The Bootstrap and Multiple Comparisons Procedures as Remedy on Doubts about Correctness of ANOVA Results

Izabela CHMIEL 1,* and Maciej GORKIEWICZ 2

1 Jagiellonian University of Krakow, Health Sciences Faculty, ul. Michałowskiego 12, 31-126 Krakow, Poland.
2 Jagiellonian University of Krakow, Health Sciences Faculty, ul. Zaruskiego 9 / 57, 43-000 Bielsko- Biała, Poland.
E-mail(s): izabela_chmiel@wp.pl; gorkiewicz@poczta.fm

* Author to whom correspondence should be addressed; Tel.: +48(12)6343397; Fax: +48(12)6324881

Received: 16 November 2012 / Accepted: 29 February 2012 / Published online: 10 March 2012

Abstract
Aim: To determine and analyse an alternative methodology for the analysis of a set of Likert responses measured on a common attitudinal scale when the primary focus of interest is on the relative importance of items in the set - with primary application to health-related quality of life (HRQOL) measures. HRQOL questionnaires usually generate data that manifest evident departures from fundamental assumptions of Analysis of Variance (ANOVA) approach, not only because of their discrete, bounded and skewed distributions, but also due to significant correlation between mean scores and their variances. Material and Methods: Questionnaire survey with SF-36 has been conducted among 142 convalescents after acute pancreatitis. The estimated scores of HRQOL were compared with use of the multiple comparisons procedures under Bonferroni-like adjustment, and with the bootstrap procedures. Results: In the data set studied, with the SF-36 outcome, the use of the multiple comparisons and bootstrap procedures for analysing HRQOL data provides results quite similar to conventional ANOVA and Rasch methods, suggested at frames of Classical Test Theory and Item Response Theory. Conclusions: These results suggest that the multiple comparisons and bootstrap both are valid methods for analysing HRQOL outcome data, in particular at case of doubts with appropriateness of the standard methods. Moreover, from practical point of view, the processes of the multiple comparisons and bootstrap procedures seems to be much easy to interpret by non-statisticians aimed to practise evidence based health care.

Keywords: Quality of life; Questionnaire; SF-36; Statistical analysis.

Introduction

Our research has been primarily motivated by a problem of choosing strategies of rehabilitation at convalescents after hard disease on the base of the health-related quality of life (HRQOL) examination. Usually, a rehabilitation at convalescents is managed as a minimally invasive intervention, desirable non-pharmacological therapy, with focus on patient’s self-management skills [1], and psycho-educational actions addressed to convalescents and their kin [2]. So, in practice there arises a great difficulty in defining exactly the active ingredients of an intervention. It may be impossible to single out which particular parts are effective, since one component may not work without another [3]. The HRQOL examinations make available a holistic insight on this, how an
individual person perceived his own physical and mental health at the moment, and over time [4], create opportunity to better understanding how an illness interferes with a person’s day-to-day life, [5]. From other point of view, HRQOL questionnaires were designed as psychometric scales, so the classical test theory and the modern item response theory offers there many well-known statistical procedures, [6]. For these reasons the HRQOL measures are increasingly used in rehabilitation practice as primary outcome measures. Alas, HRQOL data often manifest evident departures from fundamental assumptions of the basic parametric procedures, in these of the Analysis of Variance (ANOVA) approach [7,8]. From practical point of view, this occurrence intensifies the usual difficulties with harnessing the value of scientific evidence within individualised patient care [9], [10], [11]. Thus, the pragmatic motivation as well as statistical prudence would suggest non-parametric methods is used to analyse HRQOL data.

The SF-36 is the most frequently used multi-item HRQOL instruments [12]. The obligatory scope of the analyses of any HRQOL data set obtained with Version 2 of the SF-36 Health Survey has been defined by owners of SF-36 methodology in the manual [12]. In the thesis [11], these recommendations [12] have been rigorously respected; therefore the whole analysis was executed at the frame of the classical test theory. Nevertheless, the suitability of the other methodologies was proved in some associated studies [13-15].

