STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:4055–4066
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2395

Permutation based methods for comparing quality of life between observed treatments

Beatrijs Moerkerke[1,*,†], Els Goetghebeur[1,2], Kristel Van Steen[2], Simon Van Belle[3] and Veronique Cocquyt[3]

[1]*Department of Applied Mathematics and Computer Science, UGent, Ghent University, Krijgslaan 281-S9, B-9000 Ghent, Belgium*
[2]*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.*
[3]*Medical Oncology, University Hospital Ghent, De Pintelaan 185, B-9000 Ghent, Belgium*

SUMMARY

Quality of life is becoming an important outcome for the comparison of aggressive therapies. To measure quality of life (QOL), questionnaires have been designed that ask patients about symptoms and functionality in several aspects of daily life. Primary analyses of such questionnaires typically focus on a summary statistic, such as a sum score or a single global question. This avoids inflated type I errors or loss of power due to multiple testing of individual items. In return, specific questions and answers that initially mattered to the patient may unfortunately get buried. To avoid reduced specificity and interpretability for both patients and physicians, we propose to also analyse all original questions. In this paper, we seek to detect items of the QOL questionnaire that differ significantly over observed treatments even in the face of multiple testing. We sequentially build a model that combines features which additionally discriminate between treatments. To achieve this, we draw on insights gained in the field of statistical genetics where one is often confronted with a vast amount of predictors, e.g. of a genotypic nature. Specifically, we adopt a permutation based approach to evaluate the null distribution of the maximum of many correlated test statistics and use it to build a regression model that explains QOL differences between treatment arms. We apply the new methodology to analyse QOL data in an observational study of four different treatments of breast cancer. We discover that a single question captures most of the observed treatment differences in this population. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: quality of life; permutation tests; multiple testing; breast cancer; sequential selection

*Correspondence to: B. Moerkerke, Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281-S9; B-9000 Ghent, Belgium.
†E-mail: beatrijs.moerkerke@ugent.be

Received 15 November 2004
Accepted 20 May 2005

1. INTRODUCTION AND PROBLEM SETTING

As societies become more prosperous and medical practice more ambitious, the definition of health has been broadened to include quality of life (QOL). Measuring QOL is rapidly becoming an integral part of observational studies as well as clinical trials evaluating aggressive therapies. In general, quality of life is multi-dimensional and hard to define as it is in part subjective and a consequence of the patient's values and perception. Typically QOL is assessed using self-report questionnaires containing items (questions) which record on a binary or ordinal scale (e.g. poor, moderate, good) how well the patient functions in daily life and how the patient feels about a number of things. The European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30 is one such 30-item questionnaire developed for cancer patients in general. It was initially tested in a population of lung cancer patients [1], and subsequently in patients with breast cancer [2] and other cancers [3, 4]. It was found to meet the required standards of validity (measuring what it is intended to measure), reliability (measuring with sufficient precision) and responsiveness (ability to detect changes).

Whether scores are assessed repeatedly over time or not, it is common practice to consider summary measures [5] for QOL assessments. The focus on a summary capturing a relevant latent trait which influences responses to several items or a single global QOL measure can help to gain statistical power [6]. Analysis of global QOL measures also helps overcome the technical problem of multicollinearity created by QOL variables which measure related features and produces valid and useful inference [7]. Summary and global measures thus have become popular (not only in the field of QOL) to handle the problem of multiple outcomes because they reduce the number of hypotheses tests for treatment comparisons, are easy to implement and simplify the reporting of results.

When attempting to personalize treatments, it remains a challenge to evaluate the original finer QOL scales and study individual items. As the original QOL items ask pertinent questions about aspects of daily life that matter to individual patients, it is of value to compare them directly between treatments. Indeed, the global score loses relevant information. Similar concerns are expressed by Ribaudo and Thompson [8]. Fairclough [9] points to the weakness of summary measures in studies that aim to identify the aspects of QOL that are impacted by the disease or a particular therapy. Hence the need for complementary treatment comparisons which recognize the potential value of each item, guarding against inflated type I errors due to multiple testing.

Permutation methods have widely been studied and applied in many fields as they require relatively few distributional assumptions. Because they allow to protect the type I error of multiple correlated tests relatively easily [10], these procedures also received much attention in the area of statistical genetics. To select appropriate threshold values in the face of multiple testing in the context of QOL, we will borrow from (sequential) permutation strategies designed in this field [11, 12].

2. METHODOLOGY

Consider a random sample of n patients ($i = 1, \ldots, n$) receiving a questionnaire at a fixed point in time. We adopt the following notation:

- n_k is the number of patients in treatment arm k ($k = 1, \ldots, K$).

