Stability of Normative Data for the SF-36

Results of a Three-year Prospective Study in Middle-aged Canadians

Wilma M. Hopman, MA¹

Claudie Berger, MSc²

Lawrence Joseph, PhD³

Tanveer Towheed, MD⁴

Elizabeth vandenKerkhof, DrPH⁵

Tassos Anastassiades, MD⁶

Ann Cranney, MD⁶

Jonathan D. Adachi, MD⁷

George Ioannidis, MSc⁷

Suzette Poliquin, BSc⁸

Jacques P. Brown, MD⁹

Timothy M. Murray, MD¹⁰

David A. Hanley, MD¹¹

Emmanuel A. Papadimitropoulos, PhD¹²

Alan Tenenhouse, MD⁷

and the CaMos Research Group¹³

ABSTRACT

Background: The SF-36 is widely used to assess health-related quality of life (HRQOL), but with few longitudinal studies in healthy populations, it is difficult to quantify its natural history. This is important because any measure of change following an intervention may be confounded by natural changes in HRQOL. This paper assesses mean changes in SF-36 scores over a 3-year period in men and women between the ages of 40 and 59 years at baseline.

Methods: Subjects were randomly selected from nine Canadian cities. Mean SF-36 changes over a 3-year period (1996/1997-1999/2000) were calculated for each gender within 5-year age categories. Multiple imputation was used to correct for potential bias due to missing data.

Results: The baseline cohort included 1,974 women and 975 men between 40 and 59 years. Mean changes in HRQOL tended to be small. Women demonstrated small average declines in 22 of the 32 age and domain groupings (4 age groups, 8 SF-36 domains) while men showed declines in 14/32. Most participants stayed within 10 points of their original baseline score.

Interpretation: Mean SF-36 scores change only slightly over three years in middle-aged Canadians, although there is much individual variation. It may be necessary to adjust for the natural evolution of SF-36 scores when interpreting results from longitudinal studies.

La traduction du résumé se trouve à la fin de l'article.

1. Clinical Research Centre, Kingston General Hospital and Department of Community Health and Epidemiology, Queen's University, Kingston, ON
2. CaMos Methods Centre, McGill University, Montreal, QC
3. Department of Epidemiology and Biostatistics, McGill University
4. Division of Rheumatology, Department of Medicine; Department of Community Health and Epidemiology, Queen's University
5. Department of Anesthesiology; Department of Community Health and Epidemiology, Queen's University
6. Division of Rheumatology, Department of Medicine, Queen's University
7. Department of Medicine, McMaster University, Hamilton, ON
8. CaMos National Coordinating Centre, McGill University
9. Laval University, Ste-Foy, QC
10. University of Toronto, Toronto, ON
11. University of Calgary, Calgary, AB
12. Eli Lilly Canada Inc, Toronto, and University of Toronto, Faculty of Pharmacy
13. See Acknowledgements for Complete List

Correspondence and reprint requests: Wilma Hopman, Research Facilitator, Clinical Research Centre, Kingston General Hospital, Kingston, ON K7L 2V7, E-mail: hopmanw@kgh.kari.net

Acknowledgements: CaMos was funded by the Senior's Independence Research Program through the National Health Research and Development Program (project no. 6605-4003-OS), the Medical Research Council of Canada-Pharmaceutical Manufacturers Association of Canada (MRC-PMAC) Health Program, Merck Frosst Canada Inc., Eli Lilly Canada Inc., Procter & Gamble Pharmaceuticals Canada, Inc., the Dairy Farmers of Canada, The Arthritis Society.

The authors thank all the participants in the Canadian Multicentre Osteoporosis Study.

