP/WONCA Charts	295	0.56	0.67	-	0.48	0.43	0.51	0.64	-
		189	0.57	0.65	-	0.58	0.53	0.60	0.72	-
Dorman *et al.*[38]	EuroQol	130	0.60	0.64	-	0.30	-	-	0.45	-
	median overall		0.69	0.58	0.52	0.49	0.47	0.54	0.58	0.55
	median $n \geq 50$		0.71	0.67	0.49	0.48	0.50	0.60	0.68	0.66
	median $n < 50$		0.67	0.46	0.58	0.52	0.30	0.51	0.39	0.40

[a] PF - physical functioning (EORTC QLQ-C30, SF-36), ambulation (SIP), daily living (QL-Index), physical fitness (COOP/WONCA Charts), mobility (EuroQol, HUI-2)
[b] RF - role functioning (EORTC QLQ-C30), role limitations due to physical health (SF-36), household management (SIP), activity (QL-Index), daily activities (COOP/WONCA Charts), usual activities (EuroQol)
[c] CF - cognitive functioning (EORTC QLQ-C30), attention cognitive function (QOLIE), alertness behavior (SIP), cognition (HUI-2)
[d] EF - emotional functioning (EORTC QLQ-C30), mental health (SF-36), emotional behavior (SIP), outlook (QL-Index), feelings (COOP/WONCA Charts), anxiety/depression (EuroQol), emotion (HUI-2)
[e] SF - social functioning (EORTC QLQ-C30, SF-36), social interaction (SIP), support (QL-Index), social activities (COOP/WONCA Charts)
[f] OH - overall health/QL (EORTC QLQ-C30), general health perceptions (SF-36), health (QL-Index), overall health (COOP/WONCA Charts)
[g] PA - pain (EORTC QLQ-C30, COOP/WONCA Charts, HUI-2), bodily pain (SF-36), pain/discomfort (EuroQol)
[h] FA - fatigue (EORTC QLQ-C30), vitality (SF-36)
[i] Intraclass correlation coefficient, if not available Pearson *r* or (weighted) *k*
[j] Results obtained at two points in time
[k] Results obtained for two patient groups
[l] Results obtained for two types of significant others

Comparisons of patient and significant other mean scores for specific HRQL domains were reported in most studies. Two studies did not report differences in mean scores at all,[16,21] and three studies did so for overall scores but not for specific domains.[29,33,38] Two studies did report mean scores, but standardized differences (effect size *d*) could not be calculated because standard deviations were absent.[26,35] Patient and significant other mean scores deviated from each other in both positive and negative directions (Table 1). Most typically, however, significant others tended to rate the patients as having slightly lower levels of functioning, health and quality of life, and slightly more symptomatology than did the patients themselves, as indicated by median differences of about -1 to -7 points on a 0-100 scale and median standardized differences of about -0.10 to -0.30 in the majority of comparisons.

Table 3. Standardized differences in mean scores of HRQL reports provided by patients and significant others for various domains

			PF[a]	RF[b]	CF[c]	EF[d]	SF[e]	OH[f]	PA[g]	FA[h]
Study	Instrument	*n*	*d*[i]	*d*	*d*	*d*	*d*	*d*	*d*	*d*
Sigurdardottir *et al.*[17]	EORTC QLQ-C30/C36	30	-0.09	-0.05	-	-0.06	-0.22	0.14	-0.22	-0.04
Sneeuw *et al.*[18]	EORTC QLQ-C30	103	-0.26	-0.38	-0.38	-0.20	-0.24	-0.17	-0.13	-0.56
Sneeuw *et al.*[19; j]	EORTC QLQ-C30	307	-0.26	-0.17	0.01	-0.45	-0.12	-0.33	-0.23	-0.34
		224	-0.24	0.04	0.00	-0.46	-0.17	-0.29	-0.26	-0.32
Sneeuw *et al.*[20]	EORTC QLQ-C30	72	-0.34	-0.31	-0.05	-0.17	-0.04	-0.07	0.02	-0.08
Hays *et al.*[24]	SF-36/QOLIE	292	-0.01	0.09	0.30	0.06	0.02	-0.14	0.01	-0.08
Wu *et al.*[25]	SF-36/MOS-HIV	41	0.06	0.00	-0.03	-0.62	-0.24	-0.28	-0.32	-0.33
Pierre *et al.*[27; k]	SF-36	22	-0.42	0.00	-	-0.15	-0.32	-0.21	0.00	-0.21
		19	-0.80	0.10	-	-0.26	-0.24	0.59	0.07	-0.24
Forjaz *et al.*[28; l]	SF-36	33	-0.33	-0.12	-	-0.36	-0.44	-0.17	-0.32	-0.17
		16	-0.26	0.16	-	-0.15	-0.42	0.24	0.22	-0.23
Sneeuw *et al.*[30]	SIP	228	-0.01	-0.32	-0.28	-0.18	-0.45	-	-	-
Mathias *et al.*[32]	HUI-2	33	0.00	-	0.23	0.05	-	-	-0.38	-
Sneeuw *et al.*[36; j]	COOP/WONCA Charts	295	-0.18	-0.20	-	-0.30	-0.14	-0.20	-0.20	-
		189	-0.30	0.00	-	-0.33	-0.23	-0.22	-0.25	-
	median overall		-0.26	-0.03	-0.02	-0.20	-0.24	-0.17	-0.21	-0.23
	median $n \geq 50$		-0.25	-0.19	-0.03	-0.25	-0.16	-0.20	-0.20	-0.32
	median $n < 50$		-0.26	0.00	0.10	-0.15	-0.28	-0.02	-0.22	-0.22

[a-h] See table 2

[i] Standardized difference (effect size) *d* = mean difference/standard deviation of difference
Note: For the functioning domains and overall health, a negative *d* means that proxies report worse functioning/overall health
For pain and fatigue, a negative *d* means that proxies report more pain/fatigue

[j] Results obtained at two points in time

[k] Results obtained for two patient groups

[l] Results obtained for two types of significant others

Table 3 displays standardized differences in patient and significant other mean scores for the eight specific HRQL domains. In total, 104 standardized differences could be calculated across the eight domains, with a median of -0.20. Based on all studies combined, very small median standardized differences were observed for role and cognitive functioning (-0.03 and -0.02, respectively). For all other domains, median standardized differences ranged between -0.17 and -0.26. In larger studies, the median standardized difference for role functioning was within the latter range (-0.19). In general, results were quite similar for smaller and larger studies, although a broader range of standardized differences was noted for small studies (-0.80 to 0.59) as compared to large studies (-0.56 to 0.30). Substantial disagreement in mean scores, as indicated by standardized differences below -0.50, was found on only three occasions (3%). The two largest discrepancies (-0.80 and -0.62) were observed in studies employing small patient-proxy samples (19 and 40 pairs, respectively).[25,27]

Agreement between patients and health care providers

Of the ten studies assessing agreement between patients and health care providers, correlation coefficients (or kappa statistics) were provided for all but two studies.[23,31] As shown in Table 1, four studies reported moderate to good levels of agreement (median correlations between 0.43 and 0.68),[33,34,36,37] and the other four showed poor to fair agreement (median correlations between 0.16 and 0.37).[16,17,22,27] In six studies patient-proxy agreement was examined not only for health care providers, but also for significant others.[16,17,27,33,36,37] However, in only two of these studies, were results of both proxy raters based on the same patient group.[33,36] Both of these latter studies enabled a direct head-to-head comparison of physicians and significant others as proxy raters of patients' HRQL at two points in time. Differences between these two types of proxy raters were minimal. Both studies showed slightly higher levels of patient-proxy agreement at the follow-up assessment point, irrespective of the type of proxy rater.

Table 4 shows correlations between patient and health care provider ratings of the eight specific HRQL domains. In total, 73 correlations were reported across the eight domains, with a median of 0.41. Correlations ranged from -0.25 to 0.80, and 34 of the 73 correlations (47%) were below 0.40. Based on all studies combined, median correlations for specific domains differed substantially. Poor agreement was noted for fatigue and social functioning (0.16 and 0.19, respectively), and moderate agreement for all other domains (ranging from 0.41 for emotional functioning and overall health to 0.57 for pain). Given the relatively small number

of studies, no attempt was made to differentiate between studies with smaller versus larger sample sizes. We would note, however, that in the only study including more than 100 patients,[36] only 2 correlations were below 0.40. Specifically, fair agreement was observed between patient and physician ratings of emotional and social functioning at the initial assessment point (correlations of 0.37 and 0.32), which improved to a moderate level of agreement at the follow-up assessment point (correlations of 0.47 and 0.50).