- T_i indicates the treatment of the ith patient ($T_i = 1,\ldots,K$). T_{ik} is the dummy indicator of whether the ith patient belongs to treatment arm k ($T_{ik}=1$) or not ($T_{ik}=0$).
- S_{ij} is a numeric code for the response of the ith patient to question j ($j=1,\ldots,J$), S_{ij} may be continuous or not. In the QOL context, responses are usually classified into a small number of ordered response categories.
- $\mathbf{X}_{n\times m}$ represents an $(n\times m)$-matrix where each row i contains the values of m covariates for patient i.

We assume $(T_i, S_{ij}, \mathbf{X}_i)$ are i.i.d. ($\mathbf{X}_i^{\mathrm{T}}$: ith row of $\mathbf{X}_{n\times m}$).

2.1. A permutation approach

To determine whether treatment affects response to question j ($j=1,\ldots,J$), we consider the working model:

$$S_{ij} = \alpha_{1j} + \sum_{k=2}^{K} \alpha_{kj} T_{ik} + \varepsilon_{ij} \text{ with } E(\varepsilon_{ij}|T_i) = 0 \tag{1}$$

where treatment arm $k=1$ is used as the reference treatment arm. We use the multivariate Wald statistic to test the null hypothesis of no treatment effect on item j, $\alpha_{2j} = \cdots = \alpha_{Kj} = 0$. Since QOL responses are often discrete and in most cases skewed, the normality assumption for the residuals in (1) is not reasonable implying that the null distribution of the Wald statistics is generally no longer a χ^2-distribution. This makes permutation tests particularly attractive in this setting as they construct a non-parametric null distribution for the test statistic, conditional on the observed data.

When observations are exchangeable under the null hypothesis of no treatment effect, a random permutation of treatment (i.e. the T_i values) over the observed patients can be seen as a random draw under the null hypothesis of no association between treatment and response. Redistributing treatment indicators T_i over the different patients while retaining their outcomes S_{ij} therefore gives us the conditional null distribution of the Wald statistic. When the observed test statistic is extreme with respect to this non-parametric permutation distribution, it is declared significant at the corresponding level. When treatment is randomly allocated, exchangeability under the null is ensured by design. In observational settings however, permutation tests must be justified by making distributional assumptions or by achieving (approximate) exchangeability after correction for confounding covariates.

We correct for multiple ($J\times$) testing when judging the strength of all associations simultaneously by constructing the permutation distribution of the maximum Wald statistic over all questions. By comparing observed values of the original test statistics with the null distribution of this maximum, the familywise error rate or global type I error is controlled. This permutation procedure respects the correlation structure in the data.

For large n, it is computationally prohibitive to exhaust all $n!$ permutations. We follow the approach of Nettleton and Doerge [13] who show that it is sufficient to obtain a confidence interval for a quantile based on a modestly sized random sample of permutations. Alternative approaches can be found in References [14, 15], for instance. Let $M_{(1)} \leqslant \cdots \leqslant M_{(N)}$ denote the ordered maximum test statistics obtained by N permutations. An approximate $(1-\gamma)\times 100$ per cent confidence interval for the empirical threshold value at significance level α is then $[M_{(L)}, M_{(U)}]$ with $L = \lceil \mathrm{N}(1-\alpha) - \Phi^{-1}(1-(\gamma/2))\sqrt{\mathrm{N}(1-\alpha)\alpha}\rceil$ and $U = \lceil \mathrm{N}(1-\alpha) + \Phi^{-1}(1-(\gamma/2))\sqrt{\mathrm{N}(1-\alpha)\alpha}\rceil$. Φ represents the cumulative standard normal distribution

function and $\lceil x \rceil$ the smallest integer larger than or equal to x. L and U are based on the normal approximation for the binomial distribution and hence valid when $N\alpha \geqslant 5$. This suggests that at least $\lceil 5/\alpha \rceil$ permutation steps should be performed. Let W_j $(j = 1, \ldots, J)$ represent the observed test statistics based on model (1). Accounting for multiple testing by controlling the global α-level, the treatment effect is declared (non) significant for question j at level α when $W_j \geqslant M_{(U)}$ $(W_j < M_{(L)})$. The number of permutations must be increased when results are inconclusive for some questions $(M_{(L)} \leqslant W_j < M_{(U)})$.

