· 论著 ·

中国正畸专家对错殆畸形严重程度的主观判断一致性研究

刘思琦[1],沈 刚[2],白 丁[3],周 洪[4],厉 松[5],陈文静[6],王大为[7],李巍然[1],耿 直[8],许天民[1△]
(1.北京大学口腔医学院·口腔医院正畸科,北京 100081;2.上海交通大学医学院·第九人民医院口腔正畸科,上海 200011;3.四川大学华西口腔医学院·口腔医院正畸科,成都 610041;4.西安交通大学医学院·口腔医院正畸科,西安 710049;5.首都医科大学口腔医学院·北京口腔医院正畸科,北京 100006;6.南京医科大学口腔医学院·江苏省口腔医院正畸科,南京 210029;7.中山大学光华口腔医学院·口腔医院正畸科,广州 510055;8.北京大学数学科学学院,北京 100871)

[摘 要] **目的**:分析中国正畸专家主观判断错殆畸形严重程度的一致性。**方法**:从参与中国正畸疗效评价标准研究的6所院校所提供的具有完整病例资料的完成正畸治疗的2 383例病例中以院校来源和安氏分类作为分层因素随机抽取120例病例,由全国69位正畸专家,根据病例治疗前的模型、头颅侧位片、面像、曲面断层片和病历基本资料主观判断每例错殆畸形的严重程度,选取轻度、中等偏轻、中等、中等偏重和重度5个严重程度级别之一作为每例的判断结果,并对主观判断结果进行专家自身可靠性和专家之间评价一致性的分析。**结果**:采用加权Kappa检验分析专家自身可靠性,8.33%的专家自身一致性达到极好水平(Kappa≥0.81);78.33%的专家自身一致性达到好及以上水平(Kappa≥0.61);96.67%的专家自身一致性达到中等及以上水平(Kappa≥0.41)。采用组内相关系数(intra-class correlation coefficient,ICC)检验分析专家之间评价一致性($r=0.989$,$P<0.01$),专家之间评价一致性非常好。**结论**:中国正畸专家依据病例治疗前的模型、头颅侧位片、面像、曲面断层片和病例基本信息对错殆畸形严重程度进行主观判断的专家自身可靠性好,且专家之间评价一致性高,这为建立错殆畸形严重程度的客观分级系统奠定了良好的基础。

[关键词] 错殆;疾病严重程度;判断;可重复性,结果;正畸学
[中图分类号] R783.5 [文献标志码] A [文章编号] 1671-167X(2012)01-0098-05
doi:10.3969/j.issn.1671-167X.2012.01.021

Consistency of the subjective evaluation of malocclusion severity by the Chinese orthodontic experts

LIU Si-qi[1], SHEN Gang[2], BAI Ding[3], ZHOU Hong[4], LI Song[5], CHEN Wen-jing[6], WANG Da-wei[7], LI Wei-ran[1], GENG Zhi[8], XU Tian-min[1△]
(1. Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing 100081, China; 2. Department of Orthodontics, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200011, China; 3. Department of Orthodontics, West China School / Hospital of Stomatology Sichuan University, Chengdu 610041, China; 4. Department of Orthodontics, Xi'an Jiaotong University Stomatology Hospital, Xi'an 710049, China; 5. Department of Orthodontics, Beijing Stomatological Hospital, Capital Medical University, Beijing 100006, China; 6. Department of Orthodontics, Jiangsu Stomatological Hospital Affiliated Nanjing Medical University, Nanjing 210029, China; 7. Department of Orthodontics, Guanghua School of Stomatology, Hospital of Stomatology, Sun Yan-sen University, Guangzhou 510055, China; 8. School of Mathematical Sciences, Peking University, Beijing 100871, China)

