Breast Cancer Res Treat (2008) 109:515–526
DOI 10.1007/s10549-007-9682-9

CLINICAL TRIAL

Symptom measurement in the Breast Cancer Prevention Trial (BCPT) (P-1): psychometric properties of a new measure of symptoms for midlife women

David Cella · Stephanie R. Land · Chih-Hung Chang · Richard Day · Joseph P. Costantino · Norman Wolmark · Patricia A. Ganz

Received: 29 June 2007 / Accepted: 9 July 2007 / Published online: 13 September 2007

Abstract *Purpose*: To evaluate scalability of a symptom scale administered to women enrolled in the Breast Cancer Prevention Trial (BCPT) (P-1) conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP). *Patients and methods*: Responses of 11,064 women recruited into a study of 20 mg daily tamoxifen versus placebo to prevent breast cancer in high-risk women were analyzed. Exploratory factor analyses of the 12 month data were conducted on a random subset of 4,000 women to estimate the factor structure. Baseline data on these same 4,000 women were analyzed to confirm the structure. The remaining sample was divided randomly into two data sets. Data on each set were then grouped by age (35–49, 50–59, or ≥60 years) and treatment (tamoxifen or placebo) to corroborate these analyses. Correlations between the obtained symptom clusters and two standard instruments (SF-36 and CES-D) were examined. Content analysis of open-ended responses was also conducted.*Results*: Eight clinically-interpretable clusters of symptoms were identified and confirmed: Cognitive symptoms, musculoskeletal pain, vasomotor symptoms, nausea, sexual problems, bladder problems, body image, and vaginal symptoms. Scoring for each scale represented by these eight clusters is provided. Content analysis of open-ended responses suggested four items that are additional candidates: fatigue, back problems, abdominal pain, and leg/foot cramps or pain. *Conclusions*: Symptoms associated with hormone therapy for breast cancer can vary. Nevertheless, the BCPT Eight Symptom Scale (BESS) can be clustered into clinically relevant and reproducible factors that may be useful in future outcomes research.

Keywords Breast cancer · Breast cancer prevention · Symptoms · Scale · Eight · Menopausal · Endocrine · NSABP · NHLBI (PEPI) · BCPT · P-1

Trial registration: Cancer.gov
URL for BCPT/P-1: http://www.cancer.gov/search/ResultsClinicalTrialsAdvanced.aspx?protocolsearchid = 3071847

D. Cella · S. R. Land · C.-H. Chang · R. Day · J. P. Costantino · N. Wolmark · P. A. Ganz
National Surgical Adjuvant Breast and Bowel Project (NSABP), Pittsburgh, PA, US

D. Cella (✉)
Center on Outcomes, Research and Education, Evanston Northwestern Healthcare 1001 University Place, Suite 101, Evanston, IL 60201, US
e-mail: d-cella@northwestern.edu

D. Cella · C.-H. Chang
Institute for Healthcare Studies and Robert H. Lurie Comprehensive Cancer Center, Northwestern University Medical School, Evanston, IL, US

S. R. Land · J. P. Costantino
NSABP Biostatistical Center, Pittsburgh, PA, US

S. R. Land · R. Day · J. P. Costantino
Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, US

C.-H. Chang
Buehler Center on Aging, Health & Society, Feinberg School of Medicine, Northwestern University, Chicago, IL, US

N. Wolmark
Allegheny General Hospital, Pittsburgh, PA, US

P. A. Ganz
Schools of Medicine and Public Health, and the Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA, US

Introduction

Breast cancer is the most common malignancy diagnosed in women, and the second leading cause of cancer-related death among women [1]. Its prevention is an important public health issue. The National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial (BCPT) (P-1) demonstrated that tamoxifen was an effective preventive agent, reducing breast cancer risk by 49% in women 35 years or older with an increased risk of breast cancer [2]. Critical to the adoption of a preventive therapy is the balance between benefits and risks as perceived by those for whom the therapy is targeted. For this reason, a comprehensive assessment of health-related quality of life (HRQL) was included as part of the P-1 trial [3] to monitor safety of the therapy as well as to facilitate subsequent dissemination of the outcomes to potential users. In particular, there was concern that tamoxifen, an anti-estrogen, might contribute to an increase in vasomotor or depressive symptoms among women taking the active agent in this placebo-controlled trial. Thus, in the design of the HRQL battery, a checklist was used to monitor symptoms that could be occurring in normal healthy aging women and/or might be exacerbated by tamoxifen. Use of tamoxifen in women with breast cancer is associated with symptoms consistent with anti-estrogen activity [4, 5], but interpretation is confounded by the coexistence of disease or premature menopause induced by chemotherapy. Symptoms associated with the normal menopausal transition, such as vaginal dryness, sexual dysfunction, irritability, headache and cognitive complaints [6], also require consideration. A similar approach was used by Fallowfield et al. [7] in the Royal Marsden breast cancer prevention trial. Thus, it was critical that the assessment of outcomes in the P-1 trial be able to discriminate between side effects and symptoms attributable to the treatment vs. those that would have occurred without treatment in a healthy aging population of midlife women. To do this, a reproducible and valid assessment of symptoms associated with hormone therapy is required.

In preparation for the P-1 trial, the NSABP investigators searched for an appropriate questionnaire that would capture a range of symptoms that occur in midlife women, including the symptoms known to be associated with tamoxifen therapy for breast cancer (e.g., hot flashes, vaginal discharge). We selected the symptom checklist developed for the Post-menopausal Estrogen/Progestin Interventions (PEPI) trial of the National Heart Lung and Blood Institute (NHLBI), because this checklist included an array of symptoms similar to those deemed relevant for our trial, and it had been used in a clinical trial with healthy postmenopausal women [8]. With input from NHLBI staff (S. Shumaker et al. personal communication), the checklist was modified to expand the response options for each symptom from a "yes/no" checklist, to add items (e.g., "problems with vaginal discharge" and "unhappy with body appearance") and to eliminate some items that were not appropriate or were deemed redundant.

