· 论著 ·

正畸疗效满意度主观评价一致性的探讨

宋广瀛,李巍然,耿 直,许天民△
(北京大学口腔医学院·口腔医院正畸科,北京 100081)

[摘 要] **目的**:研究正畸医师对正畸疗效满意度主观评价的一致性情况及其影响因素。**方法**:从北京大学口腔医学院正畸科2004年7月至2008年8月之间治疗完成的、具有完整临床资料的806个病例中,以安氏分类作为分层因素,随机抽取48个病例,将其治疗后模型、头颅侧位片和面像作为实验材料,进行随机临床试验。仍以安氏分类作为分层因素,将48个病例随机分为4组,请12位正畸专家依据所给治疗后的各项临床资料或者多项资料的组合,对每组12个病例依据治疗结果满意度高低进行排序判断。**结果**:本试验共得到Spearman相关系数1 584对,Spearman相关系数平均值为0.565 3±0.239 9。不同组别的疗效评价一致性差异有统计学意义($P<0.05$)。评价治疗后模型和头颅侧位片时,医师的一致性最高($P<0.001$),其余5项实验项目之间差异无统计学意义($P>0.05$)。在评价疗效满意度较好和较差的样本时,医师之间的一致性明显高于疗效满意度中等组($P<0.05$)。疗程在2.5年以上的样本得分的一致性高于疗程介于1.5年至2.5年之间的样本($P<0.05$)。不同安氏分类、不同年龄段和拔牙或不拔牙的样本得分的一致性差异无统计学意义($P>0.05$)。**结论**:医师之间对正畸疗效满意度主观评价呈现中度的一致性,各项因素中,样本的组别、实验项目、疗效满意度水平和疗程是影响医师主观评价一致性的主要因素,样本的安氏分类、年龄和是否拔牙对医师主观评价的一致性无统计学影响。

[关键词] 正畸学;治疗结果;统计学;医师
[中图分类号] R783.5 [文献标志码] A [文章编号] 1671-167X(2012)01-0103-05
doi:10.3969/j.issn.1671-167X.2012.01.022

Agreement analysis of subjective evaluation of orthodontic treatment outcome

SONG Guang-ying, LI Wei-ran, GENG Zhi, XU Tian-min△
(Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing 100081, China)

ABSTRACT **Objective**: To investigate the agreement of subjective evaluation of orthodontic treatment outcome and to analyze possible factors that may be related to it. **Methods**: As a randomized clinical trial, with Angle's classification as a stratification factor, our study contained 48 cases with integrity data, which were randomly extracted from 806 orthodontic treatment cases in Peking University School and Hospital of Stomatology during July 2004 and August 2008, and gathered post-treatment study casts, cephalometrics and photographs of the 48 cases as the research subjects. Similarly with Angle's classification as a stratification factor, the 48 cases were randomly divided into 4 groups. According to the monomial and combined subjects, 12 clinicians were asked to act as the raters to rank the 12 cases in each group. **Results**: Overall, there were 1 584 pairings between the raters in the examination of evaluation. The mean Spearman r was 0.565 3±0.239 9. Grouping factor was related to the agreement of subjective evaluation ($P<0.05$). In the third trial item-Post-M+P, the correlations were the greatest among the judge-pairs ($P<0.001$). The other five items were at the same agreement level. The level of orthodontic treatment outcome was a factor that influenced the agreement level of subjective evaluation ($P<0.05$). The score stability of the patients, whose treatment duration was longer than 2.5 years, was significantly higher than that of the patients whose treatment duration was between 1.5 years and 2.5 years ($P<0.05$). The following factors, such as Angle's classification, age of patients and whether the teeth was extracted or not, were the insignificant factors ($P>0.05$). **Conclusion**: The average correlations present a moderate agreement level. Grouping, experimental item, the length of treatment duration and the level of orthodontic treatment outcome are the factors that affect the agreement of subjective evaluation. Several factors including Angle's classification, age of patients and whether the teeth is extracted or not, do not affect the agreement of subjective evaluation.