ABSTRACT **Objective**: To assess the consistency of the subjective evaluation of malocclusion severity by the Chinese orthodontic experts. **Methods**: Sixty-nine Chinese orthodontic experts subjectively evaluated the malocclusion severity for 120 cases which were selected randomly from 6 University orthodontic clinics by checking each case's pretreatment records including study cast, lateral head film, panoramic radiograph, facial photographs and patient chart. Each orthodontist was asked to independently rate the severity of every case into five grades: mild, mildly moderate, moderate, severely moderate and severe. Rating data was finally gathered to evaluate the intra-judge's reliability and the inter-judges' consistency. **Results**: Weighted Kappa test revealed that 8.33% orthodontists showed excellent intra-judge's reliability (Kappa≥0.81), 78.33% orthodontists showed good intra-judge's reliability (Kappa≥0.61) and 96.67% specialists displayed general intra-judge's reliability (Kappa≥0.41). And intra-class correla-

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:2012-1-4 14:26:29 **网络出版地址**:http://www.cnki.net/kcms/detail/11.4691.R.20120104.1426.001.html

tion coefficient demonstrated a high level of inter-judges' consistency ($r = 0.989$, $P < 0.01$). **Conclusion**: Good intra-judge's reliability and inter-judges' consistency can be demonstrated in the subjective evaluation of malocclusion severity by the Chinese orthodontic experts, which could the basis for establishing the objective grading system of malocclusion severity.
KEY WORDS Malocclusion; Severity of illness index; Judgment; Reproducibility of results; Orthodontics

由于已知或未知的、先天或后天的原因引起的殆、颌及颅面的畸形称之为错殆畸形[1]。错殆畸形严重程度是指错殆畸形偏离理想颅颌面结构和关系的程度,并可进一步表示为其在口腔形态和功能上给患者带来损伤的程度[2-3]。通过评价错殆畸形严重程度能反映正畸医师对不同错殆畸形特征的认识。

在日常正畸临床工作中,正畸医师不仅要对病例的错殆畸形严重程度做出判断,并且还要在矫治方案的设计过程中评价正畸病例的治疗难度或者复杂程度。"正畸治疗难度"是指治疗过程中为建立理想颅颌面关系所需要付出的努力[3-4];而"正畸治疗复杂程度"则被定义为能导致治疗结果满意度下降的多种因素的组合体[5]。Richmond 等[6]认为可以将正畸治疗难度和正畸治疗复杂程度看作同义词,它们均是对正畸医师在治疗中运用技术和所付出努力的衡量。与错殆畸形严重程度的评价相比而言,对正畸治疗难度或者复杂程度的评价受到正畸治疗过程中多种因素影响,如诊疗环境和水平、患者依从性、矫治器选择等[7],而非仅仅反映病例本身的错殆畸形特征。

大部分研究人员认为殆特征在评价过程中起主要作用[8],因此在研究中只用模型评价病例的严重程度和难度[3,5,9],然而,错殆畸形可以有多种临床表现,其可简单地表现为牙齿大小、形态及排列异常,或者上下牙弓的殆关系异常,也可进一步表现为上下颌骨大小、形态及相互关系异常等[1]。Gramling[10]分别总结了治疗成功和失败的安氏Ⅱ类病例的特征,形成由 5 个头影测量指标和相应正常值范围组成的可能性指数(probability index),与全牙列间隙分析一并形成 Merrifield 的个体化诊断系统(differential diagnosis)用于临床诊断。可见,头颅侧位片提供的信息也能够有效地反映错殆畸形的特征,这在 Pae 等[3]的研究中得到了证实。除此以外,如果在模型的基础上添加其他临床参考资料也将有助于完善对错殆畸形严重程度的主观判断研究[2,9]。

虽然目前临床工作中针对殆特征、骨面型特征和软组织特征的各项单独测量指标可用于评价某一方面错殆畸形特征偏离正常的程度,但仅以颅、颌、面任何一方面特征的正常标准作为错殆畸形严重程度的判断标准难以整体地反映错殆畸形的严重程度。本研究邀请中国正畸专家依据病例的模型、头颅侧位片、面像、曲面断层片和病历基本信息对病例的错殆畸形严重程度做出主观判断,以期反映中国正畸专家对错殆畸形严重程度的主观评价一致性。

