ARTHRITIS & RHEUMATISM
Vol. 56, No. 5, May 2007, pp 1676–1684
DOI 10.1002/art.22580

Impact of Oral Cyclophosphamide on Health-Related Quality of Life in Patients With Active Scleroderma Lung Disease

Results From the Scleroderma Lung Study

Dinesh Khanna,[1] Xiaohong Yan,[2] Donald P. Tashkin,[2] Daniel E. Furst,[2] Robert Elashoff,[2] Michael D. Roth,[2] Richard Silver,[3] Charlie Strange,[3] Marcy Bolster,[3] James R. Seibold,[4] David J. Riley,[5] Vivien M. Hsu,[5] John Varga,[6] Dean E. Schraufnagel,[6] Arthur Theodore,[7] Robert Simms,[7] Robert Wise,[8] Fredrick Wigley,[8] Barbara White,[8] Virginia Steen,[9] Charles Read,[9] Maureen Mayes,[10] Ed Parsley,[10] Kamal Mubarak,[11] M. Kari Connolly,[12] Jeffrey Golden,[12] Mitchell Olman,[13] Barri Fessler,[13] Naomi Rothfield,[14] Mark Metersky,[14] and Philip J. Clements,[2] for the Scleroderma Lung Study Group

Objective. **To assess the impact of cyclophosphamide (CYC) on the health-related quality of life (HRQOL) of patients with scleroderma after 12 months of treatment.**

Methods. **One hundred fifty-eight subjects participated in the Scleroderma Lung Study, with 79 each randomized to CYC and placebo arms. The study evaluated the results of 3 measures of health status: the Short Form 36 (SF-36), the Health Assessment Questionnaire (HAQ) disability index (DI), and Mahler's dyspnea index, and the results of 1 preference-based measure, the SF-6D. The differences in the HRQOL between the 2 groups at 12 months were calculated using a linear mixed model. Responsiveness was evaluated using the effect size. The proportion of subjects in each treatment group whose scores improved at least as much as or more than the minimum clinically important difference (MCID) in HRQOL measures was assessed.**

Results. **After adjustment for baseline scores, differences in the HAQ DI, SF-36 role physical, general health, vitality, role emotional, mental health scales, and SF-36 mental component summary (MCS) score were statistically significant for CYC versus placebo ($P < 0.05$). Effect sizes were negligible (<0.20) for all of the**

Supported by the NIH (National Heart, Lung, and Blood Institute grant U01-HL605). Dr. Khanna's work was supported by a Physician Scientist Development Award from the Arthritis Foundation and the Scleroderma Foundation, and by an NIH Building Interdisciplinary Research Careers in Women's Health (BIRCWH K12) Award (grant HD-051953). Dr. Tashkin's work was supported by NIH grants U01-HL-60587 and U01-HL-60606. Dr. Connolly's work was supported by the NIH (National Heart, Lung, and Blood Institute grant 5-U01-H-660587).

[1]Dinesh Khanna, MD, MSc: University of Cincinnati, Cincinnati, Ohio; [2]Xiaohong Yan, MSc, Donald P. Tashkin, MD, Daniel E. Furst, MD, Robert Elashoff, PhD, Michael D. Roth, MD, Philip J. Clements, MD, MPH: University of California, Los Angeles; [3]Richard Silver, MD, Charlie Strange, MD, Marcy Bolster, MD: Medical University of South Carolina, Charleston; [4]James R. Seibold, MD: University of Michigan, Ann Arbor; [5]David J. Riley, MD, Vivien M. Hsu, MD: University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, New Brunswick; [6]John Varga, MD, Dean E. Schraufnagel, MD: University of Illinois, Chicago; [7]Arthur Theodore, MD, Robert Simms, MD: Boston University, Boston, Massachusetts; [8]Robert Wise, MD, Fredrick Wigley, MD, Barbara White, MD: Johns Hopkins School of Medicine, Baltimore, Maryland; [9]Virginia Steen, MD, Charles Read, MD: Georgetown University, Washington, DC; [10]Maureen Mayes, MD, MPH, Ed Parsley, DO: University of Texas–Houston Medical School; [11]Kamal Mubarak, MD: Wayne State University, Detroit, Michigan; [12]M. Kari Connolly, MD, Jeffrey Golden, MD: University of California, San Francisco; [13]Mitchell Olman, MD, Barri Fessler, MD: University of Alabama, Birmingham; [14]Naomi Rothfield, MD, Mark Metersky, MD: University of Connecticut Health Center, Farmington.

Dr. Simms has received consulting fees, speaking fees, and honoraria (less than $10,000) from Actelion. Dr. Mayes has received honoraria (less than $10,000 each) from Actelion and Encysive. Dr. Mubarak has received consulting fees, speaking fees, and honoraria (less than $10,000 each) from Actelion, Encysive, Myogen, Pfizer, InterMune, and United Therapeutics.

Address correspondence and reprint requests to Dinesh Khanna, MD, MSc, Division of Immunology, Department of Medicine, University of Cincinnati, PO Box 670563, Cincinnati, OH 45267-0563. E-mail: Dinesh.Khanna@uc.edu.

Submitted for publication October 23, 2006; accepted in revised form January 29, 2007.

scales of the SF-36, HAQ DI, and SF-6D at 12 months. In contrast, a higher proportion of patients who received CYC achieved the MCID compared with placebo in the HAQ DI score (30.9% versus 14.8%), transitional dyspnea index score (46.4% versus 12.7%), SF-36 MCS score (33.3% versus 18.5%), and SF-6D score (21.3% versus 3.8%).