KEY WORDS Orthodontics; Treatment outcome; Statistics; Physicians

正畸治疗涉及到牙齿及颌面部的美观,应该如何评价其疗效,国内尚缺乏客观的标准。从为数不

基金项目:卫生公益性行业科研专项(200802056)资助 Supported by Specific Research Project of Health Pro Bono Sector, Ministry of Health, China(200802056)
△Corresponding author's e-mail, tmxuortho@gmail.com
网络出版时间:2011-9-20 15:27:05 **网络出版地址**:http://www.cnki.net/kcms/detail/11.4691.R.20110920.1527.007.html

多的国外标准来看,如 1992 年欧洲建立的治疗标准指数(peer assessment rating,PAR)[1-2]、1998 年美国正畸医师委员会(American Board of Orthodontics,ABO)建立的客观评分标准(objective grading system,OGS)[3]以及 1998 年欧美九国联合建立的正畸治疗难度、结果、需要指数(index of complexity, outcome and need,ICON)[4],专家主观评价的均值都被当做权威的参照标准决定各项客观指标的纳入与否及权重大小[5-10]。本研究作为建立中国正畸疗效评价标准的第一步,主要探讨多个专家主观评价结果之间的一致性及其相关影响因素,从而为后续全国性大规模研究客观评价指标的建立提供一定的理论依据。

1 资料与方法

1.1 样本来源

从北京大学口腔医学院正畸科 2004 年 7 月至 2008 年 8 月之间治疗完成的、具有完整临床资料(包括病历记录,治疗前后的模型、X 线片和面殆像,其中病历记录由病案室提供,X 线片由放射科提供,治疗前后的模型和面殆像由正畸科提供)的 806 个病例中,随机抽取 48 个病例,保证所有病例治疗前后头颅侧位片是在同一台 X 光机上拍摄(确保治疗前后头颅侧位片可以进行重叠测量)。随机抽取样本的过程中以安氏分类作为分层因素,48 例样本由安氏Ⅰ类、安氏Ⅱ类和安氏Ⅲ类的患者各 16 名组成。所有入选病例均为汉族且无遗传性疾病。

本研究中患者年龄分布从 11 岁到 47 岁(成年患者 18 名,未成年患者 30 名),其中男性患者 16 名,女性患者 32 名,Ⅲ类的手术患者 5 名,非手术患者中拔牙病例 20 例,不拔牙病例 23 例,疗程从 9 个月到 36 个月不等。矫治医师涵盖正畸科 20 位医师,医师职称从住院医师到主任医师。

收集每位患者治疗后的记存模型、治疗后的头颅侧位片、治疗后的面像(包括正面像、侧面像和正面笑像)作为实验材料。

1.2 评判医师入选条件

评判医师为北京大学口腔医学院正畸科在职医师 12 名,均为受过正畸研究生专业培训或具有研究生导师资格的副高级职称以上的医师,且从事专业正畸医师工作 10 年以上。

1.3 研究方法

以安氏分类为分层因素,将 48 名患者随机分成 4 组,每组 12 名患者中包括安氏Ⅰ类、安氏Ⅱ类和安氏Ⅲ类的患者各 4 名,将患者的模型、头颅侧位片和面像编号按 4 组分别放置。原始资料的编号 1 至 48 分别对应了分组后的一个随机号,排列为 A1 ~ 12、B1 ~ 12、C1 ~ 12、D1 ~ 12。

设置 6 组实验:(1)单独评价治疗后模型(post-treatment models,Post-M);(2)单独评价治疗后头颅侧位片(post-treatment cephalometrics,Post-C);(3)单独评价治疗后面像(post-treatment photographs,Post-P);(4)评价治疗后模型和头颅侧位片(post-treatment models and cephalometrics,Post-M + C);(5)评价治疗后模型和面像(post-treatment models and photographs,Post-M + P);(6)评价治疗后模型、头颅侧位片和面像(post-treatment models,cephalometrics and photographs,Post-M + C + P)。

请专家依据所给治疗后的各项临床资料或者多项资料的组合判断正畸治疗结果满意度,将每组 12 个样本依据治疗结果满意度高低排序,“1”为治疗结果最满意,“12”为治疗结果最不满意。

