STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3865–3884 (DOI: 10.1002/sim.1358)

Multicollinearity in prognostic factor analyses using the EORTC QLQ-C30: identification and impact on model selection[‡]

Kristel Van Steen[1,*,†], Desmond Curran[2], Jocelyn Kramer[3], Geert Molenberghs[1], Ann Van Vreckem[4], Andrew Bottomley[3] and Richard Sylvester[3]

[1] *Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus, B 3590 Diepenbeek, Belgium*
[2] *Icon Clinical Research, South County Business Park, Leopardstown, Dublin 18, Ireland*
[3] *EORTC Data Center, Avenue E. Mounier, 83, Bte 1, B 1200 Brussels, Belgium*
[4] *Oncology Division Europe BMS Waterloo, Park Dreve Richelle 161 Building j, B 1410 Waterloo, Belgium*

SUMMARY

Clinical and quality of life (QL) variables from an EORTC clinical trial of first line chemotherapy in advanced breast cancer were used in a prognostic factor analysis of survival and response to chemotherapy. For response, different final multivariate models were obtained from forward and backward selection methods, suggesting a disconcerting instability. Quality of life was measured using the EORTC QLQ-C30 questionnaire completed by patients. Subscales on the questionnaire are known to be highly correlated, and therefore it was hypothesized that multicollinearity contributed to model instability. A correlation matrix indicated that global QL was highly correlated with 7 out of 11 variables. In a first attempt to explore multicollinearity, we used global QL as dependent variable in a regression model with other QL subscales as predictors. Afterwards, standard diagnostic tests for multicollinearity were performed. An exploratory principal components analysis and factor analysis of the QL subscales identified at most three important components and indicated that inclusion of global QL made minimal difference to the loadings on each component, suggesting that it is redundant in the model. In a second approach, we advocate a bootstrap technique to assess the stability of the models. Based on these analyses and since global QL exacerbates problems of multicollinearity, we therefore recommend that global QL be excluded from prognostic factor analyses using the QLQ-C30. The prognostic factor analysis was rerun without global QL in the model, and selected the same significant prognostic factors as before. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: multicollinearity; prognostic factor analysis; quality of life data; bootstrap

[*] Correspondence to: Kristel Van Steen, Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus, B 3590, Diepenbeek, Belgium.
[†] E-mail: kristel.vansteen@luc.ac.be
[‡] Presented at the International Society for Clinical Biostatistics, Twenty-first International Meeting, Trento, Italy, September 2000.

Received January 2001
Accepted June 2002

1. INTRODUCTION

Prognostic factor analyses are used in oncology to identify variables that are 'independent' predictors of outcome, and that therefore should be used to stratify patients in the design and analysis of clinical trials, to assist in the interpretation of the data generated by such trials, to aid the clinical management of individual patients and to increase efficiency. Since the advent of methods for measuring health-related quality of life (QL), several studies have been published in which QL variables derived from visual analogue scales and patient-completed questionnaires have been identified as important prognostic factors in addition to clinical factors [1–14]. This finding has considerable importance, particularly in advanced disease where treatment is generally palliative and the aim is to optimize QL.

We therefore undertook prognostic factor analyses for response and survival using both clinical and QL variables from an EORTC study (10923) of single agent paclitaxel versus doxorubicin as first line therapy in advanced breast cancer. The two drugs were expected to yield no substantial differences in response and survival, and therefore QL was considered an important endpoint.

Details of the prognostic factor analyses from this trial in advanced breast cancer have already been reported in Kramer *et al.* [15]. However, our findings and our knowledge of the structure of QL questionnaires alerted us to the possibility that *harmful multicollinearity* was present: a situation that arises in multiple regression when two or more predictor variables are so highly correlated that non-sensical results are obtained, such that the analysis yields parameter estimates of incorrect magnitude, incorrect sign etc. These difficulties can lead to incorrect model selection or can make it impossible to determine the direction or magnitude of effects of the predictor variable on the response variable, even with a correct model specification (Cramer [16], Slinker and Glantz [17] and Sithisarankul [18]). It is self-evident that these issues need to be addressed before drawing conclusions from a prognostic factor analysis.

Although intercorrelation between variables measured using QL questionnaires has been observed before (Coates [12] and Aaronson [19]), neither proper identification tools nor the impact of multicollinearity on prognostic factor analyses using QL variables have been explored. We consider this a potentially important and previously unreported context in which multicollinearity can occur, which will be dealt with in this paper.

In Sections 2 and 3 we will respectively introduce the data and describe the models that will be the subject of further investigation. In Section 4, we will identify harmful multicollinearity using standard diagnostic techniques. Fully aware of the limited use of such techniques in the light of ordinal categorical data (such as produced by the QLQ-C30), we use a bootstrap technique in Section 5 and investigate model (in)stability. In Section 6, we comment upon ways to avoid or circumvent the problems of multicollinearity. Finally, we conclude with practical recommendations in Section 7.

2. THE DATA

The data were taken from an EORTC study (10923) of single agent paclitaxel versus doxorubicin as first line therapy in advanced breast cancer. Information about eligibility criteria can be found in Kramer *et al.* [15]. Characteristics of all eligible patients are reported in

Table I. The EORTC QLQ-C30 vl.0 is a 30-item questionnaire that consists of five function scales, three symptom scales, six single item scales and a global health status/quality of life (QL) scale.

Scales in the EORTC QLQ-C30 Description	Code
Function scales	
physical functioning	PF
role functioning	RF
emotional functioning	EF
cognitive functioning	CF
social functioning	SF
Symptom scales	
fatigue	FA
nausea/vomiting	NV
pain	PA
Single item scales	
dyspnoea	DY
sleep disturbance	SL
appetite loss	AP
constipation	CO
diarrhoea	DI
financial impact of disease/treatment	FI
Global health status indicator	
quality of life	QL

Paridaens *et al.* [20]. Since the two drugs were expected to yield no substantial differences in response and survival, QL was considered an important endpoint.

Of 249 available eligible patients only 187 completed baseline QL evaluations (compliance rate of 64 per cent). Baseline QL was assessed using the EORTC QLQ-C30 vl.0 questionnaire [19]. This is a 30-item questionnaire that consists of five function scales, three symptom scales, six single item scales and a global health status/quality of life (QL) scale (Table I). Items are scored and then scaled to values ranging from 0–100, with higher values representing better function and global QL or more severe symptoms [21]. Since the number of possible categories for a scale range from three (RF), over four (all single item scales) to more than four, up to 13 (for example, QL) [21], we may treat all scale scores as if continuous, hereby using caution in interpreting analyses results.

In order to limit the number of variables under consideration, three questionnaire items (CO, DI and FI) were not included in our prognostic factor analyses: constipation and diarrhoea were present in only small proportions of patients at baseline and therefore the power to detect an effect of these variables was limited; financial impact was considered difficult to interpret as a prognostic factor in a multinational clinical trial.

3. THE MODELS

Similar to the prognostic factor analyses carried out in Kramer *et al.* [17], we used the Cox proportional hazards model with stratification for treatment arm for both univariate and multivariate analyses of survival. The logistic regression model with treatment arm included in the model was applied for both univariate and multivariate analyses of response. Scale scores were dichotomized at the median (CF was dichotomized at a score of 70, see Kramer *et al.* [15]) and variables were used in binary form.