Table 4. Level of agreement between HRQL reports provided by patients and health care providers for various domains

				PF[a]	RF[b]	CF[c]	EF[d]	SF[e]	OH[f]	PA[g]	FA[h]
Study	Instrument	Proxy	*n*	ICC[i]	ICC	ICC	ICC	ICC	ICC	ICC	ICC
Blazeby *et al.*[16]	EORTC QLQ-C30	Physician	52	0.43	0.55	0.35	0.41	0.14	0.33	0.61	0.53
Sigurdardottir *et al.*[17]	EORTC QLQ-C30/C36	Nurse	40	0.23	-0.25	-	0.10	-0.01	0.30	-	0.08
Berlowitz *et al.*[22]	SF-36	Physician	69	0.59	0.31	-	0.26	0.24	0.15	-	0.21
		Nurse	69	0.55	0.20	-	0.45	0.07	0.28	-	0.11
Pierre *et al.*[27; j]	SF-36	Nurse	41	0.38	0.08	-	0.41	0.01	0.36	0.42	0.60
		Nurse	38	0.45	0.09	-	0.41	0.11	0.43	0.39	0.11
Grassi *et al.*[33; k]	QL-Index	Physician	49	0.76	0.67	-	0.21	0.47	0.40	-	-
		Physician	49	0.75	0.81	-	0.33	0.27	0.68	-	-
Sainfort *et al.*[34]	QL-Index	Physician	37	0.60	0.14	-	0.43	0.10	0.43	-	-
Sneeuw *et al.*[36,37; k]	COOP/WONCA Charts	Physician	295	0.53	0.63	-	0.37	0.32	0.51	0.53	-
		Physician	189	0.50	0.61	-	0.47	0.50	0.52	0.71	-
		Nurse	90	0.38	0.58	-	0.43	0.43	0.41	0.66	-
		median overall		0.52	0.43	-	0.41	0.19	0.41	0.57	0.16

[a-h] See table 2
[i] Intraclass correlation coefficient, if not available Pearson *r* or (weighted) *k*
[j] Results obtained for two patient groups
[k] Results obtained at two points in time for physicians

Comparison of patient and health care provider mean scores for specific HRQL domains was reported in only few studies. Two studies did not show differences in mean scores at all,[16,31] and two studies reported mean differences for overall scores but not for specific domains.[33,34] The remaining six studies yielded many conflicting results,[17,22,23,27,36,37] with mean scores of health care providers deviating from those of patients in both positive and negative directions (Table 1). As was found for significant others, health care providers tended to rate the patients as having slightly lower levels of functioning and health, and slightly more symptomatology than did the patients themselves, as indicated by median (standardized) differences below 0 in the majority of comparisons.

Table 5 shows standardized differences in patient and health care provider mean scores for the eight specific HRQL domains. In total, 65 standardized differences could be calculated across the eight domains, with a median of -0.17. Inconsistent results were noted for role functioning, overall health and pain. As compared to studies of significant others, a broader range of standardized differences was observed (-1.33 to 0.83), with substantial disagreement (standardized differences below -0.50 or above 0.50) on 12 of 65 occasions (18%).

Table 5. Standardized differences in mean scores of HRQL reports provided by patients and health care providers for various domains

				PF[a]	RF[b]	CF[c]	EF[d]	SF[e]	OH[f]	PA[g]	FA[h]
Study	Instrument	Proxy	*n*	*d*[i]	*d*	*d*	*d*	*d*	*d*	*d*	*d*
Sigurdardottir *et al.*[17]	EORTC QLQ-C30/C36	Nurse	40	0.19	0.05	-	0.09	-0.16	0.45	0.83	0.46
Berlowitz *et al.*[22]	SF-36	Physician	69	-0.44	0.07	-	-0.50	-0.27	-0.88	-	-0.28
		Nurse	69	-0.53	-0.09	-	-0.71	-0.20	-1.33	-	-0.83
Meers *et al.*[23]	SF-36	Physician	30	0.10	-0.12	-	-0.69	-0.69	-0.65	-0.18	-0.31
		Nurse	30	-0.17	-0.29	-	-0.46	-0.34	-0.69	-0.16	-0.42
Pierre *et al.*[27; j]	SF-36	Nurse	41	-0.55	-0.17	-	0.10	-0.11	-0.18	-0.09	-0.19
		Nurse	38	-0.56	0.32	-	0.15	-0.19	0.04	-0.02	0.00
Sneeuw *et al.*[36,37; k]	COOP/WONCA Charts	Physician	295	-0.18	0.00	-	-0.18	-0.21	0.20	0.40	-
		Physician	189	-0.18	0.10	-	-0.20	-0.17	0.30	0.38	-
		Nurse	90	-0.08	0.08	-	-0.10	0.07	-0.10	0.08	-
		median overall		-0.18	0.03	-	-0.19	-0.20	-0.14	0.03	-0.28

[a,b] See table 2

[i] Standardized difference (effect size) *d* = mean difference/standard deviation of difference
Note: For the functioning domains and overall health, a negative *d* means that proxies report worse functioning/overall health
For pain and fatigue, a negative *d* means that proxies report more pain/fatigue

[j] Results obtained for two patient groups

[k] Results obtained at two points in time for physicians

DISCUSSION

Patient-proxy agreement at individual level

The majority of the studies reviewed have compared patients' HRQL ratings with those provided by their significant others. In general, moderate to good levels of patient-proxy agreement were reported in these studies. As in our previous review,[1] a clear trend emerged indicating lower levels of agreement predominantly in studies employing a small sample size. Mixed results were reported in studies comparing HRQL ratings provided by patients with those of their health care providers. However, most of the latter studies employed a relatively small sample size (i.e., 30-70 patient-proxy pairs).

A key question is how strong measures of agreement need to be before one can conclude that agreement is satisfactory. One might suggest that correlations above 0.60, representing good agreement, are satisfactory, and that moderate correlations (i.e., between 0.40 and 0.60) are unacceptable. It is important to consider, however, that even when moderate correlations are found, proxy raters appear to provide identical or quite similar ratings to those of the patients in the vast majority of cases. That is, it has been demonstrated that even when moderate patient-proxy correlations are found for specific HRQL domains, responses to individual questionnaire items are identical or different by only one response category (for questions with four or five response categories) in 75% to 95% of all patient-proxy comparisons.[18,36] Large discrepancies between patient and proxy responses tend to occur in a minority of all comparisons made.

Importantly, not only is the overall level of patient-proxy agreement moderate to good, correlations between patient and proxy scores for corresponding HRQL domains have also been shown to generally be higher than those for diverging domains.[19,24,26] This latter finding suggests that, although proxy raters may sometimes provide different information than the patients themselves, they are capable of making clear distinctions between various aspects of patients' HRQL.

Patient-proxy agreement at group level

When patients and proxy raters disagree, these discrepancies have been found to occur in both directions (i.e., under- and overestimations). However, in line with our previous review, we have noted a tendency for both significant others and health care providers to report lower levels of functioning, health and quality of life, and more symptoms than the patients themselves. Employing the patient's rating as the point of reference, this tendency has usually been interpreted as an underestimation of patients' HRQL. The use of proxy ratings may thus introduce some bias in mean HRQL scores.

Again, a key question is how small differences between mean scores need to be before one can conclude that agreement is satisfactory. One might suggest that differences in mean scores are unacceptable when these differences are statistically significant. However, statistical significance of mean differences is, in part, dependent on sample size. Therefore, we consider the use of standardized differences (effect sizes) more appropriate. The current findings indicate that the magnitude of standardized differences between patient and proxy mean scores are, in general, small. Substantial disagreement between patient and proxy mean scores was found in only a minority of all comparisons made.

Nevertheless, researchers might wish to adjust proxy ratings for systematic bias.[3,25,30] One approach to reducing bias is to use calibrated proxy scores when the patient response is missing. Calibration can be based on the subset of cases for which both patient and proxy responses are available. This strategy has been applied in a study among stroke survivors of the relationship between stroke type and HRQL as measured by the SIP.[30] While use of adjusted (calibrated) proxy scores necessarily resulted in different mean scores as compared to those based on unadjusted (raw) proxy ratings, this did not impact on the association observed between HRQL and stroke type or on the interpretation of the study findings.

Psychometric considerations

Several psychometric factors may exert a profound impact on the level of patient-proxy agreement, and should therefore be considered when evaluating the extent of agreement. Obviously, the reliability of the HRQL ratings obtained from patients and proxy raters provides a frame of reference for interpreting patient-proxy agreement. That is, high levels of agreement between two methods of assessment cannot reasonably be expected when either one would provide ratings with compromised reliability.[58,59] Thus, it might be appropriate to correct for the reliability of patient and proxy ratings. For example, in a recent study comparing HRQL ratings provided by children and their parents,[60] the observed intraclass correlations were corrected by dividing these correlations by the square root of the product of the reliability of the two methods of assessment. While observed patient-proxy correlations in the latter study ranged from 0.37 to 0.70, corrected correlations varied between 0.48 and 0.83. In general, reliability estimates for HRQL instruments tend to vary between 0.70 and 0.90. When assuming a median reliability of 0.80 for both patient and proxy ratings, the median correlation of 0.56 as observed between patients and significant others in the current review would be corrected upward to 0.70.

Another psychometric factor which is often overlooked is the variability of the responses provided by patients and proxy raters. HRQL ratings are generally skewed toward the positive end of the scale. That is, symptoms or problems are most often reported to be absent. Theoretically, correlation coefficients are dependent on the range and variability of responses. Low reliability estimates and patient-proxy correlations are more likely to be found for measures with a low prevalence or truncated distributions, and for single-item measures, all of which can result in limited score variability.[59] For example, in a study among brain cancer patients and their significant others,[18] a substantial difference was noted in the level

of patient-proxy agreement for appetite loss and shortness of breath (ICC = 0.60 and 0.31, respectively). Both symptoms were rated on a four-point response scale ranging from 'not at all' to 'very much' . The percentages of patients without these symptoms were similar (about 65%), and the level of patient-proxy agreement was also similar (about 68% exact agreement). The only difference was that none of the patients had shortness of breath severe enough to be rated as 'very much'.

Especially in studies employing small sample sizes, the full range of scores may not always be observed. Additionally, the more extreme scores are likely to be reported by only a few patients. The extent of patient-proxy agreement for these few patients may have a strong impact on the magnitude of patient-proxy correlations observed for the total patient sample. This may be an important reason why smaller studies not only tend to report a greater range of agreement results than larger studies (i.e., as a consequence of larger standard errors), but also a lower overall level of patient-proxy agreement. In summary, low patient-proxy correlations cannot always be attributed to a lack of agreement. They can also reflect a range of interrelated methodological limitations, including insufficient sample size, limited score reliability, and lack of score variability.

