Quality of Life Research (2005) 14: 1251–1261
DOI 10.1007/s11136-004-6673-8

Comparing SF-36 scores across three groups of women with different health profiles

Kathleen J. Yost[1], Mary N. Haan[2], Richard A. Levine[3], & Ellen B. Gold[4]
[1]*Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare Research Institute (E-mail: kyost@enh.org);* [2]*Department of Epidemiology, University of Michigan;* [3]*Department of Mathematics and Statistics, San Diego State University;* [4]*Department of Epidemiology and Preventive Medicine, University of California at Davis*

Accepted in revised form 15 November 2004

Abstract

Background: The widespread use of the Medical Outcomes Study (MOS) 36-item Short-Form Health Survey (SF-36) facilitates the comparison of health-related quality of life (HRQL) across independent studies. *Objectives*: To compare the scores of eight scales and two summary scales of the SF-36 across participants in the Women's Healthy Eating and Living (WHEL) trial, the Women's Health Initiative-Dietary Modification trial (WHI-DM), and the MOS, and to illustrate the use of effect sizes for interpreting the importance of group differences. *Methods*: WHEL and WHI-DM are both multi-center dietary interventions; only data from the UC Davis sites were used in our study. WHEL participants had a recent history of breast cancer, WHI-DM participants were healthy, postmenopausal women, and women in the MOS had a history of hypertension, diabetes, heart disease, or depression. General linear models were used to identify statistically significant differences in scale scores. Meaningful differences were determined by effect sizes computed using a common within-group standard deviation (SD) and SDs from normative data. *Results*: After adjusting for age and marital status, SF-36 scores for the WHI-DM and WHEL samples were similar and both had statistically significantly higher scores than the MOS sample. Relative to the WHEL or WHI-DM studies, MOS scores for scales related to the physical domain were clearly meaningfully lower whereas scale scores related to the mental health domain were potentially meaningfully lower. *Conclusions*: The HRQL of breast cancer survivors is comparable to that of healthy women and better than that of women with chronic health conditions, particularly with respect to physical health. This study illustrated the use of ranges of effects sizes for aiding the interpretation of SF-36 scores differences across independent studies.

Key words: Breast cancer, Effect sizes, Meaningful difference, Quality of life, SF-36

Abbreviations: ANCOVA – analysis of covariance; HRQL – health-related quality of life; MANCOVA – multivariate analysis of covariance; MCS – mental component summary scale; MOS – Medical Outcomes Study; PCS – physical component summary scale; SF-36 – 36-item short-form health survey; SD – standard deviation; WHEL – Women's Healthy Eating and Living trial; WHI-DM – Women's Health Initiative-Dietary Modification trial

Introduction

The Medical Outcomes Study (MOS) was aimed at evaluating the effects of cost containment on patient outcomes [1]. The most renowned outcome measurement instrument to emerge from the MOS was the 36-item short form (SF-36) health survey, which was developed to assess the physical and

mental health components of general health status in individuals with chronic conditions (i.e., hypertension, diabetes, heart disease, and depression) [2, 3]. The SF-36 measures eight health concepts: physical functioning, role limitations due to physical problems (role-physical), bodily pain, general health perceptions, vitality, social functioning, role limitations due to emotional problems (role-emotional), and mental health [3]. Scores for these eight scales can be combined into two summary scales: the physical component summary scale (PCS) and the mental component summary scale (MCS) [4]. The physical functioning, role-physical and bodily pain scales have a strong association with the physical health component, and the mental health, role-emotional and social functioning scales have a strong association with the mental health component. Vitality and general health are moderately associated with both the physical and mental health components, and social functioning is also moderately associated with physical health [3].

Despite being designed to assess general health status, the SF-36 has often been interpreted as a measure of health-related quality of life (HRQL) [3]. Because of its popularity and acceptance as a measure of HRQL, the SF-36 has been administered to a variety of individuals, including samples from the general population [3], disadvantaged elderly [5], and people with particular health conditions, including cancer [6–8].

The widespread use of the SF-36 facilitates comparisons of results across different groups of individuals. Ganz et al. [9] found that breast cancer survivors who were assessed 2 and 3 years post-diagnosis scored higher (i.e., better HRQL) on all SF-36 scales when compared to participants in the MOS longitudinal study. However, the statistical and clinical significance of these differences were not discussed. Anderson et al. [6] compared SF-36 scores of AIDS patients to those of patients with cancer or other serious illnesses. Patients with cancer or other illnesses consistently scored significantly higher than AIDS patients on social functioning, role-emotional, and vitality. Baseline bodily pain scores were significantly higher (less pain) for cancer patients compared to AIDS patients ($p = 0.024$), but the difference was only 3.0 points on a scale ranging from 0 to 100 [6], which is unlikely to be clinically or socially meaningful [3]. In a secondary analysis of patients with 13 different health conditions, Sprangers et al. [8] found that cancer patients ranked sixth overall in their SF-36 scores. That is, patients from five other illness groups, including those with cardiovascular conditions, scored higher than cancer patients, whereas seven patient groups, including those with chronic respiratory diseases, scored lower [8]. Sprangers et al. used a difference in group means of 2 points on a scale of 0 to 100 to identify meaningful group differences, although they acknowledge that this criterion was not based on empirical evidence.