Investigators in the CaMos Research Group are: McGill University, Montreal General Hospital, Montreal, Quebec (Coordinating Centre): Alan Tenenhouse (principal investigator), Suzette Poliquin (national coordinator), Suzanne Godmaire (research assistant), Lawrence Joseph (study statistician), Claudie Berger (statistician); Memorial University, St. John's, Newfoundland and Labrador: Carol Joyce (centre director), Emma Sheppard (centre coordinator); Dalhousie University, Halifax, Nova Scotia: Susan Kirkland, Stephanie Kaiser (centre co-directors), Barbara Stanfield (centre coordinator); Laval University, Quebec City, Quebec: Jacques P. Brown (centre director), Evelyne Lejeune (centre coordinator); Queen's University, Kingston, Ontario: Tassos Anastassiades (centre director), Barbara Matthews (centre coordinator); University of Toronto, Toronto, Ontario: Nancy Kreiger (study epidemiologist), Timothy M. Murray (centre director), Barbara Gardner-Bray (centre coordinator); McMaster University, Hamilton, Ontario: Jonathan D. Adachi (centre director), Laura Pickard (centre coordinator); University of Saskatchewan, Saskatoon, Saskatchewan: Wojciech P. Olszynski (centre director), Jola Thingvold (centre coordinator); University of Calgary, Calgary, Alberta: David A. Hanley (centre director), Jane Allan (centre coordinator); University of Alberta, Edmonton, Alberta: Kerry Siminoski (Quality Assurance); University of British Columbia, Vancouver, British Columbia: Jerilynn C. Prior (centre director), Brian Lentle (Radiology study consultant), Yvette Vigna (centre coordinator).

The Medical Outcomes Study 36-item short form (SF-36)^{1,2} is widely used to compare the health-related quality of life (HRQOL) of general and specific populations, estimate the burden of disease, and examine the benefits of treatment interventions.³⁻⁵ Cross-sectional data have shown that the SF-36 is reliable and able to detect differences between groups defined by age and sex,^{6,7} as well as differences between a number of clinical conditions.^{3,6} However, few reports used repeated measures of the SF-36 in populations other than patients, making it difficult to assess the natural history of HRQOL. This is important because any measure of change following an intervention, particularly if the follow-up is of several years duration, may be confounded by changes in HRQOL as a result of population aging.

Two studies using cross-lagged panel analyses and structural equation modeling demonstrated a high degree of stability in the SF-36 domains⁸ and summary scores.⁹ A longitudinal study of a population-based sample (mean follow-up of 36 months)⁶ noted that within-person declines with increasing age were greater than declines

TABLE I

Sample Size and Loss to Follow-up at Year 3, by Age Group and Gender

Age Group (Years)	Women			Men		
	Baseline Total 1996/1997	3-Year Total 1999/2000	Lost to Follow-up	Baseline Total 1996/1997	3-Year Total 1999/2000	Lost to Follow-up
40 to 44	135	120	15 (11.1 %)	89	74	15 (16.9 %)
45 to 49	487	458	29 (6.0 %)	284	247	37 (13.0 %)
50 to 54	623	554	69 (11.1 %)	302	249	53 (17.5 %)
55 to 59	729	652	77 (10.6 %)	300	265	35 (11.7 %)
Total	1974	1784	190 (9.6%)	975	835	140 (14.4 %)

TABLE II

Percent of Women Who Improved by 10 or More Points, Declined by 10 or More Points, or Stayed Within 10 Points of Their Baseline Score (1996/1997) at Three Years (1999/2000) (unadjusted values)