To build the final multivariate models, clinical variables were identified that were significant predictors of survival and response. More specifically, the following patient and disease variables were included: WHO performance status [22]; age; disease-free interval (DFI, the time between diagnosis of breast cancer and diagnosis of advanced disease) and dominant anatomical site of disease according to UICC criteria [23]. Only a subgroup of 177 patients was used for the multivariate analyses, since those were the patients with complete information on all QLQ-C30 variables. The significant clinical variables selected by the final multivariate models were fixed into models to which QLQ-C30 variables were then added using forward selection (selection entry criterion $= 0.01$) and backward elimination (selection stay criterion $= 0.01$). In adding the QL variables, the same final models were obtained for survival irrespective of the selection method used. In the models for response, different final models were obtained with the two selection methods.

In practice, the entry criterion is often less restrictive than the stay criterion. Hence, we fixed the criterion for variables to enter into the model at 0.05 instead of 0.01 and noticed that the kind of instability in the models for response seemed to have disappeared.

4. IDENTIFICATION OF PROBLEMATIC MULTICOLLINEARITY

4.1. Introduction

Slinker and Glantz [17] observed that the usual source of multicollinearity in physiological data is inability to manipulate all predictor variables independently. In the case of questionnaires measuring QL, multicollinearity is inherent in the questionnaire itself, since all variables are designed to measure putative components of QL. Moreover, the greater the number of items, the higher the risk of multicollinearity.

A number of diagnostic tests can be used to determine the degree to which multicollinearity might be a problem in our data set when all QL variables are included in the analysis [24]. One may be warned by (i) considering correlation matrices, (ii) by regressing each explanatory variable on other explanatory variables, (iii) by investigating the (in)stability of regression models with QL variables to predict the global QL, or (iv) by comparing principal components analyses or factor analyses with and without the variable global QL included etc. Standard diagnostic techniques include calculating variance inflation factors (VIF) and considering tolerance and condition indices. While allowing to let the QL scaled scores range from 0 to 100, we take a closer look at these techniques as a first approach to testing for multicollinearity and highlight the potential of a second approach in which a bootstrap procedure is implemented to investigate the stability of models in prognostic factor analyses (Section 5).

4.2. Alarming signals

4.2.1. Pairwise correlations between QL variables. In our data set of 177 women with advanced breast cancer and complete QLQ-C30 data available at baseline, Spearman's correlation

Table II. Correlation matrix for EORTC QLQ-C30 function variables, symptoms and global quality of life (absolute values of Spearman correlation coefficients, $n = 177$).

	PF	RF	EF	CF	SF	FA	NV	PA	DY	SL	AP
RF	0.706										
EF	0.151	0.196									
CF	0.268	0.236	0.361								
SF	0.496	0.549	0.322	0.408							
FA	0.657	0.573	0.359	0.464	0.584						
NV	0.353	0.339	0.110	0.151	0.308	0.421					
PA	0.495	0.457	0.189	0.294	0.453	0.511	0.315				
DY	0.375	0.303	0.288	0.246	0.345	0.489	0.294	0.142			
SL	0.044	0.201	0.301	0.116	0.186	0.203	0.117	0.201	0.037		
AP	0.374	0.344	0.208	0.356	0.361	0.505	0.489	0.318	0.324	0.104	
QL	0.578	0.523	0.332	0.319	0.674	0.709	0.363	0.535	0.422	0.253	0.413

r was used to identify the degree of correlation between QL variables (Table II). The highest correlations were observed between PF and RF, and FA and global QL (Spearman's correlation $|r| = 0.71$). One-third of the correlations in the matrix have $|r| > 0.4$, with FA, global QL and SF having the greatest number of strong correlations. Although these intercorrelations do not reach the extreme levels of $|r| > 0.9$ reported in studies of multicollinearity in biological and physicochemical data [17, 18, 25, 26], it has been observed that pairwise correlations in absolute value greater than 0.70–0.80 between two predictor variables signify a harmful multicollinearity between the two variables [17]. Moreover, important multicollinearities among three or more predictor variables can exist even though pairwise correlations are small. We hypothesize that small correlation coefficients do not exclude the possibility that harmful multicollinearity has an influence on model selection involving variables derived from patient-completed QL questionnaires.

4.2.2. Regression analyses. If the assumptions underlying multiple linear regression are met, the usual initial indications of the presence of harmful multicollinearity are unexpected magnitudes or signs of parameter estimates [17]. This is shown particularly well if slight changes in model structure result in considerable changes in the magnitude or sign of parameter estimates. Since the QLQ-C30 variable global QL has been shown in published studies to be a significant prognostic factor [12, 13], and since intuitively it is the variable expected to be most affected by multicollinearity, it was chosen as an illustration of the problem. Table III shows the sensitivity of the parameter estimates to slight changes in predictive models for global QL, particularly for those variables that are highly correlated with FA. Of interest is the sign change of the parameter estimate for CF from a positive sign in the univariate model to a negative sign in the multivariate model.

Next, global QL was entered as dependent variable in a stepwise selection model (criterion for variables to enter into the model set to 0.05, criterion for variables to stay in the model = 0.01) for multiple regression in which all other QLQ-C30 variables (except CO, DI and FI) were entered as predictors. The replication stability of the final model predicting global QL was investigated using a bootstrap resampling technique [27, 28]. A total of 1000 samples of size $n = 177$ each were generated by randomly selecting patients with replacement. The frequency of inclusion of the component variables in the resulting models using stepwise

Table III. Multiple linear regression model for predicting global QL. The scaled variables are allowed to range from 0–100. Column 1, QL covariate; column 2, parameter estimate of QL variable in predicting global QL; column 3, parameter estimate of QL variable in a multiple regression model with the QL variable and FA as only covariates; column 4, parameter estimate of QL variable in a multiple regression model with the QL variable, FA and SF as covariates; column 5, similar to columns 3 and 4, now using the additional covariates FA, SF and PA.

Variable in model	Parameter estimates (p-values) Univariate		+FA		+FA+SF		+FA+SF+PA	
PF	0.49	(<0.001)	0.19	(0.001)	0.13	(0.018)	0.10	(0.064)
RF	0.34	(<0.001)	0.13	(0.002)	0.06	(0.153)	0.04	(0.342)
EF	0.30	(<0.001)	0.09	(0.110)	0.04	(0.417)	0.04	(0.390)
CF	0.33	(<0.001)	−0.02	(0.811)	−0.08	(0.198)	−0.08	(0.151)
SF	0.53	(<0.001)	0.29	(<0.001)	—	—	—	—
FA	−0.58	(<0.001)	—	—	—	—	—	—
NV	−0.52	(<0.001)	−0.18	(0.025)	−0.12	(0.102)	−0.11	(0.138)
PA	−0.40	(<0.001)	−0.19	(<0.001)	−0.14	(0.002)	—	—
DY	−0.36	(<0.001)	−0.09	(0.065)	−0.07	(0.128)	−0.09	(0.044)
SL	−0.19	(0.001)	−0.09	(0.031)	−0.09	(0.028)	−0.07	(0.073)
AP	−0.30	(<0.001)	−0.05	(0.252)	−0.02	(0.707)	−0.01	(0.825)

selection can be considered to be indicative of the importance of the variables other than global QL entered as predictors. In line with the correlation observed in Table II, both SF (94.4 per cent) and FA (97.3 per cent) show the highest occurrence frequencies in predictive models for global QL. The QL variable PA gives an inclusion frequency of 59.8 per cent. Therefore, inclusion or exclusion of PA in the model seems to be fairly unpredictable. The observed manifestations of sensitivity and sign change confirmed the presence of potentially harmful multicollinearity.