These examples illustrate several points. First, the widespread use of the SF-36 facilitates comparisons of HRQL across patient groups. Second, the HRQL of cancer patients and survivors, as measured by the SF-36, is fairly good relative to patients with other chronic illnesses. Third, comparisons of SF-36 scores across groups are limited in that the magnitude of group differences are not interpreted at all, or interpretation relied on statistical significance and unsubstantiated criteria.

We were interested in capitalizing on the widespread use of the SF-36 to compare existing HRQL data for three distinct groups of women with different health profiles: (1) women with a recent history of breast cancer, (2) post-menopausal women without a history of breast or colorectal cancer or heart disease, and (3) women with a history of hypertension, heart disease, diabetes, or depression. The objectives of our study were to compare baseline scores for the eight SF-36 scales and two summary scales for three samples of women with different health profiles, and to illustrate the use of effect sizes for interpreting the importance of group differences.

Methods

Study participants

We performed a secondary analysis of baseline SF-36 data for three groups of women: participants at the University of California at Davis (UC Davis) site of the Women's Healthy Eating and Living trial (WHEL), participants at the UC Davis site of the Women's Health Initiative-Dietary Modification trial (WHI-DM), and female participants in

the MOS. WHEL is a multi-center, randomized dietary intervention trial of women diagnosed and treated for breast cancer [Stage I ($\geqslant$ 1 cm), II, or IIIA] within four years of enrollment to the study and between the ages of 18 and 70 years at diagnosis [10]. WHEL was designed to determine the effects of a low fat, high fiber diet on breast cancer recurrence and disease-free survival [11]. The intervention group of the WHEL study reduced their daily calories from fat to 15–20% and consumed daily 5 servings of vegetables, 16 ounces of vegetable juice, 3 servings of fruit, and 30 grams of fiber. The control arm followed the National Cancer Institute's "5-A-Day" program, which recommends eating five servings of fruits and vegetables each day [11].

The WHI is another multi-center study of women that has several clinical trials and an observational study [12]. In the dietary modification arm of the clinical trial (WHI-DM), post-menopausal women were randomized to a dietary intervention to determine the effects of a low fat diet on the incidence of breast cancer, colorectal cancer, and heart disease [12, 13]; thus, women were ineligible if they had a history of these diseases. WHI-DM participants were between 50 and 79 years of age at baseline. The intervention group reduced total fat intake to 20% and saturated fat to less than 7% of daily calories, and the control group had no dietary modification [12, 13]. Only the UC Davis site of the WHI-DM was included in our study.

The MOS had both cross-sectional and longitudinal components, and it included men and women aged 18–98 years living in Boston, Chicago, or Los Angeles who had prevalent and treatable chronic conditions particularly, hypertension, heart disease, diabetes, and/or depression [14]. Persons with a history of any cancer (except skin cancer) within the past 3 years were excluded from the longitudinal component [15]. Methods for sampling patients for the longitudinal component of the MOS are described in detail elsewhere [15]. The SF-36 was administered to 3588 participants in the longitudinal component of the MOS [14], which measured change in chronic conditions over time and evaluated outcomes with respect to systems of care, provider specialty, and styles of practice.

At the time this secondary data analysis was initiated, baseline SF-36 data were available for 427 women enrolled at the UC Davis WHEL site. Over 500 women were eventually enrolled at that site. Baseline SF-36 data were obtained for 791 women randomized to the dietary modification trial at the UC Davis WHI-DM site. Recruitment for WHI-DM began in 1993 and was completed at the time of our study. Information on study design and participant recruitment are presented elsewhere for the WHI-DM [12, 13] and WHEL [11]. Data from the MOS longitudinal study (Radius data set #30–34) were purchased from Sociometric Corporation and included baseline SF-36 data. Only the 2,180 female MOS participants in the longitudinal component of the MOS were included in our study.

Participants who were missing data for one or more SF-36 scales or for the demographic variables age, education, marital status and race were excluded as these individuals would not contribute information to the analyses. For WHEL participants, 420 (98.4%) women had complete data. Complete data were available for 764 (96.6%) WHI-DM participants, and for 1741 (79.9%) of the female MOS participants.

During preliminary analyses, we conducted tests for homogeneity of variances (Levene test and Brown & Forsythe test, SAS PROC GLM, HOVTEST option [16]). The tests were statistically significant for all scales (all $p < 0.001$ except for vitality where $p < 0.01$ and bodily pain where $p < 0.05$) indicating variances were not equal across the three studies. When the sample sizes of each group are equal, statistical tests are robust to violations in the assumption of homogeneity of variance across groups for an analysis of variance (ANOVA) [16, 17] and to violations in the assumption of homogeneity of variance–covariance matrices for a multivariate analysis of variance (MANOVA) [18]. To safeguard against the heteroscedasticity in our data, all samples were limited to the size of the smallest group, which was the WHEL sample. Samples of similar size ($n = 420$) were selected from the WHI-DM and MOS studies using simple random sampling without replacement. We evaluated the representativeness of the random samples by comparing SF-36 scores (median, 1st and 3rd quartile, mean, standard deviation) and demographic

characteristics of the WHI-DM and MOS samples (n = 420 each) to scores and characteristics of their respective sampling frames (i.e., n = 764 WHI-DM or n = 1741 MOS participants). Because the samples and sampling frames were not independent, formal statistical tests were not conducted to assess representativeness.