Conclusion. **One year of treatment with CYC leads to an improvement in HRQOL in patients with scleroderma lung disease.**

The Scleroderma Lung Study (SLS) (1) is a double-blind, randomized, placebo-controlled trial of the effects of oral cyclophosphamide (CYC), administered for 1 year, on the course of forced vital capacity (FVC) (% predicted) in patients with evidence of active systemic sclerosis (SSc; scleroderma)–related interstitial lung disease (ILD), or scleroderma lung disease (SLD). The study showed that CYC produced a statistically significant, albeit modest, improvement in the 12-month FVC % predicted relative to placebo. In addition, results of the SLS showed a beneficial effect of CYC on the health-related quality of life (HRQOL). We have previously described in detail the baseline correlates of HRQOL from the SLS (2); the present report describes the impact of treatment on the change in the HRQOL of participants in the SLS from baseline to 12 months following initiation of the study drug by treatment assignment.

Study of HRQOL has originated from 2 fundamentally different approaches: health status assessment and health value/preference/utility assessment (3,4). Generally, health status measures describe a person's functioning in 1 or more domains (e.g., physical functioning and mental well-being). The SLS evaluated results obtained using 3 measures of health status: the Short Form 36 (SF-36), a generic measure; the Health Assessment Questionnaire (HAQ) disability index (DI), a musculoskeletal-targeted measure; and Mahler's dyspnea index, a dyspnea measure.

In contrast, health value/preference/utility measures assess the value or desirability of a state of health against an external metric (5) and summarize HRQOL using a single number. Preference-based measures can be determined either directly via a face-to-face interview with a subject or indirectly based on an individual's responses to a health status questionnaire. There are 2 major families of utilities: direct and indirect (or multiattribute) (4). The direct preference–based scores can be assessed using the standard gamble (the risk of death, usually, that one would be willing to take to improve a state of health), the time tradeoff, and the rating scale (6). Indirect preference–based scores use direct utility scores from a representative general population sample for a particular health state, and these scores are then applied to different health states captured by health status measurement instruments to derive a single score (7). The SF-6D, an example of an indirect measure (8), derives preference-based scores from the SF-36.

A recent advancement in the study of HRQOL is the estimation of minimum clinically important difference (MCID), the smallest improvement in the score of a HRQOL measurement instrument that patients perceive as beneficial and that may lead to a change in disease management (9). MCID can provide a benchmark for future design of SSc clinical trials by helping researchers and clinicians understand whether differences in HRQOL scores between 2 treatment groups are significant, or if changes within 1 group over time are clinically meaningful (4,10).

PATIENTS AND METHODS

Patient selection. Patients with SSc, as defined by the American College of Rheumatology (formerly, the American Rheumatism Association) classification criteria (11), and a disease duration of ≤7 years (with onset defined as the date of the first typical non–Raynaud's phenomenon manifestation) were included in the current study. Other inclusion criteria were as follows: FVC ≤85% of predicted, dyspnea on exertion (grade 2 or higher on the magnitude of task component of Mahler's baseline dyspnea index [BDI]), and evidence of alveolitis on bronchoalveolar lavage (neutrophils ≥3%, eosinophils ≥2%) and/or ground-glass opacification on high-resolution computed tomography (2). The complete inclusion and exclusion criteria were recently published (1).

Health status measurement instruments. Health status was assessed using SF-36 version 2, the HAQ DI, and Mahler's dyspnea index. The SF-36 is a generic measure of HRQOL consisting of 8 scales assessing 36 items (12,13). In addition, it includes a single item that assesses health transition. The 8 SF-36 scales can be summarized into physical component summary (PCS) and mental component summary (MCS) scores. The 8 scales and summary scores are standardized to responses from the US general population, for which the mean score is 50 and the standard deviation is 10 (14). We used the SF-36 with a standard (4-week) recall period. The MCID for the SF-36 summary scores is between 2.5 and 5.0 in different arthritides (15–17) (Table 1).

The HAQ DI is a disease-specific, musculoskeletal-targeted measure designed to assess functional ability in arthritis (18). It is a self-administered 20-question instrument designed to assess a patient's level of upper and lower extremity functioning. The HAQ DI score is determined by summing the highest score in each of the 8 domains and dividing the sum by 8, yielding a score ranging from 0 (no disability) to 3 (severe disability). In the original HAQ DI, an additional grade of

Table 1. Minimum clinically important difference (MCID) scores for the HRQOL measurement instruments*

Instrument	MCID score
SF-36 summary scores	2.5–5.0
Health Assessment Questionnaire disability index	
Systemic sclerosis	0.14
Rheumatoid arthritis	0.22
Mahler's dyspnea index	1.0
SF-6D	0.041

* HRQOL = health-related quality of life; SF-36 = Short Form 36.

difficulty was added for patients using assistive/adaptive devices (such as canes or walkers). For consistency with other recent studies of SSc (19,20), the patients' responses were not modified for use of assistive/adaptive devices. An improvement of $\geq$0.14 and $\geq$0.22 in the HAQ DI score is considered to be the MCID in patients with SSc (10) and rheumatoid arthritis (21), respectively.