Factors affecting patient-proxy agreement

While a number of variables has been shown to exert an influence on the extent of response agreement,[1-3,8] findings have generally been inconsistent across studies. A recent study examining the relative effect of various factors on the extent of patient-proxy agreement concluded that the HRQL domain under consideration and the health status of the patient is quantitatively a more important source of disagreement than the type of proxy rater or raters' background characteristics.[37]

The current review also indicates that the levels of patient-proxy agreement may vary as a function of the HRQL domains under consideration. In studies examining agreement between patients and their significant others, the highest level of agreement was found for physical functioning. This finding is in line with the previously observed trend that proxy ratings are more accurate when the information sought is concrete and observable.[1,8] The extent of agreement (about 0.70) is similar to that observed in studies focusing exclusively on functional status (e.g., activities of daily living or functional independence) in large samples of the elderly[61-64] or persons with traumatic brain injury[65,66]. For all other HRQL domains, moderate levels of agreement have been noted in the current review. However, when focusing on larger studies, the level of agreement for other physical domains, such as role functioning, overall health, and physical symptoms, is fairly similar to that of physical functioning. For psychosocial domains, including emotional, social

and cognitive functioning, the extent of agreement is more limited (about 0.50). Yet, the difference between agreement levels for physical and psychosocial issues appears to be less dramatic than has been suggested in some earlier studies.[29,67,68] This has also been the conclusion drawn by other research groups.[69-71]

Based on the current review, one might be tempted to conclude that the level of agreement between patients and health care providers is lower than that between patients and significant others. Such a conclusion would be in line with that of a frequently cited study.[72] However, the overall quality of the evidence relating to comparisons between patients and health care providers is rather poor. Most studies employed small patient samples and there was diversity in the findings. Only two studies have incorporated a head-to-head comparison of health care providers and significant others as proxy raters of patients' HRQL,[33,36] showing only minimal differences between the two types of proxy raters. Minor differences have also been reported in a study comparing functional status ratings of a large sample of chronically dependent elderly and their informal and professional caregivers.[64] It should be noted that additional methodological issues need to be addressed in studies examining agreement between HRQL ratings of patients and health care providers. That is, such studies typically include a limited number of professional caregivers and a sample of patients distributed across those caregivers. This issue of nested (dependent) data may require the use of multi-level models. However, a minimum of approximately 30 cases in the highest level (i.e., number of caregivers) would be needed to make appropriate use of such multi-level techniques.[73] In summary, while it is not unreasonable to hypothesize that professional caregivers may be less aware of patients' psychosocial concerns than are close relatives, as they may have more limited access to patients' thoughts and feelings, the currently available data do not lend strong support to this position.[74,75]

Relevance of patient-proxy agreement studies

It may be argued that the relevance of studies comparing patient and proxy ratings is limited for at least two reasons. First, by definition, comparisons can only be made for patients who are able to provide information about themselves. Generalizing the results of these studies to situations in which patients cannot provide their own responses is questionable. Though not a solution to this problem, several studies have examined trends in patient-proxy agreement as a function of patients' health status, so that findings can potentially be extrapolated to more impaired patient subgroups. In general, the level of patient-proxy agreement appears to be dependent, in part, on the patients' health status, but the nature of this

relationship is unclear. Recent studies suggest a U-shaped relationship.[19,37] That is, large discrepancies tend to occur most frequently among patients with a slightly or moderately impaired health status, and less frequently among patients with either a very good or very poor health status. This pattern can also be understood intuitively, given the smaller potential for response discrepancies in patients with either a very good or very poor health status. While for such patients the answers to many questions will be evident (i.e., either at the top or bottom end of the scale), ratings are more likely to diverge for patients with an intermediate health status.

Second, observed discrepancies between patient and proxy ratings should not be interpreted, a priori, as evidence of the inaccuracy or biased nature of proxy information. Patients' self-reports are often taken as a gold standard to which proxy ratings should conform. However, like proxy ratings, patient ratings are not perfectly reliable. As noted earlier, reliability estimates for HRQL instruments tend to vary between 0.70 and 0.90. Moreover, patients' self-reports are also subject to several forms of bias.[69,76-78] For example, patients who report relatively low levels of functional problems or symptoms may do so in an effort to avoid presenting themselves as a burden to others. In diseases affecting the brain, where patients may lack the cognitive abilities to accurately interpret the questions at a given moment in time, the value of patients' self-report may be questionable. Thus, it seems appropriate to go beyond examination of patient-proxy agreement by exploring alternative ways of establishing the value of proxy information.

Alternative approaches to establishing the value of proxy ratings

Two alternative methods have recently been described for determining the usefulness of proxy ratings. A rather straightforward approach is to examine the reliability and validity of proxy ratings in the same way as is common for patient self-reports. Reliability and validity estimates of proxy ratings can then be compared with those of the patients' own ratings. A number of the reviewed studies have reported a head-to-head comparison of reliability estimates based on both patient and proxy ratings.[18,19,22,24,27,30,35,36] In general, the findings indicate that the reliability of proxy-generated data is similar or slightly better than that of the patient. The validity of proxy ratings, relative to that of patient ratings, has been examined in several ways.[19,29,30,36,69,79] For example, the responsiveness to changes over time of HRQL ratings provided by patients and proxy raters has been compared.[19,36] Results indicated that patients and proxy scores were all responsive to changes over time in specific HRQL domains. Importantly, the reliability and validity of proxy ratings can also be determined when patients' self-reports have not been directly obtained.[80-82]

A second alternative is to examine the impact of using proxy ratings instead of patients' self-reports on the specific study outcomes under consideration.[82-85] For example, in two randomized trials of palliative treatment for patients with lung cancer, involving over 700 patients, ratings from patients and their physicians were obtained with respect to eleven physical symptoms.[85] Results indicated that the between-treatment comparisons warranted similar conclusions, regardless of whether the source of data was the patients or the physicians.

Balance of benefits and limitations of proxy ratings

While the present findings lend support to the usefulness of proxy ratings in HRQL research, they also indicate that ratings obtained from significant others or health care providers will not always be identical to those that would have been obtained from patients themselves. The practical goal in many HRQL studies, however, is to provide a representative analysis of the research question under consideration. Thus, possible bias due to imperfections in proxy ratings need to be balanced against the bias introduced by exclusion of highly relevant subgroups of patients.

Recently, two studies have suggested that the benefits of using proxy ratings for individuals who would otherwise have been excluded from study participation outweigh their limitations.[30,86] For example, a study among stroke survivors, in which proxy raters were employed for 25% of the patients,[30] indicated that, if the analyses had been based on available patients' self-reports only, the level of HRQL problems would have been underestimated, and questionable conclusions would have been drawn about the impact of several stroke types on the patients' HRQL.

CONCLUSIONS

Based on the current evidence, our previous conclusion that the concordance between patient and proxy HRQL ratings is far from optimal[1] needs to be tempered. The accumulated results suggest that judgements made by significant others and health care providers concerning various aspects of patients' HRQL are reasonably accurate. Substantial discrepancies between patient and proxy ratings occur in only a minority of cases. Moreover, when low patient-proxy correlations are observed, this may not always be due to lack of agreement, but also to a number of interrelated methodological weaknesses, such as insufficient sample size, suboptimal reliability, and lack of score variability. As concluded previously, significant others and health

care providers tend to report more HRQL problems than do patients themselves, and proxy ratings tend to be in greater agreement with those of patients for physical HRQL domains as compared to psychosocial domains. However, the current review also suggests that these latter differences may be smaller than has often been assumed.

The need for additional studies of proxy ratings of patients' HRQL in which the focus is primarily, if not exclusively, on the level of agreement between patient and proxy ratings may be limited. We recommend that future research focus on: (a) the reliability and validity of proxy ratings according to common psychometric standards, and (b) the added value of employing proxy ratings, particularly where the alternative would be to exclude relevant subgroups of patients from HRQL investigations.

REFERENCES

1. Sprangers MAG, Aaronson NK: The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992; 45: 743-760.
2. Zimmerman SI, Magaziner J: Methodological issues in measuring the functional status of cognitively impaired nursing home residents: the use of proxies and performance-based measures. Alzheimer Dis Assoc Disord 1994; 8 Suppl 1: S281-S290.
3. Magaziner J: Use of proxies to measure health and functional outcomes in effectiveness research in persons with Alzheimer disease and related disorders. Alzheimer Dis Assoc Disord 1997; 11 Suppl 6: 168-174.
4. Aaronson NK: Methodologic issues in assessing the quality of life of cancer patients. Cancer 1991; 67: 844-850.
5. De Haan R, Aaronson NK, Limburg M, Hewer RL, Van Crevel H: Measuring quality of life in stroke. Stroke 1993; 24: 320-327.
6. Murrell R: Quality of life and neurological illness: a review of the literature. Neuropsychol Rev 1999; 9: 209-229.
7. Wallander JL, Schmitt M, Koot HM: Quality of life measurement in children and adolescents: issues, instruments, and applications. J Clin Psychol 2001; 57: 571-585.
8. Magaziner J: The use of proxy respondents in health studies of the aged. In: The epidemiologic study of the elderly. Wallace RB, Woolson RF, eds. New York: Oxford University Press, 1992; 120-129.
9. Chin MH, Jin L, Karrison TG, et al: Older patients' health-related quality of life around an episode of emergency illness. Ann Emerg Med 1999; 34: 595-603.
10. McCusker J, Bellavance F, Cardin S, Belzile E: Validity of an activities of daily living questionnaire among older patients in the emergency department. J Clin Epidemiol 1999; 52: 1023-1030.
11. De Haan RJ, Limburg M, van der Meulen JHP, Jacobs HM, Aaronson NK: Quality of life after stroke: impact of stroke type and lesion location. Stroke 1995; 26: 402-408.
12. Hackett ML, Duncan JR, Anderson CS, Broad JB, Bonita R: Health-related quality of life among long-term survivors of stroke: results from the Auckland Stroke Study, 1991-1992. Stroke 2000; 31: 440-447.

