Applied Medical Informatics
Original Research Vol. 30, No. 1 /2012, pp: 9-15

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-Biala, 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 Jagiellonian 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 H_0 with known significance P; a hypothesis H_0 should be rejected, if $P \leq P_0$; where: $P_0 = 0.05$ assumed significance level. Now, let us consider K logically related null hypotheses H_{0k}; k = 1,2, ..., K; with known significance estimates P(k); k = 1,2, ..., K. Put the P(k)-values in descending order yielding: $P(1) \geq P(2) \geq \ldots \geq P(K-1) \geq P(K)$; a hypothesis H_{0k} should be rejected, if $P(k) \leq P_0(k)$; where reference levels of significance $P_0(k)$ are defined as either $P_{0B}(k)$, or $P_{0H}(k)$, or $P_{0R}(k)$, accordingly to a previously chosen testing procedure:

i) for basic Bonferroni procedure: $P_{0B}(k) = 1 / K$; k = 1,2, ..., K;

ii) for Hommel-Bonferroni procedure: $P_{0H}(k) = 1 / k$; k = 1,2, ..., K;

iii) for Rom-Bonferroni procedure: for k > 10: $P_{0R}(k) = P_{0H}(k) = 1 / k$; k = 11, K;

but for k ≤ 10 read the values of $P_{0R}(k)$ from Table 3, column P_{0R}; 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) without any compound calculations. The non-parametric bootstrap, applied in our study, can be considered as a non-parametric alternative for common parametric procedures, helpful in solving same questions, but without many troublesome assumptions [7]. This procedure draws each value of 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 are drawn several thousands times, aiming to obtain needed reliability of the estimates, [8, 22].

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 1. Descriptive statistics of quality of life of a study group

Domain of quality of life	Mean	SD	Median	Skew	Kurtozis
PF – physical functioning	0.65	0.27	0.70	-0.52	-0,52
RLP – role limitation physical	0.59	0.31	0.56	-0.11	-1,08
RLM – role limitation mental	0.62	0.32	0.58	-0.25	-1,12
SF – social functioning	0.60	0.26	0.63	-0.23	-0,67
MH – mental health	0.56	0.23	0.55	-0.20	-0,59
EV – energy/vitality	0.53	0.17	0.50	-0.28	0,43
P – physical pain	0.61	0.27	0.56	-0.09	-0,80
HP – general health perceptions	0.43	0.19	0.45	0.22	-0,19
CIH – change in health			0.50	0.24	-0,07
			max	0.24	0,43
SD^ = -0.043+0.524*Mean; R= 0.705; F = 6.9; p = 0.03.			min	-0.52	-1.12

SD – estimate of standard deviation, calculated as usual using given set of measurements of the domains of quality of life; SD^ = -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 where 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 $P_0 = 0.05$. 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

k	domain	groups	mean_1	mean_2	p(t)	P_{0B}	P_{0H}	P_{0R}
1	RLP	g61 - g41	0.57	0.63	0.19	0.01	0.05	0.05
2	RLM	g61 - g40	0.65	0.55	0.08	0.01	0.025	0.025
3	CIH	g41 - g40	0.52	0.44	0.08	0.01	0.017	0.017
4	MH	g61 - g40	0.61	0.51	0.03	0.01	0.013	0.013
5	SF	g61 - g41	0.55	0.55	0.03	0.01	0.010	0.010
6	P	g41 - g40	0.53	0.69	0.002	0.01	0.008	0.009
7	PF	g61 - g41	0.58	0.77	0.0001	0.01	0.007	0.007

p(t) – significance of the Student t test for two independent samples;
P_{0B}, P_{0H}, P_{0R} – 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 (group61, 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 significance of each separate domain in column p(ANOVA), but in columns P_{0B}, P_{0H} and P_{0R} 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 $P_0 = 0.05$ for PF (physical functioning) domain only, having a significance level p(ANOVA) = 0.002 less then reference level $P_{0B} = 0{,}006$.

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

k	Domain	p(ANOVA)	P_{0B}	P_{0H}	P_{0R}
1	HP	0.910	0.006	0.050	0.050
2	RLP	0.670	0.006	0.025	0.025
3	EV	0.460	0.006	0.017	0.017
4	CIH	0.320	0.006	0.013	0.013
5	RLM	0.290	0.006	0.010	0.010
6	SF	0.130	0.006	0.008	0.009
7	MH	0.080	0.006	0.007	0.007
8	P	0.026	0.006	0.006	0.006
9	PF	0.002	0.006	0.006	0.006

p(ANOVA) – significance of the one-way ANOVA for three group of patients;
P_{0B}, P_{0H}, P_{0R} – 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:

[RLM] > [RLP] > [PF = HP] > [CIH = MH] > [EV = SC = P].