Scale scores

The SF-36 scoring algorithms were used to compute scores for the eight scales [3] and the two summary measures [4]. Although the WHI-DM [19] and WHEL studies use the RAND-36 scoring algorithms, we used the SF-36 algorithms to facilitate comparisons with population normative data. A scale was considered missing if greater than 50% of the items comprising the scale were missing [3]. If 50% or less of the items were missing, the scale was calculated as the average of the observed items. This approach has been shown to be an acceptable method of imputing missing data for other HRQL instruments when at least 50% of items comprising a scale are observed [20]. Individuals with missing scale scores for any of the eight scales were excluded. Scale scores range from 0 to 100, with higher scores representing better HRQL.

The bodily pain scale consists of two items. One of these items, measuring severity of pain during the past 4 weeks (BP1), has six response options ranging from 1 = 'none' to 6 = 'very severe'. However, this item was incorrectly typed on the questionnaire administered to WHI-DM and WHEL participants in that only five response options were provided (the 'very severe' option was inadvertently omitted). For these two studies, BP1 was scored the same way as the second bodily pain item (BP2), which addresses pain interfering with normal work and has a 5-category response scale. Specifically, BP1 was scored as follows: 'none' = 6, 'very mild' = 4.75, 'mild' = 3.5, 'moderate' = 2.25 and 'severe' = 1. Other than this minor deviation, the scoring algorithms were followed as prescribed by the developers of the SF-36 to convert the two items into the bodily pain scale with possible scores ranging from 0 to 100 [3, 4]. PCS and MCS scores were calculated as weighted linear combinations of the eight standardized SF-36 scales, where the weights were based on factor loadings from a principal component analysis [4]. In the general population, PCS and MCS have mean scores of 50 and SDs of 10, with higher scores representing better physical and mental health [4].

Data analysis

Multivariate analysis of covariance (MANCOVA, SAS PROC GLM [21]) was used to compare the eight SF-36 scale scores across the WHI-DM, WHEL, and MOS studies. The SF-36 scores were the dependent variables and study (WHI-DM, WHEL, MOS) was the independent variable. Because the PCS and MCS are linear combinations of the eight SF-36 scales, they were assessed in a separate MANCOVA. Numerous studies have shown that HRQL varies by demographic characteristics [22–26]; therefore, the covariates age, education ($\leqslant$high school vs. > high school), marital status (married or living as married vs. other), and race (white vs. non-white) were assessed for statistical significance, which was defined as Wilks' Lambda $p < 0.05$. If the null hypothesis (i.e., the vectors of average scale scores for all three groups were the same) was rejected, separate ANCOVAs were used to identify the individual scales on which the groups differed significantly. *Post hoc* analyses were performed to determine which groups contributed to the significant differences in the ANCOVA and to estimate least squares adjusted mean score differences among the three studies. Significance levels for the *post hoc* analyses were adjusted using Tukey's Honestly Significant Difference, with the experiment-wise error held at $\alpha = 0.05$. Least squares adjusted mean score differences were computed as the mean difference in SF-36 scores between studies evaluated at the mean of each of the covariates in the model.

Identifying meaningful differences

In the present study, we used effect sizes to assess the meaningfulness of group differences in SF-36 scale scores. Effect sizes have been widely used for the purpose of interpreting HRQL score differences [27–31]. The effect size of a mean difference was calculated by dividing the difference in group means by a reference SD; effect size = $(\overline{X_1} - \overline{X_2}/SD$ [32, 33]. We employed two sources

for the SD. The first was the root mean square error (RMSE) for each scale or summary scale from the ANCOVA, which is an estimate of the common within-group SD. The second source was an external reference population; the general U.S. population SD for women [3]. Cohen provides guidelines for interpreting effect sizes where 0.2 is 'small', 0.5 is 'medium', and 0.8 is 'large' [33]. For our purposes, we defined group differences as very small and clearly not meaningful if they were associated with effect sizes less than 0.2. Group differences with effect sizes greater than 0.5 were considered clearly meaningful [31], and differences with effect sizes between 0.2 and 0.5 were considered potentially meaningful.

Results

Study populations

Table 1 provides information on the characteristics of the WHEL, WHI-DM, and MOS samples. Random samples of 420 WHI-DM and 420 MOS participants were representative of their sampling frames with the following exceptions. The WHI-DM sample had slightly fewer married (66.2% vs. 68.2%) and more highly educated (post-high school 85.5% vs. 83.3%) women than the sampling frame of 764 women. The MOS sample had slightly lower mean role-emotional scores (62.8 vs. 65.7) and slightly higher mean general health scores (62.0 vs. 60.1) than the sampling frame of 1741 women. All other mean scores for the MOS sample differed by approximately 1 point or less from those of the sampling frame.