13. Lee J, Koh D, Ong CN: Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med 1989; 19: 61-70.
14. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.
15. Cohen J: Statistical power analysis for the behavioral sciences. Hillsdale, New Yersey: Lawrence Erlbaum Associates, 1988; 19-74.
16. Blazeby JM, Williams MH, Alderson D, Farndon JR: Observer variation in assessment of quality of life in patients with oesophageal cancer. Br J Surg 1995; 82: 1200-1203.
17. Sigurdardottir V, Brandberg Y, Sullivan M: Criterion-based validation of the EORTC QLQ-C36 in advanced melanoma: the CIPS questionnaire and proxy raters. Qual Life Res 1996; 5: 375-386.
18. Sneeuw KC, Aaronson NK, Osoba D, et al: The use of significant others as proxy raters of the quality of life of patients with brain cancer. Med Care 1997; 35: 490-506.
19. Sneeuw KC, Aaronson NK, Sprangers MA, Detmar SB, Wever LD, Schornagel JH: Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. J Clin Epidemiol 1998; 51: 617-631.
20. Sneeuw KC, Albertsen PC, Aaronson NK: Comparison of patient and spouse assessments of health related quality of life in men with metastatic prostate cancer. J Urol 2001; 165: 478-482.
21. Segal ME, Schall RR: Determining functional/health status and its relation to disability in stroke survivors. Stroke 1994; 25: 2391-2397.
22. Berlowitz DR, Du W, Kazis L, Lewis S: Health-related quality of life of nursing home residents: differences in patient and provider perceptions. J Am Geriatr Soc 1995; 43: 799-802.
23. Meers C, Hopman W, Singer MA, MacKenzie TA, Morton AR, McMurray M: A comparison of patient, nurse, and physician assessment of health-related quality of life in end-stage renal disease. Dial Transpl 1995; 24: 120-124.
24. Hays RD, Vickrey BG, Hermann BP, et al: Agreement between self reports and proxy reports of quality of life in epilepsy patients. Qual Life Res 1995; 4: 159-168.
25. Wu AW, Jacobson DL, Berzon RA, et al: The effect of mode of administration on medical outcomes study health ratings and EuroQol scores in AIDS. Qual Life Res 1997; 6: 3-10.

26. Rogers J, Ridley S, Chrispin P, Scotton H, Lloyd D: Reliability of the next of kins' estimates of critically ill patients' quality of life. Anaesthesia 1997; 52: 1137-1143.
27. Pierre U, Wood-Dauphinee S, Korner-Bitensky N, Gayton D, Hanley J: Proxy use of the Canadian SF-36 in rating health status of the disabled elderly. J Clin Epidemiol 1998; 51: 983-990.
28. Forjaz MJ, Guarnaccia CA: Hematological cancer patients' quality of life: self versus intimate or non-intimate confidant reports. Psychooncology 1999; 8: 546-552.
29. Rothman ML, Hedrick SC, Bulcroft KA, Hickam DH, Rubenstein LZ: The validity of proxy-generated scores as measures of patient health status. Med Care 1991; 29: 115-124.
30. Sneeuw KC, Aaronson NK, De Haan RJ, Limburg M: Assessing quality of life after stroke: the value and limitations of proxy ratings. Stroke 1997; 28: 1541-1549.
31. Brunelli C, Costantini M, Di Giulio P, et al: Quality-of-life evaluation: when do terminal cancer patients and health-care providers agree? J Pain Symptom Manage 1998; 15: 151-158.
32. Mathias SD, Bates MM, Pasta DJ, Cisternas MG, Feeny D, Patrick DL: Use of the Health Utilities Index with stroke patients and their caregivers. Stroke 1997; 28: 1888-1894.
33. Grassi L, Indelli M, Maltoni M, Falcini F, Fabbri L, Indelli R: Quality of life of homebound patients with advanced cancer: assessments by patients, family members, and oncologists. J Psychosoc Oncol 1996; 14: 31-45.
34. Sainfort F, Becker M, Diamond R: Judgments of quality of life of individuals with severe mental disorders: patient self-report versus provider perspectives. Am J Psychiatry 1996; 153: 497-502.
35. Moinpour CM, Lyons B, Schmidt SP, Chansky K, Patchell RA: Substituting proxy ratings for patient ratings in cancer clinical trials: an analysis based on a Southwest Oncology Group trial in patients with brain metastases. Qual Life Res 2000; 9: 219-231.
36. Sneeuw KC, Aaronson NK, Sprangers MA, Detmar SB, Wever LD, Schornagel JH: Value of caregiver ratings in evaluating the quality of life of patients with cancer. J Clin Oncol 1997; 15: 1206-1217.

37. Sneeuw KC, Aaronson NK, Sprangers MA, Detmar SB, Wever LD, Schornagel JH: Evaluating the quality of life of cancer patients: assessments by patients, significant others, physicians and nurses. Br J Cancer 1999; 81: 87-94.
38. Dorman PJ, Waddell F, Slattery J, Dennis M, Sandercock P: Are proxy assessments of health status after stroke with the EuroQol questionnaire feasible, accurate, and unbiased? Stroke 1997; 28: 1883-1887.
39. 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 1993; 85: 365-376.
40. Osoba D, Aaronson N, Zee B, Sprangers M, te Velde A: Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. Qual Life Res 1997; 6: 103-108.
41. Aaronson NK, Ahmedzai S, Bullinger M, et al: The EORTC Core Quality of Life Questionnaire: interim results of an international field study. In: Effect of cancer on quality of life. Osoba D, ed. Boston: CRC Press, 1991; 185-203.
42. Osoba D, Aaronson NK, Muller M, et al: The development and psychometric validation of a brain cancer quality-of-life questionnaire for use in combination with general cancer-specific questionnaires. Qual Life Res 1996; 5: 139-150.
43. Albertsen PC, Aaronson NK, Muller MJ, Keller SD, Ware JE: Health-related quality of life among patients with metastatic prostate cancer. Urology 1997; 49: 207-216.
44. Ware JE, Sherbourne CD: The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30: 473-483.
45. McHorney CA, Ware JE, Raczek AE: The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993; 31: 247-263.
46. Dodds TA, Martin DP, Stolov WC, Deyo RA: A validation of the functional independence measurement and its performance among rehabilitation inpatients. Arch Phys Med Rehabil 1993; 74: 531-536.

47. Devinsky O, Vickrey BG, Cramer J, et al: Development of the quality of life in epilepsy inventory. Epilepsia 1995; 36: 1089-1104.
48. Wu AW, Revicki DA, Jacobson D, Malitz FE : Evidence for reliability, validity and usefulness of the Medical Outcomes Study HIV Health Survey (MOS-HIV). Qual Life Res 1997; 6: 481-493.
49. Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981; 19: 787-805.
50. Tamburini M, Rosso S, Gamba A, Mencaglia E, De Conno F, Ventafridda V : A therapy impact questionnaire for quality-of-life assessment in advanced cancer research. Ann Oncol 1992; 3: 565-570.
51. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q: Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. Med Care 1996; 34: 702-722.
52. Spitzer WO, Dobson AJ, Hall J, et al: Measuring the quality of life of cancer patients: a concise QL- index for use by physicians. J Chronic Dis 1981; 34: 585-597.
53. Wood-Dauphinee S, Williams JI: The Spitzer Quality of Life Index: its performance as a measure. In: Effect of cancer on quality of life. Osoba D, ed. Boston: CRC Press, 1991; 169-184.
54. Van Weel C: Functional status in primary care: COOP/WONCA charts. Disabil Rehabil 1993; 15: 96-101.
55. Nelson E, Wasson J, Kirk J, et al: Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. J Chronic Dis 1987; 40 Suppl 1: 55S-69S.
56. The EuroQol Group: EuroQol - a new facility for the measurement of health-related quality of life. Health Policy 1990; 16: 199-208.
57. Dorman PJ, Waddell F, Slattery J, Dennis M, Sandercock P: Is the EuroQol a valid measure of health-related quality of life after stroke? Stroke 1997; 28: 1876-1882.
58. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 8476: 307-310.
59. Nunnally JC, Bernstein IH: Psychometric theory. New York: McGraw-Hill, 1994.

