A parametric analysis of ordinal quality-of-life data can lead to erroneous results

Elke Kahler\textsuperscript{a}, Anja Rogausch\textsuperscript{b}, Edgar Brunner\textsuperscript{a}, Wolfgang Himmel\textsuperscript{b,*}

\textsuperscript{a}Department of Medical Statistics, University of Göttingen, Humboldtallee 32, D-37073 Göttingen, Germany
\textsuperscript{b}Department of General Practice, University of Göttingen, Humboldtallee 38, D-37073 Göttingen, Germany

Accepted 30 May 2007

Abstract

Objective: Measurements from health-related quality-of-life (HRQoL) studies, although usually of an ordered categorical nature, are typically treated as continuous variables, allowing the calculation of mean values and the administration of parametric statistics, such as \textit{t}-tests. We investigated whether parametric, compared to nonparametric, analyses of ordered categorical data may lead to different conclusions.

Study Design and Setting: HRQoL data were obtained from patients with a diagnosis of asthma (\textit{n} = 192) and chronic obstructive pulmonary disease (COPD; \textit{n} = 88) at two time points. The impact of the group factor (asthma vs. COPD) and the time factor (t1 vs. t2) on HRQoL was analyzed with a metric approach (repeated measures ANOVA) and two ordinal approaches (each with a nonparametric repeated measures ANOVA).

Results: Using the metric approach, a significant effect of “group” (\textit{P} = 0.0061) and “time” (\textit{P} = 0.0049) on HRQoL was found. The first ordinal approach (ranked total score) still showed a significant effect for “group” (\textit{P} = 0.0033) with a worse HRQoL for patients suffering from COPD. In the second approach (ranks for each HRQoL item and summed ranks), there were no significant effects.

Conclusion: Applying simple parametric methods to ordered categorical HRQoL scores led to different results from those obtained with nonparametric methods. In these cases, an ordinal approach will prevent inappropriate conclusions.

Keywords: Quality of life; Methods; Nonparametric statistics; Data interpretation; Comparative study; Psychometrics

1. Introduction

Health-related quality of life (HRQoL) is a strongly recommended and widely used measure. It is used to assess the health status of patients as the personal burden of illness cannot be described adequately by measures of disease status such as tumor load or forced expiratory volume \textsuperscript{[1]}. The need to incorporate patients’ opinions, values, and preferences is what distinguishes HRQoL from all other measures of health \textsuperscript{[2]}. In recent years, it has become increasingly clear that questionnaires can provide accurate evidence of outcomes from the patient’s perspective \textsuperscript{[3]}. Skepticism and confusion remain as to how HRQoL should be measured and analyzed. Most HRQoL questionnaires (“instruments”) consist of items with a Likert-scaled format (e.g., “Do you have any trouble taking a long walk?”) with four or more response categories provided: “not at all,” “a little,” “quite a bit,” “very much” \textsuperscript{[4]}. Several items are often pooled together to generate a score such as a physical functioning score, a mental health score, or a social functioning score \textsuperscript{[5,6]}. Sometimes the items are weighted before pooling. For an easy analysis, the measurements are typically treated as continuous variables, using standard linear models and corresponding estimators of statistics (e.g., allowing the calculation of simple sums or mean values and \textit{t}-tests or ANOVAs).

Although this approach is frequently used, it is criticized because methods for analyzing continuous data are applied to ordered categorical (“discrete”) variables, and measurement scales with an ordinal structure are treated as metric variables \textsuperscript{[7,8]}. The ordinal structure of the data does not allow the interpretation of differences and means. Instead, HRQoL data may require different techniques of analysis that take into account the ordinal character of the data, such as the methods described by Akritas and coworkers \textsuperscript{[9,10]}, Brunner and Langer \textsuperscript{[11]}, or Agresti \textsuperscript{[12]}. Otherwise, the statistical validity of the results may be doubtful \textsuperscript{[13]}. Munzel and Bandelow \textsuperscript{[14]} discussed this problem for
psychiatric studies, and Singer et al. [7] illustrated it with an example of scores used in dentistry.