The NSABP modification used a two-part response format in which women first indicated the presence or absence of the symptom, then rated the severity ("bother") on a five-point scale, where "0" = "not at all" and "4" = "extremely." This format was drawn from the experience of the University of Rochester Community Clinical Oncology Program (G. Morrow, personal communication). Items were reviewed and approved by a 10-member expert quality of life panel, including some authors in this report (PAG and DC). A total of 42 symptoms was retained in the final questionnaire, which has been widely known as the BCPT Symptom Checklist (SCL). The resulting questionnaire was therefore untested but believed by NSABP investigators to include the important symptoms relevant to the trial population and the preventive agent being studied. The BCPT SCL had not been subjected to systematic psychometric analysis in this patient population. We report here the psychometric evaluation of participant responses to this questionnaire. The analyses presented here provide information that will permit scoring guidelines and recommendations for use and interpretation of trial results in the prevention setting—to allow for more complex analyses than were feasible in the initial reports from the P-1 Trial [9, 10, 3]. A systematic approach to scaling helps to identify clinically meaningful and psychometrically sound subscales that will improve the interpretability and ultimate value of the symptom data collected on this trial and in future studies.

Methods

Participants and administration

These trials were approved by local human investigations committees or institutional review boards in accordance with assurances filed with and approved by the Department of Health and Human Services. Written informed consent was required for participation in each trial.

Data were drawn from 11,064 women enrolled in the first NSABP breast cancer prevention trial (BCPT) [9, 10, 3]. Symptom data were collected at baseline (pre-treatment), 3, 6, 12, 24, and 36 months. The BCPT SCL data were collected in conjunction with the Medical Outcomes Study 36-item Short Form (SF-36; [11, 12]) questionnaire and the Center for Epidemiologic Studies—Depression scale (CES-D; [13, 3]). Baseline and 12 month data from the subset of women (N = 11,064) enrolled into the trial from June 1,

1992 to May 31, 1994 were used to establish and confirm the scale structure. Analyses were restricted to this cohort of participants, representing 82.6% of the total P-1 accrual (N = 13,388), to assure that subjects would have at least 36 months of complete follow-up prior to the study unblinding in the spring of 1998. Demographic information (e.g., age, ethnicity, education, etc.), medical history (e.g., hysterectomy, oophorectomy, menopausal status, prior malignancy, heart disease, cerebrovascular disease, osteoporosis, diabetes, etc.) and treatment assignment were recorded. Demographic and clinical characteristics of this cohort are reported in Table 1. More detailed information on this cohort can be found in Day [10].

Administration of BCPT SCL

On the BCPT SCL, women were asked if they had experienced any of the 42 listed symptoms and, for each endorsed symptom, its severity on a 5-point Likert-type scale. This produced a six-level scale for analysis: 0 = did not report symptom; 1 = reported symptom with no bother; 2 = slightly bothered by symptom; 3 = moderately bothered by symptom; 4 = quite a bit bothered by symptom; 5 = extremely bothered by symptom. The baseline and 12-month data were selected for scaling analyses. Compliance rates for evaluation on the BCPT SCL were 99.3% (10,982 of 11,064 randomized) at baseline and 88.1% (9,747 of 11,064 randomized) at the 12 month assessment. There were no differences in form completion rate between placebo and tamoxifen groups.

Data preparation

The underlying factor structure of the 42-item BCPT SCL data (yes/no and severity combined) from baseline (N = 11,064) and 12-month follow-up data were investigated via both exploratory and confirmatory factor analyses for scale construction. Figure 1 details the steps taken to systematically divide the sample and derive interpretable symptom clusters.

From the baseline sample, 4,000 participants were randomly selected; 2,000 were then randomly selected and the remaining 2,000 were split according to treatment arm, with 1,025 women in the tamoxifen group and 975 women in the placebo group. Exploratory factor analysis (EFA) was first performed on the 12-month follow-up data followed by confirmatory factor analysis (CFA) of the baseline data for each of three age groups (35–49, 50–59, and ≥60 years). The remaining (unselected) baseline sample (N = 7,064) was first split by the two treatment arms and then divided into the three age groups within each arm, resulting in 6 subsets of samples. Samples in each of these 6 subsets were then randomly divided into two smaller groups (referred to as samples A and B in Fig. 1), with approximately equal numbers of women in each group. An EFA was completed on both the baseline and the12-month follow-up data in the first half of the sample in each group. The stability of the derived scales across the three different age groups and two treatment arms was then evaluated via CFA of the second half of the samples in each of the 6 stratified groups.

Exploratory factor analysis (EFA)

In the initial scale development stage, we employed factor analysis to explore the underlying structure of these 42 symptoms. Because the baseline responses were recorded *before* administration of any drug, we chose the 12 month data for these initial exploratory analyses. Treatment-associated symptom clusters, perhaps not apparent at baseline, could thus be identified. Our strategy of exploring the factor structure at 12 months, followed by confirmation with baseline data, permitted us to ensure that the derived scales would perform well in untreated women.

By grouping responses according to the underlying structure of a correlation matrix, EFA [14] enabled us to identify any redundancies in the 42 responses to the questions, and to derive more efficient scales of correlated items, or factors. We employed principal components analysis (PCA); [15, 16], wherein linear combinations of the observed variables are formed. The components are estimated one at a time to represent the overall variability in the data matrix. The first principal component accounts for the largest amount of variance, and the second component explains the next largest amount of variance and is uncorrelated to the first one. Each component has an eigenvalue, expressing the amount of variance accounted for by the component. To decide on a reasonable number of factors for EFA, one common rule, the Kaiser-Guttman criterion, is to retain for rotation any eigenvalue (or latent root) greater than 1.0. The number of factors to retain can be further evaluated using Cattell's scree plot, which plots the incremental variance (y-axis) accounted for by each successive factor (x-axis) to determine the point at which the explained variance levels out [17]. We used this scree plot technique to explore multiple optimal factor solutions. We used orthogonal rotation using the Varimax procedure [18], in which factors were kept uncorrelated, to produce a simple structure as interpretable and meaningful as possible. The items with the highest factor loadings were tpdel selected for relevant scales. Items with small (below 0.30) factor loadings were omitted from that factor. Scales were then constructed for each distinctive factor, with items selected using the criteria as described above. The SPSS statistical package was used to facilitate these steps. Factor structures obtained from