The three samples differed significantly with respect to age, race, marital status and education (all $p < 0.001$). The MOS sample had the youngest mean age and broadest age range. The samples were predominantly white, but the MOS study had a higher proportion of non-white participants. WHEL and WHI-DM participants were more likely to be married and have a post-high school education than MOS participants.

Figure 1 compares average crude (unadjusted) SF-36 scale scores for the three samples and normative data for females from the general U.S. population (n = 1412) [3]. The eight scales are ordered on this figure from left to right based on the degree to which they measure physical or mental health [3]. WHEL, WHI-DM and norm scores were similar and higher than those for the MOS. There was no clear pattern to suggest that group differences between scales on the left side of the figure that are more closely related to physical health were consistently larger or smaller than differences between scales on the right side of the figure that are more closely related to mental health.

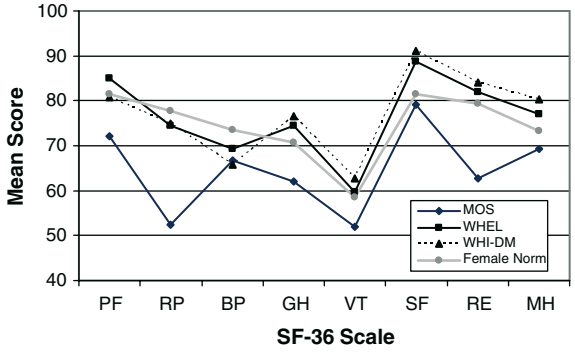

Figure 1. Comparison of SF-36 scores for WHEL, WHI-DM, MOS studies and General U.S. Population Norms for Females.WHEL: Women's Healthy Eating and Living Study WHI-DM: Women's Health Initiative-Dietary Modification trial MOS: Medical Outcomes Study PF: physical functioning, RP: role-physical, BP: bodily pain, GH: general health, VT: vitality, SF: social functioning, RE: role-emotional, MH: mental health.

General linear models

The MANCOVA revealed overall differences in the eight SF-36 scales ($p < 0.001$) and two summary scales ($p < 0.001$) across the WHEL, WHI-DM, and MOS samples. The covariates age and marital status were statistically significant in the MANCOVA for SF-36 scales and summary scales, but race and education were not significant in either analysis. ANCOVAs were conducted for each scale and summary scale and included the covariates age and marital status. Results of the ANCOVA indicated that the three studies differed significantly in their scores for scales and summary scales; all $p < 0.001$ except for bodily pain ($p = 0.32$).

Adjusted mean differences for the eight scales, MCS, and PCS were estimated and are presented in Table 2. Because the mean differences were adjusted for age, we decided to use the age-specific population norm SDs rather than the SD for all

Table 1. Characteristics of the three study samples

Variable	WHEL $N = 420$		WHI-DM $N = 420$		MOS $N = 420$	
Age [mean (SD)]	53.9	(8.8)	62.8	(6.8)	52.5	(16.0)
Age [n (%)]						
< 45	63	(15.0)	0	(0)	143	(34.1)
45–54	187	(44.5)	50	(11.9)	78	(18.6)
55–64	111	(26.4)	194	(46.2)	80	(19.1)
65–74	59	(14.1)	157	(37.4)	93	(22.1)
75+	0	(0)	19	(4.5)	26	(6.2)
Race/ethnicity [n (%)]						
White	372	(88.6)	372	(88.6)	323	(76.9)
Black	5	(1.2)	12	(2.9)	71	(16.9)
Hispanic	14	(3.3)	13	(3.1)	11	(2.6)
Other	29	(6.9)	23	(5.5)	15	(3.6)
Marital status [n (%)]						
Married/living as married	311	(74.1)	278	(66.2)	214	(51.0)
Single/never married	35	(8.3)	7	(1.7)	61	(14.5)
Divorced/separated	53	(12.6)	65	(15.5)	75	(17.9)
Widowed	21	(5.0)	70	(16.7)	70	(16.7)
Education [n (%)]						
Less than high school	4	(1.0)	8	(1.9)	51	(12. 1)
High school/GED	57	(13.6)	54	(12.9)	137	(32.6)
Post HS training/some college	181	(43.1)	221	(52.6)	135	(32.1)
College degree or higher	178	(42.4)	137	(32.6)	97	(23.1)

MOS: Medical Outcomes Study
WHEL: Women's Healthy Eating and Living Study
WHI-DM: Women's Health Initiative-Dietary Modification trial
Some categories of WHEL and MOS data were collapsed (e.g., divorced and separated) to be comparable to WHI-DM data categories.