两种排序方法如下:方法(一)为直接提取排序法,适用于实验(2)和实验(3),即按照治疗结果满意度由好到差的顺序依次提取样本。方法(二)为分组排序法,适用于模型及含有模型的组合资料,如实验(1)、实验(4)、实验(5)、实验(6)。模型属于三维实验材料,评判时需要从唇颊侧、咬合面和舌侧三个方向观察牙齿排列及咬合关系,专家很难一次性将 12 个样本排序。方法(二)分为两个步骤:首先,从 12 个样本中选出最佳的 4 个病例放在最满意组,然后选出最差的 4 个放在最不满意组,剩余的 4 个样本则属治疗效果中等组;其次,分别将治疗效果“好”组、治疗效果“差”组和治疗效果“中等”组中的 4 个样本按照满意度从最好到最差进行组内排序,这时允许跨组调整(例如:认为“好”组中的第 4 个样本满意度低于“中等”组中的第 1 个样本,可以进行调换)。

1.4 统计方法

采用 SPSS 16.0 统计软件,用 Spearman 秩相关分析法和 Kendall 和谐系数法研究医师之间主观评价的一致性情况;用两因素方差分析和卡方检验研究不同因素对正畸疗效主观评价一致性的影响。

2 结果

本实验针对 12 位医师对正畸疗效满意度评判的一致性进行统计分析,对于每项试验中的每组样本而言,12 位医师的评判结果两两之间进行 Spearman 相关分析,共得到 66 个相关系数。对于六项实验,四组样本而言,共得到 Spearman 相关系数 1 584 对,最大值为 0.99,最小值为 -0.48。相关系数变异范围比较大:大于 0.7(高度相关)的有

515 对,占 32.51%;介于 0.4 ~0.7(中度相关)之间的有 712 对,占 44.95%;介于 0 ~0.4(低度相关)之间的有 313 对,占 19.82%;小于 0(负相关)或等于 0(无相关性)的有 44 对,占 2.78%。所有组别 6 项实验的相关系数均值为 0.565 3 ±0.239 9,Kendall 和谐系数均值为 0.606 8 ±0.130 1,显示为中度相关。表 1 和表 2 分别列出了每组各项实验 Spearman 相关系数均值和 Kendall 和谐系数。

表 1　每组各项实验 Spearman 相关系数均值

Table 1　Mean value of Spearman *r*

Group	Post-M	Post-C	Post-P	Post-M + C	Post-M + P	Post-M + C + P
A	0.49 ±0.19	0.57 ±0.25	0.74 ±0.11	0.64 ±0.16	0.65 ±0.14	0.63 ±0.16
B	0.54 ±0.22	0.50 ±0.22	0.35 ±0.27	0.69 ±0.14	0.45 ±0.24	0.58 ±0.17
C	0.67 ±0.16	0.68 ±0.13	0.70 ±0.17	0.74 ±0.16	0.72 ±0.13	0.77 ±0.13
D	0.53 ±0.19	0.34 ±0.25	0.48 ±0.19	0.42 ±0.26	0.34 ±0.26	0.28 ±0.26

表 2　每组各项实验 Kendall 和谐系数

Table 2　Kendall's W value of each group

Group	Post-M	Post-C	Post-P	Post-M + C	Post-M + P	Post-M + C + P
A	0.54	0.61	0.76	0.67	0.68	0.66
B	0.59	0.55	0.41	0.71	0.50	0.61
C	0.70	0.71	0.73	0.77	0.75	0.79
D	0.58	0.40	0.53	0.47	0.40	0.34

Kendall's W, $P<0.05$.

对于不同组别的样本,疗效评价一致性差异有统计学意义,其中 C 组和 A 组一致性较高,B 组和 D 组一致性较低($P<0.001$)。各组别相关系数分布情况卡方检验结果如表 3 所示。表 4 显示了每组各项实验中 Spearman 相关系数为负值的配对个数,负值数目越多,一致性越低。

表 3　各组别相关系数分布情况*

Table 3　Distribution of Spearman *r* of each group*

Group	$r<0.4$	$0.4\leqslant r<0.7$	$r\geqslant 0.7$
A	52	192	152
B	109	204	83
C	12	146	238
D	184	170	42

* The unit of the following column stands for the count of Spearman *r*. Chi-square test, $\chi^2=392.628$, $P<0.001$.