4.2.3. Principal component analysis. To assess the effects of including global QL in the variable pool of prognostic factor analyses, we first considered all QL variables (except CO, DI and FI, as usual) in a principal components analysis, based on the variance-covariance matrix. The aim of principal components analysis is to reduce the dimensionality of the data by creating new orthogonal (independent) variables from linear combinations of the original variables and selecting only a few of these. Since the newly defined variables are uncorrelated, the problem of multicollinearity is circumvented. The method can be legitimately used in a non-normal setting, provided no inferences are being made.

The more desirable situation in a principal component analysis is the one where one eigenvalue is ‘large’ with the remaining ones very ‘small’. Since it is hard to judge what is large and what is small based on the absolute numbers, the proportion of explained variability may be an option. However, there is no fixed gold standard and the meaning of a ‘acceptable’ percentage highly depends on the application field. For the data under study, the first six principal components account for 84 per cent of the total variability.

Alternatively, a Monte Carlo study performed by Guadagnoli and Velicer [29] highlights the importance of the absolute magnitude of the loadings (loadings pertain to the Pearson correlation coefficients between the variables and the principal component, that is, the particular linear combination of the variables) and the absolute sample size. Bearing in mind

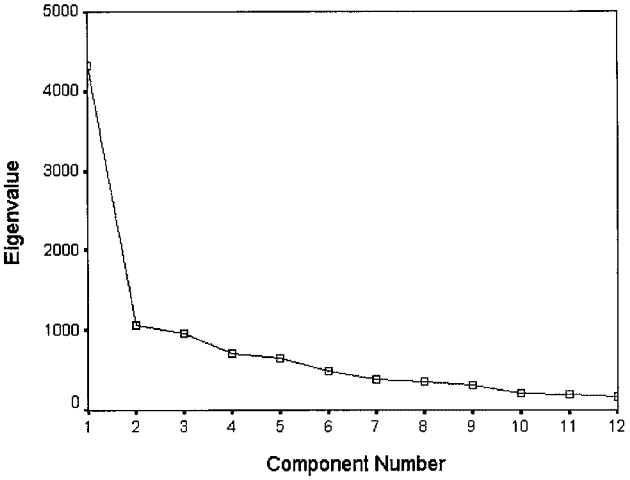

Figure 1. Scree plot.

that a component with only a few important loadings (in terms of absolute value) defeats the purpose, and since the principal component analysis is used in an exploratory fashion, we may restrict attention to the first three or even two components. This choice can also be justified by looking at the so-called scree plot (Figure 1). It shows that there are at most two eigenvalues (corresponding with the first two principal components) before the scree plot starts to level off.

The question remains which loadings should be used for interpretation, or in other words, which variables make up the selected components. One approach is to test each loading for significance at alpha $=0.01$ (two-tailed test). Setting the alpha level for each separate test more stringently is prompted by the use of multiple testing. Stevens [30] provides a table with critical values for a correlation coefficient for alpha $=0.01$ and a two-tailed test. A good approximation for the actual sample size of $n=177$ is found by interpolating between $n=140$ and $n=180$ in the latter table, resulting in a critical value of $c=0.002$. A rough check for assessing whether or not a loading is statistically significant is then given by doubling this critical value. Hence, using this technique, only loadings of 0.388 in absolute value will be declared statistically significant.

It appears that the first principal component has significant loadings on 11 of the 12 variables, which makes it difficult to interpret. It should be noted though that PF, RF, QL and FA seem to dominate this component. The second principal component seems to be dominated by SL (the latter having a loading of 0.889, which largely exceeds the loadings for the other variables).

If we exclude QL from the variable pool, we obtain similar results. Remarkably, restricting attention again to the first two principal components, the loadings hardly change (results not shown). Hence, there seems to be little use in adding global QL to the list of remaining variables (PF, RF, EF, CF, SF, FA, NV, PA, DY, SL and AP) in order to increase the explanatory capacity of the model. This is entirely in line with expectations, since QL was designed to be a global score. Moreover, inclusion of QL might cause unnecessary

Table IV. Factor analysis results: extraction method = principal components analysis; rotation method = varimax with Kaiser normalization.

	Rotated factor pattern					
	QL included			QL not included		
	FACTOR1	FACTOR2	FACTOR3	FACTOR1	FACTOR2	FACTOR3
RF	0.89572	−0.16034	−0.04207	0.89920	−0.17875	−0.05891
PF	0.81611	−0.26985	0.13596	0.81372	−0.28080	0.12661
QL	0.61169	−0.46462	−0.21463			
SF	0.59799	−0.42766	−0.12395	0.58645	−0.43204	−0.12606
PA	−0.66003	0.17197	0.31908	−0.65267	0.17994	0.32405
AP	−0.17310	0.78221	0.14116	−0.16685	0.79100	0.14947
DY	−0.24535	0.70110	−0.17883	−0.23605	0.70214	−0.17988
FA	−0.59170	0.60447	0.12974	−0.57785	0.60898	0.13184
NV	−0.23270	0.50349	0.05611	−0.22482	0.50682	0.05892
CF	0.21261	−0.50169	−0.14455	0.20935	−0.50918	−0.14960
SL	−0.11507	−0.01424	0.92100	−0.10485	−0.01080	0.92325
EF	0.05324	−0.44412	−0.47217	0.04266	−0.44604	−0.47186

multicollinearity problems. Especially RF, EF and FA seem to be forming a group together with QL. Also note that RF was one of the parameters that gives a counterintuitive sign in the normal regression analysis of QL on the remaining variables.

4.2.4. Factor analysis. When performing an exploratory factor analysis, we used the mineigen criterion to specify the smallest eigenvalue for which a factor is retained. In order to determine the appropriate number of factors to include in the model, we considered the point at which including additional factors did not substantially increase the variance explained by the common factors. The appropriate number of factors to fit appeared to be three. The estimated factor loadings obtained after an orthogonal varimax rotation with Kaiser normalization are displayed in Table IV. The data show that the function scales PF, RF, SF and QL are contrasted with PA in one factor, AP, DY, FA, NV are contrasted with CF in a second factor, and SL primarily loads on the third factor. Relatively large loadings are obtained for EF on more than one factor, causing some ambiguity in the interpretation of the factors. Applying a Harris–Kaiser oblique rotation to allow for correlation between the factors, and forcing a large loading on only one factor for each variable, gives similar interpretations of the factors as before. The variable EF now loads high on two of the three factors, making it less clear whether AP, DY, NV, FA should be contrasted against EF and CF together or only against CF.

Conclusions were unchanged when global QL was excluded from the analyses (Table IV).