women combined to compute standardized effect sizes, which allowed us to derive a range of plausible standardized effects sizes to aid the interpretation of group differenced. RMSEs were similar to age-specific population norm SDs, although they tended to be on the lower end of the ranges. Effect sizes computed using the RMSE were also comparable to those computed using the population norm SDs. Scale score difference between the WHI-DM sample compared to the WHEL sample were very similar and statistically significant only for the role-physical and general health scales; however, these differences were no longer significant after adjusting for multiple comparisons. Effect sizes for the WHI-DM vs. WHEL differences were all less than 0.20 and clearly not meaningful. Average scores for the MOS samples were significantly lower than scores for the WHI-DM and WHEL samples for all scales except bodily pain. The largest differences in adjusted scale scores were for role-physical and role-emotional, and scores were the most similar for bodily pain and mental health (Table 2). Differences between the WHI-DM and MOS studies and between the WHEL and MOS studies were statistically significant and clearly meaningful (effect sizes > 0.5) for physical functioning, role-physical, general health, role-emotional and PCS, whereas the statistically significant score differences for vitality, social functioning, mental health and MCS were potentially meaningful.

Discussion

WHI-DM participants had comparable HRQL to WHEL participants and both samples had better HRQL than MOS participants. Adjusting for age and marital status tended to decrease the magnitude of the differences between groups. WHEL participants had a recent history of Stage I, II, or IIIA breast cancer, the treatment of which involves one or more of the following: surgery, radiation therapy, chemotherapy, and hormone therapy

Table 2. Mean differences for SF-36 Scales Summary Scales, adjusted for age and marital status

			Group Comparison											
			WHEL - MOS				WHI-DM - MOS				WHI-DM - WHEL			
			Effect Size				Effect Size				Effect Size			
SF-36 Scale	RMSE	Range of Norm SDs[a]	Adjusted Mean Difference	*p*-value	RMSE[b]	Stand[c]	Adjusted Mean Difference	*p*-value	RMSE[b]	Stand[c]	Adjusted Mean Difference	*p*-value	RMSE[b]	Stand[c]
Physical Functioning	20.9	17.7–29.0	12.8	< 0.001 *	0.61	0.44–0.72	14.5	< 0.001 *	0.70	0.50–0.82	1.7	0.26	0.08	0.06–0.10
Role-Physical	36.0	28.0–42.5	21.5	< 0.001 *	0.60	0.51–0.77	26.7	< 0.001 *	0.74	0.63–0.95	5.3	0.05	0.15	0.12–0.19
Bodily Pain	23.8	20.9–27.1	2.5	0.14	0.11	0.09–0.12	1.0	0.59	0.04	0.04–0.05	−1.5	0.37	−0.06	−0.06–−0.07
General Health	18.8	17.2–23.4	12.2	< 0.001 *	0.65	0.52–0.71	14.9	< 0.001 *	0.79	0.64–0.87	2.7	0.05	0.14	0.12–0.16
Vitality	20.8	19.7–23.5	6.8	< 0.001 *	0.33	0.29–0.34	9.2	< 0.001 *	0.44	0.39–0.47	2.5	0.10	0.12	0.11–0.13
Social Functioning	20.3	20.8–27.7	8.5	< 0.001 *	0.42	0.31–0.41	9.8	< 0.001 *	0.48	0.35–0.47	1.3	0.39	0.06	0.05–0.06
Role-Emotional	33.8	31.3–39.7	17.5	< 0.001 *	0.52	0.44–0.56	17.4	< 0.001 *	0.51	0.44–0.55	−0.1	0.95	0.00	0.00–0.00
Mental Health	16.4	16.8–19.9	6.5	< 0.001 *	0.40	0.33–0.39	6.6	< 0.001 *	0.40	0.33–0.39	0.1	0.92	0.01	0.01–0.01
SF-36 Summary Scale														
Physical Component	9.8	7.7–11.6	4.7	< 0.001 *	0.48	0.41–0.61	5.7	< 0.001 *	0.58	0.49–0.74	1.0	0.15	0.10	0.09–0.13
Mental Component	9.6	9.5–10.5	3.6	< 0.001 *	0.38	0.34–0.38	3.7	< 0.001 *	0.39	0.35–0.39	0.1	0.86	0.01	0.01–0.01

RMSE: root mean square error from the ANCOVA, which is an estimate of the common within-group SD for WHEL, WHI-DM and MOS. WHEL: Women's Healthy Eating and Living Study WHI-DM: Women's Health Initiative-Dietary Modification trial MOS: Medical Outcomes Study.

[a]Range of SDs from age-specific normative data for females [3,4]

[b]Effect size calculated using RMSE.

[c]Range of standardized effect sizes calculated using age-specific population norm SDs for females.

*Significantly different ($p < 0.05$) after Tukey adjustment for multiple comparisons.