60. Verrips GH, Vogels AG, den Ouden AL, Paneth N, Verloove-Vanhorick SP: Measuring health-related quality of life in adolescents: agreement between raters and between methods of administration. Child Care Health Dev 2000; 26: 457-469.
61. Magaziner J, Simonsick EM, Kashner TM, Hebel JR: Patient-proxy response comparability on measures of patient health and functional status. J Clin Epidemiol 1988; 41: 1065-1074.
62. Magaziner J, Bassett SS, Hebel JR, Gruber-Baldini A: Use of proxies to measure health and functional status in epidemiologic studies of community-dwelling women aged 65 years and older. Am J Epidemiol 1996; 143: 283-292.
63. Long K, Sudha S, Mutran EJ: Elder-proxy agreement concerning the functional status and medical history of the older person: the impact of caregiver burden and depressive symptomatology. J Am Geriatr Soc 1998; 46: 1103-1111.
64. Santos-Eggimann B, Zobel F, Berod AC: Functional status of elderly home care users: do subjects, informal and professional caregivers agree? J Clin Epidemiol 1999; 52: 181-186.
65. Tepper S, Beatty P, DeJong G: Outcomes in traumatic brain injury: self-report versus report of significant others. Brain Inj 1996; 10: 575-581.
66. Cusick CP, Gerhart KA, Mellick DC: Participant-proxy reliability in traumatic brain injury outcome research. J Head Trauma Rehabil 2000; 15: 739-749.
67. Clipp EC, George LK: Patients with cancer and their spouse caregivers. Cancer 1992; 69: 1074-1079.
68. Farrow DC, Samet JM: Comparability of information provided by elderly cancer patients and surrogates regarding health and functional status, social network, and life events. Epidemiology 1990; 1: 370-376.
69. Magaziner J, Zimmerman SI, Gruber-Baldini AL, Hebel JR, Fox KM: Proxy reporting in five areas of functional status: comparison with self-reports and observations of performance. Am J Epidemiol 1997; 146: 418-428.
70. Bassett SS, Magaziner J, Hebel JR: Reliability of proxy response on mental health indices for aged, community-dwelling women. Psychol Aging 1990; 5: 127-132.
71. Epstein AM, Hall JA, Tognetti J, Son LH, Conant Jr L: Using proxies to evaluate quality of life: can they provide valid information about patients' health status and satisfaction with medical care? Med Care 1989; 27: S91-S98

72. Slevin ML, Plant H, Lynch D, Drinkwater J, Gregory WM: Who should measure quality of life, the doctor or the patient? Br J Cancer 1988; 57: 109-112.
73. Goldstein M: Multilevel models in educational and social research. New York: Oxford University Press, 1987; 1-31.
74. Lampic C, Sjoden PO: Patient and staff perceptions of cancer patients' psychological concerns and needs. Acta Oncol 2000; 39: 9-22.
75. Sprangers MA, Sneeuw KC: Are healthcare providers adequate raters of patients' quality of life - perhaps more than we think? Acta Oncol 2000; 39: 5-8.
76. Kempen GI, van Heuvelen MJ, van den Brink RH, et al: Factors affecting contrasting results between self-reported and performance-based levels of physical limitation. Age Ageing 1996; 25: 458-464.
77. Kempen GI, Jelicic M, Ormel J: Personality, chronic medical morbidity, and health-related quality of life among older persons. Health Psychol 1997; 16: 539-546.
78. Gotay CC: Patient-reported assessment versus performance-based tests. In: Quality of life and pharmacoeconomics in clinical trials. Spilker B, ed. Philadelphia: Lippincott-Raven publishers, 1996; 413-420.
79. Von Dras DD, Siegler IC, Williams RB, Clapp-Channing N, Haney TL, Mark DB: Surrogate assessment of coronary artery disease patients' functional capacity. Soc Sci Med 1997; 44: 1491-1502.
80. le Coq EM, Boeke AJ, Bezemer PD, Bruil J , van Eijk JT: Clinimetric properties of a parent report on their offspring's quality of life. J Clin Epidemiol 2000; 53: 139-146.
81. Albert SM, Del Castillo-Castaneda C, Sano M, et al: Quality of life in patients with Alzheimer's disease as reported by patient proxies. J Am Geriatr Soc 1996; 44: 1342-1347.
82. Iezzoni LI, McCarthy EP, Davis RB, Siebens H: Mobility problems and perceptions of disability by self-respondents and proxy respondents. Med Care 2000; 38: 1051-1057.
83. Todorov A, Kirchner C: Bias in proxies' reports of disability: data from the National Health Interview Survey on disability. Am J Public Health 2000; 90: 1248-1253.
84. Semaan S: Impact of proxy-reported data on the relationship between income and severity of functional impairment among impaired elderly. J Applied Gerontol 1994; 13: 341-354.

85. Stephens RJ, Hopwood P, Girling DJ, Machin D: Randomized trials with quality of life endpoints: are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? Qual Life Res 1997; 6: 225-236.
86. Corder LS, Woodbury MA, Manton KG: Proxy response patterns among the aged: effects on estimates of health status and medical care utilization from the 1982-1984 long-term care surveys. J Clin Epidemiol 1996; 49: 173-182.

Summary

The purpose of this thesis is to investigate systematically the value and limitations of proxy ratings of patients' health-related quality of life (HRQL). The patient is generally considered as the primary source of information on his or her HRQL. For those patients unable or unwilling to provide HRQL information themselves, however, their significant others (e.g., spouses, parents, relatives, friends) or health care providers (e.g., physicians, nurses) might be employed as alternative sources of such information. The selective use of such proxy respondents may contribute to resolving the problem of missing HRQL data for highly relevant patient subgroups in clinical studies. In clinical practice, proxy judgements of patients' HRQL can and often do play a role, at least implicitly, in decisions regarding treatment and patient care. Both the problem of missing data in clinical studies and the factoring of HRQL considerations into the clinical decision making process lead to the same basic question: to what extent are health care providers and other individuals involved in the care of patients able to assess accurately the patients' quality of life?

Chapters 2 to 4 of this thesis describe the results of a study employing several strategies to examine the extent to which significant others and health care providers can provide useful information on the HRQL of a heterogeneous group of cancer patients. The study sample was composed of 320 cancer patients under active treatment with chemotherapy, their significant others (most often spouses), their treating physicians, and nurses for those receiving inpatient chemotherapy. Patients and significant others completed two standardized multidimensional HRQL questionnaires, the COOP/WONCA charts (7 global health status questions) and the EORTC QLQ-C30 (30 cancer-specific questions). Physicians and nurses completed the COOP/WONCA charts only. The respondents completed the questionnaires at two points in time, during an early phase of treatment and three months later.

Chapter 2 presents a head-to-head comparison of COOP/WONCA chart ratings provided by all patients, significant others and physicians at two points in time. This included not only examination of the level of patient-proxy agreement, but also assessment of the relative validity (i.e., responsiveness to changes over time) of patient- versus proxy-generated information. At baseline, all sources of information were available for 295 (92%) of 320 participating patients. Complete follow-up data were obtained for 189 patient-proxy triads. Comparison of mean scores on the COOP/WONCA charts revealed close agreement between patient and proxy ratings. At the individual patient level, exact or global agreement was observed in the majority of cases (73%-91%). Corrected for chance agreement, fair to good

intraclass correlations were noted (ICC=0.32 to 0.72).[1] Patient, physician and informal caregiver COOP/WONCA scores were all responsive to changes over time in specific HRQL domains, but differed in their relative performance. In comparison to the patients, the physicians were more efficient in detecting changes over time in physical fitness and overall health, but less so in relation to social function and pain.

Chapter 3, focusing on the subgroup of 90 inpatients for whom nurse COOP/WONCA chart ratings were obtained as well, describes the relative effects of the (three) types of proxy raters, the (seven) types of questions/HRQL domains, the patients' clinical status, and several background characteristics of all raters on the level of patient-proxy agreement. With few exceptions, mean scores of the proxy raters were equivalent or similar to those of the patients. Most patient-proxy correlations varied between 0.40 and 0.60, indicating a moderate level of agreement at the individual level. Of all comparisons made, 41% were in exact agreement and 43% agreed within one response category (global agreement), leaving 17% more profound patient-proxy discrepancies. Disagreement was not dependent on the type of proxy rater, nor on raters' background characteristics, but was influenced by the HRQL domain under consideration and the clinical status of the patient. Better patient-proxy agreement was observed for more concrete questions (daily activities, pain) and for patients with either a very good (ECOG 0) or poor (ECOG 3) performance status.

Chapter 4 describes the level and pattern of agreement between patients' and significant others' EORTC QLQ-C30 ratings, the reliability and validity of both types of information, and the influence of several factors on the extent of agreement. Complete baseline and follow-up data were obtained for 307 and 224 patient-proxy pairs, respectively. Comparison of mean scores revealed small but systematic differences between patient and proxy ratings, with a tendency of significant others to rate patients as having a poorer quality of life than the patients themselves. Agreement at the individual level was moderate to good (ICC=0.42 to 0.79). Multitrait-multimethod analysis demonstrated good convergence and discrimination of specific HRQL domains. The maximum level of disagreement was found at intermediate levels of quality of life, with smaller discrepancies noted for patients with either a relatively poor or good quality of life. Both patient and proxy HRQL ratings were reliable and responsive to changes over time. Several characteristics of the patients and their significant others (e.g., patients' tendency

[1] Guidelines for the ICC as a measure of the strength of agreement were labeled as follows: ≤ 0.40, poor to fair agreement; 0.41 - 0.60, moderate agreement; 0.61 - 0.80, good agreement, and 0.81 - 1.00, excellent agreement.

toward a socially desirable response set, significant others' own health/HRQL) were found to be associated with the level of agreement, but explained less than 15% of the variance in patient-proxy differences.

Chapters 5 to 7 describe the results of three clinical studies among specific patient populations, whereby proxy HRQL ratings were collected in addition to patients' self-report. Two of these studies were conducted among cancer patients employing the EORTC QLQ-C30 and cancer-specific questionnaire modules. The remaining study was conducted among stroke survivors employing the SIP.