2.2. Selecting a set of questions: a sequential approach

Once the strongest association is declared significant, one can search for the next question that adds differentiation between treatment groups by following a sequential residual empirical threshold (RET) approach as in Reference [12]. If question ℓ shows the highest significant association with T_i, then for each remaining question $j \neq \ell$, we fit regression model

$$S_{ij} = \beta_{0j\ell} + \beta_{1j\ell} S_{i\ell} + \varepsilon'_{ij\ell} \text{ with } E(\varepsilon'_{ij\ell} | S_{i\ell}) = 0 \tag{2}$$

The permutation procedure described in Section 2.1 is repeated with the estimated residuals $\varepsilon'_{ij\ell}$ as new scores for the remaining questions. We thus estimate the effect of treatment on the part of response j that is not yet explained by response ℓ. The procedure is repeated until no significant associations are left.

Eventually, *post hoc* pairwise permutation tests can be applied to the significant questions to identify the particular pair(s) of treatments on which the QOL question differs. We performed each of these at the nominal significance level α.

2.3. Accounting for covariates

When evaluating treatment effects based on observational data, one must avoid confounding and adjust for covariates associated with response as well as treatment. We cannot simply permute residuals from the regression of responses on the covariate(s), but must also correct treatment for these covariates to retain exchangeability under the null hypothesis of the permuted variables. To see this, consider the model:

$$S_{ij} = \gamma_{1j} + \sum_{k=2}^{K} \gamma_{kj} T_{ik} + \boldsymbol{\gamma}^{\mathrm{T}}_{(K+1)j} \mathbf{X}_i + \varepsilon^*_{ij} \text{ with } E(\varepsilon^*_{ij} | T_i, \mathbf{X}_i) = 0 \tag{3}$$

where $\mathbf{X}_i^{\mathrm{T}}$ represents the ith row of the unadjusted covariate matrix $\mathbf{X}_{n \times m}$. Equation (3) corrects the treatment–response relationship for covariates contained in $\mathbf{X}$. It becomes clear that reshuffling the responses over the n patients breaks not only their relationship with treatment but also with the covariates. Similarly, randomly reshuffling the treatment indicators breaks the relationship of treatment with the covariates. Therefore, we propose to estimate first the residuals ε^{**}_{ij} for each question j from:

$$S_{ij} = \lambda_{0j} + \boldsymbol{\lambda}^{\mathrm{T}}_{1j} \mathbf{X}_i + \varepsilon^{**}_{ij} \text{ with } E(\varepsilon^{**}_{ij} | \mathbf{X}_i) = 0 \tag{4}$$

and to regress them on corresponding treatment residuals. This makes sense because:

$$\varepsilon^{**}_{ij} = S_{ij} - \lambda_{0j} - \boldsymbol{\lambda}^{\mathrm{T}}_{1j} \mathbf{X}_i$$

$$= S_{ij} - \gamma_{1j} - \sum_{k=2}^{K} \gamma_{kj} P(T_{ik} = 1|\mathbf{X}_i) - \boldsymbol{\gamma}_{(K+1)j}^{\mathrm{T}} \mathbf{X}_i$$

$$= \sum_{k=2}^{K} \gamma_{kj} (T_{ik} - P(T_{ik} = 1|\mathbf{X}_i)) + \varepsilon_{ij}^{*}$$

In practice, we estimate residuals $T_{ik} - P(T_{ik} = 1|X_i)$ from a multinomial logistic regression model where for each $k \neq 1$:

$$\log\left(\frac{P(T_{ik} = 1|\mathbf{X}_i)}{P(T_{i1} = 1|\mathbf{X}_i)}\right) = \theta_{1k} + \boldsymbol{\theta}_{2k}^{\mathrm{T}} \mathbf{X}_i \tag{5}$$

and hence $T_{ik} - P(T_{ik} = 1|X_i)$ $(k \neq 1)$ is estimated using

$$\hat{P}(T_{ik} = 1|X_i) = \frac{\exp(\hat{\theta}_{1k} + \hat{\boldsymbol{\theta}}_{2k}^{\mathrm{T}} \mathbf{X}_i)}{1 + \sum_{k'=2}^{K} \exp(\hat{\theta}_{1k'} + \hat{\boldsymbol{\theta}}_{2k'}^{\mathrm{T}} \mathbf{X}_i)} \tag{6}$$

The permutation based null distributions for Wald statistics, evaluating treatment effect after adjusting both variables for covariates, can now be determined by permuting the treatment residuals. When working with *estimated* residuals, exact permutation tests are not possible due to the non-exchangeability of these residuals [16]. We developed the residual approach following the suggestions of Anderson and Robinson [17] who performed a comparative simulation study for several approximate permutation tests in linear models. We supplemented this with simulations for our specific conditioning approach and found that the achieved significance level was close to the *a priori* significance level.