(e.g., Tamoxifen). Because of these treatments, one might expect the HRQL of these women to be lower than that of women in the general population or women without a history of cancer. Although the HRQL of women with breast cancer declines during treatment, it improves during the first year after diagnosis and fluctuates only slightly 2 and 3 years after diagnosis [9, 34]. Helgeson et al. assessed the trajectories of change in PCS and MCS scores from 4 to 55 months after breast cancer diagnosis and found that scores for the majority of women either stayed roughly the same or improved slightly over that time period [35]. Only 2.1% of women exhibited a decreasing trend in PCS, and 12.2% of women had a decreasing trend in MCS scores [35]. The average time from breast cancer diagnosis to enrollment in the WHEL study was approximately 2 years; therefore, HRQL for the majority of women in the WHEL sample may have recovered (i.e., approached or even surpassed pre-cancer HRQL levels) by the time they were enrolled. Furthermore, breast cancer survivors often report better HRQL than the general population [9, 36, 37].

The similarity in HRQL among the WHEL and WHI-DM studies may also be due in part to similarities in sample characteristics. WHEL and WHI-DM participants in our study live in roughly the same area of northern California, both studies are dietary interventions, participants have similar educational and race/ethnic backgrounds, and participants in both studies are highly motivated. However, the lack of statistically significant and meaningful differences between the WHI-DM and WHEL studies may be due to the general nature of the SF-36 instrument, which may not be sensitive to subtle differences in HRQL due to breast cancer diagnosis and treatment. For example, arm pain, sexual functioning, body image and fear of recurrence are not measured on the SF-36. These have been shown to be related to HRQL in breast cancer patients, but unlike other HRQL domains, they are less likely to improve over time [9, 24, 38]. It is possible that differences between WHEL and WHI-DM participants in domains such as arm pain, sexual functioning, body image, or fear of recurrence may be statistically significant and meaningful.

It has been suggested that a difference of 5 points between a group mean and a fixed norm, such as a general population norm, is 'clinically and socially relevant' ([3], p. 7:12). Wywrich et al. identified meaningful intra-individual differences for SF-36 scales based on one standard error of measurement [39]. They reported a range of meaningful differences from 7.7 points for the physical functioning scale to 14.2 points for the social functioning scale [39]. We observed a group difference in the role-physical scale of 5.3 points (see Table 2) that we concluded was clearly not meaningful due to small effect sizes (0.12–0.19). Similarly, a score difference as large as 9.8 points for the social functioning scale was deemed only potentially meaningful (effect sizes 0.35–0.47). Our results and those of Wywrich et al. suggest that the variability across SF-36 scales may simply be too great to rely on a single number for interpreting score differences.

We used two sources of SDs to calculate effect sizes; the RMSE from the ANCOVA models and normative data. For most SF-36 scales, the RMSE tended to fall in the center or towards the low end of the range of norm SDs. Because the magnitude of a meaningful difference may vary slightly across patient groups, reporting a range rather than a single number has been recommended [29, 40]. By using two sources of SDs and computing a range of plausible effect sizes, we have met this recommendation.

Many of the statistically significant differences we observed in SF-36 scores were associated with small to moderate effect sizes and were considered only potentially meaningful. In a recent study of HRQL in the WHI Hormone Replacement Therapy (WHI-HRT) trial, Hays et al. [19] compared RAND-36 scores between the estrogen plus progestin and the placebo groups. They reported statistically significant differences for physical functioning (0.8 points) and bodily pain (1.9–2.0 points) scales, but concluded these differences were not meaningful based on the small effect sizes associated with them. As with our study, Hays et al. [19] incorporated effect sizes to aid in the interpretation of HRQL comparisons across groups.

Effect sizes for the WHEL vs. MOS or the WHI-DM vs. MOS score differences were larger and more likely to be clearly meaningful for scales more strongly associated with the physical health domain, (namely physical function, role-physical, general health and the PCS) than for scales more

strongly associated with the mental health domain (namely vitality, social functioning, mental health and the MCS). By incorporating effect sizes we were able to detect a pattern not readily apparent by comparing the magnitude of unadjusted scores (Figure 1) or the magnitude of adjusted scale score differences (Table 2). That we observed larger differences in the physical health domain of these samples is understandable given that three of the four conditions present in the MOS participants were physical (hypertension, diabetes, and heart disease). Comorbid conditions, such as those experienced by the MOS sample, have been shown to be related to decreased HRQL [3, 23, 41].

WHEL and WHI-DM women were self-selected to participate in dietary intervention trials, and therefore, were likely to have better overall health status, even though WHEL participants had been diagnosed and treated for breast cancer within the 4 years prior to enrollment. The WHEL and WHI-DM protocols are somewhat demanding; participants are highly functioning, and they are motivated to adopt the dietary changes. These characteristics may further explain why women in the WHEL and WHI-DM samples had better HRQL than the MOS sample.