4.3. Standard diagnostic tool kit

4.3.1. Variance inflation factors and tolerances. Slinker and Glantz [17] report on a variety of diagnostic checks that can be used to evaluate the severity, number and structure of multivariable multicollinearities. For instance, the severity of multicollinearity may be suggested by the magnitude of the variance inflation factors (VIF) for the regression of each predictor

Table V. Diagnostic tests for multicollinearity between predictor variable.

Variable	Tolerance	Variance inflation
PF	0.3730	2.6813
RF	0.4036	2.4776
EF	0.7494	1.3344
CF	0.7087	1.4111
SF	0.5469	1.8286
FA	0.3537	2.8274
NV	0.6779	1.4752
PA	0.6139	1.6289
DY	0.6631	1.5082
SL	0.7986	1.2522
AP	0.5985	1.6709

on all remaining predictor variables. The variance inflation factor measures the inflation in the variance of the parameter estimate due to collinearity between the explanatory variable and other variables. For the jth dependent variable, the variance inflation factor is defined as

$$\mathrm{VIF}_j = \frac{1}{(1 - R_j^2)}$$

where R_j^2 is the coefficient of determination for the regression of the jth independent variable on all other independent variables. A $\mathrm{VIF} > 10$ is generally considered indicative of harmful multicollinearity, but this value is arbitrary, and relatively small VIFs may still be unstable (Myers [31]). Some authors (for example, Slinker and Glantz [17]) accept a $\mathrm{VIF} > 4$ (corresponding to an auxiliary regression R^2 of 0.75). These values are based on the situation of normally distributed data, but data derived from patient-completed questionnaires are categorical and often not normally distributed, and the problem might exist in this special case even when the usual criteria are not fulfilled. The VIFs we observed when regressing each predictor variable on all the other predictor variables are shown in Table V. The variables FA, PF and RF produced the highest VIFs, although none was $\geqslant 4$. We considered that these three variables might nevertheless produce harmful multicollinearity in the present context. The tolerance factor is another statistic that measures the strength of interrelationships among the explanatory variables in the model. Tolerances close to zero indicate a strong linear association or collinearity among the explanatory variables. From Table V it can be seen that FA, PF and RF produced the lowest tolerance factors. These results are consistent with the ones obtained by using VIFs before.

4.3.2. Condition indices. Analysis of the structure of relationships among a set of variables, and hence diagnosis of harmful multicollinearity, may be pursued by examination of the structure of $X^{\mathrm{T}}X$ (for example, eigenvalues) where X is the design matrix. The design matrix contains all dependent variables included in the analysis. The literature suggests that eigenvalues $\leqslant 0.01$ indicate a serious problem [17]. The square root of the ratio of the largest to each individual eigenvalue is known as the condition index, and allows one to assess the relative magnitudes of the eigenvalues. The largest condition index (square root of the ratio of the largest to the smallest eigenvalue) is the condition number, and when this number is large the data are said to be ill-conditioned. Criteria for a condition number to signify serious multicollinearity are arbitrary, with values of 30 to 100 often quoted [31]. However, it has also been suggested that a condition number larger than 10 probably indicates harmful multicollinearity [17]. Clearly, insight into the particular problem at hand should guide the diagnosis rather than strict adherence to a rule. In addition, for each variable, the proportion of variance of its estimate accounted for by each component can be evaluated. A problem is indicated when an eigenvalue with a high condition index is associated with high values for the variance proportion indicating near linear dependence between those variables. Belsley *et al.* [24] suggest the following approach: identify eigenvalues having condition numbers >30, then variables with variance proportions >0.5 for each of those eigenvalues are considered to be involved in the near linear dependency that produces the large condition numbers.

The structure of relationships among the variables can be analysed by using raw (non-centred) variables or by scaling and centring them. Practically, this is achieved by respectively including or not including an intercept term in the model statement. Hence, in the case of QL variables that are derived from ordered categorical variables, already scaled, it does not make sense to exclude the intercept while basing the analysis on scaled and centred variables. Table VI shows selected collinearity diagnostics obtained when the intercept is included and global QL is the dependent variable. The condition number is 25.81. In the row corresponding to this value (Table VI) relatively large variance proportions are observed for FA, CF and PF, suggesting that these variables are involved in a near linear dependency. The latter statement should not be overemphasized, since large condition numbers are likely to be produced if the origin lies outside the range of the data. Table VI only lists condition indices >10 and suggests near linear dependency between RF and EF, between PF and SF, but also between PF, RF and CF.

4.4. Comments

In summary, the diagnostic techniques investigating pairwise correlations, sensitivity and sign change suggest the presence of harmful multicollinearity in the data set, and this is supported by investigating variance inflation factors, the condition number and variance proportions. In Section 5, we will use a bootstrap resampling technique as a promising diagnostic tool in the present context. QL variables will be used in dichotomized form, since it is common practice in prognostic factor analyses to do so. Moreover, the technique allows investigating the impact of multicollinearity on prognostic factor analyses.

5. INFLUENCE OF MULTICOLLINEARITY ON MODEL STABILITY: THE BOOTSTRAP RESAMPLING TECHNIQUE

The problem of multicollinearity demonstrated in the prediction of global QL from other QLQ-C30 variables using multiple linear regression, can also arise in Cox and logistic regression. It may be that neither of two correlated factors will be identified as statistically significant although both have an influence on the outcome [27]. Therefore, investigation of the stability of the chosen regression model is important. Stability is defined by the replication stability for the choice of variables included in the model and the predictive ability of the model itself [28].

Table VI. Multicollinearity diagnostics, using the raw (not centred) variables. Eight eigenvalues are not displayed. They have associated condition indices ranging from 1.00000 to 7.79876.

Condition			Variance proportion										
Eigenvalue	Index	Intercept	PF	RF	EF	CF	SF	FA	NV	PA	DY	SL	AP
0.06563	11.12328	0.0046	0.0127	0.2283	0.6374	0.1600	0.1575	0.0007	0.0024	0.0022	0.0130	0.1420	0.0031
0.05718	11.91682	0.0051	0.4048	0.1280	0.0025	0.0367	0.5154	0.0040	0.0097	0.0111	0.0021	0.0330	0.0018
0.04611	13.26969	0.0000	0.2924	0.3510	0.0306	0.4653	0.2391	0.0012	0.0019	0.0075	0.0000	0.0154	0.0003
0.01219	25.81170	0.9888	0.2503	0.0005	0.0920	0.3018	0.0781	0.3066	0.0028	0.0867	0.0320	0.0027	0.0161

Table VII. Frequency of inclusion of QL variables in multivariate models of survival and response based on 1000 simulated data sets containing the clinical variables disease-free interval and dominant site of disease. The minus sign indicates that in the majority of the cases, the variable entered the model with a minus sign. In turn, a percentage without a sign indicates that in the majority of the cases, the variable entered with a positive sign. The value is italicized if the sign is consistent throughout all selections.