Women Age Group at Baseline	SF-36 Domain									
	PF	RP	BP	GH	VT	SF	RE	MH	PCS*	MCS*
40 to 44										
Improved by ≥ 10 points	10.0	20.0	30.8	18.3	24.2	31.7	15.0	12.5	16.7	16.7
Change of < 10 points	67.5	60.0	25.8	54.2	46.7	41.7	67.5	65.8	63.3	64.2
Declined by ≥ 10 points	22.5	20.0	43.3	27.5	29.2	26.7	17.5	21.7	20.0	19.2
45 to 49										
Improved by ≥ 10 points	20.5	20.5	33.6	22.5	31.0	29.9	17.7	24.9	20.3	27.7
Change of < 10 points	58.1	61.8	31.4	51.8	43.7	48.2	69.7	60.0	54.1	56.8
Declined by ≥ 10 points	21.4	17.7	34.9	25.8	25.3	21.8	12.7	15.1	25.6	15.5
50 to 54										
Improved by ≥ 10 points	15.9	15.0	28.9	23.1	27.1	27.4	13.4	20.2	18.8	22.6
Change of < 10 points	58.5	67.2	35.2	52.0	45.7	49.8	74.0	62.3	58.5	59.9
Declined by ≥ 10 points	25.6	17.9	35.9	24.9	27.3	22.7	12.6	17.5	22.7	17.5
55 to 59										
Improved by ≥ 10 points	19.5	17.8	30.8	24.5	25.8	25.1	15.0	21.3	18.9	24.4
Change of < 10 points	56.0	63.6	36.0	53.2	48.9	54.8	73.2	62.1	58.4	59.0
Declined by ≥ 10 points	24.5	18.6	33.1	22.2	25.3	20.1	11.8	16.6	22.7	16.6

Note: PF = Physical Function; RP = Role Physical; BP = Bodily Pain; GH = General Health Perceptions; VT = Vitality; SF = Social Function; RE = Role Emotional; MH = Mental Health; PCS = Physical Component Summary; MCS = Mental Component Summary

* For the PCS and the MCS, 5 was used instead of 10 as the cut-off for the computation of percents. The first row therefore represents the percentage of those who improved by 5 or more points, the middle row represents those who had a change of less than 5 points in either direction, and the last row represents those who declined by 5 points or more.

Percentages within each cell do not always equal 100 due to rounding.

estimated by cross-sectional data alone, and emphasized the need for repeated measures of the SF-36 in population-based studies.

The initiation of the Canadian Multicentre Osteoporosis Study (CaMos) in 1996 provided the opportunity to develop age- and sex-standardized normative data for Canadians⁷ and assess regional differences.¹⁰ Three years following initiation of CaMos, participants aged 40 to 60 years at baseline were interviewed for a second time. This made it possible to address the research question of whether mean SF-36 scores change over a three-year period. This may have important implications for interpretation of minimal clinically important differences in quality of life measures.¹¹⁻¹³

METHODS

CaMos is a prospective cohort study of 9,423 non-institutionalized, randomly selected men and women aged 25 years and older, drawn from a 50-kilometre radius of nine Canadian cities (St. John's,

Halifax, Quebec City, Toronto, Hamilton, Kingston, Saskatoon, Calgary and Vancouver). Details of the purpose, methodology and sampling framework are presented elsewhere.^{7,14} Briefly, households in each region were selected by random draws of listed telephone numbers, and one randomly selected household member 25 years of age or older was asked to participate. Ethics approval was obtained through the Review Boards of each participating centre.

Baseline assessments took place between February 1996 and September 1997, with the 3-year follow-up taking place between February 1999 and September 2000 for those between the ages of 40 and 60 years at baseline. Age categories of 40 to 44, 45 to 49, 50 to 54 and 55 to 59 years were examined to allow a detailed evaluation of the changes by age group. Those aged 60 years at baseline were not included since they represented a single year of age (e.g., only one year of the 60 to 64 year age group).

The SF-36 was administered via in-person interview at both time points. The

SF-36 yields 8 domain scores including physical functioning (PF), role physical (RP), role emotional (RE), bodily pain (BP), general health perceptions (GH), vitality (VT), social function (SF) and mental health (MH),¹ as well as a Physical Component Summary (PCS) and a Mental Component Summary (MCS).² Although a more recent SF-36 version is now available,³ version 1.0 was used to allow comparisons between the two time points.