Table 1 Demographic and clinical characteristics of the HRQL sample (N = 11,064) from the NSABP BCPT

Characteristic	Placebo		Tamoxifen		Total	
	No.	%	No.	%	No.	%
Age						
Mean	53.83(SD ± 9.167)		53.82(SD ± 9.184)		53.83(SD ± 9.175)	
Median	52(35–79)		52(35–78)		52(35–79)	
Ethnicity						
White	5290	95.54	5282	95.57	10572	95.55
Hispanic	63	1.14	49	0.89	112	1.01
Black	88	1.59	95	1.72	183	1.65
Asian	35	0.63	37	0.67	72	0.65
Other	47	0.84	39	0.71	86	0.78
Missing	14	0.25	25	0.45	39	0.35
Education						
Grade school	61	1.10	66	1.19	127	1.15
Some high school	248	4.48	218	3.94	466	4.21
High school graduate	1003	18.11	1009	18.26	2012	18.19
Vocational school	593	10.71	614	11.11	1207	10.91
Some college	1180	21.31	1194	21.60	2374	21.46
Associate degree	349	6.30	349	6.31	698	6.31
College and beyond	2069	37.36	1936	36.93	4110	37.15
Missing	34	0.61	36	0.65	70	0.63
Employment						
Unemployed	239	4.32	229	4.14	468	4.23
Retired	925	16.71	938	16.97	1863	16.84
Full-time homemaker	660	11.92	670	12.12	1330	12.02
Employed full-time	2713	49.00	2682	48.53	5395	48.76
Employed part-time	880	15.89	878	15.89	1758	15.89
Other	106	1.91	104	1.88	210	1.90
Missing	14	0.25	26	0.47	40	0.36
Marital status						
Never married	398	7.19	394	7.13	792	7.16
Presently married	3843	69.41	3876	70.43	7719	69.77
Marriage-like	139	2.51	125	2.26	264	2.39
Divorced	748	13.51	707	12.79	1455	13.15
Widowed	395	7.13	399	7.22	794	7.18
Missing	14	0.25	26	0.47	40	0.36
Relative Risk of BC						
1–2	416	7.51	416	7.53	832	7.52
2–3	929	16.78	865	15.65	1794	16.21
3–5	2074	37.46	2154	38.97	4228	38.21
5–10	1618	29.22	1605	29.04	3223	29.13
10+	500	9.03	487	8.81	987	8.92

separate analysis were compared. We did not expect perfect conformity of item composition across the exploratory comparisons, so we planned to select the factor solution that was most consistently observed across analyses for further refinement.

Confirmatory factor analysis (CFA)

Several CFA analyses were conducted to determine the extent to which the derived factor model fit the sub-samples of women described in Fig. 1. CFA allowed us to test

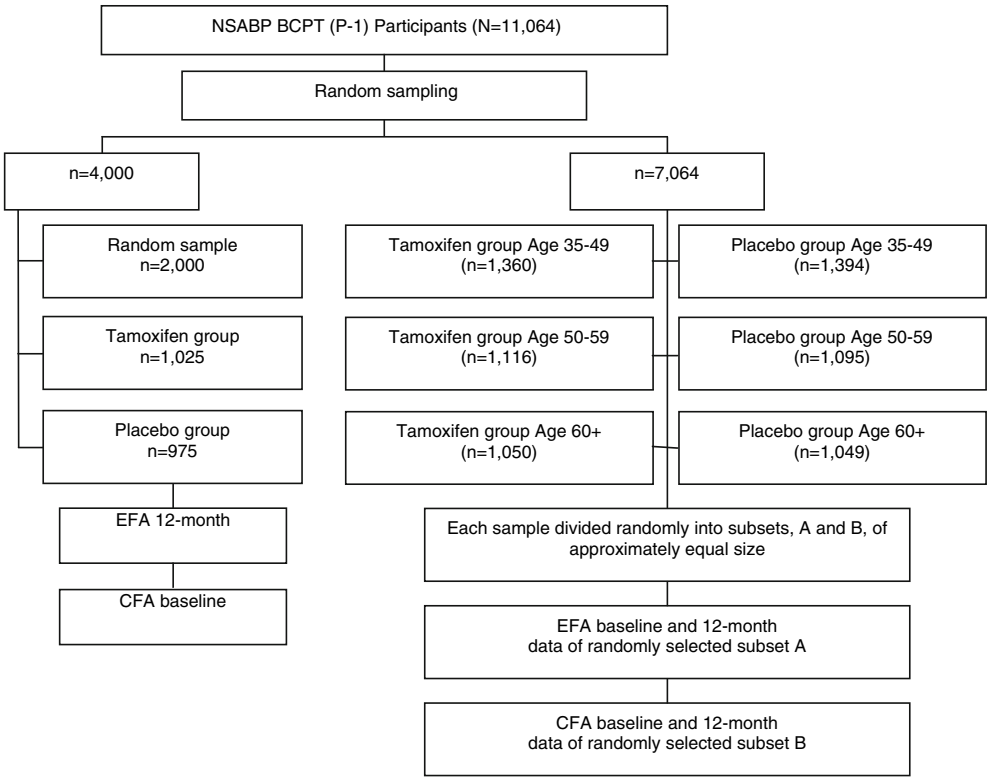

Fig. 1 Sample selection and analysis flow for the BCPT Symptom Check List

whether the scales identified by multiple EFA procedures on independent samples indeed tapped the latent dimensions thought to underlie this array of self-reported symptoms. The CFA model was estimated with the LISREL 8 computer program using the maximum likelihood method based on the sample covariance matrix. The analysis was based on an a priori specification that certain factors characterized the data, and the model investigated corresponded exactly to the expected structure. That is, each item was constrained to load on one and only one factor. This allowed us to empirically validate the factor structure of the BCPT SCL items when applied to different groups of samples.