表 4　每组各项实验 Spearman 相关系数为负值的个数*

Table 4　Count of negative Spearman *r**

Group	Post-M	Post-C	Post-P	Post-M + C	Post-M + P	Post-M + C + P
A	0	2	0	0	0	0
B	1	0	9	0	4	0
C	0	0	0	0	0	0
D	0	6	0	5	6	11

* The unit of the following column stands for the count of negative Spearman *r*.

卡方检验结果显示,评价治疗后模型和头颅侧位片时,医师的一致性相对较高($P<0.05$),其余 5 项实验项目之间差异无统计学意义($P>0.05$),各项实验相关系数分布情况卡方检验的结果如表 5 所示。疗程在 2.5 年以上的样本得分的一致性高于疗程介于 1.5 年至 2.5 年之间的样本($P<0.05$)。在好、中、差 3 个疗效满意度水平组中,卡方检验结果显示,在评价"好"组和"差"组的样本时,医师之间的一致性显著高于中等组($P<0.05$)。此外,在评价 3 种安氏分类的样本、不同年龄段的样本、拔牙和不拔牙病例时,医师之间的一致性均相同($P>0.05$)。

表 5　各项实验相关系数分布情况*

Table 5　Distribution of Spearman *r**

	$r<0.4$	$0.4\leqslant r<0.7$	$r\geqslant 0.7$
Post-M	60	127	77
Post-C	70	129	65
Post-P	58	117	89
Post-M + C	44	107	113
Post-M + P	65	115	84
Post-M + C + P	60	117	87

* The unit of the following column stands for the count of Spearman *r*. Chi-square test, $\chi^2=23.973$, $P<0.05$.

3　讨论

3.1　评价一致性的统计学方法回顾

3.1.1　Kappa 一致性分析　在 Kappa 一致性分析的相关研究中,1975 年 Bowden 等[11]和 Helm [12]的研究表明,正畸医师对患者正畸治疗需要的主观评价一致性的 Kappa 值在 0.8 左右,显示正畸医师在治疗需要性的问题上观点比较稳定且能够达成一定的共识。1992 年 Richmond 等[1]对 PAR 指数的研究表明,医师自身评价 Kappa 值为 0.39 ~0.87,自

身一致性的变异较大;医师之间的 Kappa 值为 0.43 ~0.58,具有中度的一致性。1998 年,Richmond 等[13]研究了 ICON 指数,97 位医师对"正畸治疗效果能否接受"这一问题的主观评价自身一致性均值为 0.50 ±0.26,对"改善程度大小"这一问题的主观评价自身一致性均值为 0.40 ±0.18。2003 年,Savastano 等[14]也对 ICON 指数进行了分析,结果显示,在治疗结果能否接受的问题上,医师自身评价 Kappa 值仅为 0.07,医师之间的 Kappa 值为 0.18;在改善程度的问题上,医师自身评价 Kappa 值为 0.47,医师之间的 Kappa 值为 0.04 。

3.1.2 相关分析评价一致性 在相关分析的一致性研究中,1987 年 Evans 等[15]曾采用相关分析的方法评价医师对样本美观情况主观评价的一致性,5 名正畸医师的 Spearman 相关系数介于 0.81 ~0.97 之间。同年,Woolass 等[16]的研究结果显示,3 位评判者的自身相关系数分别为 0.725、0.856 和 0.754,三者测评均值的自身相关系数为 0.892。1995 年,DeGuzman 等[2]对 PAR 指数的进一步研究表明,两次评价结果的自身相关系数高达 0.98。2009 年,许天民等[17]的研究结果显示,1 980 对医师间相关系数均为正值,相关系数的变异较大,从 0.004 到 0.960,中位数为 0.540,其中 18.7% 低于 0.4,41.0% 低于 0.5,68.8% 低于 0.6,91.6% 低于 0.7,只有 8.4% 高于 0.7。在评价中国患者时,中国医师之间的一致性高于美国医师,差异具有统计学意义;在评价美国患者时,美国医师的一致性稍高于中国医师,但差别不具有统计学意义。刘妍等[18]的研究以每一位患者作为研究对象,结果显示对于中国患者,两国医师评分均值的相关系数为 0.86;对于美国患者,两国医师评分均值的相关系数为0.92。