An alternative to regression for covariate adjustment, especially useful when their nature is discrete, is stratification. Let the covariates divide the sample into $\mathscr{S}$ independent strata $(s = 1, \ldots, \mathscr{S})$, each with sample size n_s $(\sum_{s=1}^{\mathscr{S}} n_s = n)$. With $\boldsymbol{\gamma}_{js}$ the stratum-specific vector of regression coefficients for $(T_{i2}, \ldots, T_{iK})$ in (3) and $\mathbf{V}_{js}$ the corresponding variance–covariance matrix for question j, the stratified Wald statistic becomes

$$W_j = \boldsymbol{\gamma}_j^{\mathrm{T}} (\mathbf{V}_j)^{-1} \boldsymbol{\gamma}_j \tag{7}$$

where

$$\boldsymbol{\gamma}_j = \sum_{s=1}^{\mathscr{S}} \frac{n_s}{n} \boldsymbol{\gamma}_{js}$$

and

$$\mathbf{V}_j = \sum_{s=1}^{\mathscr{S}} \left(\frac{n_s}{n}\right)^2 \mathbf{V}_{js} \tag{8}$$

Permutation tests can be performed as before provided observations are permuted within their stratum before calculating (7). As for the regression approach, this adjustment reduces bias while the more homogeneous strata are expected to yield more precise treatment comparisons. When the maximum value of the J test statistics shows a significant treatment effect, a stratified RET approach can continue to detect remaining associations.

3. APPLICATION IN AN OBSERVATIONAL BREAST CANCER STUDY

3.1. Description of the study

We obtained survey data of a study on 277 women who underwent one of four possible forms of surgery for the treatment of breast cancer at the university hospital of Ghent. Breast preserving treatment (B+) was received by 184 patients, mastectomy with immediate (M+now), delayed (M + late) and without (M−) reconstruction by 34, 9 and 50 patients, respectively.

The two questionnaires used in this study were the QLQ-C30 and QLQ-BR23 from the EORTC and can be found in Reference [18]. They contain in total 53 items on QOL: the QLQ-C30 for cancer patients in general contains 30 items and the QLQ-BR23 designed for breast cancer patients contains 23 items. We numbered the questions from 1 to 53 starting with the items from the QLQ-C30. Two questions (question 29 'How would you rate your overall health during the past week?' and question 30 'How would you rate your overall quality of life during the past week?') are more global in nature [1]. The other 51 questions reflect specific aspects of daily QOL. Due to the structure of the questionnaire, some of these questions are often combined as they are targeting similar aspects of QOL (e.g. physical, role, cognitive, emotional and social scales). But our approach is to consider each aspect or symptom in turn by examining how answers on separate items differ over treatment groups. For ease of comparison, all answers were rescaled from 0 to 100 in such a way that higher scores consistently imply a more negative QOL. Permutation tests do not require this rescaling but it facilitates interpretation.

Not all patients had complete score information. The overall percentage of missing scores was 1.36 per cent with an average of 3.75 missing scores per question. Although missing scores cannot enter the association measure, they can be involved in the permutation procedure. If these cases are permuted, each permutation substitutes a different randomly chosen outcome for the missing one and excludes the contribution of a randomly chosen patient instead. This is valid when scores are missing at random (MAR). A person with some missing scores could still have valuable answers for other questions whereas patients with unknown treatment added no information and were not included in the 277 patients under consideration. This is the approach of software package Map Manager QT [19] where permutation tests are used for testing genetic marker-trait associations for many markers. We switched the role of markers and traits as we are interested in the effect of one treatment on many questions.

3.2. Results

In a first step, it is informative to analyse the global questions 29 and 30 of the QLQ-C30. Wald statistics based on model (1) were used as association measures examined with B+ as the reference category. Both questions, each analysed separately on the 5 per cent significance level, showed a significant association with treatment with 0.008 and 0.02 as the respective permutation based p-values. Permutation distributions are shown in Figure 1. *Post hoc* permutation tests revealed that the M + now and the M + late groups scored best as can be seen on the box plots in Figure 2.