In recent years, there has been an increasing interest in the use of the family of Rasch models and, more generally, the item response theory (IRT) that offers some advantages compared to traditional approaches. These models provide the means for constructing interval measures from raw data, even if these are nominal or ordinal [15]. Moreover, IRT models yield estimates that do not vary with the characteristics of the population with respect to the underlying trait [16]. That means the single person, according to their response, as well as the individual items, according to their difficulty, are conceptualized to lie on the same interval scale so that parametric statistical methods are considered appropriate for analysis. In IRT models, the measured proficiency or attitude of a person does not depend on who else takes part in the measurement or the “difficulty” level of the items (parameter separation) [15]. One major disadvantage of IRT, however, is that it requires software designed for specialists, which is often inconvenient to clinicians and even to other researchers [16].

In this paper, we applied both a standard parametric analysis and two additional different nonparametric approaches to compare the quality-of-life assessment made by two groups of patients with airway obstruction at a baseline survey and 6 months later. Using different methods to analyze the HRQoL data, we studied whether these patients differed in their HRQoL scores according to their group and/or the time point of the survey. We were especially interested to determine whether a metric approach, compared to ordinal approaches, may lead to different results—and the danger of potentially inappropriate conclusions.

We hypothesized that both groups did not differ in their HRQoL and that neither of the groups had a better—or worse—HRQoL after 6 months. The time span seemed much too short for such an effect, especially as no intervention to improve the patients’ quality of life had taken place. The sample was drawn from general practices so that we did not expect an overrepresentation of severe cases in any of the two groups. This was also the case in a large Spanish representative sample of the general population [17].

2. Methods

2.1. Sample and procedure

Data were taken from the MedViP project (“Medizinische Versorgung in der Praxis” [Medical Care in General Practice]; www.medvip.uni-goettingen.de). The study protocol was approved by the Research Ethics Committee of the University of Göttingen. Design and recruitment have been described in detail elsewhere [18].

In brief, general practitioners were invited to provide routinely collected electronic medical data. Electronic patient records (EPRs) were extracted via a standardized interface.

We identified pseudonymized patient codes with an electronically documented diagnosis of asthma or chronic obstructive pulmonary disease (COPD) made on the EPR.

With informed patient consent, a trained study nurse then asked patients to fill out questionnaires on health care utilization, HRQoL, and depression at baseline (t1). Diagnoses of obstructive lung diseases were validated by spirometry before and after inhalation of 200 µg Salbutamol. In the case of either an absent (FEV1/FVC >70% and FEV1 ≥ 80% of the predicted value or FEV1/FVC ≥ 70% and 60% ≥ FVC < 80%) or a reversible obstruction (FEV1/FVC < 70% and FEV1 < 80% and change in FEV1 after broncholysis ≥15%), patients were regarded as suffering from asthma. With a nonreversible obstruction (FEV1/FVC < 70% and FEV1 < 80% and change in FEV1 after broncholysis <15%), patients were defined as COPD sufferers [19,20]. After 6 months of this initial baseline assessment, patients were again asked to complete the questionnaire on HRQoL (t2).

2.2. Questionnaires

To assess HRQoL, we used the Saint George’s Respiratory Questionnaire (SGRQ) [21]. It consists of 50 items, often in a Likert-scaled format. The number of response options varies from two to five. For example, one of the “Symptoms”-items (item no. 1) reads: “Over the last years I have coughed …,” providing the answers: “Most days a week,” “Several days a week,” “A few days a month,” “Only with chest infections”, “Not at all.” An example from the “Impact”-items (no. 9) reads: “How would you describe your chest condition?”, providing the following answers: “The most important problem I have,” “Causes me quite a lot of problems,” “Causes me a few problems,” “Causes no problem.” The answers are weighted and computed to three scales: symptoms (eight items), activity (16 items), and impact (26 items). Additionally, a total score is computable. The transformed values range from 0 to 100, with lower values representing a better HRQoL.