3.2 评价一致性的影响因素

正畸医师对 3 种不同的临床资料进行判断时,一致性存在较大的变化范围,本研究结果显示,医师之间对治疗满意度的判断总体上显示为中度一致性,其中比较明显的影响因素主要有以下 4 个方面。

3.2.1 面像评价一致性的影响因素 从 Spearman 相关系数的分析中可以看出,B 组面像评价的一致性较差。进一步对 B 组 12 个样本的面像进行分析,发现医师在对其中 4 个样本进行评价时一致性较差(图 1 ~4)。其中,图 1(B2)的女患者肤色偏黑,侧貌颏部形态较为突出;图 2(B5)的女患者发型等特征导致其女性化特征不明显,且正面微笑像稍显不自然;图 3(B6)为男性患者,其面部男性化特征较为明显,将其选入"好"组的 4 位医师中,有 3 位是女性,将其选入"差"组的 4 位医师中,有 3 位是男性,图 4(B7)的女患者正面像比较具有吸引力,但是侧貌明显下颌后缩,推测将其选入"好"组的医师是将其正面特征作为主要评判依据,而将其选入"差"组的医师则可能是将其侧貌特征作为主要评判依据。从正畸的专业角度考虑,在评判正面像时,医师主要关注面部对称性和颅颌面高度、宽度的比例关系;评判正面笑像时,医师主要关注唇齿龈的关系是否协调;评判侧面像时,医师主要关注面型特征、上下唇软组织形态及颏部形态等,但当将 3 张面像放在一起判断时,医师对这 3 张面像认知权重的差异可能会影响医师之间的一致性。

图 1 B2 患者面像 **图 2** B5 患者面像
图 3 B6 患者面像 **图 4** B7 患者面像
Figure 1 Post-P of patient B2 **Figure 2** Post-P of patient B5
Figure 3 Post-P of patient B6 **Figure 4** Post-P of patient B7

3.2.2 头颅侧位片评价一致性的影响因素 从

Spearman 相关系数的分析中可以看出,D 组头颅侧位片评价的一致性较差,进一步对 D 组 12 个样本的头颅侧位片进行分析,发现该组中有 4 个样本是正畸正颌联合治疗的病例,医师在对这 4 个样本进行评价时一致性较差。在评价正畸正颌联合治疗病例的头颅侧位片时,有些医师认为即使患者治疗后的状态较为理想,但其面型及骨骼形态的改善主要是正颌手术带来的,并不是正畸治疗的结果,因此将手术病例放入疗效不满意组;有些医师则认为以治疗结束的各项指标作为依据,该患者的确获得了令人满意的疗效,因而将手术病例放入疗效满意组。因此,正畸正颌联合治疗的病例,头颅侧位片的评价一致性较低。

3.2.3　组合资料评价一致性的影响因素　组合资料评价的一致性以其中各个单项资料评价的一致性为基础,受到各个组合项目权重大小的影响。模型、头颅侧位片和面像 3 种资料中,面像的评价受到诸多主客观因素的影响[17-18],难以形成统一的标准。模型和头颅侧位片是专业性资料,医师在对这两项资料进行评价的时候主要是从专业的角度出发,在受教育程度、临床工作经验及社会背景等方面因素相似的情况下,专业医师的评判具有较为一致的标准。本研究统计结果显示,在评价治疗后模型和头颅侧位片时,医师之间的相关系数大多数分布在高度相关和中度相关的水平,一致性显著高于其余 5 项实验。除外正畸正颌联合治疗样本较多的 D 组,其余 3 组的模型、头颅侧位片和面像 3 种组合资料的评价项目中,医师评判的一致性也比较高,说明在 3 种组合资料同时出现时,医师从专业的角度出发,将权重主要集中在模型和头颅侧位片两项资料上,面像只作为一种辅助资料,并不起主要作用。

3.2.4　其他影响因素　研究结果显示,患者年龄、错𬌗分类、是否采用了拔牙治疗手段等因素并不影响专家判断的一致性,但在满意度水平被判断为"好"组和"差"组中,专家判断的一致性明显好于中等水平组。

参考文献

[1] Richmond S, Shaw WC, Buchanan IB, et al. The development of the PAR index (Peer Assessment Rating): reliability and validity [J]. Eur J Orthod, 1992, 14(2): 125-139.