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

	P	SC	EV	MH	CIH	HP	PF	RLP	RLM
P		0.82	0.84	0.99	0.98	1.00	1.00	1.00	1.00
SC	0.18		0.50	1.00	0.97	1.00	1.00	1.00	1.00
EV	0.16	0.50		1.00	0.97	1.00	1.00	1.00	1.00
MH	0.01	0.00	0.00		0.77	1.00	1.00	1.00	1.00
CIH	0.02	0.03	0.03	0.23		0.96	0.98	1.00	1.00
HP	0.00	0.00	0.00	0.00	0.04		0.74	1.00	1.00
PF	0.00	0.00	0.00	0.00	0.02	0.26		0.94	1.00
RLP	0.00	0.00	0.00	0.00	0.00	0.00	0.06		0.94
RLM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	

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 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

1. Barlow J, Wright C, Sheaby J, Turner A, Hainsworth J. (2002). Self-management approaches for people with chronic conditions: a review. Patient Educ Couns 2002;48(2):177-87.
2. Chan C. Psychoeducational intervention, a critical review of systematic analyses. Clin Eff Nurs 2005;9:101-11.
3. Campbell MJ. Complexity theory and complex interventions. ISCB'26 Annu Conf Abstr. 2005:38.
4. Zou G. Quantifying responsiveness of quality of life measures without an external criterion. Qual Life Res 2005;14(6):1545-52.
5. Halonen K, Pettila V, Leppaniemi A, Kemppainen E, Puolakkainen P, Haapiainen R. Long-term health-related quality of life (HRQL) in survivors acute pancreatitis. Intensive Care Med 2003;29:782-6.
6. Martin M, Kosinski M, Bjorner JB, Ware JE, MacLean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res 2007;16:647-60.
7. Van Der Laan P, Verdooren LR. Classical Analysis of Variance methods and nonparametric counterpart. Biom J 1987;29(6):635-55.
8. Walters SJ, Campbell MJ. The use of bootstrap methods for estimating sample size and analyzing health-related quality of life outcomes. Stat Med 2005;24(7):1075-102.

9. Zieffler A, Garfield J, delMas R, Reading C. A framework to support research on informal inferential reasoning. Statistics Education Research Journal [Internet] 2008 [cited 2010 Nov 14];7(2):40-58. Available from: URL: http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Zieffler.pdf
10. EBM Resource Center Web Page. The Section of Evidence Based Health Care of the New York Academy of Medicine [Internet]. 2011 [updated 2011 Nov 03; cited 2011 Nov 14]. Available from: http://www.nyam.org/fellows-members/ebhc/
11. Chmiel I. Determinants of quality of life following acute pancreatitis (in Polish: Determinanty jakości życia rekonwalescentów po przebytym ostrym zapaleniu trzustki) [PhD thesis]. Jagiellonian University of Krakow (Poland); 2011.
12. Ware JE, Kosinski M, Dewey JE. How to Score Version 2 of the SF-36 Health Survey. Lincoln, RI: Quality Metric Inc; 2000.
13. Chmiel I, Górkiewicz M, Czupryna A, Brzostek T. Age and gender as predictors of the physical ability among convalescents after acute pancreatitis. In: Fidecki W, Wysokiński M (Eds.) Selected problems of the aging population. Radomska Szkoła Wyższa: Radom, 2009;279-291.
14. Chmiel I, Górkiewicz M, Czupryna A, Brzostek T. Multiple comparisons procedures in analysis of health-related quality of life outcomes. In: Balcerar-Nicolau H, Bobrowski L, Doroszewski J, Kulikowski C. (Eds). Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, Warszawa, 2008;62-7.
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.
16. Lim TS, Loh WY. A comparison of tests of equality of variances. Computational Statistics & Data Analysis [Internet]. 1996 [cited 2011 Nov 14];22:287-301. Available from: URL: www.elsevier.com/locate/csda
17. StatSoft, Inc. (2011). Electronic Statistics Textbook. Tulsa: OK: StatSoft.; [online] 2011 [cited 2011 Nov 14] chapter ANOVA/MANOVA; Assumptions and Effects of Violating Assumptions. Available from: http://www.statsoft.com/textbook/.
18. Wilcox RR. Pairwise comparisons of dependent groups based on medians. Comput Stat Data Anal 2006;50:2933-41.
19. Hommel G. Bonferroni procedures for logically related hypotheses. J Stat Plan Inference 1999;82:119-28.
20. Rom DM. A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 1990;77:663-6.
21. Hommel G. Aesthetics and Power in Multiple Testing - a Contradiction?. 5th International Conference on Multiple Comparison Procedures 2007;55.
22. Martin MA. Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. Comput Stat Data Anal 2007;51:6321-42.
23. Kanuk L, Berenson C. Mail surveys and response rates: A literature review. J Marketing Res 1975;12:440-53.
24. Rowland ML, Forthofer RN. Adjusting for nonresponse bias in a health examination survey. Public Health Rep 1993;108:380-6.
25. Hochberg Y, Weiss G, Hart S. On graphical procedures for multiple comparisons. J American Statistical Association 1982;380:767-72.
26. Bolboacă SD. Assessment of Random Assignment in Training and Test Sets using Generalized Cluster Analysis Technique. Appl Med Inform 2011;2:9-14.