In the current study, the HRQOL data from [11] were reanalysed with use of the multiple comparisons procedures and the bootstrap procedures, because of some doubts on suitability of the conventional ANOVA methods in the case. It should be noted that, basing only on the features of the separate domains of the HRQOL, one can conclude that data under consideration met the basic assumptions of ANOVA method [7]. Indeed, the postulated normality of distributions was supported with the moderate values of skew and of kurtosis; the postulated homogeneity of variances was supported with the similarity of the estimated standard deviations SD [16]. Nevertheless, a fatal special case arose here because of significant linear regression between means and standard deviations of scores [17].

The current study has twofold objectives, both motivated with doubts on appropriateness of the standard methods to uncover relationship among HRQOL domains. The empirical purpose was to provide the additional support for findings in the matter, obtained in thesis [11] with exclusive use of the standard methods of HRQOL data analyses. The methodological (educational) objective was to demonstrate that in case of the doubts on suitability of the ANOVA approach one can apply the straightforward alternative procedures, instead either of the very hard to interpretation nonparametric counterparts [7], or very sophisticated verification procedure [17]. It seems, that the first purpose has its importance for only limited group of individuals interesting in rehabilitation after acute pancreatitis [5,11,13], but the second purpose can be inspiring for many medical professionals, interesting in individualised patient care and in evidence based medicine [10].

Material and Method

The initial sample included all of 422 patients hospitalised for acute pancreatitis at the 1st Department of General Surgery in Jagellonian University of Krakow (Poland) from 2000 to 2006 years. The four exclusion criteria were used: age: < 18 years or > 70 years (66 excluded); death (34 excluded); non complete clinical data (20 excluded), complication with other illness (36 excluded). The standard Polish version of SF-36 questionnaire with proper instructions was mailed to all of 266 non-excluded survivors. The standard procedure for mail survey was applied with proper thoroughness. A covering letter accompanying each questionnaire included also the explanation of the survey purpose and of the possible health benefits for the respondent. The phone consultation in completing the form, if needed, was offered. Nevertheless, the N = 124 participants didn’t return a worthy answer, but N=142 survivors (81 men and 61 women) return the enough complete forms.

The mean scores of HRQOL at the three groups of participants, defined with criterion of the same clinical type of the disease, were compared for each HRQOL domain separately, using the pair-wise comparisons methodology, with Bonferroni-like adjustment [18-21], for results obtained with two basic parametric tests, t-test and one-way ANOVA. Then, the comparisons between mean
values of all nine HRQOL domains were made with use of bootstrap approach [22]. All calculations were made with Statistica v.8 software [17].

The all three adjustment procedures applied in this study, i.e. basic Bonferroni procedure, Hommel’s and Rom’s Bonferroni procedures work in a quite similar way [18]. Let us consider a single null hypothesis H₀ with known significance P; a hypothesis H₀ should be rejected, if P ≤ P₀ where: P₀ = 0.05 assumed significance level. Now, let us consider K logically related null hypotheses H₀:k; k = 1,2, ..., K; with known significance estimates P(k); k = 1,2, ..., K. Put the P(k)-values in descending order yielding: P(1) ≥ P(2) ≥ ... ≥ P(K-1) ≥ P(K); a hypothesis H₀:k should be rejected, if P(k) ≤ P₀(k); where reference levels of significance P₀(k) are defined as either P₀(k), or P₀(k), or P₀(k), accordingly to a previously chosen testing procedure:

i) for basic Bonferroni procedure: P₀(k) = 1 / K; k = 1,2, ..., K;
ii) for Hommel-Bonferroni procedure: P₀(k) = 1 / k; k = 1,2, ..., K;
iii) for Rom-Bonferroni procedure: for k > 10: P₀(k) = P₀(k) = 1 / k; k = 11, K;
but for k ≤ 10 read the values of P₀(k) from Table 3, column P₀(k); cited from [18].

Any bootstrap procedure imitate a real process of drawing a set of statistical data X from some population with fixed distribution F(X), but in contrary to real inquiries, the artificial ones must operate with F(X) known in advance. The parametric bootstrap, also named Monte Carlo method, uses F(X) defined analytically. This creates opportunity to estimate some characteristics of F(X) from a whole real-data sample under considerations, sometimes properly standardised [22]. Thus, in the non-parametric sample any value from the original sample can either be absent or occur more than once. Usually, bootstrap samples of the same size as an original real-data sample were standardised to interval from the possible the worst score HRQOL = 0 to the best score HRQOL = 1. The set of N=142 individual scores was summarised in the Table 1.