Mahler's dyspnea index allows patients to assess their own level of dyspnea (22). Dyspnea at baseline is assessed using the BDI. Scores on the BDI depend on ratings of 3 different categories: functional impairment, magnitude of task, and magnitude of effort. Limitation of ability in each of these 3 categories is graded from 0 (severe) to 4 (unimpaired). The ratings of the 3 categories are added to produce the total baseline score, ranging from 0 (severe) to 12 (no dyspnea). The transitional dyspnea index (TDI) assesses the change in dyspnea in each of the 3 categories, with scores ranging from -3 (major deterioration) to $+3$ (major improvement) for each domain, with the TDI focal score being the sum of scores in the 3 domains (-9 to $+9$). An improvement of 1 unit in the TDI is considered to be the MCID (23).

Dyspnea was also assessed using the breathing visual analog scale (VAS; 0–100 mm), in which patients assessed, on a continuous scale, their own degree of difficulty in performing daily activities due to shortness of breath.

Preference-based measure. In the SF-6D (24), preference-based scores are derived from the SF-36. The SF-36 was revised into a health state classification system consisting of 6 dimensions (physical function, role limitation, social function, pain, mental health, and vitality), with a single HRQOL score derived from 11 items of the SF-36, covering the 6 dimensions. To assess preferences for the multi-attribute health states defined by the SF-6D, Brazier et al (24) used an interviewer-administered standard gamble in a representative sample from the UK. The weights for the US population have not been developed. The SF-6D scores ranged from 0.29 to 1.00, with a score of 1.00 indicating perfect or full health (24). An improvement of 0.041 in the SF-6D score was considered to be the MCID (25).

Statistical analysis. We analyzed the HRQOL data using the approach proposed by Osoba and colleagues (26). Baseline descriptive statistics for the SF-36, HAQ DI, BDI, and SF-6D scores, and the percentage with floor and ceiling effects, were calculated. Floor and ceiling effects are the percentages of respondents scoring at the lowest and highest possible scale levels. These effects can influence responsiveness because they may limit a change in score over time (27).

The baseline scores between the 2 groups were compared using Student's *t*-test. The instruments were administered every 3 months; we assessed the proportion of subjects in the CYC and placebo groups who completed the instruments at months 0, 3, 6, 9, and 12. Internal consistency reliability for multi-item scales was estimated using Cronbach's alpha (28), with $\alpha \geq 0.70$ considered satisfactory for group comparisons (28). We calculated the differences in the HRQOL measures between the 2 groups at 12 months using a linear mixed model and adjusting for the baseline value (29,30). Unadjusted strengths of association between HRQOL measures and physiologic impairment (FVC % predicted and diffusing capacity for carbon monoxide [DLCO] % predicted) at baseline were assessed using Pearson's correlation coefficients and interpreted as proposed by Franzblau (31): 0.0–0.20 = no correlation, 0.21–0.40 = low degree of correlation, 0.41–0.60 = moderate degree, 0.61–0.80 = marked degree, and 0.81–1.00 = high degree.

Because the number of patients in a study can affect statistical significance, we examined clinical significance in 2 additional ways, by calculating effect size and the proportion of patients whose change was greater than the MCID (32). Responsiveness to change was evaluated using the effect size (27). Effect size was assessed using the following formula:

$$M_{CYC} - M_{PLAC}/SD_{PLAC} \text{ at baseline,}$$

where $M_{CYC} - M_{PLAC}$ is the average difference at month 12 between the CYC and placebo arms. Cohen's guide for interpreting effect size for HRQOL data is that a value of 0.20–0.49 represents a small change, 0.50–0.79 a medium change, and $\geq$0.80 a large change (20,33).

We assessed the proportion of patients whose SF-36 summary score improved by 5 units from baseline to month 12, the proportion of patients whose HAQ DI score improved by $\geq$0.14 and $\geq$0.22, the percentage of patients whose TDI changed by $\geq$1 unit, and the proportion of patients whose SF-6D score improved by $\geq$0.041. The aforementioned cut points were selected a priori (Table 1). Based on the proportion of patients whose scores improved at least as much as or more than the MCID, we calculated the number needed to treat (NNT) (34) in order to achieve, on average, 1 patient with improved HRQOL:

$$NNT = 1/([Improved_{CYC}/Total_{CYC}] - [Improved_{PLACEBO}/Total_{PLACEBO}]) \times 100.$$

As suggested by Osoba and colleagues (35), we did not use the Bonferroni adjustment because it assumes that the variables being tested are completely independent of each other; however, some SF-36, HAQ DI, and Mahler's dyspnea index scales are moderately correlated (13,36,37). Statistical analyses were performed using SAS software, version 8.02 (SAS Institute, Cary, NC). *P* values less than 0.05 were considered significant.

RESULTS

The main findings of the SLS have recently been published (1). Briefly, the mean $\pm$ SD age of the study population was 48.5 $\pm$ 12.3 years, most of the participants were women (71.0%), and the patients had a mean $\pm$ SD disease duration of 3.1 $\pm$ 2.1 years, mild functional disability as assessed by the HAQ DI (0.82 $\pm$