Our study is not without limitations. We excluded participants due to missing SF-36 scale data or demographic data. It is possible that participants with missing data differ in their HRQL from those with complete data; thus, excluding participants with missing data may have introduced selection bias. The random sample of the 420 WHI-DM participants was not representative of sampling frame 0f 764 women from which it was taken with respect to demographic characteristics, and the random sample of 420 MOS participants was not representative of its sampling frame of 1741 participants with respect to mean role-emotional and general health scales. To ensure these discrepancies did not markedly affect the results, the analysis was repeated with the all 2925 participants with complete data from the three studies (420 WHEL + 764 WHI-DM + 1741 MOS; data not shown). The results for the large samples and the random samples were the same for general health scale. The mean differences for the role-emotional scale comparisons were slightly smaller for the large samples and the effect sizes were in the potentially meaningful range for both the WHEL – MOS comparison (0.37–0.46) and WHI-DM – MOS comparison (0.38–0.48) rather than the clearly meaningful range (0.44–0.56 and 0.44–0.55, respectively) for the random samples. The overall conclusions, however, were the same in that HRQL for the WHEL and WHI-DM samples was better than that for the MOS, particularly with respect to scales related to physical health.

The item measuring the severity of bodily pain that was administered to the WHEL and WHI-DM participants was missing the 'very severe' response option, and it could not be scored as intended by the SF-36 scoring algorithms [3]. Although this error did not affect comparisons of the bodily pain scale between the WHEL and WHI-DM samples, it might have affected the comparisons between these two samples and the MOS sample and population norms; bodily pain was the only scale that did not differ between the MOS and the WHEL or WHI-DM samples.

Because women enrolled in WHEL and WHI-DM were motivated and self-selected to participate in these dietary intervention studies, the results of the current study may not be generalizable to the general population, to breast cancer survivors, or to other patient groups. Furthermore, only participants at the UC Davis sites of the WHEL and WHI-DM studies were included in our study. As these are both multi-center studies, the results of our study may not reflect the HRQL of the WHEL or WHI-DM participants in general. Finally, WHEL and WHI-DM participants were from the same geographic area in northern California, whereas MOS participants were from Boston, Chicago, and Los Angeles, which may account for some of the observed differences in HRQL between MOS and either the WHEL or WHI-DM.

We conclude that the HRQL of the healthy postmenopausal women and breast cancer survivors was similar, whereas both of these groups had better HRQL than women with chronic health conditions. Statistically significant group differences between women with chronic health conditions and either breast cancer survivors or healthy women were more likely to be clearly meaningful (i.e., effect sizes greater than 0.5) for the physical health domain of HRQL as measured by the SF-36 (e.g., physical functioning, role-physical,

general health, PCS) than for the mental health domain (e.g., vitality, social functioning, mental health, MCS), a pattern that may not have been detected without the aid of effect sizes. This study illustrated the use of internal (RMSE) and external (U.S. population norms for females) sources of SDs for deriving ranges of effect sizes to aid the interpretation of multiple group comparisons. These results should be useful to investigators interested in estimating and interpreting SF-36 score differences across independent studies.

Acknowledgements

The Women's Healthy Eating and Living Study (WHEL) was funded by a Walton Family Foundation grant and National Cancer Institute grant CA69375. The Women's Health Initiative (WHI) was funded by the National Heart, Lung and Blood Institute, National Institutes of Health, Department of Health and Human Services. The authors would like to thank Amy Peterman, PhD, David Eton, PhD and Elizabeth Hahn, M.A. for their thoughtful reviews and helpful suggestions.