Results

The raw individual scores of the K=9 domains of HRQOL, as measured with SF-36 questionnaire, were standardised to interval from the possible the worst score HRQOL = 0 to the best score HRQOL = 1. The set of N=142 individual scores was summarised in the Table 1. Descriptive statistics of quality of life of a study group.

<table>
<thead>
<tr>
<th>Domain of quality of life</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF – physical functioning</td>
<td>0.65</td>
<td>0.27</td>
<td>0.70</td>
<td>-0.52</td>
<td>-0.52</td>
</tr>
<tr>
<td>RLP – role limitation physical</td>
<td>0.59</td>
<td>0.31</td>
<td>0.56</td>
<td>-0.02</td>
<td>-1.08</td>
</tr>
<tr>
<td>RLM – role limitation mental</td>
<td>0.62</td>
<td>0.32</td>
<td>0.58</td>
<td>-0.25</td>
<td>-1.12</td>
</tr>
<tr>
<td>SF – social functioning</td>
<td>0.60</td>
<td>0.26</td>
<td>0.63</td>
<td>-0.23</td>
<td>-0.67</td>
</tr>
<tr>
<td>MH – mental health</td>
<td>0.56</td>
<td>0.23</td>
<td>0.55</td>
<td>-0.20</td>
<td>-0.59</td>
</tr>
<tr>
<td>EV – energy/vitality</td>
<td>0.53</td>
<td>0.17</td>
<td>0.50</td>
<td>-0.07</td>
<td>0.43</td>
</tr>
<tr>
<td>P – physical pain</td>
<td>0.61</td>
<td>0.27</td>
<td>0.56</td>
<td>-0.09</td>
<td>-0.80</td>
</tr>
<tr>
<td>HP – general health perceptions</td>
<td>0.43</td>
<td>0.19</td>
<td>0.45</td>
<td>0.22</td>
<td>-0.19</td>
</tr>
<tr>
<td>CIH – change in health</td>
<td>0.50</td>
<td>0.24</td>
<td>0.24</td>
<td>-0.07</td>
<td>0.43</td>
</tr>
<tr>
<td>SD^2 = -0.043 + 0.524*Mean; R = 0.705; F = 6.9; p = 0.03,</td>
<td>min</td>
<td>-0.52</td>
<td>-1.12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SD – estimate of standard deviation, calculated as usual using given set of measurements of the domains of quality of life; SD^2 = -0.043 + 0.524*Mean equation of linear regression estimated given set of pairs (SD, Mean); F – statistics F, calculated as usual; p – significance of hypothesis F = 1; R – coefficient of correlation between mean scores (Mean) and their standard deviation (SD).
The three clinical types of the disease were represented at the study sample with three groups, group g61 of N = 61 patients, group g41 of N = 41, and group g40 of N = 40. The seven pairs of the group mean scores for separate domains of quality of life chosen to paired comparisons with Student t test, see Table 2. Significances of pair-wise comparisons with t test. It can be seen at Table 2 that four domains, namely PF (physical functioning), P (pain), SF (social functioning) and MH (mental health), had the significance scores p(t) estimated with Student t test less or equal to reference level P₀ = 0.05. On the other hand, under any applied here Bonferroni-like adjustment, only the two domains, PF (physical functioning) and P (pain), had the significance scores p(t) less than adjusted reference levels of significance.

Table 2. Significances of pair-wise comparisons with t test

<table>
<thead>
<tr>
<th>k</th>
<th>domain</th>
<th>groups</th>
<th>mean_1</th>
<th>mean_2</th>
<th>p(t)</th>
<th>P₀01</th>
<th>P₀05</th>
<th>P₀010</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RLP</td>
<td>g61 - g41</td>
<td>0.57</td>
<td>0.63</td>
<td>0.19</td>
<td>0.01</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>RLM</td>
<td>g61 - g40</td>
<td>0.65</td>
<td>0.55</td>
<td>0.08</td>
<td>0.01</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>3</td>
<td>CIH</td>
<td>g41 - g40</td>
<td>0.52</td>
<td>0.44</td>
<td>0.08</td>
<td>0.01</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>4</td>
<td>MH</td>
<td>g61 - g41</td>
<td>0.61</td>
<td>0.51</td>
<td>0.05</td>
<td>0.01</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>5</td>
<td>SF</td>
<td>g61 - g41</td>
<td>0.55</td>
<td>0.55</td>
<td>0.05</td>
<td>0.01</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>g41 - g40</td>
<td>0.53</td>
<td>0.69</td>
<td>0.002</td>
<td>0.01</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>7</td>
<td>PF</td>
<td>g61 - g41</td>
<td>0.58</td>
<td>0.77</td>
<td>0.0001</td>
<td>0.01</td>
<td>0.007</td>
<td>0.007</td>
</tr>
</tbody>
</table>