Chapter 5 reports on the level of response agreement between 103 patients with brain cancer and their significant others on the EORTC QLQ-C30 and a brain cancer-specific questionnaire module. Statistically significant differences in mean scores were noted for several HRQL domains, with significant others tending to rate the patients as having a poorer quality of life than the patients themselves. However, with the exception of fatigue ratings, this response bias was of a limited magnitude. For most HRQL domains assessed, moderate to good agreement was found (ICC=0.41 to 0.74). Low patient-proxy correlations (ICC<0.40) were observed for a few (mostly single-item) symptom measures with low frequencies of occurence, suggesting that such low correlations can be attributed to a lack of score variability rather than a lack of agreement. Less agreement was noted for more impaired patients, particularly those exhibiting mental confusion. This finding was confirmed by longitudinal analyses, which indicated lower levels of patient-proxy agreement at follow-up for those patients whose physical or neurological condition had deteriorated over time.

Chapter 6 compares the responses of 72 men with metastatic prostate cancer and their spouses on the EORTC QLQ-C30 and a prostate cancer-specific questionnaire module. The two questionnaires combined assess a wide range of symptoms and functional limitations and form, in total, 21 HRQL outcomes. For 5 of the 21 patient-proxy comparisons, systematic differences in mean scores were noted, with spouses rating the patients as having more impairments than the patients themselves. For 19 of the 21 HRQL domains, moderate to good agreement was found (ICC=0.47 to 0.75). Low correlations (ICC<0.40) were found only for the measures of sexual functioning and satisfaction, which were most likely due to (a combination of) low score reliability, low score variance or outliers.

Chapter 7 reports on a study among 437 patients who had suffered a stroke six months earlier. HRQL was assessed by means of the SIP. For one-quarter of the patients, who were not able to provide self-report ratings, SIP ratings were provided by their significant others. For 228 of the remaining patients, both patient

and significant other SIP ratings were obtained. In addition to evaluating the level of patient-proxy agreement, this study estimated the impact of using proxy HRQL ratings for one-quarter of the patient sample on the results pertaining to one of the research question under investigation (i.e., the relationship between stroke type and quality of life). When comparing mean SIP scores for patients with both self-report and proxy-derived data available, the proxy mean scores were generally in close agreement with those of the patients. Yet, systematic differences were noted for several SIP scales, with significant others rating patients as having more impairments than the patients themselves. Intraclass correlations were moderate to excellent for the SIP subscales (average ICC=0.63), the physical (ICC=0.85) and psychosocial dimension (ICC=0.61), and the total SIP score (ICC=0.77). The proxy SIP scores were sensitive to differences in patients' functional health, which supports the validity of these ratings. For all patients combined, more impairments were found for patients with supratentorial (sub)cortical infarctions and hemorrhages as compared to patients with lacunar infarctions and infratentorial strokes. The results indicated that, if the analyses had been based on available patients' self-reports only, the level of impairments would have been underestimated, and questionable conclusions would have been drawn about the impact of several stroke types on patients' quality of life.

Chapter 8 provides a quantitative analysis of the results of the 6 studies described in this thesis and 17 other recent studies examining patient-proxy agreement for well-known multidimensional HRQL instruments. In general, moderate to high levels of patient-proxy agreement were reported. Lower levels of agreement were found predominantly in studies employing a small sample size (about 50 patient-proxy pairs or less). In larger studies comparing patients and their significant others, median correlations were between 0.60-0.70 for physical HRQL domains and about 0.50 for psychosocial domains. Mixed results were reported in studies comparing patients and their health care providers, but most of these studies employed a relatively small sample size. The accumulated results suggest that judgements made by significant others and health care providers about various aspects of patients' HRQL are reasonably accurate. Substantial discrepancies between patient and proxy ratings occur in a minority of cases. Moreover, when low patient-proxy correlations are observed, this may not always be due to lack of agreement, but also to a number of interrelated methodological limitations, such as insufficient sample size, suboptimal reliability, and lack of score variability. As concluded previously, significant others and health care providers tend to report more HRQL problems than do patients themselves, and proxy ratings tend to be in greater agreement with those of patients for physical HRQL domains as compared to

psychosocial domains. However, the current review also suggests that these latter differences may be smaller than has often been assumed. The need for additional studies of proxy ratings of patients' HRQL in which the focus is primarily, if not exclusively, on the level of agreement between patient and proxy ratings may be limited. We recommend that future research focus on: (a) the reliability and validity of proxy ratings according to common psychometric standards, and (b) the added value of employing proxy ratings, particularly where the alternative would be to exclude relevant subgroups of patients from HRQL investigations.

Samenvatting

Proxy-metingen van kwaliteit van leven

Ziekten en behandelingen kunnen de kwaliteit van leven van mensen sterk beïnvloeden. In klinisch onderzoek wordt kwaliteit van leven (KvL) steeds vaker gemeten om de effecten van ziekte en behandeling te evalueren. In de afgelopen decennia is hiervoor een groot aantal KvL-vragenlijsten ontwikkeld. Hoewel er geen algemene definitie van kwaliteit van leven bestaat, wordt in de meeste KvL-vragenlijsten vastgesteld hoe patiënten een aantal uiteenlopende aspecten van hun gezondheid en welbevinden ervaren, waaronder hun lichamelijk, psychisch en sociaal functioneren, klachten en symptomen, en hun algehele gezondheid. Meer recentelijk is aandacht besteed aan een mogelijke rol van KvL-vragenlijsten in de dagelijkse klinische praktijk. Artsen en andere zorgverleners maken impliciet vaak al een inschatting van 'hoe het met de patiënt gaat'. Door het meten van kwaliteit van leven en het presenteren van de resultaten hiervan krijgen de zorgverleners systematische informatie over de kwaliteit van leven van hun patiënten en kan de communicatie over KvL-onderwerpen worden bevorderd. Op deze wijze kunnen KvL-overwegingen expliciet een rol spelen bij beslissingen over de behandeling.

Bij het meten van kwaliteit van leven is de patiënt zelf de meest geschikte informatiebron. Er zijn echter situaties waarin de patiënt geen (valide) informatie over zijn of haar eigen kwaliteit van leven kan geven. Problemen met het invullen van KvL-vragenlijsten kunnen ontstaan als de cognitieve of communicatieve vaardigheden van patiënten niet toereikend zijn, als er (tijdelijk) sprake is van ernstige klachten of symptomen, en als het beantwoorden van een vragenlijst lichamelijk of emotioneel te belastend is. In dergelijke gevallen is het mogelijk om gebruik te maken van zogenaamde proxy-metingen, waarbij naasten van de patiënt (zoals de partner, kinderen, ouders, familie en vrienden) of zorgverleners (zoals artsen en verpleegkundigen) fungeren als alternatieve informatiebron. In klinisch onderzoek kan door het gebruik van proxy-metingen worden voorkomen dat KvL-gegevens voor een juist zeer relevant deel van de patiënten ontbreken. In de klinische praktijk kunnen proxy-oordelen over de kwaliteit van leven van patiënten een belangrijke rol spelen bij beslissingen omtrent de behandeling, vooral bij patiënten die niet in staat zijn hun eigen toestand te beoordelen of te rapporteren.

Het doel van dit proefschrift is de bruikbaarheid en beperkingen van proxy-metingen van kwaliteit van leven te onderzoeken. De centrale vraagstelling luidt: in hoeverre zijn naasten en zorgverleners in staat om betrouwbare en valide informatie te leveren over de kwaliteit van leven van de patiënt?

Hoofdstuk 2 tot 4 van dit proefschrift beschrijft de resultaten van een onderzoek waarin op verschillende manieren is bepaald in hoeverre naasten en zorgverleners bruikbare informatie kunnen verschaffen over de kwaliteit van leven van een heterogene groep kankerpatiënten. De onderzoekspopulatie bestond uit 320 kankerpatiënten die op de polikliniek of een verpleegafdeling van het ziekenhuis behandeld werden met chemotherapie, hun naaste (meestal de partner), hun behandelend arts en, indien er sprake was van opname in het ziekenhuis, een verpleegkundige. Door patiënten en naasten werden twee gestandaardiseerde multidimensionele KvL-vragenlijsten ingevuld: de COOP/WONCA kaarten (7 globale vragen) en de EORTC QLQ-C30 (30 kankerspecifieke vragen). Door artsen en verpleegkundigen werden alleen de COOP/WONCA kaarten ingevuld. De respondenten vulden de vragenlijsten twee keer in: vlak na het begin van de behandeling (voormeting) en drie maanden later (nameting).

Hoofdstuk 2 geeft een vergelijking van de antwoorden van patiënten, naasten en artsen op de COOP/WONCA kaarten. Hierbij is niet alleen gekeken naar de mate van patiënt-proxy overeenstemming, maar ook naar de relatieve validiteit van de informatie van patiënten, naasten en artsen in termen van responsiviteit (gevoeligheid voor verandering in kwaliteit van leven). Bij de voormeting waren de drie informatiebronnen beschikbaar voor 295 van de 320 deelnemende patiënten. Bij de nameting werd informatie geleverd door 189 drietallen. Vergelijking van de gemiddelde scores op de COOP/WONCA kaarten liet een redelijk goede overeenstemming zien tussen patiënten, naasten en artsen. Bij individuele vergelijking van de antwoorden van patiënten met die van hun naasten en artsen was er in de meeste gevallen (73%-91%) sprake van exacte overeenstemming of globale overeenstemming (een verschil van 1 antwoordcategorie bij 5 antwoordmogelijkheden). Uitgedrukt in intraclass correlaties varieerde de mate van overeenstemming van gering tot goed (ICC=0.32 tot 0.72).[1] De antwoorden van patiënten, naasten en artsen op de COOP/WONCA kaarten waren in alle gevallen gevoelig voor veranderingen in de diverse kwaliteit van leven aspecten tussen voor- en nameting (indien de patiënt tijdens de nameting bij retrospectie een verbetering of verslechtering aangaf, dan was deze verandering terug te vinden in de COOP/WONCA scores van de voor- en nameting). Wel waren er enkele verschillen in relatieve validiteit. Bij lichamelijke fitheid en algehele gezondheid waren de veranderingscores van de artsen het meest onderscheidend, bij sociaal functioneren en pijn die van de patiënt zelf.