QLQ-C30 variable	Frequency of inclusion in multivariate model (%)			
	Survival		Response	
	Forward	Backward	Forward	Backward
PF	9.8	4.7	−9.6	−7.2
RF	−6.6	−2.8	7.3	3.5
EF	−11.5	−4.4	*72.8*	*49.9*
CF	*−36.5*	*−21.0*	−10.8	−4.8
SF	−14.7	−10.6	−11.1	−5.7
QL	*33.3*	*31.2*	−17.0	−9.6
DY	17.0	*8.4*	*−77.7*	*−59.9*
FA	23.8	15.3	*−50.4*	*−39.1*
NV	17.3	10.8	*−34.9*	*−20.7*
PA	*50.7*	*39.9*	−6.5	−3.7
SL	−13.2	*−5.8*	−8.9	−2.2
AP	16.8	10.3	−15.7	*−9.2*

The replication stability of the final multivariate models predicting survival and response was checked using a bootstrap resampling technique [27, 28]. A total of 1000 samples each of the same size as the complete baseline data set ($n = 177$) were generated by randomly selecting patients and replacing them before selecting the next patient. The frequency of inclusion of the component variables in models built on these simulated data sets using both forward selection and backward elimination gives an indication of the prognostic importance of the variable. The results are shown in Table VII for a significance level for inclusion set at $p = 0.05$ and a significance level for variables staying in the model set at $p = 0.01$.

For most QLQ-C30 variables, the frequency of inclusion in the multivariate model differed by < 10 per cent according to whether forward or backward selection was used. The exceptions for the survival model were CF and PA and for the response model EF (more than 20 per cent difference), DY, FA and NV. Sauerbrei and Schumacher [27] suggested a cutpoint of 30 per cent for a low frequency of inclusion in the model, and 60–70 per cent as a high cutpoint.

For survival, the frequency of inclusion of QLQ-C30 variables in the model is relatively low, with PA (the most frequent) included in fewer than 55 per cent and global QL included in fewer than 35 per cent. For response, the inclusion frequency for DY is more than 60 per cent. Two other variables, EF and FA, have inclusion frequencies ranging between about 50 –70 per cent and 40–50 per cent, respectively, which account for the instability (according to the selection method used) of the model for response.

In addition, we draw attention to the behaviour of PF in both the survival and response model. Despite its low inclusion frequency, we observe that once PF is selected in the response model, it most frequently enters into the model with a minus sign and it enters most often

Table VIII. Frequency of inclusion of QL variables (global QL excluded) in multivariate models of survival and response based on 1000 simulated data sets containing the clinical variables disease-free interval and dominant site of disease. The minus sign indicates that in the majority of the cases, the variable entered the model with a minus sign. In turn, a percentage without a sign indicates that in the majority of the cases, the variable entered with a positive sign. The value is italicized if the sign is consistent throughout all selections.

QLQ-C30 variable	Frequency of inclusion in multivariate model (%)			
	Survival		Response	
	Forward	Backward	Forward	Backward
PF	14.4	8.0	−11.2	−7.5
RF	−7.3	−3.0	6.6	3.2
EF	−10.0	−4.0	*71.9*	*49.4*
CF	*−38.7*	*−21.3*	−11.1	−4.4
SF	−9.5	−4.2	−13.1	−6.3
DY	18.9	*9.9*	*−79.1*	*−61.6*
FA	36.4	*28.8*	*−54.5*	*−42.4*
NV	17.5	9.1	*−33.2*	*−21.3*
PA	*54.0*	*44.6*	−6.1	−3.1
SL	−12.8	−4.7	−10.3	−1.7
AP	14.5	10.4	−15.7	*−9.8*

with a positive sign in the survival models. Intuitively one would expect high scores on the function variables and global QL to be associated with better survival and response, and high scores for symptoms to be associated with a poor outcome. Hence, apart from rather low inclusion VII frequencies, the figures in Table VII give further evidence of instability.

We repeated the bootstrap analyses with global QL deleted from the predictor variables, to see whether this improves model stability. The results are presented in Table VIII and show similar differences in inclusion frequencies as before. For the survival and response models, again, respectively, CF and PA, and EF, DY, FA and NV give rise to the largest differences in inclusion frequencies. It therefore seems that global QL is redundant as a predictor variable. It also appears to interact with FA (note the higher inclusion probabilities for FA in the survival model, compared to the ones obtained in Table VII), which is consistent with the finding that they are the most highly correlated variables in Table II.

Finally, Tables IX and X give an impression of the size of the parameter estimates for the selected variables, based on the 1000 simulated data sets. For every variable in the variable pool (global QL excluded), we averaged the (maximal 1000) obtained parameter estimates, together with their corresponding standard errors. The Cox multivariate model (Table IX, column 2), stratifying by treatment arm and adding QLQ-C30 variables to the fixed clinical variables multiple sites of visceral disease (DS) and DFI $\leqslant 2$ years, indicated that poor survival was associated with DS (estimated $= 0.660$, p-value $= 0.003$), DFI $\leqslant 2$ years (estimated $= 0.385$, p-value $= 0.026$) and pain PA (estimated $= 0.505$, p-value $= 0.003$). The backward elimination method produced the same results. On the basis of the simulation study, an average PA parameter estimate of 0.598 (p-value $= 0.001$) was obtained with the forward selection method (Table IX, column 3), and an even larger average PA parameter estimate of

Table IX. The Cox multivariate model, stratifying by treatment arm and retaining the clinical variables disease-free interval (DFI) and dominant site of disease (DS). Column 1, variable pool (global QL is excluded); column 2, parameter estimates and corresponding standard errors via forward selection (identical results are obtained via backward selection); column 3, mean parameter estimates and mean standard errors for multivariate models of survival based on 1000 simulated data sets and forward selection; column 4, similar to column 3, now with backward selection.

Survival variable	Cox multivariate model		
		Forward	Backward
	Estimate (SE)	Estimate (SE)	Estimate (SE)
DFI	0.385 (0.173)	0.473 (0.183)	0.449 (0.180)
DS	0.660 (0.225)	0.609 (0.237)	0.630 (0.235)
PF		0.519 (0.195)	0.592 (0.189)
RF		−0.201 (0.215)	−0.260 (0.216)
EF		−0.414 (0.189)	−0.528 (0.190)
CF		−0.612 (0.210)	−0.678 (0.211)
SF		−0.473 (0.213)	−0.478 (0.207)
DY		0.511 (0.190)	0.583 (0.187)
FA		0.668 (0.199)	0.709 (0.184)
NV		0.541 (0.202)	0.665 (0.195)
PA	0.505 (0.169)	0.598 (0.186)	0.644 (0.179)
SL		−0.446 (0.186)	−0.568 (0.188)
AP		0.523 (0.196)	0.594 (0.192)

0.644 (p-value<0.001) was obtained relying on backward elimination (Table IX, column 4). Note that PA was selected in 54.0 per cent of the cases using forward selection and in 44.6 per cent of the cases using backward selection (Table VIII). Consequently, the average PA parameter estimates via forward or backward selection are based on averaging, respectively, 540 or 446 parameter estimates.

The final logistic regression model for response (Table X, column 2) with treatment arm (ARM) included in the model in addition to QLQ-C30 variables predicted a poor response with multiple sites of visceral disease (estimated$=-1.420$, p-value$=0.028$), DFI$\leqslant 2$ years (estimated$=-1.169$, p-value$=0.004$), bad emotional functioning EF (estimated$=1.093$, p-value$=0.008$), dyspnoea DY (estimated$=-1.262$, p-value$=0.003$) and fatigue FA (estimated$=-1.268$, p-value$=0.004$). The same results were produced using the backward elimination method. Although the signs are consistent throughout the analyses (Table X, within rows), the magnitude of the estimates may vary substantially. For example, focusing on FA, we observe an average parameter estimate of -1.555 (p-value$=0.002$) in the forward procedure and -1.775 (p-value <0.001) using the backward elimination method.