Table 2. Baseline characteristics of the participants*

	All (n = 158)	CYC (n = 79)	Placebo (n = 79)
Age, years	48.5 ± 12.3	48.8 ± 12.2	48.2 ± 12.4
% female	71.0	76.5	64.6
SSc duration, years	3.1 ± 2.1	3.1 ± 2.3	3.1 ± 1.9
% with diffuse SSc	58.6	61.4	55.7
FVC, % predicted	68.1 ± 11.3	67.8 ± 12.9	68.4 ± 12.1
Skin score (range 0–51)	14.7 ± 10.9	15.4 ± 11.3	14.0 ± 10.5
Diffuse SSc	21 ± 10.3	21.6 ± 9.4	20.4 ± 9.8
Limited SSc	5.7 ± 3.5	5.8 ± 3.3	5.6 ± 3.4
SF-36			
Physical function	34.43 ± 11.32	33.2 ± 11.0	35.7 ± 11.6
Role physical	36.17 ± 12.38	34.3 ± 11.5	38.1 ± 13.0
Body pain	41.97 ± 10.55	41.4 ± 10.6	42.5 ± 10.5
General health	35.71 ± 10.32	35.7 ± 10.3	35.7 ± 10.3
Vitality	40.24 ± 11.31	39.3 ± 10.04	41.2 ± 12.5
Social function	43.94 ± 12.23	43.8 ± 12.3	44.1 ± 12.2
Role emotional	45.26 ± 12.69	44.1 ± 12.8	46.5 ± 12.6
Mental health	48.74 ± 10.24	47.3 ± 10.2	50.2 ± 10.1
SF-36 PCS	33.43 ± 10.74	32.6 ± 10.8	34.2 ± 10.7
SF-36 MCS	49.65 ± 10.53	48.7 ± 10.5	50.6 ± 10.5
HAQ DI	0.82 ± 0.69	0.95 ± 0.66†	0.71 ± 0.70
Mahler's BDI focal score (0–12)	5.7 ± 1.89	5.7 ± 1.81	5.7 ± 1.96
Functional impairment score (0–4)	1.74 ± 0.84	1.72 ± 0.78	1.76 ± 0.89
Magnitude of task score (0–4)	1.99 ± 0.68	1.99 ± 0.72	1.99 ± 0.65
Magnitude of effort score (0–4)	1.97 ± 0.70	1.99 ± 0.69	1.96 ± 0.71
Breathing VAS	28.35 ± 26.20	27.39 ± 24.91	29.29 ± 27.54
SF-6D	0.63 ± 0.10	0.63 ± 0.10	0.64 ± 0.09

* Except where indicated otherwise, values are the mean ± SD. For the Short Form 36 (SF-36), Mahler's dyspnea index score, and SF-6D, higher scores indicate better health, better breathing, and greater desirability of one's current health state, respectively. For the Health Assessment Questionnaire (HAQ) disability index (DI), a higher score indicates greater functional disability. CYC = cyclophosphamide; SSc = systemic sclerosis; FVC = forced vital capacity; PCS = physical component summary; MCS = mental component summary; BDI = baseline dyspnea index; VAS = visual analog scale.
† $P = 0.03$ versus placebo.

0.69), and moderate dyspnea as assessed by Mahler's BDI (5.68 ± 1.89) (Table 2). On average, the participants rated their physical and mental health as 1.7 SD and 0.1 SD, respectively, below that of the adjusted general US population (4). The SF-6D score was 0.63; in other words, on average, the participants in the SLS rated their health as 63% of perfect health.

Comparison of HRQOL scores in the CYC and placebo groups at baseline. At baseline, the mean HAQ DI score was significantly higher in the CYC group (0.95 ± 0.66) compared with the placebo group (0.71 ± 0.70) ($P = 0.03$), whereas the SF-36 role physical (38.1 ± 13.0) and mental health (50.2 ± 10.1) scores were numerically higher in the placebo group compared with the CYC group (34.3 ± 11.5 and 47.3 ± 10.2; $P = 0.05$ and $P = 0.07$), respectively. Otherwise, there were no statistically significant differences in the HRQOL measures between the 2 groups (Table 2).

Reliability. Floor effects were more common in the HAQ DI (13.1%) than the BDI (1.24%); no floor effect was seen for the SF-36, SF-36 PCS, or SF-36 MCS scores. There were no ceiling effects for the 3 instruments. Internal consistency reliability scores were acceptable for all 3 measures: Cronbach's alpha was >0.90 for the total HAQ DI score, 0.81 for the total BDI score, and ranged from 0.89 for the social functioning scale to 0.90 for the mental health scale of the SF-36.

Change in the HRQOL over 12 months. The percentage of patients who completed the HRQOL measures (calculated as the percentage of patients with FVC % predicted data available at each followup visit) in the CYC and placebo groups, respectively, was 88.9% versus 90.1% at 3 months, 88.9% versus 88.9% at 6 months, 83.9% versus 80.3% at 9 months, and 85.2% versus 80.3% at 12 months. The difference was not statistically significant at any time point and therefore no statistical modeling was performed for missing data (38).

Association between HRQOL and physiologic measure scores at 12 months. SF-36 PCS and breathing VAS showed a low degree of correlation with FVC % predicted and DLco % predicted, SF-36 MCS showed no correlation with FVC % predicted or DLco % pre-

Table 3. Correlation coefficients between health-related quality of life (HRQOL) and physiologic measure change scores at 12 months*

	HRQOL measures						Physiologic measures	
	SF-36 PCS	SF-36 MCS	HAQ DI	TDI	Breathing VAS	SF-6D	FVC % predicted	DLCO % predicted
SF-36 PCS	1.00	−0.14	−0.42†	0.37†	−0.27†	0.45†	0.25†	0.24†
SF-36 MCS		1.00	0.12	0.19†	−0.26†	0.52†	0.13	0.03
HAQ DI			1.00	0.29†	0.33†	−0.24†	−0.34†	−0.18
TDI				1.00	−0.30†	0.37†	0.41†	0.31†
Breathing VAS					1.00	−0.23†	−0.32†	−0.24†
SF-6D						1.00	0.28†	0.32†
FVC % predicted							1.00	0.52†
DLCO % predicted								1.00

* SF-36 = Short Form 36; PCS = physical component summary; MCS = mental component summary; HAQ DI = Health Assessment Questionnaire disability index; TDI = transitional dyspnea index; VAS = visual analog scale; FVC = forced vital capacity; DLCO = diffusing capacity for carbon monoxide.
† $P < 0.05$.

dicted, and TDI had a moderate degree of correlation with FVC % predicted and a low degree with DLCO % predicted (Table 3).