To retrieve more information about the impact of treatment on different aspects of QOL, the 51 specific questions of the QLQ-C30 and QLQ-BR23 have been analysed one at a time. First, we used the permutation approach unadjusted for covariates while accounting for multiple testing. Figure 3 shows the original Wald statistics estimated based on model (1) with B+

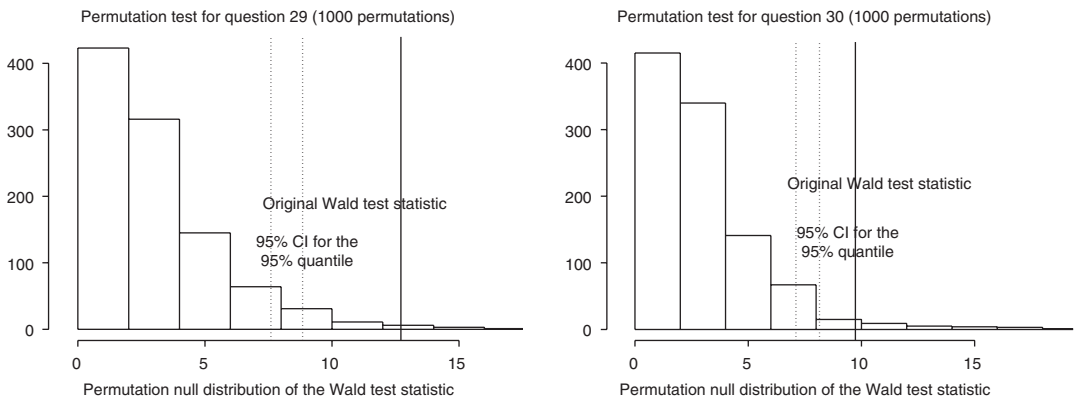

Figure 1. Results for global questions 29 and 30 of the QLQ-C30, analysed separately.

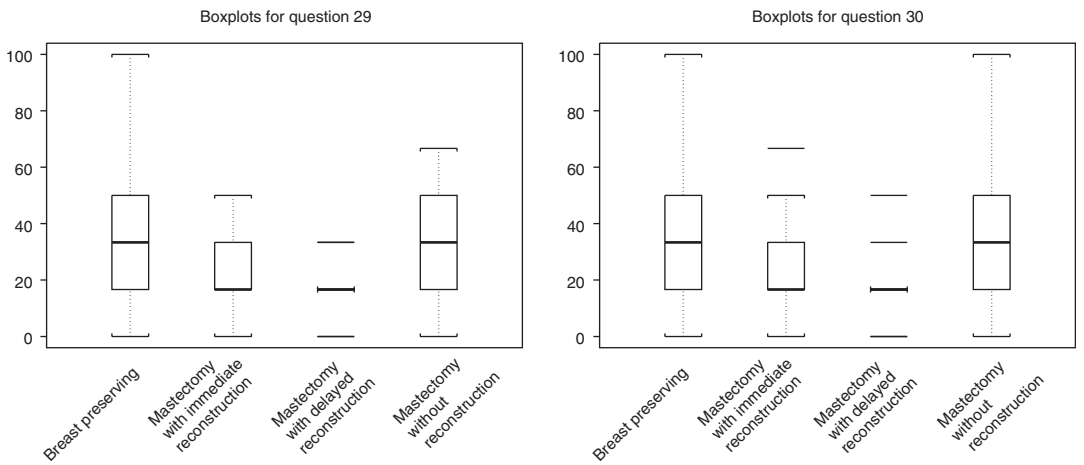

Figure 2. Box plots for answers to global questions 29 and 30 of the QLQ-C30.

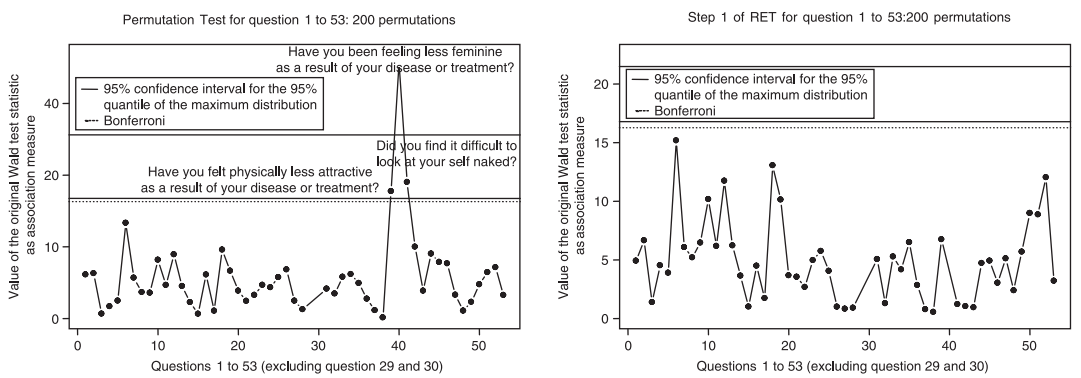

Figure 3. Result of the global permutation test analysing the four treatment groups.