The LISREL 8 program provides a chi-square test of the overall goodness of fit of a model; a non-significant chi-square indicates excellent fit between a proposed model and the data at hand. Because of the known limitations of this statistic as a measure of model fit (see Marsh [19] for a discussion of the effect of sample size on χ^2), it is important, however, to evaluate other fit criteria when deciding that a model adequately fits a dataset. Three supplemental fit indexes are available in LISREL 8—the normed fit index (NFI), the non-normed fit index (NNFI), and the comparative fit index (CFI). Values for these fit statistics typically range from 0 to 1, with values greater than 0.90 indicating that the tested model fit the data well [20].

Reconciliation of item composition of scales from EFA and CFA

We set out to determine, after applying EFA and CFA, which items to retain in each derived scale. The items dropped during this reconciliation were retained for later study, as they could either be re-added into a scale for conceptual or clinical reasons (usually with minimal loss to reliability and precision), or they could be set aside and analyzed separately for clinical interest.

In summary, the development of the scales began with EFA, followed by CFA and conceptual reconciliation of the subscale composition. We expected many similarities across analyses and determined that any remaining inconsistencies could be evaluated and rectified on conceptual (i.e., clinical) grounds. The reduction of the BCPT SCL data into meaningful and usable components, important for its usefulness as a clinical and research tool, was thereby achieved.

Determining validity

Planned concurrent and content validity analyses were conducted after the above factor analysis-based scaling procedures were conducted and subscales were defined. Concurrent validation was possible by virtue of the

concurrent collection of health status data (SF-36; [12, 11]), and depression symptom data (CES-D [13]). After scale construction was completed, concurrent validity of the newly created SCL scales was examined using the 8-scale and 2 summary component scores of the SF-36 to define different groups. Scores on the CES-D depression scale were also used to evaluate concurrent validity of the newly derived scales.

To assess item coverage of the new scales, we performed content analysis of a single open-ended item asking patients to list "other" symptoms. This item was included in the checklist completed by all participants, but the content analysis was based on a stratified random sample of 54 subjects who reported being "quite a bit" or "extremely" bothered by "other" symptoms after the start of treatment. The stratification was by assessment time point and treatment arm. For the selected subjects, all submitted questionnaires were examined. Respondents' descriptions of the symptoms were transcribed and categorized as either redundant with existing items or distinct. The distinct items were examined with respect to severity and frequency reported. Items that listed personal circumstances (such as "I am worried about my husband's job loss") or medical conditions (e.g., diabetes) were eliminated. A symptom was considered a candidate for future instrument development if it was reported by at least 5 participants among the 54 examined. Formal estimation of the symptom prevalence based on the content analysis is not possible because it would require an estimation of the under-reporting that is expected for symptoms not included in the questionnaire.

Results

Factor analysis results are summarized in Table 2. Symptom items were retained in the same cluster when two criteria were met: 1) Its factor loading exceeded .40; and 2) The percent agreement among analyses of different sets of samples was over 50%. As a result, eight interpretable

Table 2 Exploratory factor analysis results

Symptom	12 Months		12 Months						Baseline						% times item loads on factor
			35–49		50–59		60+		35–49		50–59		60+		
	T	P	T	P	T	P	T	P	T	P	T	P	T	P	
1. Cognitive symptoms															
Difficulty concentrating	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Easily distracted	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Forgetfulness	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Tendency to take nap; stay in bed	●	●				●		●							29
Avoidance of social affairs	●	●				●					●	●			36
Tendency toward accident															0
2. Musculoskeletal pain															
Joint pains	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Muscle stiffness	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
General aches and pains	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Swelling of hands or feet						●						●			14
Numbness, tingling													●		7
3. Vasomotor symptoms															
Night sweats	●	●	●	●	●	●	●		●	●	●	●	●		**86**
Hot flashes	●	●	●	●	●	●	●		●	●	●	●	●		**86**
Cold sweats	●		●	●	●	●			●	●			●		**57**
Early awakening															0
4. Gastrointestinal symptoms															
Vomiting	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Nausea	●	●	●	●	●	●	●	●	●		●	●	●	●	**93**
Diarrhea	●		●	●	●	●		●	●	●		●		●	**71**
5. Dyspareunia															
Vaginal dryness	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Pain with intercourse	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**

Table 2 continued

Symptom	12 Months		12 Months						Baseline						% times item loads on factor
			35–49		50–59		60+		35–49		50–59		60+		
	T	P	T	P	T	P	T	P	T	P	T	P	T	P	
6. Bladder control															
Difficulty with bladder control (when laughing or crying)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
Difficulty with bladder control (at other times)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	**100**
7. Weight concerns															
Weight gain	●	●	●	●			●	●	●	●	●	●	●	●	**86**
Unhappy with the appearance of my body	●	●	●	●		●		●	●	●	●	●	●	●	**86**
Weight loss	●		●		●	●	●								36
Decreased appetite					●		●								14
8. Gynecologic symptoms															
Vaginal discharge	●	●	●	●	●	●	●	●	●		●				**71**
Genital itching/ irritation	●		●	●	●	●	●		●			●			**57**
Vaginal bleeding or spotting	●	●		●		●		●			●	●			**50**
Difficulty breathing	●	●	●					●	●		●		●	●	57
Feeling of suffocation	●	●	●					●			●		●	●	50
Chest pains	●	●	●					●	●						36
Dry mouth															0
Ringing in ears		●		●		●		●	●	●					43
Blind spots, fuzzy vision			●	●						●				●	29
Constipation	●										●	●			21
Dizziness, faintness		●				●		●	●	●	●	●		●	57
Headaches	●		●			●						●			29
Cramps	●	●		●					●						29
Breast sensitivity/ tenderness	●	●		●					●						29
Short temper													●	●	14
Excitability													●	●	14

● A dot in a box indicates an item factor loading ≥0.40

Last column summarizes the percentage of times the item in that row had a factor loading ≥0.40

Items in boldface comprise symptom clusters

T = Tamoxifen; P = Placebo

clusters of symptoms across age groups, treatment arms and time of assessment were identified. Each factor contained two or three items. Twenty-one of 42 items (50%) could therefore be retained in one of these eight newly created multi-item scales, the BCPT Eight Symptom Scale (BESS).