6. AVOIDING THE PROBLEM OF MULTICOLLINEARITY

Because of the interdependent nature of variables derived from the QLQ-C30, it is not possible to overcome the problem of multicollinearity when undertaking a prognostic factor analysis since most variables need to be entered into the model. One technique that has been proposed

Table X. The logistic regression model for response, with treatment arm (ARM) included in the model, and retaining the clinical variables disease-free interval (DFI) and dominant site of disease (DS). Column 1, variable pool (global QL is excluded); column 2, parameter estimates and corresponding standard errors via forward selection (identical results are obtained via backward selection); column 3, mean parameter estimates and mean standard errors for multivariate models of response based on 1000 simulated data sets and forward selection; column 4, similar to column 3, now with backward selection.

Variable	Logistic regression model		
		Forward	Backward
	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept	−0.623 (0.652)	−0.572 (0.705)	−0.676 (0.679)
Arm	0.568 (0.372)	0.641 (0.401)	0.638 (0.392)
DFI	−1.169 (0.409)	−1.265 (0.441)	−1.207 (0.431)
DS	−1.420 (0.645)	−1.828 (4.423)	−1.774 (4.647)
PF		−0.971 (0.459)	−1.406 (0.446)
RF		1.107 (0.517)	1.491 (0.564)
EF	1.093 (0.411)	1.495 (0.469)	1.579 (0.465)
CF		−1.197 (0.532)	−1.534 (0.535)
SF		−1.255 (0.503)	−1.545 (0.501)
DY	−1.262 (0.419)	−1.452 (0.458)	−1.605 (0.448)
FA	−1.268 (0.442)	−1.555 (0.503)	−1.775 (0.486)
NV		−1.473 (0.541)	−1.774 (0.544)
PA		−0.466 (0.446)	−0.467 (0.442)
SL		−0.662 (0.440)	−1.031 (0.443)
AP		−1.208 (0.458)	−1.470 (0.444)

for mitigating the problem is to collect additional data over a wider range or over more than one experimental condition [17]. The original validation study for the QLQ-C30 reported pretreatment Pearson correlations in over 300 patients with inoperable lung cancer [19]. In our study of patients with advanced breast cancer, many of the findings were similar, except for higher correlation coefficients for PF, RF and SF with most other variables. Coates *et al.* [12] reported Spearman correlations between the five function scales of the QLQ-C30 in over 700 patients with advanced malignancies. Our findings were remarkably similar, except for lower correlations between EF and all other variables in our study. Coates *et al.* [12] unfortunately did not report correlations for the symptom scales. It seems that there are some differences between our correlation findings and those of published studies, but that most pairwise correlations are similar. It is notable that the variables we identified as problematic are also those where differences from published data were observed.

One of the remedial measures to deal with multicollinearity (in a linear regression setting, for example, regressing global QL on all other QL variables of interest, scaled scores allowed to range from 0 to 100 again) is to perform ridge regression (Draper *et al.* [32]). It is an alternative to ordinary least squares and is one of several biased regression estimators that have been proposed. The ridge standardized regression estimators can be determined by introducing a biasing constant c in the usual least squares normal equations, resulting in

$$(r_{XX} + cI)b^R = r_{YX}$$

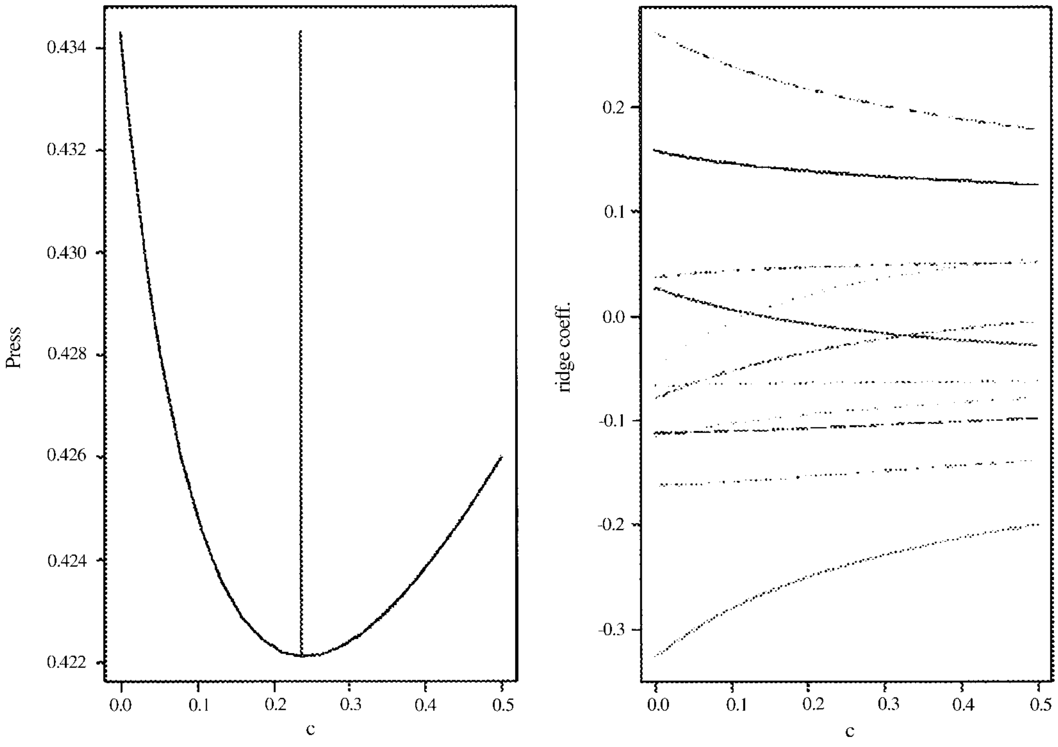

Figure 2. Left shows plot of the minimal press score versus the biasing constant c in an attempt to find the minimal press score. Right shows ridge trace plot of estimated standardized regression coefficients for different values of c, here containing amongst others 20 values ranging from 0 to 0.5 (equally spaced). The idea is to select the value of c for which the estimated standardized regression coefficients in the ridge trace first appear to be stable and the VIFs have become sufficiently small.

where r_{XX} is the correlation matrix of the X-variables, r_{YX} represents the vector of coefficients of simple correlation between Y and each X variable separately, where I denotes the identity matrix of appropriate order and in which b^R is the vector of standardized ridge regression coefficients. One method to select an appropriate biasing constant is based on the so-called ridge trace (Figure 2). A more objective choice of c can be made based on the press score.

In our data set, the minimal press score was obtained for $c = 0.237$. The result of applying ridge regression with the latter choice of biasing give estimated regression coefficients that are more consistent with expectations (Table XI). In particular, the improper sign on the estimates for RF and AP is eliminated (in contrast to an ordinary least squares regression).

Ridge estimators can also be used in logistic regression to improve the parameter estimates and to diminish error made by further predictions (Le Cessie *et al.* [33]).