[2] DeGuzman L, Bahiraei D, Vig KW, et al. The validation of the Peer Assessment Rating index for malocclusion severity and treatment difficulty [J]. Am J Orthod Dentofacial Orthop, 1995, 107(2): 172-176.

[3] Casko JS, Vaden JL, Kokich VG, et al. Objective grading system for dental casts and panoramic radiographs. American Board of Orthodontics [J]. Am J Orthod Dentofacial Orthop, 1998, 114(5): 589-599.

[4] Daniels C, Richmond S. The development of the Index of Complexity, Outcome and Need (ICON) [J]. J Orthod, 2000, 27(2): 149-162.

[5] Louwerse TJ, Aartman IH, Kramer GJ, et al. The reliability and validity of the Index of Complexity, Outcome and Need for determining treatment need in Dutch orthodontic practice [J]. Eur J Orthod, 2006, 28(1): 58-64.

[6] Summers CJ. The occlusal index: a system for identifying and scoring occlusal disorders[J]. Am J Orthod, 1971, 59(6): 552-567.

[7] Pangrazio-kulbersh V, Kaczynski R, Shunock M. Early treatment outcome assessed by the Peer Assessment Rating index [J]. Am J Orthod Dentofacial Orthop, 1999, 115(5): 544-550.

[8] Deans J, Playle R, Durning P, et al. An exploratory study of the cost-effectiveness of orthodontic care in seven European countries [J]. Eur J Orthod, 2009, 31(1): 90-94.

[9] 李巍然, 胡　炜, 谷　岩, 等. 正畸病例治疗结果的初步评价[J]. 口腔正畸学, 2008, 15(3): 104-107.

[10] Airton OA. Occlusal indexes as judged by subjective opinions [J]. Am J Orthod Dentofacial Orthop, 2008, 134(5): 671-675.

[11] Bowden DE, Davies AP. Inter- and intraexaminer variability in assessment of orthodontic treatment need [J]. Community Dent Oral Epidemiol, 1975, 3(4): 198-200.

[12] Helm S. Intra-examiner reliability of epidemiologic registrations of malocclusion [J]. Acta Odontol Scand, 1977, 35 (3): 161-165.

[13] Richmond S, Daniels CP. International comparisons of professional assessments in orthodontics: Part 2-treatment outcome [J]. Am J Orthod Dentofacial Orthop, 1998, 113(3): 324-328.

[14] Savastano NJ, Firestone AR, Beck FM, et al. Validation of the complexity and treatment outcome components of the Index of Complexity, Outcome, and Need (Icon)[J]. Am J Orthod Dentofacial Orthop, 2003, 124 (3): 244-248.

[15] Evans R, Shaw W. Preliminary evaluation of an illustrated scale for rating dental attractiveness [J]. Eur J Orthod, 1987, 9 (4): 314-318.

[16] Woolass KF, Shaw WC. Validity and reproducibility of rating dental attractiveness from study casts [J]. Br J Orthod, 1987, 14(3): 187-190.

[17] Xu TM, Edward LK, Liu Y, et al. Facial attractiveness: Ranking of end-of-treatment facial photographs by pairs of Chinese and US orthodontists[J]. Am J Orthod Dentofacial Orthop, 2008, 134(1): 74-84.

[18] Liu Y, Korn EL, Oh HS, et al. Comparison of Chinese and US orthodontists' averaged evaluations of "facial attractiveness" from end-of-treatment facial photographs [J]. Am J Orthod Dentofacial Orthop, 2009, 135(5): 621-634.

(2011-07-14 收稿)
(本文编辑:赵　波)