[1] Richtlijnen voor de ICC als maat voor overeenstemming waren alsvolgt:
≤ 0.40, geringe overeenstemming; 0.41 - 0.60, matige overeenstemming; 0.61 - 0.80, goede overeenstemming; 0.81 - 1.00, uitstekende overeenstemming.

Hoofdstuk 3 richt zich op de subgroep van 90 patiënten die op een verpleegafdeling van het ziekenhuis behandeld werden met chemotherapie. Bij deze patiënten werden de COOP/WONCA kaarten tevens ingevuld door de verpleegkundige, zodat per patiënt drie proxy-respondenten beschikbaar waren. In dit deelonderzoek werd met name aandacht besteed aan de relatieve effecten van (drie) verschillende proxy-respondenten, (zeven) verschillende KvL-aspecten, de conditie van de patiënt en een aantal achtergrondkenmerken op de mate van patiënt-proxy overeenstemming. In de meeste gevallen kwamen de gemiddelde scores van de proxy-respondenten redelijk goed overeen met die van de patiënten. De meeste patiënt-proxy correlaties (ICC) varieerden tussen 0.40 en 0.60, hetgeen matige overeenstemming weerspiegelt. Bij directe vergelijking van de antwoorden van patiënten met die van hun naasten, artsen en verpleegkundigen was er in 41% van de gevallen sprake van exacte overeenstemming en bij 43% was er een verschil van 1 antwoordcategorie (globale overeenstemming). In 17% van de gevallen was er sprake van een (redelijk) groot patiënt-proxy verschil (2 of meer antwoordcategorieën). Zowel dit percentage grote verschillen als de patiënt-proxy correlaties waren redelijk overeenkomstig voor de drie verschillende proxy-respondenten. De mate van patiënt-proxy overeenstemming was wel verschillend voor de zeven KvL-aspecten, waarbij goede overeenstemming werd gevonden voor pijn en dagelijkse activiteiten en geringe overeenstemming voor sociale activiteiten. Het percentage grote verschillen was nauwelijks geassocieerd met leeftijd, geslacht en opleiding (van zowel de patiënt als proxy-respondent), maar was wel afhankelijk van de conditie van de patiënt. Relatief weinig grote verschillen (10%) werden geobserveerd bij patiënten met een zeer goede (ECOG 0, normale activiteit) of juist een slechte conditie (ECOG 3, grootste deel van de dag bedlegerig). Relatief veel grote verschillen (24%) kwamen voor bij patiënten met een matige conditie (ECOG 2, zelfredzaam maar deels bedlegerig).

Hoofdstuk 4 beschrijft een vergelijking van de antwoorden van patiënten en hun naasten op de EORTC QLQ-C30. Het onderzoek richt zich op de mate van patiënt-proxy overeenstemming, het patroon van overeenstemming als functie van de kwaliteit van leven van de patiënt, de relatieve betrouwbaarheid en validiteit van de informatie van patiënten en naasten, en de invloed van diverse factoren op de mate van overeenstemming. Bij de voor- en nameting waren beide informatiebronnen beschikbaar voor respectievelijk 307 en 224 patiënten. Bij vergelijking van gemiddelde scores bleek er een klein maar systematisch verschil te zijn tussen de scores van patiënten en naasten, waarbij in de meeste gevallen de naasten een slechtere kwaliteit van leven aangaven dan de patiënten zelf. Bij vergelijking op individueel niveau was de overeenstemming matig tot goed (ICC=0.42 tot 0.79). Een correlatiematrix van de scores van patiënten en naasten op alle KvL-aspecten

liet duidelijk hogere correlaties zien bij vergelijking van overeenkomstige KvL-aspecten dan bij vergelijking van verschillende KvL-aspecten (goede convergente en discriminante validiteit). De patiënt-proxy overeenstemming was relatief goed bij patiënten met een goede of juist slechte kwaliteit van leven. Een relatief geringe mate van overeenstemming werd gevonden bij patiënten met een matige kwaliteit van leven. De betrouwbaarheid (interne consistentie) en validiteit (gevoeligheid voor veranderingen in kwaliteit van leven) van de scores van naasten was vrijwel gelijk aan die van de patiënten. Diverse kenmerken van de patiënten en hun naasten (zoals de neiging van patiënten om sociaal wenselijke antwoorden te geven of de gezondheid en kwaliteit van leven van de naasten zelf) waren geassocieerd met de mate van patiënt-proxy overeenstemming, maar verklaarden slechts minder dan 15% van de variantie in de geobserveerde verschillen.

Hoofdstuk 5 tot 7 beschrijft de resultaten van drie klinische onderzoeken bij specifieke patiëntengroepen, waarin niet alleen KvL-gegevens werden verzameld bij de patiënten maar ook bij hun naasten. Twee onderzoeken vonden plaats bij kankerpatiënten, waarbij in beide gevallen gebruik werd gemaakt van de EORTC QLQ-C30 en een tumorspecifieke vragenlijstmodule. De derde studie werd verricht bij patiënten die een beroerte (herseninfarct of hersenbloeding) hebben doorgemaakt, waarbij gebruik werd gemaakt van de SIP.

Hoofdstuk 5 toont een vergelijking van de antwoorden van 103 patiënten met een hersentumor en hun naasten op de EORTC QLQ-C30 en een hersentumor-specifieke vragenlijstmodule. Voor een aantal KvL-aspecten werden statistisch significante verschillen in gemiddelde scores waargenomen, waarbij in de meeste gevallen de naasten een slechtere kwaliteit van leven aangaven dan de patiënten zelf. Rekening houdend met de variantie in de scores waren deze systematische verschillen echter van relatief geringe betekenis, met uitzondering van die voor vermoeidheid. Voor de meeste KvL-aspecten was de mate van patiënt-proxy overeenstemming matig tot goed (ICC=0.41 tot 0.74). Lage correlaties (ICC<0.40) werden waargenomen bij scores op enkele weinig voorkomende klachten en symptomen (meestal gemeten met één item). In deze gevallen worden de lage correlaties wellicht eerder veroorzaakt door een gebrek aan variantie dan door een gebrek aan overeenstemming tussen de patiënt en naaste. De mate van overeenstemming was lager bij patiënten met een relatief slechte conditie, vooral bij patiënten die volgens de behandelend neuroloog tekenen van enige verwardheid vertoonden. Ook bij longitudinale analyse bleek dat de mate van overeenstemming verslechterde bij patiënten die tussen de voor- en nameting een lichamelijke of neurologische achteruitgang vertoonden.

Hoofdstuk 6 geeft een vergelijking van de antwoorden van 72 mannen met uitgezaaide prostaatkanker en hun partners op de EORTC QLQ-C30 en een prostaatkanker-specifieke vragenlijstmodule. In de twee vragenlijsten worden in totaal 21 KvL-aspecten gemeten. Bij 5 van de 21 patiënt-proxy vergelijkingen werden geringe maar systematische verschillen in gemiddelde scores waargenomen, waarbij de partners een slechtere kwaliteit van leven aangaven dan de patiënten zelf. Op individueel niveau was bij 19 van de 21 KvL-aspecten sprake van matige tot goede patiënt-proxy overeenstemming (ICC=0.47 tot 0.75). Lage correlaties (ICC<0.40) werden alleen gevonden bij scores van de patiënten en hun partners voor sexueel functioneren, die deels verklaard konden worden door (een combinatie van) gebrek aan betrouwbaarheid, gebrek aan variantie en enkele extreme verschillen. Naast vergelijking van de KvL-aspecten is ook gekeken naar het percentage overeenstemming bij de 32 (van in totaal 41) vragen, waarbij de patiënten en partners gevraagd werden de ernst van specifieke klachten en functionele beperkingen aan te geven (4 antwoordmogelijkheden). Bij in totaal 2286 directe vergelijkingen van de antwoorden van patiënten en partners was er in 63% van de gevallen sprake van exacte overeenstemming en bij 31% was er een verschil van 1 antwoordcategorie. Bij een kleine minderheid (6%) was er sprake van duidelijke verschillen.

Hoofdstuk 7 doet verslag van een onderzoek bij 437 patiënten die zes maanden eerder een beroerte (herseninfarct of hersenbloeding) hadden doorgemaakt. Kwaliteit van leven werd gemeten met behulp van de SIP. Een kwart van de patiënten was niet in staat om zelf KvL-gegevens te verstrekken. Voor deze patiënten werd de SIP ingevuld door een naaste (meestal de partner). Voor 228 van de overige patiënten werden KvL-gegevens verzameld bij zowel de patiënten als hun naasten. Bij deze groep patiënten werd de mate van patiënt-proxy overeenstemming bepaald. Daarnaast werd in dit onderzoek een inschatting gemaakt van de mate waarin het gebruik van proxy-metingen bij een kwart van de patiënten invloed had op de uitkomsten van één van de onderzoeksvraagstellingen (de samenhang tussen het type beroerte en kwaliteit van leven). De gemiddelde SIP-scores van de 228 patiënten en hun naasten kwamen in de meeste gevallen redelijk goed overeen. Toch waren de waargenomen verschillen meestal statistisch significant, waarbij de naasten een slechtere kwaliteit van leven aangaven dan de patiënten zelf. De intraclass correlaties waren matig tot uitstekend voor de diverse SIP-schalen (gemiddelde ICC=0.63), de lichamelijke (ICC=0.85) en psychosociale dimensie (ICC=0.61), en de totale SIP-score (ICC=0.77). De validiteit van de scores van naasten, in termen van de gevoeligheid voor verschillen in de conditie van de patiënt, kwam overeen met die van de patiënten. Bij analyse van de totale groep patiënten (met proxy-metingen voor een kwart van de populatie) bleken patiënten

met een supratentoriaal (sub)corticaal infarct of een hersenbloeding een slechtere kwaliteit van leven te hebben dan patiënten met een lacunair infarct of een infratentoriaal infarct. Indien de analyse beperkt zou zijn tot patiënten die zelf in staat waren KvL-gegevens te verstrekken (drie-kwart van de populatie) dan zou deze conclusie over de samenhang tussen het type beroerte en kwaliteit van leven niet getrokken kunnen worden en zou er sprake zijn van een algehele onderschatting van de beperkingen van patiënten met een beroerte.