p(t) – significance of the Student t test for two independent samples; P₀01, P₀05, P₀10 – Bonferroni-like adjusted reference levels of significance RLP – role limitation physical; RLM – role limitation mental; CIH – change in health; MH – mental health; SF – social functioning; P – physical pain

In this study, because of doubts on two-way ANOVA, the one-way ANOVA for three groups under consideration (g61, group40 and group41) were made separately for each of nine quality of life domains estimated with SF-36 questionnaire. The Table 3 Significances of ANOVA results, showed significances of each separate domain in column p(ANOVA), but in columns P₀01, P₀05 and P₀10 it shows the reference levels for basic Bonferroni procedure, Hommel’s and Rom’s Bonferroni procedures respectively. It can be seen at Table 3 that under any applied here Bonferroni-like adjustment the null hypothesis should be rejected at the significance level equal to P₀ = 0.05 for PF (physical functioning) domain only, having a significance level p(ANOVA) = 0.002 less then reference level P₀05 = 0.006.

Table 3. Significances of ANOVA results for three groups of participants

<table>
<thead>
<tr>
<th>k</th>
<th>Domain</th>
<th>p(ANOVA)</th>
<th>P₀01</th>
<th>P₀05</th>
<th>P₀10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HP</td>
<td>0.910</td>
<td>0.006</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td>2</td>
<td>RLP</td>
<td>0.670</td>
<td>0.006</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>3</td>
<td>EV</td>
<td>0.460</td>
<td>0.006</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>4</td>
<td>CIH</td>
<td>0.320</td>
<td>0.006</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>5</td>
<td>RLM</td>
<td>0.290</td>
<td>0.006</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>6</td>
<td>SF</td>
<td>0.130</td>
<td>0.006</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>7</td>
<td>MH</td>
<td>0.080</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>0.026</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>9</td>
<td>PF</td>
<td>0.002</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
</tr>
</tbody>
</table>

p(ANOVA) – significance of the one-way ANOVA for three group of patients; P₀01, P₀05, P₀10 – Bonferroni-like adjusted reference levels of significance; HP – general health perceptions; RLP – role limitation physical; EV – energy/vitality; CIH – change in health; RLM – role limitation mental; SF – social functioning; MH – mental health; P – physical pain; PF – physical functioning

The Table 4. Probabilities of relations: mean(i-th HRQOL domain) > mean(j-th HRQOL domain), based on K = 2,000 bootstrap simulations for each pair of the i-th and j-th domain under
examination, where: i = 1, 2, …, 9 refer to domains at the columns of the Table 4, and j = 1, 2, …, 9 refer to domains at the rows of the Table 4. The relations showed at the Table 4 can be summarized with ordering of the five clusters of the HRQOL domains: 

\[ \text{RLM} > \text{RLP} > \text{PF} = \text{HP} > \text{CIH} = \text{MH} > \text{EV} = \text{SC} = \text{P}. \]

Table 4. Probabilities of relations: mean(i-th HRQOL domain) > mean(j-th HRQOL domain)

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>SC</th>
<th>EV</th>
<th>MH</th>
<th>CIH</th>
<th>HP</th>
<th>PF</th>
<th>RLP</th>
<th>RLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.82</td>
<td>0.84</td>
<td>0.99</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>SC</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EV</td>
<td>0.16</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td>0.01</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIH</td>
<td>0.02</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RLP</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RLM</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P – physical pain; EV – energy/vitality; MH – mental health; CIH – change in health; HP – general health perceptions; PF – physical functioning; RLP – role limitation physical; RLM – role limitation mental;

Discussion

In the current study, the statistical data from thesis [11] were reanalysed with use of the multiple comparisons procedures and the bootstrap procedures, because of some doubts on suitability of the conventional methods in the case [15,17]. The objective of this paper was to provide empirical evidence that the use of the t-test and ANOVA with the F-test is both methodological and empirical. The methodological objective was defined as follows: to demonstrate usefulness of an alternative methodology for the analysis of a set of Likert responses in a situation when the serious doubts on appropriateness of the standard methods. The empirical objective was to prove, if the proposed alternative methodology provides an additional support to the findings in thesis [11], and in consequence, some indirect support to practical recommendations in thesis [11], regarding to the needed scope of rehabilitation in the study sample of convalescents.