We assessed the change in the HRQOL using 3 different statistical approaches: statistical significance, responsiveness index, and the proportion of patients whose scores improved more than the MCID.

Statistical significance. After adjustment for baseline scores and application of a linear mixed model, the difference in the HAQ DI score, SF-36 role physical, general health, vitality, role emotional, and mental health scores, and the SF-36 MCS score between the CYC and placebo arms was statistically significant at 12 months ($P < 0.05$) (Table 4). Since the TDI itself represents a change score and therefore was not administered at baseline, we assessed the impact of CYC on the TDI using the generalized estimating equation (39). CYC had a statistically significant favorable impact on the TDI ($P < 0.0001$) (Table 4).

Responsiveness index. In comparison, the effect sizes were negligible (<0.20) for all of the scales and summary scores of the SF-36 and the HAQ DI at 12 months. The effect size was not calculated for the TDI.

Table 4. Differences in the HRQOL scores between CYC-treated and placebo-treated patients at 12 months*

	CYC (month 12 − baseline score)	Placebo (month 12 − baseline score)	Difference	*P*
SF-36				
Physical function	−0.56	−0.57	0.009	0.99
Role physical	3.11	−1.03	4.4	<0.001
Body pain	0.60	−0.43	1.03	0.21
General health	−0.89	0.58	−1.48	0.03
Vitality	2.22	0.31	1.91	0.009
Social function	0.75	−0.22	0.97	0.31
Role emotional	1.99	−1.52	3.52	0.005
Mental health	1.51	−0.56	2.07	0.006
SF-36 PCS	0.08	−0.07	0.15	0.84
SF-36 MCS	2.12	−0.56	2.67	0.003
HAQ DI	−0.07	0.11	−0.17	0.0001
Mahler TDI focal score (−9 to +9)†	1.38	−1.25	2.63	<0.0001
Functional impairment (−3 to +3)†	0.37	−0.34	0.71	0.001
Magnitude of task (−3 to +3)†	0.62	−0.44	1.06	0.0001
Magnitude of effort (−3 to +3)†	0.39	−0.47	0.86	0.0001
Breathing VAS	−2.60	−3.22	0.62	0.90
SF-6D	0.02	0.01	−0.01	0.54

* For the SF-36 scales and summary scores and the BDI, a positive score indicates improvement, and for the HAQ DI, a negative score indicates improvement. The numbers in the difference column may not be exact due to rounding to the nearest decimal. See Table 2 for definitions.
† The generalized estimating equation model was used to assess statistical significance.

Table 5. Proportion of patients in the CYC and placebo groups whose scores improved more than the MCID*

	CYC, %	Placebo, %	*P*	NNT
SF-36 PCS score ≥5 units	21.0	18.5	0.65	40.00
SF-36 MCS score ≥5 units	33.3	18.5	0.09	6.75
HAQ DI score ≥0.14	30.9	14.8	0.05	6.25
HAQ DI score ≥0.22	30.9	14.8	0.05	6.25
TDI score ≥1.0 unit	46.4	12.7	0.0001	3.00
SF-6D score ≥0.041	21.3	3.8	0.001	5.71

* CYC = cyclophosphamide; MCID = minimal clinically important difference; NNT = number needed to treat; SF-36 = Short Form 36; PCS = physical component summary; MCS = mental component summary; HAQ DI = Health Assessment Questionnaire disability index; TDI = transitional dyspnea index.

Proportion of patients whose scores improved more than MCID. Although the effect size did not show any difference in the average scores between the CYC and placebo groups, a higher proportion of patients who took CYC compared with placebo achieved MCID in the SF-36 MCS score (33.3% versus 18.5%), HAQ DI score (30.9% versus 14.8%), and TDI score (46.4% versus 12.7%) (Table 5). Also, a greater proportion of patients who took CYC showed improvement in the health transition scale of the SF-36, a single-item measure determined by response to the question, "Compared to one year ago, how would you rate your health in general now?" In response, 47.0% of the patients who took CYC rated their health as somewhat better or much better, compared with 18.2% in the placebo group ($P < 0.001$). In comparison, only 24.2% in the CYC group rated their health somewhat worse or much worse, compared with 38.7% in the placebo group ($P = 0.01$).

The proportion of patients whose scores improved more than the MCID was then used to assess the NNT (34,40) in order to achieve, on average, 1 patient with improved HRQOL. For the HAQ DI, this was calculated as follows (Table 5):

$$\text{NNT} = 1/(31\% - 15\%) \times 100 = 6.25.$$

In other words, one would need to treat 6 or 7 patients with CYC to have 1 person experience a clinically meaningful improvement in his or her functional ability greater than that expected with placebo. Similarly, one would need to treat 3 patients with CYC to have 1 person experience a clinically meaningful difference in dyspnea compared with placebo ($\text{NNT} = 1/[46\% - 13\%] \times 100 = 3$).