Hoofdstuk 8 geeft een kwantitatief overzicht van de resultaten van de 6 bovenstaande patiënt-proxy studies en 17 andere recentelijk gepubliceerde onderzoeken naar de mate van patiënt-proxy overeenstemming op gestandaardiseerde multidimensionele KvL-vragenlijsten. Daarnaast wordt ingegaan op een aantal methodologische kwesties omtrent het bepalen van de bruikbaarheid van proxy-metingen van kwaliteit van leven.

De resultaten van de meeste onderzoeken duiden op een matige tot goede overeenstemming tussen de KvL-scores van patiënten en proxy-respondenten. Lage overeenstemming werd voornamelijk aangetroffen in onderzoeken met een geringe steekproefgrootte (<50 patiënt-proxy paren). In grotere onderzoeken die patiënten en hun naasten vergelijken, waren de mediane correlaties tussen de 0.60 en 0.70 voor lichamelijke KvL-aspecten (lichamelijk en dagelijks functioneren, algehele gezondheid, pijn en vermoeidheid) en rond de 0.50 voor psychosociale aspecten (psychisch, sociaal en cognitief functioneren). In studies die patiënten en hun zorgverleners vergelijken, was vaak sprake van een geringe steekproefgrootte en werden meer uiteenlopende resultaten gevonden. Op basis van de huidige resultaten kan geconcludeerd worden dat de naasten en zorgverleners van patiënten in staat zijn om redelijk goede informatie te verschaffen over diverse aspecten van de kwaliteit van leven van de patiënten. In een minderheid van de gevallen zijn er duidelijke verschillen tussen de scores van patiënten en hun naasten of zorgverleners. Lage patiënt-proxy correlaties hoeven niet altijd veroorzaakt te worden door gebrek aan overeenstemming, maar kunnen ook verklaard worden door een aantal (met elkaar samenhangende) methodologische beperkingen, zoals geringe steekproefgrootte, gebrek aan betrouwbaarheid en gebrek aan variantie.

In een eerder overzicht van de patiënt-proxy literatuur werd geconcludeerd dat naasten en zorgverleners meestal een slechtere kwaliteit van leven aangeven dan de patiënten zelf en dat de overeenstemming voor lichamelijke KvL-aspecten meestal beter is dan voor psychosociale aspecten. Het huidige overzicht bevestigt deze conclusies, maar laat tevens zien dat deze verschillen kleiner zijn dan vaak wordt verondersteld.

Hoewel op het terrein van kwaliteit van leven veel onderzoek gedaan wordt naar de mate van overeenstemming tussen de oordelen van patiënten en hun naasten of zorgverleners, zou aanvullend onderzoek naar de bruikbaarheid van proxy-metingen zich juist minder hierop moeten richten. Ten eerste is dergelijk onderzoek alleen zinvol als voldaan wordt aan een aantal methodologische vereisten (tenminste een toereikende steekproefgrootte en voldoende spreiding in de KvL-gegevens). Helaas wordt dit in veel studies in onvoldoende mate onderkend. Ten tweede is het niet altijd terecht om verschillen tussen de scores van patiënten en proxy-respondenten te interpreteren als gebrek aan validiteit van de proxy-meting. Dit geldt met name bij aandoeningen die gevolgen kunnen hebben voor de cognitieve vaardigheden van de patiënt (zoals een beroerte of hersentumor). Ten derde kunnen patiënt-proxy vergelijkingen alleen plaatsvinden bij patiënten die zelf in staat zijn KvL-gegevens te verstrekken. Het generaliseren van bevindingen naar patiënten die dat niet meer kunnen is aanvechtbaar.

Toekomstig onderzoek naar de bruikbaarheid van proxy-metingen van kwaliteit van leven kan zich richten op twee aspecten. In de eerste plaats ligt het voor de hand om de betrouwbaarheid en validiteit van proxy-metingen te bepalen volgens bekende psychometrische methoden, op dezelfde wijze als bij gegevens van patiënten zelf. Ook als er geen patiënt-gegevens beschikbaar zijn, kan de betrouwbaarheid en validiteit van proxy-metingen worden vastgesteld. Daarnaast is het van belang om onderzoek te doen naar de toegevoegde waarde van proxy-metingen in klinisch onderzoek. Scores van proxy-respondenten komen misschien niet altijd exact overeen met die van patiënten zelf. Het ontbreken van KvL-gegevens voor een juist zeer relevant deel van de patiënten heeft echter een groter effect op de representativiteit van het onderzoek.

Dankwoord

Graag wil ik iedereen bedanken zonder wiens medewerking en steun dit proefschrift niet tot stand zou zijn gekomen. Enkele mensen wil ik in het bijzonder noemen.

Als eerste wil ik mijn promotor Neil Aaronson hartelijk danken voor zijn inzet en begeleiding tijdens de uiteindelijk langer dan verwachte periode die het voltooien van het proefschrift met zich meebracht. Zijn niet aflatende kritische blik, vanaf het prille begin van het proxy-project in 1993 tot het eindprodukt in 2002, heeft in belangrijke mate bijgedragen aan de kwaliteit van het onderzoek, de artikelen en het proefschrift. Mijn copromotor Mirjam Sprangers ben ik eveneens veel dank verschuldigd. Na haar grondige analyse van de literatuur, overzichtelijk beschreven in het veelgeprezen artikel van Sprangers&Aaronson in 1992, en de vele voorbereidende werkzaamheden voor het proxy-project kon er eigenlijks niks meer fout gaan.

Het vlekkeloze verloop van het proxy-project is voor een groot deel te danken aan Symone Detmar en Lidwina Wever. Betere collega's had ik me niet kunnen wensen. Door hun inzet en praktische manier van werken was de dataverzameling eerder klaar dan verwacht. Toch een hele prestatie om dan pas zes jaar later te promoveren. Ik waardeer het dan ook zeer dat zij mij als paranimfen toch nog bij de laatste loodjes willen helpen.

Vele medewerkers van het Antoni van Leeuwenhoek ziekenhuis hebben hun steentje aan het onderzoek bijgedragen. Er was altijd een goede samenwerking met de internisten, afdelingsartsen en verpleegkundigen die diverse keren het bekende vragenlijstje met de poppetjes moesten invullen. Jan Schornagel dank ik voor zijn adviezen in de opzetfase van het onderzoek.

Voorts wil ik de vele collega's bedanken die mijn twaalfjarige carrière als onderzoeker tot een leuke en leerzame periode hebben gemaakt. Zonder anderen tekort te doen wil ik vooral Martin Muller en Harm van Tinteren bedanken, met wie ik allebei lange tijd een kamer deelde. Met hen heb ik vaak gebrainstormd over allerlei methodologische en statistische problemen. Gelukkig bleef er altijd genoeg tijd over om het te hebben over politiek, voetbal, aandelen en andere zaken die het leven leuk maken.

Tenslotte wil ik een speciaal dankwoord richten aan alle patiënten en hun partners, familieleden of vrienden die belangeloos hun medewerking verleenden aan de verschillende onderzoeken.

Kommer Sneeuw werd op 5 mei 1965 geboren te Emmeloord. In 1983 behaalde hij het eindexamen Gymnasium aan het Marnix van St.Aldegondecollege te Haarlem. Datzelfde jaar begon hij met de studie Voeding van de Mens aan de Landbouwuniversiteit te Wageningen. In 1989 behaalde hij zijn doctoraal examen met als hoofdvakken gezondheidsleer en voorlichtingskunde.

Van 1990 tot 1999 heeft hij gewerkt op de afdeling Psychosociaal Onderzoek en Epidemiologie van het Nederlands Kanker Instituut te Amsterdam, waar hij betrokken was bij onderzoek naar de kwaliteit van leven van kankerpatiënten. Vanaf 1993 werd het in dit proefschrift beschreven onderzoek uitgevoerd. Vanaf 1997 combineerde hij dit met een halftijdse aanstelling bij het Integraal Kankercentrum Amsterdam, waar hij als methodoloog/statisticus betrokken was bij onderzoek naar medische behandelingen en kwaliteit van zorg. Van 1999 tot 2002 is hij werkzaam geweest bij de divisie Jeugd van TNO Preventie en Gezondheid te Leiden, waar hij gewerkt heeft aan onderzoek op het terrein van kraamzorg en jeugdgezondheidszorg.

Vanaf 1996 is hij als vrijwilliger actief geweest in een Haarlemse wijkraad, waar hij betrokken was bij de voorbereiding van diverse projecten op het gebied van verkeer en vervoer. Sinds april 2002 is hij werkzaam als ontwerper/adviseur verkeer en vervoer bij de afdeling Openbare Werken van de gemeente Velsen.