as the reference category and the permutation based confidence interval for the critical value at the global 5 per cent significance level. This is the critical value of the null distribution of the maximum association over all 51 questions. The dotted line represents the cut-off from the χ^2-distribution with Bonferroni correction. Figure 3 reveals that treatments differed most in terms of the single question 40 (question 10 of the QLQ-BR23) 'Have you been feeling less feminine as a result of your disease or treatment?' flanked by question 39 'Have you felt physically less attractive as a result of your disease or treatment?' and question 41 'Did you find it difficult to look at yourself naked?'. Once the answer to question 40 was accounted for using the RET approach, no further discrimination between treatment arms emerged. A plot of marginal test results as in Figure 3 may reveal 'clusters' of questions reflecting the several QOL domains for which the questionnaire was developed. After conditioning on the most significant question, the questions belonging to the same cluster likely become less significant as all questions within a cluster are targeting a similar aspect of the QOL. Their effect is thus captured by a 'representative' of the cluster that is used to condition on. *Post hoc* tests for question 40 showed that the B+ and the M + late groups had significantly better answers. It is natural to look for plausible explanations for the rather surprisingly negative result for the M + now *versus* M + late group. We learned that women receiving a delayed reconstruction have lived for a while without a breast and could for that reason appreciate the improvement more. In addition, a delayed reconstruction is sometimes aesthetically more beautiful than an immediate reconstruction.

The 51 specific questions can largely be divided into two types: questions concerning symptoms and questions concerning functionality. It is of interest to examine whether any other items show up significantly besides question 40 which is concerned with functionality when restricting attention to the set of 28 symptomatic questions. Possibly less stringent corrections for multiple testing may be required. One can consider the subset of questions on functionality next. In the second step of RET, other significant associations than question 40 could appear since less comparisons may require less severe corrections. Performing these analyses on both groups separately did not alter the conclusions.

Given the reasoning above, we analysed questionnaires QLQ-C30 and QLQ-BR23 separately. Again, the only significant association (accounting for multiple testing) with treatment lay in question 40. We found no significant associations within the QLQ-C30 (excluding global questions 29 and 30).

So far, the analyses have not been adjusted for potentially important covariates. For instance, stage of disease plays a likely role in the perception of QOL and may therefore confound the treatment effect. A good indicator for disease stage is the number of tumors. In our sample we distinguished between patients with only 1 and with more than 1 tumor. For 36 of the 277 patients the number of tumors was unknown leaving a total of 241 patients for analysis. Table I shows that only one patient had more than one tumor in the already small M + late group. We therefore restricted this analysis to the three other treatments. A weighted stratified analysis on the 51 specific questions as described in Section 2.3 gave the same conclusions as the unstratified analysis. The difference between treatment groups with respect to question 40 only emerged within the group with one tumor. The unstandardized effects were comparable in both strata but variances in the second stratum were much larger due to the smaller sample size. This points to a drawback of this approach when sample sizes become small.

In practice, there is often a need to adjust for several covariates simultaneously. Dividing the data in homogeneous strata may become inefficient when data within (some) strata become

Table I. Distribution of the patients over the treatment in both tumor groups.

	1 tumor	>1 tumor	Total
Breast preserving treatment	156	11	167
Mastectomy without reconstruction	29	9	38
Mastectomy with immediate reconstruction	19	9	28
Mastectomy with delayed reconstruction	7	1	8

Table II. Distribution of the missing covariate values over the treatments.

	BMI	Bra size	Patients with missing value(s)
Breast preserving treatment	9	25	31
Mastectomy without reconstruction	0	7	7
Mastectomy with immediate reconstruction	1	8	8
Mastectomy with delayed reconstruction	0	1	1

The third column contains the total number of patients with at least one missing covariate value. For 3 (1) patients in the B + group (M + now group), both BMI and bra size were missing.

too sparse. To avoid this we used the approach of Section 2.3. In (4) and (5), we modelled additive effects of the patient's body mass index (BMI), bra size and the time in days between the end of treatment and the answering of the questionnaires. The bra size variable is coded with five outcome categories. We did not include the number of tumors as the previous analysis showed no evidence that this variable was confounding the treatment effect. For 47 of the 277 patients, either their BMI or bra size was missing leaving a total of 230 patients for analysis. Table II shows the distribution of the missing values over the different treatments. Patients with missing covariate values were completely dropped from the analysis and hence from the permutation procedure. This is appropriate when missingness is completely at random. Performing this adjusted analysis did not change our conclusion as can be seen in Figure 4.