Similar to health status measures, the SF-6D score showed a low degree of correlation with FVC % predicted and DLco % predicted. At 12 months, the average SF-6D score was 0.65 in the CYC group and 0.63 in the placebo group and there was no statistically significant difference between the groups ($P = 0.54$). The difference in effect size between the 2 treatment groups was negligible. In comparison, a higher proportion of patients who took CYC compared with placebo achieved MCID in the SF-6D score (21.3% versus 3.8%); one would need to treat 5 or 6 patients with CYC to have 1 person experience clinically meaningful improvement in his or her desirability of the current health state than expected with placebo.

DISCUSSION

The SLS is the first randomized, placebo-controlled trial to show a statistically significant, although modest, beneficial effect of CYC on the change in FVC % predicted compared with placebo after 12 months (1). The favorable effect of CYC on FVC was accompanied by parallel improvement in HRQOL measures, including functional disability, shortness of breath, and mental well-being, at the end of the 1-year study (1).

At baseline, the participants had mild functional disability as assessed by the HAQ DI (mean ± SD 0.82 ± 0.69) and moderate dyspnea as assessed by Mahler's BDI (5.68 ± 1.89) (Table 2). In addition, the participants rated their physical and mental health 1.7 SD and 0.1 SD, respectively, below that of the general US population, which is consistent with the results of another study on SSc in which patients had moderate impairment in their SF-36 PCS score but a normal or near-normal SF-36 MCS score (20). Although this was not formally assessed, the normal SF-36 MCS score may represent a psychological adjustment to the typically slow progression of SSc (41).

HRQOL experts recommend the use of both disease-specific and generic HRQOL measures in clinical trials (4,32). The HAQ DI, a musculoskeletal-targeted measure, was able to demonstrate a statistically significant and clinically meaningful difference in functional ability between the CYC and placebo arms at 12 months. In contrast, changes in the SF-36 physical functioning scale and the SF-36 PCS were not significantly different between the 2 groups at 12 months. However, changes in the SF-36 vitality, role emotional, and mental health scales and the SF-36 MCS were significantly different between the 2 groups and favored CYC.

Since improvements in the SF-36 were most

prominent in mental functioning scores, our findings suggest that improvement in the pulmonary and/or skin/musculoskeletal manifestations of ILD has a significant impact on mental well-being. This study confirms the opinion that disease-specific and generic HRQOL measures complement each other. Also, the associations between patient-reported outcomes and physiologic measures are only modest (42–44). The correlation coefficients between HRQOL measures and physiologic measures ranged from none (0.03) to a moderate degree (0.41), confirming previous observations that HRQOL and physiologic measures complement each other (42–44).

Analysis at the group level showed that treatment with CYC led to a statistically significant improvement in the HAQ DI, TDI, and SF-36 role physical, general health, vitality, role emotional, and mental health scales compared with placebo. Since statistical significance is dependent on the sample size, experts suggest also using a responsiveness index, such as effect size, to assess the clinical significance of such change (45). In contrast to the statistically significant results, effect sizes of CYC on functional disability, dyspnea, and physical and mental functioning scales were negligible (all were <0.20). At a group level, a change of 0.20–0.49 in effect size is considered a clinically meaningful improvement (46,47).

We also assessed the proportion of patients between the 2 groups whose scores improved more than the MCID (4) at 1 year. In contrast to negligible effect size results that assess average scores, the scores of a higher proportion of patients who took CYC improved more than the MCID in HRQOL measures, including the HAQ DI score, TDI score, and SF-36 MCS score, compared with placebo.

We used the proportion of patients whose scores improved more than the MCID to assess the NNT (34,40) in order to achieve, on average, 1 patient with improved HRQOL. This ranged from 3 patients for TDI to 40 patients for SF-36 PCS. In other words, one would need to treat 3 patients with CYC to have 1 person experience clinically meaningful improvement in dyspnea compared with placebo. We found these changes to be robust despite the group differences in baseline scores since ceiling effects did not occur.

Our results show a disparity between the modest physiologic benefit in SLS and the more obvious improvement in HRQOL measures in general, and breathlessness in particular. In other chronic respiratory diseases, small changes in lung function can have a significant clinical impact on symptoms and exercise tolerance (48–50). It is therefore not surprising that the modest changes in lung function parameters are accompanied by a more obvious and clinically meaningful improvement in dyspnea.

There are several possible mechanisms that might explain the disparity between the rather marked improvement in dyspnea and the small improvement in FVC in our SLS ILD subjects, such as a reduction in inflammation in the lung, leading to decreased J receptor stimulation and decreased excessive drive to breathe, thus reducing dyspnea; reduction in skin thickness that could improve chest wall compliance and the exertion required for breathing; and musculoskeletal improvement that could lead to greater mobility, with a favorable "training" effect that improves exercise tolerance.