Multiple imputation¹⁵ was used to provide a series of replacement values for each case with missing scores at year 3. This involved developing regression models to predict SF-36 scores at year 3, using the data for respondents who had complete data at both time points. The BIC model selection criterion¹⁶ was used to select best predictive models for each domain and summary score (for further information on multiple imputation, see Rubin¹⁵). Baseline data from those with year 3 missing data were then plugged into the model to predict the 3-year scores. Up to 10,000 iterations of the Gibbs sampler were run to ensure convergence, resulting in a distribu-

TABLE III

Percent of Men Who Improved by 10 or More Points, Declined by 10 or More Points, or Stayed Within 10 Points of Their Baseline Score (1996/1997) at Three Years (1999/2000) (unadjusted values)

Men Age Group at Baseline	SF-36 Domain									
	PF	RP	BP	GH	VT	SF	RE	MH	PCS*	MCS*
40 to 44										
Improved by ≥ 10 points	8.1	18.9	33.8	20.3	23.0	23.0	16.2	18.9	16.2	18.9
Change of < 10 points	83.8	73.0	36.5	60.8	56.8	63.5	74.3	63.5	75.7	66.2
Declined by ≥ 10 points	8.1	8.1	29.7	18.9	20.3	13.5	9.5	17.6	8.1	14.9
45 to 49										
Improved by ≥ 10 points	13.8	9.7	34.0	24.7	29.6	21.9	15.8	18.2	15.0	20.6
Change of < 10 points	70.8	79.0	35.2	56.7	48.6	57.5	76.5	66.4	68.4	66.4
Declined by ≥ 10 points	15.4	11.3	30.8	18.6	21.9	20.6	7.7	15.4	16.6	13.0
50 to 54										
Improved by ≥ 10 points	15.3	17.3	27.7	20.9	26.9	24.1	12.8	19.7	14.5	23.3
Change of < 10 points	62.2	71.1	39.0	53.8	46.2	58.6	75.9	66.3	64.3	57.8
Declined by ≥ 10 points	22.5	11.7	33.3	25.3	26.9	17.3	11.2	14.1	21.3	18.9
55 to 59										
Improved by ≥ 10 points	19.6	15.1	32.8	28.7	23.0	24.1	12.1	20.0	18.1	18.9
Change of < 10 points	58.1	69.1	43.8	53.2	54.0	58.5	77.7	64.5	64.9	65.3
Declined by ≥ 10 points	22.3	15.8	23.4	18.1	23.0	17.4	10.2	15.5	17.0	15.8

* For the PCS and the MCS, 5 was used instead of 10 as the cut-off for the computation of percents. The first row therefore represents the percentage of those who improved by 5 or more points, the middle row represents those who had a change of less than 5 points in either direction, and the last row represents those who declined by 5 points or more.

Percentages within each cell do not always equal 100 due to rounding.

TABLE IV

Estimated Mean Change in SF-36 Domains for Women from Baseline (1996/1997) to Year Three (1999/2000) (Bias Adjusted via Multiple Imputation)