Most patients (n = 248) also filled in a questionnaire to measure depressive symptoms (Center for Epidemiologic Studies Depression Scale, CES-D[22]) because patients with asthma [23] or COPD [24] seem to be at increased risk of developing depression compared to healthy controls. The CES-D consists of 20 items that are rated on a four-point scale (0 = “rarely or none of the time [less than 1 day],” 1 = “some or little of the time [1—2 days],” 2 = “occasionally or a moderate amount of the time [3—4 days],” and 3 = “most or all of the time [5—7 days]”). Values range from 0 to 60; a total score of 16 is regarded as indicative for depression [23].

2.3. Design and analysis strategy

The statistical set up in our trial was a repeated measures design, that is, the same subjects, patients with either
asthma or COPD (group factor), were observed at two time points (time factor).

According to the instructions of the SGRQ, the Likert-scaled items were weighted and then pooled together to generate a total score. This total score was then treated as metric data (approach 1) or ordinal data (approach 2). In both approaches, the effect of the factors “time” and “group” on the total score and the interaction of these two factors were analyzed. The interaction addressed the question whether the time profiles of the total score of the two disease groups were significantly different (the significance level was set at 5%). In statistical terms, the following hypotheses were tested:

- $H_{0, \text{group}}$: no difference between the disease groups,
- $H_{0, \text{time}}$: no difference between first and second time points,
- $H_{0, \text{interaction}}$: no difference between the time profiles of the two disease groups.

When HRQoL scores were treated as metric items (approach 1), the impact of disease group and time on the scores was examined by a parametric repeated measures ANOVA, analogous to the methods described by Brunner et al. [25] for nonparametric methods. For the ANOVA statistics in this parametric approach, the mean and standard deviation of the total score had to be calculated.

When we treated the scores as ordinal (approach 2), we applied two different methods: (i) The total SGRQ-scores were calculated for each subject in the usual way as described above according to the SGRQ instructions. Then these scores were replaced with their ranks. Using these ranks, the same repeated measures ANOVA as in the parametric approach was applied. (ii) Alternatively, ranks were applied to each SGRQ-item separately. For example, the answer of question 1 was replaced for each subject with its rank. This was done for each question of the SGRQ separately, and then these ranked items were summed up to yield one total rank sum score for each subject. With these total rank sum scores, again the same repeated measures ANOVA was performed as in the parametric approach to compare new total ranks sum scores to the disease group and the time of the survey.

3. Results

3.1. Characteristics of the sample

A total number of 327 patients with either asthma or COPD from 39 practices gave informed consent to take part in the study. Complete data concerning quality of life were obtained from 300 patients, slightly more than half of them female (54.6%). The mean age of the patients was 57.9 (range: 19.5–81.2) years. According to spirometry, 192 patients (64.0%) had a diagnosis of asthma and showed absence or reversibility of bronchial obstruction. Eighty-eight patients (29.3%) showed nonreversible obstruction, confirming the diagnosis of COPD. Twenty patients (6.7%) had to be excluded due to a restrictive disorder.

Most patients ($n = 165$) had CES-D scores in the normal range, but 83 patients (28%) had scores $>16$, indicative of depression. The scores for patients with asthma (median = 11; mean = 11.9; SD = 7.8) and those with COPD (median = 12; mean = 12.4; SD = 8.2) were nearly the same.

3.2. Quality-of-life scores and comparison of scoring procedures

In Table 1, the HRQoL scores of the two groups at the two time points are summarized. While the mean values in both disease groups increase from time point 1 to 2, the medians tend in different directions. In Table 1, the