Another way of mitigating the harmful effects of multicollinearity is to delete offending predictor variables from the regression model. This is done on the basis that one or more variables are redundant. Except perhaps for global QL, it is difficult to argue for redundancy in the present context, because despite overlap the variables are designed and assumed to measure different facets of health-related QL, and therefore all are thought to be important. Fayers

Table XI. Ridge regression coefficients were back-transformed (based on the biasing constant $c = 0.237$) to compare with ordinary least squares estimates.

Parameter	Ridge	Least squares
Intercept	53.848	59.378
PF	0.116	0.135
RF	0.017	−0.031
EF	0.046	0.036
CF	−0.030	−0.081
SF	0.179	0.230
FA	−0.200	−0.271
NV	−0.086	−0.090
PA	−0.111	−0.119
DY	−0.085	−0.089
SL	−0.070	−0.088
AP	−0.008	0.020

et al. [34, 35] attempted to identify ‘causal’ and ‘effect’ indicators for QL. They suggested that physical symptoms and side-effects can be thought of as causing changes in QL, whereas psychological variables (for example, EF) might rather reflect changes in QL and therefore be classified as effect indicators. These classifications are not entirely convincing, since it is possible to argue for CF, SF and RF being both cause and effect indicators. Although deletion of EF, CF, SF and RF on the basis of their being effect indicators would remove most of the sources of multicollinearity identified in the prediction of global QL, it would also remove important patient-assessed information and defeat the purpose of using the questionnaire for predicting outcome. When predicting survival or response to treatment, it would make more sense to delete global QL from the model since it is strongly correlated with 7 of the 11 variables studied. Also note that excluding QL from the bootstrap analyses in Section 3 did not relevantly change the results. Moreover, a finding that global QL is predictive of survival or response is not particularly helpful in the clinical context, since it is not clear how the variable could be manipulated to improve outcome. On the other hand, postulated causal indicators such as pain or dyspnoea are responsive to clinical intervention, and therefore potentially more useful prognostic factors. Of course, as before, care has to be taken with labelling an associative relation as causal.

Instead of deleting predictor variables, it has been suggested that substituting mathematically transformed variables for harmful ones or rescaling them may mitigate harmful multicollinearity [17]. Such procedures are difficult to justify in the case of numerical data derived from ordered categorical variables such as those from the QLQ-C30. However, we considered that it might be possible to reduce the number of variables by using principal components analysis to derive fewer variables as substitutes (Section 2). As mentioned before, the main idea behind any principal component analysis is to reduce the dimensionality of the data, so that manipulation of the data becomes easier. In other words, the aim is to derive only a few (new) variables that are still able to ‘explain the data’. These newly defined variables are uncorrelated, removing the intercorrelations present between the original variables. By removing the correlation structure, the information in the data is fully captured by the variance structure, the variability say. Hence, if most of the variability can be concentrated on one new variable, the other variables may be ignored, but does this solve our problem? Only at the cost of reducing the data obtained from the patient-completed questionnaire to something mathematically convenient but, in many cases, clinically difficult to interpret. Indeed, while there are several examples where such components lend themselves nicely to clinically meaningful interpretation, in a majority of cases interpretation is not straightforward or even impossible.

7. CONCLUDING REMARKS AND RECOMMENDATIONS

In a prognostic factor analysis incorporating clinical and QL variables from the QLQ-C30 in a clinical trial of patients with advanced breast cancer, we observed instability in the models predicting response and survival to chemotherapy. This led us to suspect that harmful multicollinearity between QL subscales was influencing model selection. We therefore undertook a thorough examination of the QL data used in the prognostic factor analysis, to diagnose the presence and degree of multicollinearity, and to determine the stability of the predictive models obtained.

Accepted criteria for diagnosing harmful multicollinearity (for example, variance inflation factor >10, condition number >30) may not apply in the case of QL variables. Two reasons for this are the inherent design of the questionnaire in which multicollinearity is implicit, and the structure of the variables themselves (ordered categorical, scaled from 0–100). Therefore, guided by published principles but recognizing that recommended criteria are arbitrary, we suggested how to identify and establish the impact of multicollinearity in a QL data set, in a standard way (for example, via variance inflation factors, tolerance and condition indices). Apart from these standard techniques, we used a bootstrap method to account for the fact that in clinical prognostic practice, QL variables are generally used in dichotomized form, and to obtain more insight into the stability of the models obtained. One of the major benefits of the approach is that it clearly shows which variables may or may not be important prognostic factors, by studying the inclusion frequencies. It is also an ideal instrument to put (unusual) directions of effects of predictor variables into perspective.

Patient-completed QL questionnaires constitute a previously unreported situation in which multicollinearity can occur and have practical importance. Usually, variables that are found to be significant predictors of the outcome of interest in prognostic factor analyses are described as ‘independent’ predictors. We question the appropriateness of applying this terminology to significant prognostic factors derived from patient-completed questionnaires, since these predictors are certainly not independent of each other. Global QL is particularly problematic in this respect because it is most highly correlated with all other variables on the questionnaire. We propose that global QL be excluded from the set of predictor variables when the QLQ-C30 is used in prognostic factor analyses, in order to minimize instability of the final multivariate models. Several results justify this recommendation: the correlation matrix, data reduction techniques such as principal components analysis or factor analysis which gave similar results when global QL was included and excluded from the variable pool, and the bootstrap analysis.

Multicollinearity does not invariably indicate the presence of redundant variables. QL questionnaires are a good illustration of a situation in which variables measure related concepts that have some overlap but also distinctive features. This can be seen, for example, in the function variables of the QLQ-C30. PF measures perceived difficulty in performing simple physical functions of daily living, such as walking, dressing etc. RF measures limitations in pursuing work and leisure activities, SF measures the degree of interference by the illness in one's family life and social activities, EF measures worry, tension, irritability and depression, and CF measures memory and concentration. Intuitively it is obvious that in the clinical situation these functions may be strongly interdependent, but they cannot be said to be measuring the same thing. Global QL has been described as a latent variable because it is thought to measure unobserved aspects of QL, although it presumably encompasses measured aspects as well. For this reason it is both difficult to interpret and to manipulate in the clinical situation.

In conclusion, we have demonstrated that because strong intercorrelation between variables is an inherent feature of the EORTC QLQ-C30 questionnaire, the resultant multicollinearity may influence model selection in prognostic factor analyses. Harmful multicollinearity is particularly likely to occur when global QL is included among the predictor variables. For this reason, and also because it is difficult to interpret and manipulate clinically, we suggest that it is better not to include global QL as a predictor variable in such analyses. One can always attempt to mitigate problems of multicollinearity as shown in Section 6, but since many of the variables are so highly intercorrelated, we advocate the use of bootstrap models as illustrated in Section 5 to obtain greater insight into the stability of the models obtained. Finally, our exploration of the multicollinearity that occurs in the patient-completed EORTC QLQ-C30 questionnaire highlights the need for thorough analysis and cautious interpretation of prognostic factor analyses based on such data. Factors that are identified as potentially important will ultimately need to be tested prospectively in clinical studies.