Women Age Group at Baseline	SF-36 Domain									
	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS
40 to 44 (n=127-135)*										
Mean Difference	-2.10	-1.53	-3.46	-2.31	-2.32	-1.55	-4.01	-3.28	-0.47	-1.09
SD of Mean Difference	1.17	3.42	2.02	1.23	1.52	2.12	3.53	1.42	0.63	1.00
95% CI for Mean Difference	-2.8; -1.5	-2.8; -0.3	-4.5; -2.5	-3.1; -1.6	-3.2; -1.5	-2.6; -0.6	-5.3; -2.9	-4.0; -2.6	-0.9; -0.1	-1.6; -0.6
45 to 49 (n=322-486)										
Mean Difference	-0.92	-0.93	0.04	-0.76	1.33	1.44	2.44	2.06	0.38	1.35
SD of Mean Difference	0.62	2.48	1.12	0.75	0.82	1.03	1.60	0.70	0.39	0.45
95% CI for Mean Difference	-1.2; -0.7	-1.5; -0.3	-0.4; 0.4	-1.1; -0.5	1.0; 1.7	1.1; 1.8	2.0; 2.9	1.8; 2.3	-0.6; -0.2	1.2; 1.6
50 to 54 (n=598-622)										
Mean Difference	-2.46	-2.83	-1.85	-0.98	-0.09	-0.69	-1.84	0.09	-0.85	0.43
SD of Mean Difference	0.58	1.55	0.89	0.61	0.65	0.90	1.39	0.58	0.32	0.37
95% CI for Mean Difference	-2.8; -2.2	-3.4; -2.3	-2.3; -1.4	-1.3; -0.7	-0.5; 0.3	-1.1; -0.3	-2.4; -1.3	-0.2; 0.4	-1.1; -0.6	0.2; 0.7
55 to 59 (n=612-727)										
Mean Difference	-1.93	-1.97	-0.73	0.37	-0.14	0.26	1.91	0.79	-0.72	1.07
SD of Mean Difference	0.63	1.43	0.81	0.57	0.63	0.81	1.31	0.62	0.31	0.37
95% CI for Mean Difference	-2.3; -1.6	-2.5; -1.4	-1.1; -0.3	0.1; 0.7	-0.5; 0.2	-0.1; 0.6	1.4; 2.4	0.5; 1.1	-0.9; -0.5	0.9; 1.3

* These values represent the range of the number of respondents across the domains within each age group, which vary due to the differing amounts of missing data available for the imputation models.

tion of scores for each missing year 3 item. This distribution accounts for the fact that the missing data are not observed, and so are "known" with uncertainty. When an imputed value was less than zero or exceeded 100, it was replaced by a score of zero or 100 respectively. In order to assess change over time, baseline scores were subtracted from the year 3 scores, so that positive values indicated improvement in HRQOL.

RESULTS

Loss to follow-up was 9.6% for women (190/1,974) and 14.4% for men (140/975). Table I outlines the loss to follow-up by age and gender. The majority (53.6%) of the refusals (109 women and

68 men) were participants who chose to respond to a short questionnaire rather than the full interview. Other reasons included death, not interested, too sick, moved away, unable to contact, and no time.

Mean HRQOL changes over three years tended to be small, but the standard deviations of the observations at baseline and at follow-up were generally high, suggesting that large HRQOL improvements in some participants were cancelled out by large HRQOL declines in others. To better understand the distribution of the changes, Table II (women) and Table III (men) contain the percent of participants who improved or declined by 10 or more points, and those who stayed within 10 points of their baseline score. A 10-point

change was chosen, although a 5-point change is considered to be clinically relevant,¹ to get a better understanding of the distribution of fairly large changes. A cut-off of 5 points was used for the PCS and the MCS due to the lower variability within the summary scores.

The majority of respondents stayed within 10 points of their baseline score for most domains within most age groups. However, women in all age groups reported substantial variability in BP and VT scores, with an equivalent number of respondents demonstrating increases and decreases of greater than 10 points. The three younger age groups also showed large variation in SF, with similar numbers showing increases and decreases. For men, the only domain that showed variability

TABLE V

Estimated Mean Change in SF-36 Domains for Men from Baseline (1996/1997) to Year Three (1999/2000) (Bias Adjusted via Multiple Imputation)