<table>
<thead>
<tr>
<th>Table 1</th>
<th>HRQoL scores of the two patient groups at the two time points and effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease group</strong></td>
<td><strong>HRQoL scores</strong></td>
</tr>
<tr>
<td><strong>Number of patients</strong></td>
<td>187</td>
</tr>
<tr>
<td><strong>Mean value (95% confidence interval [CI])</strong></td>
<td>31.97 (29.16 to 34.79)</td>
</tr>
<tr>
<td><strong>Standard deviation</strong></td>
<td>19.50</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>31.63</td>
</tr>
<tr>
<td><strong>Inner quartile range</strong></td>
<td>15.61 to 45.83</td>
</tr>
<tr>
<td><strong>Minimum</strong></td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Maximum</strong></td>
<td>83.86</td>
</tr>
<tr>
<td><strong>Effect sizes</strong></td>
<td>Asthma: time effect</td>
</tr>
<tr>
<td><strong>Least squares mean differences between respective groups</strong></td>
<td>$-1.23 (-7.3$ to $-0.7)$</td>
</tr>
<tr>
<td><strong>Effect size</strong></td>
<td>0.02</td>
</tr>
</tbody>
</table>

* As calculated with SAS procedure “proc mixed lsmeans.”
effect sizes for the observed time differences in both disease groups and the observed group differences at the two time points are also summarized. In Fig. 1, the distribution of the SGRQ total score is shown separately for the two time points and the two patient groups under investigation. As can be seen from these boxplots, the original values of the HRQoL scores show a slightly skewed distribution with a longer tail tending to higher HRQoL scores.

When the score was considered metric, the results of a parametric repeated measures analysis showed highly significant effects of “group” ($P = 0.0061$) and “time” ($P = 0.0049$) on HRQoL, but no interaction between “group” and “time” ($P = 0.1394$). The least squares mean difference (95% confidence interval [CI]) between time points 1 and 2, irrespective of disease group, was $-2.64\ (\pm 0.99)$. The difference between disease groups (COPD vs. asthma), irrespective of time, was $-7.13\ (\pm 2.3)$.

The distribution of the ranks of the total SGRQ-score is shown in Fig. 2. Treating the HRQoL score as ordinal, the nonparametric repeated measures still showed a significant effect of “group” on HRQoL ($P = 0.0033$) but not of “time” ($P = 0.1698$) and no interaction between these influences ($P = 0.2414$).

After the new rank sum score was calculated from the sum of each ranked item, the scores were again compared with a nonparametric repeated measures analysis. In this scenario, neither “group” ($P = 0.1004$) nor “time” ($P = 0.4157$) now showed any effect on HRQoL. Also, no interaction between “time” and “group” ($P = 0.9440$) could be detected. Fig. 3 illustrates the new rank sum score for the two patient groups at the two time points.

4. Discussion

Like other diagnostic procedures, quality-of-life measures should be valid, reliable, and sensitive over time [1]. We could demonstrate that applying simple parametric methods to ordered categorical HRQoL scores generated results deviant from those we obtained with nonparametric methods. One reason for the differences could be that applying parametric analysis to ordered categorical HRQoL data is not appropriate, because in doing so one implies that differences between ordered categorical data are interpretable similar to differences between mean values which is clearly not the case. If you consider, for example, a five-point Likert-scaled item: $0 = \text{not present}$, $1 = \text{mild}$, $2 = \text{moderate}$, $3 = \text{severe}$, and $4 = \text{very severe}$, it is clear that mild is less than severe. But although the difference between mild and moderate and between severe and very severe is both 1, it is not clear if these two differences are equal or if they can be interpreted and compared at all. Therefore, methods that take into account the ordinal structure of the data are reasonable and should be applied.

We cannot definitely exclude that there is a difference in the health-related quality of life between the two disease
groups at the two time points as the metric approach suggests. However, as we used an analysis method that considered the categorical nature of the data, it is rather likely that this analysis is superior to a metric approach. Above all, the results of the two ordinal approaches seem to not obviously contradict previous work—for two reasons:

1. As there was no study intervention or change in medication, systematic changes in HRQoL over a short period of time (6 months) are not very likely. It is well known that patients with obstructive lung diseases suffer from acute exacerbations from time to time [26,27], but that these episodes are then followed by improvements in health status. Therefore, it is to be expected that the deterioration in some patients will be cancelled out by improvement in others. This is in line with the result of our nonparametric repeated measure analysis where the effect over time was no longer significant. Furthermore, patients had been examined consecutively so that seasonal effects should also be negligible.