3.3. Role of the smaller treatment group

The M + late group contained only nine patients but given the previous results, a logical regrouping of treatment groups is impossible. To avoid an analysis with few observations in one of the arms, we excluded the M + late group. This did not change our conclusion in the unadjusted analysis and left results in the first step virtually unaltered for the adjusted analysis. Permutation intervals for the threshold values were smaller and less severe than before due to the higher variability caused by smaller group sizes.

After eliminating question 40 in the adjusted analysis, question 18 'Were you tired?' and question 52 'Was the area of your affected breast oversensitive?' became borderline significant on which the mastectomy group with immediate reconstruction (M + now) scored best. After correcting for the association with question 18, the association with question 52 was no longer significant and no other significant associations emerged. This result is not surprising

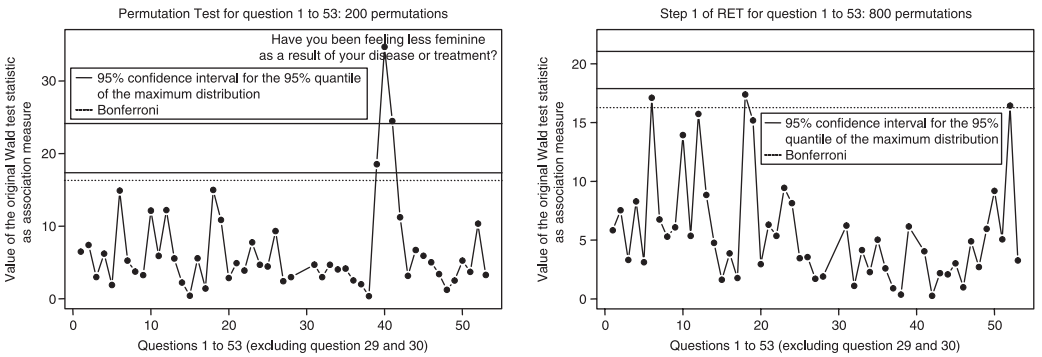

Figure 4. Result of the global permutation test adjusted for covariates analysing the four treatment groups.

as both question 18 and 52 considered physical symptoms. Patients of whom the breast was oversensitive may have been more tired, implying that part of the response to question 52 was already captured by question 18.

4. DISCUSSION

We have introduced a novel approach for analysing quality of life data. We focused on the case with one questionnaire per individual but by using multivariate procedures, the approach could be generalized to analyse responses that were gathered repeatedly over time. The methodology is more generally suited to situations where a whole series of questions, all in themselves of interest and meaningful, are asked in the form of a questionnaire. While a (weighted) sumscore yields a scalar outcome aiming to capture underlying latent QOL traits and simplifying the statistician's life, its result may mean less to the subjects concerned. Indeed, measures of QOL and IQ differ in this sense, that the answer to a particular IQ question may not matter that much per se, but for QOL they do. Hence, we have analysed QOL answers directly, in a way that enables comparisons between treatment groups while accounting for multiple testing due to the large number of questions being asked. To achieve this we have built on multiple testing lessons learned in the field of statistical genetics.

From a descriptive plot such as Figure 3, one discovers at a glance the extent to which QOL answers vary with the choice of treatment. The horizontal lines help judge whether answers that differed most over the treatments carry evidence against the global null hypothesis. The sequential nature of our approach allows to identify several questions or QOL domains in turn which differ between treatment groups, after adjusting for the QOL components that were already discovered.

In general, it is not inconceivable that the impact of treatment on question 2 is only present at a given level of question 1, calling for an interaction term in the regression part of the model. More than usual there is the concern of overtesting and it is wise to consider interactions sparingly, testing only the plausible ones. Our approach allows to incorporate such interactions in principle. It suffices to model the treatment effect in model (2) together with

an interaction term with the already established question and to investigate the permutation distribution of the treatment effect and interaction term by permuting treatment indicators and keeping scores fixed.

Working with data from an observational study, we extended our approach to correct for baseline covariates using a residual approach. Although exchangeability demands less than identically independently distributed observations, it need not perfectly hold for estimated residuals [16]. For the linear model, one often relies on the asymptotic independence of these residuals. Schmoyer [20] finds that permutation tests based on regression residuals when exchangeability is not guaranteed, are asymptotically valid under mild conditions. In a more general setting, transformations can be necessary to achieve approximate exchangeability which are characterized by Commenges [16] and are the subject of further research.

There are many other multivariate methods that also consider items separately [21]. However, the underlying motivation and target of these methods tend to differ from our approach. In particular, they often aim to construct summary measures which will form the basis of a comparison between treatment groups. The motivation may be to explore the dimensional structure of the questionnaire, to reduce the number of statistical tests or to capture latent variables. While this latter argument may remain an issue in our approach, merely considering summary measures and/or latent variables is not an option when one wants to examine individual traits or symptoms in detail.