Men Age Group at Baseline	SF-36 Domain									
	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS
40 to 44 (n=78-86)*										
Mean Difference	-0.18	6.48	0.85	1.36	0.05	1.88	6.67	0.44	0.98	0.71
SD of Mean Difference	0.86	3.74	2.11	1.33	1.66	2.21	4.01	1.61	0.74	1.01
95% CI for Mean Difference	-0.9; 0.5	5.2; 7.6	-0.6; 2.2	0.3; 2.3	-1.3; 1.2	0.5; 3.1	4.1; 9.1	-1.7; 0.8	0.5; 1.5	-0.1; 1.5
45 to 49 (n=273-284)										
Mean Difference	-1.64	-3.44	-1.76	0.26	0.84	0.41	3.22	0.40	-0.74	1.40
SD of Mean Difference	0.88	1.70	1.30	0.78	0.91	1.20	1.77	0.86	0.42	0.52
95% CI for Mean Difference	-2.1; -1.2	-4.3; -2.7	-2.4; -1.1	-0.2; 0.7	0.3; 1.4	-0.1; 0.9	2.5; 3.9	-0.1; 0.9	-1.0; -0.5	1.1; 1.7
50 to 54 (n=207-302)										
Mean Difference	-2.96	1.25	-2.51	-2.13	-0.69	3.45	1.15	1.47	-0.86	1.34
SD of Mean Difference	1.01	1.99	1.56	0.85	1.00	1.14	1.93	0.84	0.54	0.55
95% CI for Mean Difference	-3.7; -2.3	0.3; 2.2	-3.5; -1.6	-2.8; -1.5	-1.4; 0.0	2.9; 4.0	0.4; 1.9	0.9; 2.0	-1.2; -0.5	1.0; 1.7
55 to 59 (n=149-300)										
Mean Difference	-1.00	-0.85	1.69	1.60	-0.17	0.96	-0.94	0.36	0.49	0.28
SD of Mean Difference	1.10	2.03	1.26	0.81	0.85	1.14	1.72	0.80	0.44	0.47
95% CI for Mean Difference	-1.6; -0.5	-1.6; -0.1	1.1; 2.3	1.2; 2.0	-0.7; 0.3	0.5; 1.4	-1.6; -0.3	-0.1; 0.8	0.2; 0.8	0.0; 0.6

* These values represent the range of the number of respondents across the domains within each age group, which vary due to the differing amounts of missing data available for the imputation models.

within all age groups was BP, with an equivalent number showing increases and decreases of greater than 10 points. In addition, two age groups (45 to 49 and 50 to 54 years) had substantial variability in VT scores, with similar numbers of increases and decreases.

Those with follow-up data had higher mean baseline scores than those without follow-up data. For women, this was the case in 28 of the 32 domain/age groups and 7 of the 8 summary component/age groups. For men, this was the case in 29 of the 32 domain/age groups and 6 of the 8 summary component/age groups. Table IV (women) and Table V (men) therefore provide multiple imputation adjusted mean differences in scores from baseline to three years, along with the standard deviation of the mean difference and the 95% credible intervals (Bayesian analogue to frequentist confidence intervals) of the mean differences.

Women reported HRQOL declines in more domains (22 declined, 10 improved) than men (14 declined, 18 improved). This is also reflected in the summary scores, with women showing small declines in the PCS within every age group and in the MCS within one age group. For men, two of four age groups showed a decline in the PCS and no age group declined in the MCS. Declines were more common in physical than mental domains.

For women, declines were most notable in the youngest age group, with the largest drops in RE, BP and MH. The largest improvements were in women aged 45 to 49 years, in RE and MH. In men, the

largest decline was in the RP domain for the men aged 45 to 49 years, followed by the PF domain for those aged 50 to 54 years. The largest improvements were in the RP and RE domains in those aged 40 to 49 years.

It is important to note that while mean differences are small, with only two domains exceeding five points, many individuals experienced much larger, and therefore clinically relevant changes.¹ For women, the credible intervals for 20 of the domain scores and 5 of the summary scores were both negative, suggesting real declines. However, 7 domain scores and 3 summary scores had credible intervals that were positive, suggesting real improvement. Credible intervals for 5 domain scores included a zero value. For men, the credible intervals for 9 of the domain scores and 2 of the summary scores were both negative while 13 domain scores and 5 summary scores had credible intervals that were both positive. The credible intervals for 10 domain scores and 1 summary score included a zero value. While these changes are relatively small and therefore of little clinical or social relevance,¹ they do suggest that the changes accumulated over time may become relevant over a longer follow-up period.