2. With respect to group differences in the quality of life between patients with asthma vs. patients with COPD, the results from different studies are inconsistent: severe impairment and progressive worsening of patients’ health over time has been reported more for COPD sufferers than for asthma patients, especially over longer periods of time [6,28]. This observation is in line with the results obtained from our first ordinal approach, where there was a significant effect of disease. In a cross-sectional study using a metric approach, Ferrer et al. [17] did not find statistically significant differences between patients with either asthma or COPD when analyzing a large representative sample. When we treated each item as ordinal and calculated a new rank sum, differences between groups also vanished in our analysis.

According to Schünemann et al., a relevant difference of HRQoL scores is about 4 units [29]. The CIs of differences between time points (Table 1) contain these 4 units within their limits indicating that no relevant change in HRQoL scores occurred. For the time effect, the mean differences are even lower than for the group differences. These results are more in agreement with the results we got from the nonparametric analysis rather than that from the parametric analysis.

The nearly identical depression scores may be regarded as a sort of indirect evidence that the two groups in our study did not differ, because depression, or more generally, psychological mood is at least one component of HRQoL [23]. However, we should emphasize that it is, on principle, difficult to find out whether the two groups differ in their HRQoL or changed their HRQoL between the two time points. Because quality of life is a distinct concept [30] and neither a function of the severity of the disease nor the degree of depression, there is no gold standard like spirometry for a valid estimation of HRQoL. The only way is to let the patients rate their quality of life.

In contrast even to IRT, we did not make any assumptions about the distribution of the scores and, thus, our results are not based on questionable or even inappropriate assumptions. The use of parametric methods for the analysis of ordinal data is often taken for granted when the most relevant prerequisites for their use are fulfilled. These include a large sample size (~300 subjects in our case) so that the distribution of the mean resembles more a normal distribution (“central limit theorem”) [31]. Another prerequisite for the use of parametric methods is the presence of seven or more occupied categories in the HRQoL outcome, which Walters et al. [32] suggest as a general rule of thumb. This was the case for the score used in this study.

However, there are still significant differences between parametric and nonparametric analyses. A reasonable analysis of ordered categorical data should be invariant under any monotone transformation of the scale, that is, the results should not depend on the choice of the scale as long as the original order of the score is maintained. This is usually not the case when metric analysis is applied to ordered categorical data disregarding the nonmetric properties of the data. The HRQoL is not based on a metric structure that allows the interpretation of differences or means as is the case in a metric analysis. Using additional, quite arbitrary weights to calculate HRQoL summary scores further implies a metric structure that is not actually present in the data. This can lead to erroneous results.

Several authors have reported similar results concerning the analysis of ordered categorical data in instances other than quality-of-life studies. Using an impressive example from the field of Dentistry, Singer et al. [7] showed that standard metric analysis applying to scores may produce conflicting results. Munzel and Bandelow [14] investigated the use of parametric vs. nonparametric tests in psychiatric rating scales. Recommendations for the use of nonparametric analysis for ordered categorical data are also given by Walters et al. [32] and Norris et al. [13].

When there are departures from the parametric model assumptions, the nonparametric test may be more powerful and the information in the data better represented by an ordinal approach. When the assumptions are fulfilled the parametric approach is only slightly better [33]. Furthermore, the method presented in this paper is easier to apply than, for example, analyses based on IRT [34].

4.1. Conclusions

It has been pointed out quite often that the ordinal character of HRQoL scores can lead to problems when parametric methods, or other inappropriate analysis methods, are applied [7,13]. However, these warnings, often put forward by statisticians, seem rather theoretical when the consequences of the application of parametric methods to categorical HRQoL data have not been demonstrated. Our
example should vividly show the possible effects of inappropriate statistical methods for the analysis and interpretation of HRQoL scores and stimulate researchers to use categorical based statistical methods more often.

Acknowledgment

The MedViP project is supported by a research grant from the German Ministry of Research and Education (01 GK 0201).

References