Health status measurement instruments can be used to inform and monitor outcomes in clinical encounters, monitor population health, estimate the burden of different conditions, and as end points in clinical trials (3,4). However, preference-based measures serve as "quality-adjustment factors" for calculating quality-adjusted life years (QALYs) in decision and cost-effectiveness analyses that are used in resource allocation (3). A QALY takes into account both the quantity and quality of life generated by health care interventions (51). For the SF-6D, a higher proportion of patients treated with CYC had a clinically meaningful change (21%) compared with those in the placebo group (4%; $P = 0.001$). However, the difference in the average SF-6D scores at 1 year was not statistically significant (0.65 in the CYC group versus 0.63 in the placebo group; $P = 0.54$). The reason for this discrepancy is likely related to the side effect profile of CYC. If significant numbers of individuals have adverse health consequences while at the same time clinical benefit is achieved in those able to take the medication, then this seemingly paradoxical outcome is understood. Since cost-effectiveness analyses incorporate average scores, from a societal perspective this may result in an unfavorable cost-effectiveness ratio for the use of CYC in the treatment of SSc-associated ILD.

Our study has some limitations. First, we cannot determine to what extent the beneficial effects of CYC on HRQOL are due to the modest improvement in lung function as measured by the FVC % predicted and/or improvement in extrapulmonary manifestations, such as skin thickness, muscle strength, and arthritis. The improvement in HRQOL may well be due to a combination of both since the SLS showed a statistically significant improvement in both FVC % predicted and skin score (1). Whether this beneficial effect of CYC on the FVC and HRQOL measures will continue during the

second year of the SLS treatment will be determined in future analyses.

Second, Mahler's dyspnea index was developed and validated in patients with chronic obstructive pulmonary disease (22,23). Although the dyspnea index has been previously used in patients with ILD (52), it has not been subjected to psychometric rigor in this latter group of patients. While this study could have provided the platform with which to validate Mahler's dyspnea index in ILD, unfortunately no respiratory-specific questionnaire (such as St. George's Respiratory Questionnaire) was administered.

Third, floor effects were more common in HAQ DI scores (13.1%) than in other measures. In other words, 13% of the participants did not report any functional disability. This is of significance since floor effects can limit the assessment of responsiveness to change (20,27). A lower proportion of floor effects (4.5% of the participants) in the HAQ DI score was seen in another SSc clinical trial (20) in which all participants had early diffuse SSc, compared with the inclusion of subjects with both limited and diffuse SSc in the current study.

In conclusion, CYC led to an improvement in the HRQOL of patients with SLD. Patient-reported outcomes and physiologic measures complemented each other, and we present different statistical methods that can be used to assess HRQOL in a randomized clinical trial.

AUTHOR CONTRIBUTIONS

Dr. Khanna had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study design. Khanna, Tashkin, Furst, Elashoff, Roth, Silver, Bolster, Seibold, Schraufnagel, Wise, Connolly, Olman, Fessler, Metersky, Clements.

Acquisition of data. Khanna, Tashkin, Furst, Roth, Silver, Strange, Bolster, Seibold, Riley, Hsu, Vargas, Schraufnagel, Theodore, Simms, Wise, Wigley, White, Steen, Read, Mayes, Parsley, Mubarak, Connolly, Golden, Olman, Fessler, Rothfield, Metersky, Clements.

Analysis and interpretation of data. Khanna, Yan, Tashkin, Furst, Elashoff, Seibold, Wise, Steen, Mayes, Mubarak, Connolly, Olman, Metersky, Clements.

Manuscript preparation. Khanna, Tashkin, Furst, Roth, Silver, Seibold, Vargas, Schraufnagel, Wigley, White, Steen, Read, Mayes, Golden, Olman, Metersky, Clements.

Statistical analysis. Yan, Elashoff, Clements.

Medication control officer and primary rheumatology investigator. Hsu.

REFERENCES

1. Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, et al. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med 2006;354:2655–66.
2. Khanna D, Clements PJ, Furst DE, Chon Y, Elashoff R, Roth MD, et al, for the Scleroderma Lung Study Group. Correlation of the degree of dyspnea with health-related quality of life, functional abilities, and diffusing capacity for carbon monoxide in patients with systemic sclerosis and active alveolitis: results from the Scleroderma Lung Study. Arthritis Rheum 2005;52:592–600.
3. Tsevat J, Weeks JC, Guadagnoli E, Tosteson AN, Mangione CM, Pliskin JS, et al. Using health-related quality-of-life information: clinical encounters, clinical trials, and health policy. J Gen Intern Med 1994;9:576–82.
4. Khanna D. Health-related quality of life: a primer with focus on scleroderma. Scleroderma Care Res [article online]. 2006;3:3–13. URL: www.sctc-online.org/pdfs/SCARV3N2.pdf.
5. Tsevat J. What do utilities measure? Med Care 2000;38 Suppl 9:II160–4.
6. Khanna D, Ahmed M, Furst DE, Ginsburg SS, Park GS, Hornung R, et al. Health values of patients with systemic sclerosis. Arthritis Rheum 2007;57:86–93.
7. Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life [review]. J Clin Epidemiol 2003;56:317–25.
8. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 health survey. J Clin Epidemiol 1998;51:1115–28.
9. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15.
10. Khanna D, Furst DE, Hays RD, Park GS, Wong WK, Seibold JR, et al. Minimally important difference in diffuse systemic sclerosis: results from the D-penicillamine study. Ann Rheum Dis 2006;65:1325–9.
11. Subcommittee for Scleroderma Criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Arthritis Rheum 1980;23:581–90.
12. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
13. Ware JE Jr, Kosinki M, Keller S. SF-36 physical and mental health summary scales: a user's manual. Boston: The Health Institute, New England Medical Center; 1994.
14. Ware J, Kosinski M, Dewey J. How to score version two of the SF-36 health survey. Lincoln (RI): QualityMetric; 2000.
15. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE Jr. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 2000;43:1478–87.
16. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum 2001;45:384–91.
17. Strand V, Crawford B. Improvement in the health-related quality of life in patients with SLE following sustained reductions in anti-DNA antibodies. Expert Rev Pharmacoecon Outcomes Res 2005;5:317–26.
18. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
19. Clements PJ, Wong WK, Hurwitz EL, Furst DE, Mayes M, White B, et al. Correlates of the disability index of the Health Assessment Questionnaire: a measure of functional impairment in systemic sclerosis. Arthritis Rheum 1999;42:2372–80.
20. Khanna D, Furst DE, Clements PJ, Park GS, Hays RD, Yoon J, et al. Responsiveness of the SF-36 and the Health Assessment Questionnaire disability index in a systemic sclerosis clinical trial. J Rheumatol 2005;32:832–40.

21. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20: 557–60.
22. Mahler DA, Weinberg DH, Wells CK, Feinstein AR. The measurement of dyspnea: contents, interobserver agreement, and physiologic correlates of two new clinical indexes. Chest 1984;85: 751–8.
23. Witek TJ Jr, Mahler DA. Meaningful effect size and patterns of response of the transition dyspnea index. J Clin Epidemiol 2003; 56:248–55.
24. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21: 271–92.
25. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D [review]. Qual Life Res 2005;14:1523–32.
26. Osoba D, Bezjak A, Brundage M, Zee B, Tu D, Pater J. Analysis and interpretation of health-related quality-of-life data from clinical trials: basic approach of the National Cancer Institute of Canada Clinical Trials Group. Eur J Cancer 2005;41:280–7.
27. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992;1:73–5.
28. Hays RD. Reliability and validity (including responsiveness). In: Fayers P, Hays RD, editors. Assessing quality of life in clinical trials: methods and practice. 2nd ed. New York: Oxford University Press; 2005. p. 25–39.
29. Liang K, Zegar S. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–23.
30. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–74.
31. Franzblau A. A primer of statistics for non-statisticians. New York: Harcourt, Brace & World; 1958.
32. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407.
33. Cohen J. A power primer. Psychol Bull 1992;112:155–9.
34. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005;14:285–95.
35. Osoba D, Brada M, Yung WK, Prados MD. Health-related quality of life in patients with anaplastic astrocytoma during treatment with temozolomide. Eur J Cancer 2000;36:1788–95.
36. Cole JC, Motivala SJ, Khanna D, Lee JY, Paulus HE, Irwin MR. Validation of single-factor structure and scoring protocol for the Health Assessment Questionnaire-disability index. Arthritis Rheum 2005;53:536–42.
37. Cole JC, Khanna D, Clements PJ, Seibold JR, Tashkin DP, Paulus HE, et al. Single-factor scoring validation for the Health Assessment Questionnaire-disability index (HAQ-DI) in patients with systemic sclerosis and comparison with early rheumatoid arthritis patients. Qual Life Res 2006;15:1383–94.
38. Nich C, Carroll KM. Intention-to-treat meets missing data: implications of alternate strategies for analyzing clinical trials data. Drug Alcohol Depend 2002;68:121–30.
39. Jorgensen B, Tsao M. Dispersion models and longitudinal data analysis [review]. Stat Med 1999;18:2257–70.
40. Walter SD. Number needed to treat (NNT): estimation of a measure of clinical benefit. Stat Med 2001;20:3947–62.
41. Manne SL, Zautra AJ. Coping with arthritis: current status and critique [review]. Arthritis Rheum 1992;35:1273–80.
42. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life [review]. Ann Intern Med 1993;118:622–9.
43. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. JAMA 1995;273:59–65.
44. Gliddon AE, Dore CJ, Maddison PJ, and the QUINS Trial Study Group. Influence of clinical features on the health status of patients with limited cutaneous systemic sclerosis. Arthritis Rheum 2006;55:473–9.
45. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;565:395–407.
46. Sloan JA, Cella D, Hays RD. Clinical significance of patient-reported questionnaire data: another step toward consensus [editorial]. J Clin Epidemiol 2005;58:1217–9.
47. Farivar S, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in health-related quality of life scores? Expert Rev Pharmacoecon Outcomes Res 2004;4:515–23.
48. Mahler DA, Matthay RA, Snyder PE, Wells CK, Loke J. Sustained-release theophylline reduces dyspnea in nonreversible obstructive airway disease. Am Rev Respir Dis 1985;131:22–5.
49. Hay JG, Stone P, Carter J, Church S, Eyre-Brook A, Pearson MG, et al. Bronchodilator reversibility, exercise performance and breathlessness in stable chronic obstructive pulmonary disease. Eur Respir J 1992;5:659–64.
50. Bellia V, Foresi A, Bianco S, Grassi V, Olivieri D, Bensi G, et al. Efficacy and safety of oxitropium bromide, theophylline and their combination in COPD patients: a double-blind, randomized, multicentre study (BREATH Trial). Respir Med 2002;96:881–9.
51. Weinstein MC, Siegel JE, Gold MR, Kamlet M, Russell LB. Recommendations of the Panel on Cost-effectiveness in Health and Medicine [review]. JAMA 1996;276:1253–8.
52. Martinez TY, Pereira CA, dos Santos ML, Ciconelli RM, Guimaraes SM, Martinez JA. Evaluation of the short-form 36-item questionnaire to measure health-related quality of life in patients with idiopathic pulmonary fibrosis. Chest 2000;117:1627–32.