DISCUSSION

It appears that the HRQOL of Canadians aged 40 to 59 years remains relatively stable, as evidenced by the small mean changes in most SF-36 domains for all age and gender combinations. However, there

are a number of individuals who worsen and a number who improve over time, leading to high individual variability in the observations. Since the number who improve and the number who worsen are roughly equal, the impact on mean changes in HRQOL are effectively canceled out.

The finding of reasonable stability over three years is consistent with the results of the Whitehall study,⁶ which noted overall declines ranging from 0.3 to 3.0 (men) and 0.3 to 2.4 (women) in those who reported no physical disease or psychiatric comorbidity. They are also consistent with the unpublished findings of the Medical Outcomes Trust, who followed their original normative sample four years later and noted that the domain scores were fairly stable, with the exception of the PF domain, which showed a three to four point decline (Personal Communication, QualityMetric Incorporated, November 13, 2002).

Within the limited age range observed, there is evidence to suggest that the HRQOL of women, particularly the younger ones, may be more likely to deteriorate than that of men. Women showed more declines than improvements (22 versus 10) while men showed more improvements than declines (18 versus 14). Men also showed greater stability in terms of the number of cells where 50% or more of the respondents stayed within 10 points of their baseline score (26 out of 32 age/domain stratifications versus 21 out of 32 for women). Both men and women showed substantial variability in the BP

domain, with similar percentages showing declines and improvements in BP within all age stratifications.

These findings may have important implications for interpretation of minimal clinically important differences or responsiveness in quality of life measures.¹¹⁻¹³ The usefulness of instruments designed to measure changes over time is dependent on reliability and validity, which have been well documented for the SF-36,¹⁻³ and responsiveness.¹¹ The collection of longitudinal SF-36 data provides an additional reference for establishing meaningful changes over time.¹² In particular, researchers using the SF-36 for domains such as bodily pain and vitality, which demonstrated the largest variation, may need to adjust for the natural evolution of domain scores over time when interpreting minimal clinically important differences from longitudinal studies. It should be noted, however, that over three years only two cells, RE and RP for men aged 40 to 44 years, demonstrated the 5-point change considered by the developers to be clinically and socially relevant.¹ It is also important to note that these results are based on a random sample of the population, which may be different from patients selected for an intervention or inclusion in a clinical trial.

Not all subjects who provided baseline data also provided data at three years. Multiple imputation was used to adjust for bias due to missing data at three years, but like all statistical techniques, it has assumptions and limitations. We assumed that participants' baseline data were sufficiently detailed to predict average three-year changes, and that patterns of changes in SF-36 domain scores were similar in those who responded and those who did not, after adjusting for any baseline differences in responders versus non-responders at three years. Therefore, one limitation of the study is that we did not take into account the possibility that non-responders at three years may be quite different from responders, even after adjusting for their baseline characteristics.

It appears that, overall, the mean HRQOL of working-age Canadian resi-

dents remains stable over three years, although men were more likely to experience small improvements while women were more likely to experience small declines, particularly in the younger age groups. This suggests that changes in HRQOL following an intervention are probably not confounded with declines due to population aging as long as the follow-up period is of a relatively short duration (e.g., three years or less). However, if one extrapolates the three-year changes to a longer period of follow-up, HRQOL may change substantially. Future research should assess longitudinal changes in HRQOL over a longer period of time and across a larger age range.