Cluster 1 has 3 cognitive symptom items (difficulty concentrating; easily distracted; forgetfulness) and was named *cognitive symptoms*. Cluster 2 has 3 items (joint pains; muscle stiffness; general aches and pains) related to *musculoskeletal pain*. Cluster 3 has 3 items (night sweats; hot flashes; cold sweats) capturing *vasomotor symptoms*. Cluster 4 has 3 *gastrointestinal symptoms* (vomiting; nausea; diarrhea). Cluster 5 has 2 items (vaginal dryness; pain with intercourse) related to *dyspareunia*. Cluster 6 has 2 *bladder control* items (difficulty with bladder control when laughing or crying; difficulty with bladder control at other times). Cluster 7 has 2 items (weight gain; unhappy with the appearance of my body) related to *weight concerns*. Cluster 8 has 3 *gynecologic symptoms* (vaginal discharge; genital itching/irritation; vaginal bleeding or spotting).

The maximum likelihood method of factor estimation in the LISREL 8 statistical package was used to test whether the 21 retained items reflected the eight latent constructs across age groups, treatment arms and time of assessment. The NFI, NNFI, and CFI showed high goodness-of-fit indexes, indicating excellent fit to the confirmatory factor model.

Table 3 provides a summary of the proportion of women endorsing each one of the clusters at baseline and follow-up along with the mean severity levels (0–5) for each

cluster. For this analysis we included the entire sample of 11,064 as a single pre-treatment group at baseline, and then split them into tamoxifen and placebo groups at follow-up. The baseline data allows one to appreciate the degree of pre-treatment symptom severity and amount of variance observed at the baseline examination. The 12 month data provide a comparison between tamoxifen and placebo-treated patients, where 95% confidence intervals (CI) around each reported mean can be estimated by multiplying the standard error by 1.96. This enables testing of significance between treatment arms at 12 months into treatment. The data indicate that substantial numbers of women endorsed some degree of difficulty in all of the BESS-identified clusters. At 12 months, one notes significantly more symptomatology in tamoxifen-treated patients on three of the eight clusters: vasomotor, bladder and vaginal symptoms. Overall, mean severity scores are below a "moderate" threshold at 12 month follow-up, even on the tamoxifen arm.

Table 4 summarizes two related trends observed over time: The number of symptom clusters endorsed, and the severity of symptoms endorsed by number of clusters endorsed. Women tend to endorse multiple clusters, with the highest percentages scoring 2–4 clusters at baseline and 3–5 clusters at 12 month follow-up. Considerably more clusters are endorsed at 12 months compared to baseline. For example, at baseline, 26% of women report *any* degree of difficulty on 5 or more clusters; at 12 months, that percentage increases to 41% among tamoxifen-treated patients, and 33% among placebo-treated patients. At both assessments, the average symptom severity is higher among women who endorse more clusters, suggesting that symptom burden accumulates beyond an additive factor, when multiple clusters of symptoms are experienced. Nevertheless, despite these clear increases in proportions of women reporting multiple clusters, and increased severity at 12 months compared to baseline, the maximum mean severity score was 2.04 (just above a "slightly" rating), observed at 12 months among women on tamoxifen who reported some difficulty with all of the BESS clusters.

Tables 5, 6, 7 provide the correlation coefficients for the relationships among the BESS clusters and the SF-36 and CES-D data collected at the same time point. The first two SF-36 scale columns represent overall physical and mental health indices, while the next eight are the standard SF-36 subscales [12, 21]. Given the large sample size, virtually all correlations are statistically significant. However, overall only the first two clusters (cognitive symptoms and body pain) show a moderate (above 0.30) correlation with any of the SF-36 or CES-D scales. This suggests that although substantial numbers of women are endorsing symptom clusters 3–8, these endorsements are not reflected in the higher level HRQL and depression scores. Clusters 1 and 2 seem consistent with the more general quality of life data, and the moderate strength of these correlations was similar at both pre- and post-treatment time points.

In the content analysis of the open-ended "other symptoms" item, the total number of questionnaires examined was 591 for 54 participants. There were four distinct "other" symptoms that were reported by at least 5 participants at some time point during the study. These were fatigue (9 participants), back pain or problems (15 participants), abdominal pain (6 participants), and pain or cramps in the legs or feet (15 participants).

Discussion

The original (42-item) BCPT SCL was used in the P-1 Trial to assess a range of potential physical and mental symptoms associated with women's aging and an anti-estrogen hormone

Table 3 Proportion of women endorsing specific clusters and mean cluster severity scores by examination (pre- and post- treatment) and treatment arm (tamoxifen and placebo)

Cluster	Baseline (Pre-treatment) entire sample		12 Months Tamoxifen arm		12 Months Placebo arm	
	Proportion endorsing	Mean severity (Std. Error)	Proportion endorsing (change over baseline)	Mean severity (Std. Error)	Proportion endorsing (change over baseline)	Mean severity (Std. Error)
1. Cognitive	0.45	0.60 (.008)	.50 (+.043)	0.77 (.015)	.49 (+.039)	0.76 (.015)
2. Body pain	0.69	1.19 (.011)	.69 (–.004)	1.30 (.018)	.70 (+.015)	1.39 (.018)
3. Vasomotor	0.41	0.56 (.008)	**.65 (+.243)**	**1.14 (.017)**	**.49 (+.083)**	**0.74 (.014)**
4. Gastrointestinal	0.23	0.23 (.005)	.25 (+.020)	0.28 (.008)	.29 (+.059)	0.32 (.008)
5. Sexual problems	0.26	0.47 (.009)	.32 (+.056)	0.64 (.016)	.31 (+.052)	0.60 (.016)
6. Bladder	0.35	0.56 (.009)	**.46 (+.103)**	**0.85 (.016)**	**.39 (+.038)**	**0.67 (.015)**
7. Body image	0.62	1.12 (.011)	.63 (+.012)	1.26 (.018)	.65 (+.024)	1.26 (.018)
8. Vaginal	0.26	0.26 (.005)	**.45 (+.185)**	**0.52 (.011)**	**.30 (+.027)**	**0.33 (.009)**

Boldface indicates a significant difference in symptom severity (at 5% level)

Table 4 Proportion women endorsing a cumulative number of clusters and mean severity scores by examination (pre- and post- treatment) and treatment arm (tamoxifen and placebo)