References

1. Ware JE. Measures for a New Era of Health Assessment. In: Stewart AL, Ware JE (eds), Measuring Functioning and Well-Being: The Medical Outcomes Study Approach, Durham, NC: Duke University Press, 1992; 3–11.
2. Stewart A, Ware J. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach., 1992.
3. Ware J, Snow K, Kosinski M. SF-36 Health Survey: Manual and Interpretation Guide. Lincoln, RI: Quality-Metric Incorporated, 1993.
4. Ware J, Kosinski M, Keller S. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab, 1994.
5. Wolinsky FD, Wan GJ, Tierney WM. Changes in the SF-36 in 12 months in a clinical sample of disadvantaged older adults. Med Care 1998; 36: 1589–1598.
6. Anderson JP, Kaplan RM, Coons SJ, Schneiderman LJ. Comparison of the Quality of Well-being Scale and the SF-36 results among two samples of ill adults: AIDS and other illnesses. J Clin Epidemiol 1998; 51: 755–762.
7. Schlenk EA, Erlen JA, Dunbar-Jacob J, et al. Health-related quality of life in chronic disorders: A comparison across studies using the MOS SF-36. Qual Life Res 1998; 7: 57–65.
8. Sprangers MA, de Regt EB, Andries F, et al. Which chronic conditions are associated with better or poorer quality of life? J Clin Epidemiol 2000; 53: 895–907.
9. Ganz PA, Coscarelli A, Fred C, Kahn B, Polinsky ML, Petersen L. Breast cancer survivors: psychosocial concerns and quality of life. Breast Cancer Res Treat 1996; 38: 183–199.
10. Newman V, Rock CL, Faerber S, Flatt SW, Wright FA, Pierce JP. Dietary supplement use by women at risk for breast cancer recurrence. The Women's Healthy Eating and Living Study Group. J Am Diet Assoc 1998; 98: 285–292.
11. Pierce JP, Faerber S, Wright FA et al. A randomized trial of the effect of a plant-based dietary pattern on additional breast cancer events and survival: the Women's Healthy Eating and Living (WHEL) Study. Control Clin Trials 2002; 23: 728–756.
12. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control Clin Trials 1998; 19: 61–109.
13. Matthews KA, Shumaker SA, Bowen DJ, et al. Women's health initiative. Why now? What is it? What's new? [see comments]. Am Psychol 1997; 52: 101–116.
14. Umayahara M, Lang E. Medical Outcomes Study, 1986–1992: A User's Guide to the Machine-Readable Files and Documentation. Los Altos, CA: Sociometrics Corp., 1998.
15. Rogers WH, McGlynn EA, Berry SH, et al. Methods of sampling. In: Ware JE, Jr., Stewart AL, (eds), Measuring Functioning and Well-Being. Durham, NC: Duke University Press, 1992; 27–47.
16. SAS Institute Inc. SAS/STAT Online Documentation. Cary, NC: SAS Institute, Inc., 1999.
17. Rutherford A. Introducing ANOVA and ANCOVA: A GLM Approach, London: SAGE Publications, 2001.
18. Tabachnick B, Fidell L. Multivariate analysis of variance and covariance. In: Using Multivariate Statistics, New York, NY: HarperCollins College Publishers 1996: 375–440.
19. Hays J, Ockene JK, Brunner RL, et al. Effects of estrogen plus progestin on health-related quality of life. N Engl J Med 2003; 348: 1839–1854.
20. Fairclough DL, Cella DF. Functional Assessment of Cancer Therapy (FACT–G): non-response to individual questions. Qual Life Res 1996; 5: 321–329.
21. SAS Institute Inc. SAS OnlineDoc®, Version 8. Cary, NC: SAS Institute, Inc., 1999.
22. Ganz PA, Hirji K, Sim MS, Schag CA, Fred C, Polinsky ML. Predicting psychosocial risk in patients with breast cancer. Med Care 1993; 31: 419–431.
23. Ashing-Giwa K, Ganz PA, Petersen L. Quality of life of African-American and white long term breast carcinoma survivors [published erratum appears in Cancer 1999 Aug 15;86(4):732-3]. Cancer 1999; 85: 418–426.
24. King MT, Kenny P, Shiell A, Hall J, Boyages J. Quality of life three months and one year after first treatment for early stage breast cancer: Influence of treatment and patient characteristics. Qual Life Res 2001; 9: 789–800.
25. Parker PA, Baile WF, Moor Cd C, Cohen L. Psychosocial and demographic predictors of quality of life in a large sample of cancer patients. Psychooncology 2003; 12: 183–193.
26. Rustoen T, Moum T, Wiklund I, Hanestad BR. Quality of life in newly diagnosed cancer patients. J Adv Nurs 1999; 29: 490–498.
27. Cella DF, Eton DT, Fairclough DL, et al. What is a clinically meaningful change on the Functional Assessment of

Cancer Therapy - Lung (FACT-L): Results from the Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol 2002; 55: 285–295.
28. Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined Minimally Important Differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol 2004; 57: 898–910.
29. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002; 77: 371–383.
30. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989; 27: S178–S189.
31. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: The remarkable Universality of Half a Standard Deviation. Med Care 2003; 41: 582–592.
32. Hedges L, Olkin I. Estimation of a single effect size: Parametric and nonparametric methods. In: Statistical Methods for Meta-Analysis, San Diego, CA: Academic Press, 1985; 76–106.
33. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1988.
34. Michael YL, Kawachi I, Berkman LF, Holmes MD, Colditz GA. The persistent impact of breast carcinoma on functional health status: Prospective evidence from the Nurses' Health Study. Cancer 2000; 89: 2176–2186.
35. Helgeson VS, Snyder P, Seltman H. Psychological and physical adjustment to breast cancer over 4 years: Identifying distinct trajectories of change. Health Psychol 2004; 23: 3–15.
36. Ganz PA, Rowland JH, Desmond K, Meyerowitz BE, Wyatt GE. Life after breast cancer: Understanding women's health-related quality of life and sexual functioning. J Clin Oncol 1998; 16: 501–514.
37. Bower JE, Ganz PA, Desmond KA, Rowland JH, Meyerowitz BE, Belin TR. Fatigue in breast cancer survivors: Occurrence, correlates, and impact on quality of life. J Clin Oncol 2000; 18: 743–753.
38. Hartl K, Janni W, Kastner R, et al. Impact of medical and demographic factors on long-term quality of life and body image of breast cancer patients. Ann Oncol 2003; 14: 1064–1071.
39. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999; 52: 861–873.
40. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000; 18: 419–423.
41. Gijsen R, Hoeymans N, Schellevis FG, Ruwaard D, Satariano WA, van den Bos GA. Causes and consequences of comorbidity: A review. J Clin Epidemiol 2001; 54: 661–574.

Address for correspondence: Kathleen Yost, Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare Research Institute, 1001 University Place, Suite 100, Evanston, IL 60201, USA
Phone: + 1-224-364-7395
E-mail: kyost@enh.org