1　资料与方法

1.1　样本选择

本研究样本来自于中国正畸疗效评价标准研究(下文简称为"疗效评价标准研究")中的一部分。收集由北京大学口腔医学院、四川大学口腔医学院、第四军医大学口腔医学院、首都医科大学口腔医学院、南京医科大学口腔医学院和武汉大学口腔医学院 6 所院校正畸科提供的 2006 至 2009 年完成正畸治疗的具有完整临床资料(治疗前后面殆像 8 张、头颅侧位片、曲面断层片、记存模型及病历概要)的 2 383 例病例。

以院校来源和安氏分类作为分层因素,随机抽取 288 例,并由各院校根据随机抽样结果提供相应的病例资料构成初始样本,其中每所院校 48 例,分别包括安氏Ⅰ、Ⅱ、Ⅲ类病例各 16 例,将此初始样本用于各院校进行错殆畸形严重程度主观判断研究的预实验。根据预实验结果,对最终样本量进行统计推断,并兼顾疗效评价标准研究其他部分研究样本的要求,按照下述方法为最终实验样本进行重新抽样:(1)从初始样本中剔除会影响治疗后结果满意度评价的 18 例病例(包含正颌手术病例和年龄 40 岁以上病例,预实验结果显示这两类病例对治疗结果满意度评价造成偏倚);(2)以院校来源和安氏分类作为分层因素,从 270 例病例中随机抽取 108 例病例(每所院校来源病例各 18 例,分别包括安氏Ⅰ、Ⅱ、Ⅲ类病例各 6 例)。在此基础上,为保证错殆畸形严重程度主观判断研究样本的涵盖范围,重新纳入上述为兼顾治疗后结果满意度评价实验要求而剔除的 18 例样本,与此前抽样所得的 108 例病例合并形成最终样本(126 例),由于实验过程中有 6 例样本的实验资料磨损未能修复,因此,本研究实际纳入的样本量为 120 例(表 1)。

表 1 样本年龄、性别及安氏分类分布

Table 1 Age, sex and Angle's Classification distribution of the sample

	n(%)
Age	
10 – 12	26 (21.67%)
13 – 15	38 (31.67%)
16 – 18	13 (10.83%)
> 19	43 (35.83%)
Gender	
Male	32 (26.67%)
Female	88 (73.33%)
Angle's Classification	
Ⅰ	38 (31.67%)
Ⅱ	40 (33.33%)
Ⅲ	42 (35.00%)

1.2 专家甄选

本研究由课题组统一制定专家资格条件,具体如下:(1)受过正畸研究生专业培训或具有研究生导师资格;(2)大学医院正畸科从事专业正畸医师工作 10 年以上;(3)副高职称以上(双职称优先,单职称者正高优先);(4)有地区代表性,覆盖尽可能广泛的地区(院校优先)。各院校根据上述专家资格条件分别推荐 12 名专家,共 72 位专家参与本研究。

由于日程安排冲突等原因,72 位专家中有 65 位专家按照不同的地区分布分别在上述 6 所院校参与了预实验(北京地区:北京大学;华北及东北地区:首都医科大学;西南地区:四川大学;西北地区:第四军医大学;东南地区:南京医科大学;华南地区:武汉大学);69 位专家参与了在北京进行的最终实验,这其中共有 60 位专家参与了前后两次实验。

1.3 主观评价

本研究请每位正畸专家独立依据完整的临床资料(模型、头颅侧位片、面像、曲面断层片和病历基本信息)综合判断各病例的错殆畸形严重程度。主观评价开始前,研究负责人向所有参与实验的正畸专家辨析错殆畸形严重程度和正畸治疗难度的概念,然后请每位专家将本研究所提供的病例信息与自己多年的知识积累和临床经验相比较,参考本研究提供的 5 个等级严重程度(轻度、中等偏轻、中等、中等偏重和重度)判断病例的错殆畸形严重程度。本研究要求专家在判断时,应首先在轻度、中等、重度三个等级中进行选择,若无合适再考虑中等偏轻和中等偏重两个等级。