Because in this study the serious doubts be about two-way ANOVA only [17], instead this the standard t-test and one-way ANOVA were applied; and additionally the multiple comparisons [18-21], and bootstrap procedures [8,22].

It was stated, not surprisingly, that naïve use of t-test and ANOVA leads here to misguided estimates, but they can be corrected with proper adjustment procedure. In other words, with respect to question if the parametric approach can be applied in the case under study, it can be interpreted in such a way, that the all three Bonferroni-like procedures didn’t confirmed here naïve use of the t-test and ANOVA with the F-test.

It should be noted, that contrary to methods based on rank transformation [7], the methods applied in our study didn’t changed the means with medians.

Then, it was stated, not surprisingly [8], that the bootstrap procedures postulate here the ordering of the separate domains of HRQOL, with respect to their deficiencies, quite similar to orderings estimated with conventional one-way ANOVA and Rasch methods [11]. It can be interpreted in such a way, that all these methods lead here to quite similar recommendations on a needed scope of rehabilitation among convalescents after acute pancreatitis.

It should be noted, that the bootstrap has some valid advantages over one-way ANOVA used as ersatz to two-way ANOVA. First, bootstrap is more robust against departures from normality [8, 22]. Then, the bootstrap can proceed if fact quite like two-way ANOVA, taking into account three separate groups of convalescents at each HRQOL domain, but the one-way ANOVA cannot do it.

The study group can be considered as representative at least for Polish convalescents after successful clinical therapy against acute pancreatitis. The initial sample included all convalescents at
the chosen clinic and time. The standard procedure for mail survey was applied with proper
thoroughness. The three clinical types of the disease, were represented at the study sample with
appropriate proportion: 61:41:40 [11]. The response rate $RR = 142 / 266 = 53.4\%$ was enough
great for a mail survey [23]. The clinical and demographic data for non-responders and responders
were quite similar, so the adjusting for non-response was unnecessary there [24]. Nevertheless, the
actual study has some limitations, but they can be overcome in future. First, the study group was
recruited from a single clinic only. Then, in this study the advantages of graphical methods were
omitted [25]. Finally, the cluster-wise look on the study group, with respect to the needed scope of
rehabilitation, remains on an intuitive level, in spite of many recognized formal methods [26].

Conclusions

The usefulness of bootstrap and multiple comparisons procedures as remedy on doubts about
correctness of ANOVA results was proved by the comparative analyses on the exemplary data
from questionnaire survey.

From practical point of view, the findings can be interpreted in such a way, that the
convalescents after acute pancreatitis should be considered as a heterogeneous population with
respect to the needed scope of rehabilitation, particularly with regard to PF (physical functioning),
and maybe also with regard to P (physical pain).

Conflict of Interest

The authors declare that they have no conflict of interest.

Authors’ Contributions

Izabela CHMIEL defined the aim of research and the design of experiment. Maciej
GORKIEWICZ participated in the design of the study and performed the statistical analysis. All
authors read and approved the final manuscript.

References

2. Chan C. Psychoeducational intervention, a critical review of systematic analyses. Clin Eff Nurs
2005:38.
health-related quality of life (HRQL) in survivors acute pancreatitis. Intensive Care Med
can improve the measurement of physical function by combining the modified health
7. Van Der Laan P, Verdooren LR. Classical Analysis of Variance methods and nonparametric
8. Walters SJ, Campbell MJ. The use of bootstrap methods for estimating sample size and
15. Górkiewicz M. Using propensity score with receiver operating characteristics (ROC) and bootstrap to evaluate effect size in observational studies. Biocybernetics and Biomedical Engineering 2009;29(4):41-61.