In the many analyses, we found permutation intervals to come quite close to the threshold based on the Bonferroni correction. However, Bonferroni leans on the assumptions that the test statistic is χ^2-distributed and that tests are independent whereas permutation intervals rely on neither of both. When the permutation intervals are found to be more severe than the Bonferroni correction, this goes against the common belief that the Bonferroni correction is conservative. One must acknowledge that with multiple tests that may be correlated in either direction, the Bonferroni correction can still lead to an overly optimistic decision criterion.

Many modifications of the presented approach can be considered. All tools of modern day regression can be incorporated to model the impact of baseline covariates more flexibly (interactions, splines,...). One can also choose to produce several summary measures for different dimensions of QOL and analyse these jointly using the permutation approach.

We hope to have offered a flexible instrument that allows patients and physicians to evaluate evidence on the QOL impact of treatments in a way that helps decide which treatment works best for them in every day life.

ACKNOWLEDGEMENTS

The authors wish to thank all women who participated in the observational study and 'Kom op tegen Kanker—Vlaamse Liga tegen Kanker', its sponsor.

REFERENCES

1. Aaronson NK, Ahmedzai S, Bergman B *et al*. 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**(5):365–376.
2. Osoba D, Zee B, Pater J, Warr D, Kaizer L, Latreille J. Psychometric properties and responsiveness of the EORTC quality-of-life questionnaire (QLQ-C30) in patients with breast, ovarian and lung-cancer. *Quality of Life Research* 1994; **3**(5):353–364.

3. Bjordal K, Kaasa S. Psychometric validation of the EORTC core quality-of-life questionnaire, 30-item version and a diagnosis-specific module for head and neck-cancer patients. *Acta Oncologica* 1992; **31**(3):311–321.
4. Kaasa S, Bjordal K, Aaronson N, Moum T, Wist E, Hagen S, Kvikstad A. The EORTC core Quality of Life Questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy. *European Journal of Cancer* 1995; **31A**(13–14):2260–2263.
5. Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR. Quality-of-Life assessment: can we keep it simple? *Journal of the Royal Statistical Society*, *Series A—Statistics in Society* 1992; **155**(3): 353–393.
6. Fairclough DL. Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. *Statistics in Medicine* 1997; **16**(11):1197–1209.
7. Van Steen K, Curran D, Kramer J, Molenberghs G, Van Vreckem A, Bottomley A, Sylvester R. Multicollinearity in prognostic factor analyses using the EORTC QLQ-C30: identification and impact on model selection. *Statistics in Medicine* 2002; **21**(24):3865–3884. DOI: 10.1002/sim.1358
8. Ribaudo HJ, Thompson SG. The analysis of repeated multivariate binary quality of life data: a hierarchical model approach. *Statistical Methods in Medical Research* 2002; **11**(1):69–83.
9. Fairclough DL. *Design and Analysis of Quality of Life Studies in Clinical Trials.* Chapman & Hall/CRC: New York, 2002.
10. Pesarin F. *Multivariate Permutation Tests with Applications in Biostatistics.* Wiley: Chichester, 2001.
11. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. *Genetics* 1994; **138**(3): 963–971.
12. Doerge RW, Churchill GA. Permutation tests for multiple loci affecting a quantitative character. *Genetics* 1996; **142**(1):285–294.
13. Nettleton D, Doerge RW. Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics* 2000; **56**(1):52–58.
14. North BV, Curtis D, Sham PC. A note on the calculation of empirical P values from Monte Carlo procedures. *American Journal of Human Genetics* 2002; **71**(2):439–441.
15. Besag J, Clifford P. Sequential Monte-Carlo *p*-values. *Biometrika* 1991; **78**(2):301–304.
16. Commenges D. Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics* 2003; **15**(2):171–185.
17. Anderson MJ, Robinson J. Permutation tests for linear models. *Australian and New Zealand Journal of Statistics* 2001; **43**(1):75–88.
18. The EORTC Quality of Life website. http://www.eortc.be/home/qol [8 May 2005].
19. Manly KF, Olson JM. Overview of QTL mapping software and introduction to map manager QT. *Mammalian Genome* 1999; **10**(4):327–334.
20. Schmoyer RL. Permutation tests for correlation in regression errors. *Journal of the American Statistical Association* 1994; **89**(428):1507–1516.
21. Van Steen K, Molenberghs G. Multivariate and multidimensional analysis. In *Biometrics*, Wilson S (ed.). Eolss: Oxford, 2003.