REFERENCES

1. Coates A, Gebski V, Bishop JF, Jeal PN, Woods RL, Snyder R, Tattersall MH, Byrne M, Harvey V, Gill G. for the Australian-New Zealand Breast Cancer Trials Group, Clinical Oncological Society of Australia. Improving the quality of life during chemotherapy for advanced breast cancer. A comparison of intermittent continuous treatment strategies. *New England Journal of Medicine* 1987; **317**: 1490–1495.
2. Coates A, Gebski V, Signorini D, Murray P, McNeil D, Byrne M, Forbes JF. for the Australian New Zealand Breast Cancer Trials Group. Prognostic value of quality-of-life scores during chemotherapy for advanced breast cancer. *Journal of Clinical Oncology* 1992; **10**:1833–1838.
3. Coates A, Thomson D, McLeod GRM, Hersey P, Gill PG, Olver IN, Kefford R, Lowenthal RM, Beadle G. Prognostic value of quality of life scores in a trial of chemotherapy with or without interferon in patients with metastatic malignant melanoma. *European Journal of Cancer* 1993; **29A**(12):1731–1734.
4. Seidman AD, Portenoy R, Yao T-J, Lepore J, Mont EK, Kortmansky J. Quality of life in phase II trials: a study of methodology and predictive value in patients with advanced breast cancer treated with paclitaxel plus granulocyte colony-stimulating factor. *Journal of the National Cancer Institute* 1995; **87**(17):1316–1322.
5. Ganz PA, Lee JJ, Siau J. Quality of life assessment. An independent prognostic variable for survival in lung cancer. *Cancer* 1991; **67**:3131–3135.
6. Buccheri GF, Ferrigno D, Tamburini M, Brunelli C. The patient's perception of his own quality of life might have an adjunctive prognostic significance in lung cancer. *Lung Cancer* 1995; **12**:45–58.
7. Tamburini M, Brunelli C, Rosso S, Ventafridda V. Prognostic value of quality of life scores in terminal cancer patients. *Journal of Pain Symptom Management* 1996; **11**(1):32–41.
8. Earlam S, Glover C, Fordy C, Burke D, Allen-Mersh TG. Relation between tumor size, quality of life, and survival in patients with colorectal liver metastases. *Journal of Clinical Oncology* 1996; **14**:171–175.
9. De Boer MF, Van den Borne B, Pruyn JFA, Ryckman RM, Volovics L, Knegt PP, Meeuwis CA, Mesters I, Verwoerd CDA. Psychosocial and physical correlates of survival and recurrence in patients with head and neck carcinoma: results of a 6-year longitudinal study. *Cancer* 1998; **83**:2567–2579.
10. Loprinzi CL, Laurie JA, Wieand HS, Krook JE, Novotny PJ, Kugler JW, Bartel J, Law M, Bateman M, Dose AM, Etzell PS, Nelimark RA, Mailliard JA, Moertel CG. for the North Central Cancer Treatment Group. Prospective evaluation of prognostic variables from patient-completed questionnaires. *Journal of Clinical Oncology* 1994; **12**:601–607.
11. Degner LF, Sloan JA. Symptom distress in newly diagnosed ambulatory cancer patients and as a predictor of survival in lung cancer. *Journal of Pain Symptom Management* 1995; **10**:423–431.

12. Coates A, Porzsolt F, Osoba D. Quality of life in oncology practice: prognostic value of EORTC QLQ-C30 scores in patients with advanced malignancy. *European Journal of Cancer* 1997; **33**(7):1025–1030.
13. Dancey J, Zee B, Osoba D, Whitehead M, Lu F, Kaizer L, Latreille J, Pater JL. for the National Cancer Institute of Canada Clinical Trials Group. Quality of life scores: an independent prognostic variable in a general population of cancer patients receiving chemotherapy. *Quality of Life Research* 1997; **6**:151–158.
14. Tannock IF, Osoba D, Stockler MR, Ernst DS, Neville AJ, Moore MJ, Armitage GR, Wilson JJ, Venner PM, Murphy KC. Chemotherapy with mitoxantrone plus prednisone or prednisone alone for symptomatic hormone-resistant prostate cancer: a Canadian randomized trial with palliative end points. *Journal of Clinical Oncology* 1996; **14**:1756–1764.
15. Kramer JA, Curran D, Piccart M, de Haes JCJM, Bruning P, Klijn J, Van Hoorebeeck I, Paridaens R. Identification and interpretation of clinical and quality of life prognostic factors for survival and response IP treatment in first line chemotherapy in advanced breast cancer. *European Journal of Cancer* 2000; **36**(12):1498–1506.
16. Cramer EM. Multicollinearity. In *Encyclopaedia of Statistical Sciences*, vol. 2, Kotz S, Johnson NL (eds). Wiley: New York, 1985; 639–643.
17. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. *American Journal of Physiology* 1985; **249**:R1–R12.
18. Sithisarankul P, Weaver VM, Diener-West M, Strickland P. Multicollinearity may lead to artificial interaction: an example from a cross sectional study of biomarkers. *Southeast Asian Journal of Tropical Medicine and Public Health* 1997; **28**(2):404–409.
19. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality of life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute* 1993; **85**:365–376.
20. Paridaens R, Biganzoli L, Bruning P, Klijn JG, Gamucci T, Houston S, Coleman R, Schachter J, Van Vreckem A, Sylvester R, Awada A, Wildiers J, Piccart M. Paclitaxel versus Doxorubicin as first-line agent chemotherapy for metastatic breast cancer: a European Organization for Research and Treatment of Cancer Randomized Study with cross-over. *Journal of Clinical Oncology* 2000; **18**(4):724–733.
21. Fayers P, Aaronson N, Bjordal K, Curran D, Groenvold M. *EORTC QLQ-C30 Scoring Manual*. 2nd edn. EORTC Quality of Life Study Group: Brussels, 1999.
22. WHO. *WHO Handbook for Reporting Results of Cancer Treatment*. World Health Organization: Geneva, 1979; 7.
23. EORTC Breast Cancer Cooperative Group. *Manual for Clinical Research in Breast Cancer*, 3rd edn. Excerpta Medica: The Netherlands, 1998; 22–27.
24. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*, Wiley: New York, 1980.
25. Mager PP, Rothe H. Obscure phenomena in statistical analysis of quantitative structure-activity relationships. Part 1: multicollinearity of physicochemical descriptors. *Pharmazie* 1990; **45**:758–764.
26. Gunst RF, Mason RL. Advantages of examining multicollinearities in regression analysis. *Biometrics* 1997; **33**:249–260.
27. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 1992; **11**:2093–2109.
28. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine* 1989; **8**:771–783.
29. Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. *Psychological Bulletin* 1988; **103**:265–275.
30. Stevens J. *Applied Multivariate Statistics for the Social Sciences*. 3rd edn. Lawrence Erlbaum Associates Publisher: New Jersey, 1996.
31. Krzanowski WJ. *Principles of Multivariate Analysis*, *a User's Perspective*. Oxford Statistical Science Series 3: New York, 1988.
32. Draper NR, Smith H. *Applied Regression Analysis*. 2nd edn. Wiley: New York, 1981.
33. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Applied Statistics* 1992; **41**:191–201.
34. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. *Quality of Life Research* 1997; **6**:139–150.
35. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. *Quality of Life Research* 1997; **6**:393–406.