Number clusters endorsed	Baseline (Pre-treatment) entire sample		12 Months Tamoxifen arm		12 Months Placebo arm	
	Number and proportion endorsing	Mean severity (Std. error)	Number and proportion endorsing	Mean severity (Std. error)	Number and proportion endorsing	Mean severity (Std. error)
0	699 (.064)	0	178 (.037)	0	249 (.051)	0
1	1301 (.119)	0.16 (.003)	387 (.080)	0.18 (.005)	478 (.097)	0.17 (.005)
2	1858 (.169)	0.34 (.003)	556 (.115)	0.38 (.007)	697 (.142)	0.37 (.007)
3	2197 (.200)	0.53 (.004)	833 (.173)	0.58 (.008)	908 (.185)	0.57 (.007)
4	2057 (.187)	0.73 (.005)	902 (.187)	0.81 (.009)	969 (.197)	0.79 (.009)
5	1559 (.142)	0.94 (.008)	913 (.189)	1.05 (.011)	746 (.151)	1.06 (.013)
6	882 (.080)	1.22 (.012)	641 (.133)	1.34 (.015)	542 (.110)	1.30 (.017)
7	353 (.032)	1.51 (.022)	320 (.067)	1.67 (.026)	260 (.053)	1.63 (.027)
8	76 (.007)	1.79 (.055)	93 (.019)	2.04 (.047)	69 (.014)	1.90 (.069)

Table 5 Baseline (pre-treatment) examination correlations of mean severity scores on SCL clusters with SF-36 scales and CES-D scores (N = 10,982)

SF-36 Scales and CES-D											
Cluster	OP	OM	PF	RP	BP	SF	MH	RE	V	GHP	CESD
1. Psych	–0.229	**–0.419**	–0.239	–0.288	–0.271	**–0.337**	**–0.408**	**–0.388**	**–0.402**	–0.291	**0.463**
2. Pain	***–0.539***	–0.122	**–0.425**	**–0.342**	***–0.611***	–0.255	–0.220	–0.179	**–0.373**	**–0.400**	0.268
3. Vasomotor	–0.117	–0.122	–0.093	–0.106	–0.158	–0.118	–0.143	–0.095	–0.146	–0.116	0.160
4. GI	–0.176	–0.163	–0.127	–0.160	–0.198	–0.176	–0.166	–0.140	–0.213	–0.205	0.187
5. Sexual	–0.094	–0.063	–0.069	–0.087	–0.110	–0.066	–0.080	–0.069	–0.079	–0.096	0.089
6. Bladder	–0.196	–0.096	–0.196	–0.154	–0.171	–0.125	–0.115	–0.122	–0.181	–0.164	0.160
7. Body image	–0.229	–0.225	–0.215	–0.193	–0.233	–0.201	–0.237	–0.189	***–0.324***	–0.256	0.251
8. Vaginal	–0.058	–0.123	–0.027	–0.053	–0.113	–0.083	–0.121	–0.083	–0.133	–0.102	0.110
Overall	**–0.440**	**–0.325**	**–0.37**4	**–0.360**	***–0.500***	**–0.345**	**–0.373**	**–0.317**	**–0.471**	**–0.419**	**0.425**

Bold Items have correlations ≥0.30

Bold italic items have correlations ≥0.500

OP = Overall physical functioning, OM = Overall mental functioning, PF = Physical functioning, RP =Role-physical, BP = Bodily pain, SF = Social functioning, MH = Mental health, RE = Role-emotional, V = Vitality, GHP = General health perception, CES-D = Center for Epidemiological Studies-Depression scale

therapy being tested for the prevention of breast cancer [3]. This extensive exploratory and confirmatory factor analytic approach to data from several thousand women supported the long-held clinical impression that the items comprise a broad multidimensional array of symptoms. Half of these symptoms, compiled by an expert committee prior to the trial, can be aggregated into eight distinct scales (BESS) representing factors that are reproducible across women regardless of treatment received, age and menopausal status. The other half of the 42-item BCPT SCL is a mixture of unrelated symptoms that are either less relevant in this patient population, or relevant yet distinct from the symptom domains derived from the factor analysis. Importantly, BESS detected significant and expected differences in symptoms among healthy high risk women receiving tamoxifen in comparison to placebo, and thus is of value for use in the assessment of the effects of hormonal therapy in controlled clinical trials in the prevention setting. The results described here are consistent with those reported earlier from the P-1 Trial [9], but also identify subtle differences in bladder functioning which may be related to the tamoxifen therapy.

The BESS instrument, derived from the BCPT SCL, identifies unique symptom clusters that were reproducible across age groups, treatment arms, and time of assessment (Table 2), and as such has utility for future studies of healthy midlife women and women in the setting of breast cancer prevention trials.

Twenty-one items were retained, 2 or 3 in each cluster, to measure these eight distinct yet related symptoms. The remaining 21 items can be used at investigator discretion based upon clinical relevance. Among those items not included in the scales, three varied significantly between

Table 6 12 Month correlations of mean severity scores on SCL clusters with SF-36 scales and CES-D scores—tamoxifen group only (N = 5,537)

SF-36 Scales and CES-D 12 Month Scores											
Clusters	OP	OM	PF	RP	BP	SF	MH	RE	V	GHP	CESD
1 Psych	–0.254	**–0.456**	–0.268	**–0.352**	**–0.302**	**–0.364**	**–0.441**	**–0.436**	**–0.443**	**–0.331**	***0.502***
2 Pain	***–0.560***	–0.170	**–0.465**	**–0.396**	***–0.640***	–0.299	–0.246	–0.232	**–0.413**	**–0.414**	0.295
3 Vasomotor	–0.094	–0.116	–0.065	–0.105	–0.145	–0.093	–0.118	–0.100	–0.149	–0.082	0.134
4 GI	–0.245	–0.200	–0.185	–0.227	–0.277	–0.275	–0.208	–0.163	–0.260	–0.265	0.237
5 Sexual	–0.083	–0.089	–0.045	–0.087	–0.109	–0.065	–0.092	–0.072	–0.121	–0.107	0.106
6 Bladder	–0.234	–0.121	–0.224	–0.189	–0.227	–0.161	–0.136	–0.146	–0.219	–0.200	0.174
7 Body image	–0.195	–0.254	–0.193	–0.182	–0.227	–0.205	–0.255	–0.215	***–0.327***	–0.250	0.258
8 Vaginal	–0.079	–0.112	–0.059	–0.082	–0.112	–0.093	–0.121	–0.083	–0.127	–0.104	0.127
Overall	**–0.436**	**–0.362**	**–0.378**	**–0.400**	***–0.512***	**–0.372**	**–0.390**	**–0.354**	**–0.501**	**–0.424**	**0.443**