REFERENCES

- Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey Manual and Interpretation Guide. The Health Institute, New England Medical Center, Boston, 1993.
- Ware JE Jr, Kosinski M, Keller SD. SF-36 Physical and Mental Summary Scales: A User's Manual. The Health Institute, New England Medical Center, Boston, 1994.
- Ware JE Jr. SF-36 health survey update. *Spine* 2000;25:3130-39.
- Kaegi L. Medical Outcomes Trust Conference presents dramatic advances in patient-based outcomes assessment and potential applications in accreditation. *Joint Commission Journal of Quality Improvement* 1999;25:207-18.
- Garraff A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: Bibliographic study of patient assessed health outcomes measures. *Br Med J* 2002;324:1417-21.
- Hemingway H, Stafford M, Stansfield S, Shipley M, Marmot M. Is the SF-36 a valid measure of change in population health? Results from the Whitehall II study. *Br Med J* 1997;315:1273-79.
- Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, et al. and the CaMos Research Group. Canadian Normative Data for the SF-36 Health Survey. *CMAJ* 2000;163:265-71.
- Chern JY, Wan TT, Pyles M. The stability of health status measurement (SF-36) in a working population. *J Outcome Measures* 2000;4:461-81.
- Hays RD, Marshall GN, Wang EY, Sherbourne CD. Four-year cross-lagged associations between physical and mental health in the Medical Outcomes Study. *J Consulting Clinical Psychology* 1994;62:441-49.
- Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, et al. and the CaMos Research Group. Is there regional variation in the SF-36 scores of Canadian adults? *Can J Public Health* 2002;93:233-36.
- Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chron Dis* 1987;40:171-78.
- Testa MA. Interpretation of quality-of-life outcomes: Issues that affect magnitude and meaning. *Med Care* 2000;38(Supplement II):166-74.
- Liang MH. Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments. *Med Care* 2000;38(Supplement II):84-90.
- Kreiger N, Tenenhouse A, Joseph L, MacKenzie T, Poliquin S, Brown J, et al. Research Notes: The Canadian Multicentre Osteoporosis Study (CaMos): Background, Rationale, Methods. *Can J Aging* 1999;18:376-87.
- Rubin D. *Multiple Imputation for Non-response in Surveys*. New York: Wiley, 1987.
- Kass RE, Raftery AE. Bayes factors. *J Am Statistical Association* 1995;90:773-95.

Received: August 12, 2003

Accepted: April 16, 2004

RÉSUMÉ

Contexte : Le SF-36 est un court questionnaire très utilisé pour évaluer la qualité de vie liée à la santé (HRQOL). Toutefois, vu le petit nombre d'études longitudinales menées auprès de populations en bonne santé, il est difficile d'évaluer l'histoire naturelle du SF-36, ce qui est important parce que tout changement noté résultant d'une intervention peut être confondu avec des changements normaux de la qualité de vie liée à la santé. Cet article évalue les changements moyens dans les scores des SF-36 sur une période de trois ans chez les hommes et les femmes qui étaient âgés de 40 à 59 ans lors de leur visite initiale.

Méthode : Les sujets ont été choisis au hasard dans neuf villes canadiennes. Les changements moyens dans les résultats du SF-36 sur une période de trois ans (1996-1997 à 1999-2000) ont été calculés par sexe et par groupe d'âge de cinq ans. L'imputation multiple a été utilisée pour corriger tout biais résultant d'informations manquantes.

Résultats : La cohorte initiale comportait 1 974 femmes et 975 hommes âgés de 40 à 59 ans. Les changements moyens étaient plutôt légers. Chez les femmes, 22 des 32 groupes d'âge et sous-échelles (quatre groupes d'âge, huit sous-échelles SF-36) présentaient de légères diminutions moyennes, alors que chez les hommes, ce nombre était de 14 groupes sur 32. Le score de la plupart des participants est demeuré en deçà de 10 points du score initial lors de l'entrée dans l'étude.

Interprétation : Les scores moyens du SF-36 ne changent que légèrement sur une période de trois ans dans la population canadienne d'âge moyen. Toutefois, il pourrait s'avérer nécessaire de tenir compte de l'évolution naturelle des scores du SF-36 lors de l'interprétation des résultats d'études longitudinales.