OP = Overall physical functioning, OM = Overall mental functioning, PF = Physical functioning, RP = Role-physical, BP = Bodily pain, SF = Social functioning, MH = Mental health, RE = Role-emotional, V = Vitality, GHP = General health perception, CES-D = Center for Epidemiological Studies - Depression scale

Bold Items have correlations ≥0.30, Bold italic items ≥0.50

treatment arms in the BCPT [22]. Weight loss was statistically significantly greater in the tamoxifen arm, but the difference was modest (odds ratio 1.14, 95% CI 1.05–1.23 for the odds of at least slight weight loss on at least one assessment after the start of treatment with tamoxifen vs. placebo). Headaches and breast sensitivity/tenderness were reduced in the tamoxifen arm, with odds ratios of 0.83 (95% CI 0.75–0.91) and 0.68 (95% CI 0.63–0.74), respectively.

The fit indexes of the model tested by CFA suggested that the BESS model provides an adequate approximation to the data across age groups, treatment arms, and time of assessment. The results of the two types of analyses (EFA and CFA) were generally congruent (regardless of group membership), shedding important light on the theoretical nature of these symptom clusters and lending additional support to the consistency of measurement across various demographic and treatment variables.

The results from this study are also largely consistent with another recent evaluation by Stanton and colleagues of the BCPT SCL in a series of patients with breast cancer and high risk women [23], and a study of breast cancer survivors [24], suggesting further validity of these measures across women in midlife at risk for a variety of symptoms. In the Stanton study, the BCPT SCL was used in four populations of women (sample sizes 863, 577, 560 and 208) who had been previously diagnosed with breast cancer or were at risk for the disease. Exploratory and confirmatory factor analysis also yielded eight scales. Five of the scales were identical to the scales derived in the BESS analysis. However, the vasomotor and gastrointestinal scales derived in the Stanton study did not include cold sweats and diarrhea, respectively. In addition, their analysis produced an arm function scale using items that were not in the original BCPT checklist, and did not include the present gynecological scale. Participants in the Stanton study were not necessarily using tamoxifen, which might explain the differences in the scales produced. Cold sweats, diarrhea, vaginal discharge and genital itching/irritation were all significantly different between tamoxifen and placebo in the BCPT, indicating that the present scales will have greater sensitivity to tamoxifen-related treatment effects [22]. The study of breast cancer survivors by Alfano et al. used 15 of the items from the BCPT. It produced five sub-scales that were very similar to five of the scales in the present report [24].

These analyses support the use of the BESS in a variety of research and clinical settings, where the routine assessment of symptoms is deemed important. The latter is especially true in the setting of prevention, where daily symptoms may detract from adherence to a preventive regimen. Multi-item scales, even as short as 2- and 3-items long, can more reliably measure these important clusters of symptoms than single item scales. In addition, these scales when used properly might prove useful in developing appropriate strategies for providing follow-up supportive care (reduction of symptoms, etc.). Most (6) of these symptom clusters were not correlated with more general assessments of HRQL or depression, suggesting that they measure unique information that may differentiate women who are having meaningful problems associated with aging or hormonal treatment.

In future use of this instrument, we recommend a simplified response structure with Likert-like severity scores,

Table 7 12-Month correlations of mean severity scores on SCL clusters with SF-36 scales and CES-D scores— placebo group only (N = 5,527)

SF-36 scales and CESD											
Clusters	OP12	OM12	PF12	RP12	BP12	SF12	MH12	RE12	V12	GHP12	CESD12
1. Psych	–0.207	***–0.509***	–0.233	**–0.343**	–0.281	**–0.392**	**–0.483**	**–0.476**	**–0.465**	**–0.325**	***0.533***
2. Pain	***–0.551***	–0.170	**–0.461**	**–0.379**	***–0.647***	–0.298	–0.253	–0.215	**–0.410**	**–0.410**	0.301
3. Vasomotor	–0.123	–0.137	–0.084	–0.107	–0.191	–0.135	–0.158	–0.099	–0.161	–0.154	0.168
4. GI	–0.186	–0.203	–0.146	–0.182	–0.234	–0.228	–0.201	–0.178	–0.238	–0.230	0.233
5. Sexual	–0.089	–0.104	–0.066	–0.090	–0.121	–0.102	–0.108	–0.097	–0.109	–0.122	0.115
6. Bladder	–0.223	–0.123	–0.235	–0.190	–0.198	–0.154	–0.138	–0.165	–0.197	–0.199	0.164
7. Body image	–0.185	–0.251	–0.192	–0.171	–0.216	–0.217	–0.248	–0.210	**–0.325**	–0.246	0.278
8. Vaginal	–0.071	–0.134	–0.040	–0.073	–0.116	–0.098	–0.127	–0.097	–0.164	–0.131	0.122
overall	**–0.423**	**–0.392**	**–0.377**	**–0.388**	***–0.517***	**–0.398**	**–0.420**	**–0.375**	***–0.510***	**–0.448**	**0.471**

OP = Overall physical functioning, OM = Overall mental functioning, PF = Physical functioning, RP = Role-physical, BP = Bodily pain, SF = Social functioning, MH = Mental health, RE = Role-emotional, V = Vitality, GHP = General health perception, CES-D = Center for Epidemiological Studies - Depression scale

Bold Items have correlations ≥0.30, Bold italic items ≥0.50

as has been used in more recent studies (0 = Not at all; 1 = Slightly; 2 = Moderately; 3 = Quite a bit; 4 = Extremely). This is consistent with other derivatives of the original P-1 Prevention Trial questionnaire [24, 25, 23]. Administration and scoring instructions for these eight derived scales described herein are available from either the first or second author. The content analyses of the open-ended "other symptoms" item also suggested four items that might be included in future instrument revisions: fatigue, back pain or problems, abdominal pain, and pain or cramps in the legs or feet. These items were selected based on their relevance to this population, without regard to association with therapy. However, there is evidence that both a decrease in fatigue (which was assessed in the BCPT with the SF-36 vitality scale [12, 11] and an increase in leg cramps are attributable to tamoxifen [9, 7, 25].

The unique symptom clusters produced by these analyses should be useful dependent variables in outcome studies that compare the relative efficacy of pharmacologic and non-pharmacologic hormonal interventions. Further studies are needed to address the relationship between these symptom scales and other important treatment outcomes, and to determine what would be a minimally important difference or change in score over time. Use of these separate symptom clusters may help to clarify the complex interrelationships between physical symptoms and health-related behavior, such as adherence to long-term prescription of hormonal therapy to prevent or treat breast cancer. We can now use the results of this analysis to examine the longitudinal symptom data from the P-1 trial, using defined scales, for treatment arm (tamoxifen vs. placebo) differences in an intent-to-treat analysis including appropriate multivariate models and handling of missing data.

Acknowledgements Supported by Public Health Service grants U10CA-37377, and U10CA-69974 from the National Cancer Institute, Department of Health and Human Services.The authors thank Barbara C. Good, PhD, Director of Scientific Publications for the NSABP, and Wendy L. Rea, for editorial assistance.

References

1. Jemal A, Tiwari RC, Murray T, Ghafoor A, Samuels A, Ward E, Feuer E J, Thun M J (2004) Cancer statistics.CA Cancer J Clin 54:8–29
2. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Wieand S, Tan-Chiu E, Ford L, Wolmark N (1998) Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 90:1371–1388
3. Ganz PA, Day R, Ware JE Jr, Redmond C, Fisher B (1995) Baseline quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial. J Natl Cancer Inst 87:1372–1382
4. Gallicchio L, Lord G, Tkaczuk K, Danton M, Lewis LM, Lim CK, Flaws JA (2004) Association of tamoxifen (TAM) and TAM metabolite concentrations with self-reported side effects of TAM in women with breast cancer. Breast Cancer Res Treat 85:89–97
5. Yen TW, Hunt KK, Mirza NQ, Thomas ES, Singletary SE, Babiera GV, Meric-Bernstam F, Buchholz TA, Feig BW, Ross MI, Ames FC, Theriault RL, Kuerer HM (2004) Physician recommendations regarding tamoxifen and patient utilization of tamoxifen after surgery for ductal carcinoma in situ. Cancer 100:942–949
6. Greendale GA, Sowers M (1997) The menopause transition. Endocrinol Metab Clin North Am 26:261–277
7. Fallowfield L, Fleissig A, Edwards R, West A, Powles TJ, Howell A, Cuzick J (2001) Tamoxifen for the prevention of breast cancer: psychosocial impact on women participating in two randomized controlled trials. J Clin Oncol 19:1885–1892
8. Greendale GA, Reboussin BA, Hogan P, Barnabei VM, Shumaker S, Johnson S, Barrett-Connor E (1998) Symptom relief and side effects of postmenopausal hormones: results from the Post-

menopausal Estrogen/Progestin Interventions Trial. Obstet Gynecol 92:982–988
9. Day R, Ganz PA, Costantino JP, Cronin WM, Wickerham DL, Fisher B (1999) Health-related quality of life and tamoxifen in breast cancer prevention: a report from the National Surgical Adjuvant Breast and Bowel Project P-1 study. J Clin Oncol 17:2659–2669
10. Day R (2001) Quality of life and tamoxifen in a breast cancer prevention trial: a summary of findings from the NSABP P-1 study. Ann N Y Acad Sci 949:143–150
11. Ware JE Jr, Kosinski M, Keller SD (1994) SF-36 physical and mental summary scales: a user's manual. The Health Institute, New England Medical Center: Boston
12. Stewart AL, Ware JE (1992) Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Duke University Press, Durham
13. Radloff LS (1977) The CES-D Scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1:385–401
14. Harman HH (1976) Modern Factor Analysis. University of Chicago Press, Chicago. 3rd ed
15. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441
16. Spearman CH (1904) General intelligence objectively determined and measured. Am J Psychol 15:201–293
17. Cattell RB (1966) The scree test for the number of factors. Multivariate Behav Res 1:245–276
18. Kaiser HF (1958) The varimax criterion for analytic rotation in factor analysis. Psychometrika 23:187–200
19. Marsh HW, Balla JR, McDonald RP (1988) Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychol Bull 103:391–410
20. Hatcher L (1994) A Step-by-step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. SAS Institute, Inc., Cary, NC, SAS Institute
21. International resource center for health care assessment (1991) How to Score the SF-36 Health Status Survey. New England Medical Center, Boston
22. Land SR, Ganz PA (2003) Quality of life issues with endocrine chemoprevention. In: Morrow M, Jordan VC (eds). Managing Breast Cancer Risk. BC Decker, Hamilton, Ontario, Canada
23. Stanton AL, Bernaards CA, Ganz PA (2005) The BCPT symptom scales: a measure of physical symptoms for women diagnosed with or at risk for breast cancer. J Natl Cancer Inst 97:448–456
24. Alfano CM, McGregor BA, Kuniyuki A, Reeve BB, Bowen DJ, Baumgartner KB, Bernstein L, Ballard-Barbash R, Malone KE, Ganz PA, McTiernan A (2006) Psychometric properties of a tool for measuring hormone-related symptoms in breast cancer survivors. Psychooncology 15:985–1000
25. Land SR, Wickerham DL, Costantino JP, Ritter MW, Vogel VG, Lee M, Pajon ER, Wade JL III, Dakhil S, Lockhart JB Jr, Wolmark N, Ganz PA (2006) Patient-reported symptoms and quality of life during treatment with tamoxifen or raloxifene for breast cancer prevention: the NSABP Study of Tamoxifen and Raloxifene (STAR) P-2 trial. JAMA 295:2742–2751