1.4 统计分析

研究人员使用数字 1 ~ 5 依次代表错殆畸形严重等级从轻到重的 5 个级别,由 5 名研究人员在 3 个月期间协作完成 69 名专家对 120 例病例的共 8 280 条结果的录入工作。采用加权 Kappa 检验分析专家自身可靠性;采用组内相关系数(intra-class correlation coefficient,ICC)检验分析专家之间评价一致性。

2 结果

2.1 专家自身可靠性分析

本研究中参与了前后两次实验的 60 位专家及 120 例样本的地区分布见表 2,同一地区分布的专家在预实验结束后,又在最终实验中对各地区的病例进行了重复判断。为分析专家自身可靠性,本研究分别对每位专家在前后两次实验中重复判断病例的结果进行加权 Kappa 检验(表 3),并按照 Landis 等[11]制定的 Kappa 界值评价专家自身可靠性(表 4)。

表 2 专家自身可靠性分析样本情况

Table 2 Component of the sample for intra-judge's reliability assessment

	Number of specialists	Number of cases
Peking University	11	25
The Fourth Military Medical University	11	21
Wuhan University	11	20
Sichuan University	10	18
Capital Medical University	9	18
Nanjing Medical University	8	18
Total	60	120

表 3 专家自身可靠性加权 Kappa 检验结果

Table 3 Weighted Kappa test result for intra-judge's reliability

Judge's code	Weighted Kappa value	Judge's code	Weighted Kappa value	Judge's code	Weighted Kappa value
9	0.90	22	0.75	10	0.64
24	0.90	29	0.75	58	0.64
57	0.88	11	0.74	2	0.64
34	0.87	39	0.74	42	0.63
21	0.85	56	0.73	38	0.63
31	0.84	18	0.73	46	0.62
53	0.83	28	0.72	62	0.62
7	0.82	44	0.72	12	0.60
48	0.82	32	0.72	4	0.60
41	0.82	1	0.71	15	0.59
13	0.81	40	0.70	49	0.59
27	0.81	54	0.70	59	0.58
25	0.81	35	0.70	19	0.55
26	0.80	16	0.69	55	0.55
37	0.80	43	0.68	20	0.52
52	0.79	30	0.68	23	0.51
3	0.79	17	0.66	33	0.51
61	0.79	5	0.65	60	0.46
45	0.76	8	0.65	47	0.37
36	0.76	6	0.65	14	0.35

表 4　加权 Kappa 检验结果分布

Table 4　Distribution of Weighted Kappa test result for intra-judge's reliability

Kappa Value	Coincidence level	Numbers of specialists	Percentage (%)	Accumulative percentage (%)
0.81 – 1.0	Extremely excellent	11	18.33	18.33
0.61 – 0.8	Excellent	36	60.00	78.33
0.41 – 0.6	Good	11	18.33	96.67
0.21 – 0.4	General	2	3.33	100.00

2.2　专家之间评价一致性分析

应用 ICC 检验[12-14]分析预实验中各地区专家分别对本地区 48 例病例的判断一致性及最终实验中 69 位专家对 120 例病例的判断一致性(表 5)。预实验和最终实验中专家之间评价一致性均非常好($R>0.75$,$P<0.01$),且最终实验结果与预实验结果相比,专家之间的评价一致性有所提高。

表 5　专家之间评价一致性分析结果

Table 5　ICC test result for inter-judges' reliability

Study	*R*
Pilot study	
Peking University	0.957
The Fourth Military Medical University	0.952
Wuhan University	0.921
Sichuan University	0.919
Capital Medical University	0.924
Nanjing Medical University	0.925
Final study	0.989

3　讨论

3.1　专家自身可靠性

本研究在各地区分别进行的预实验和最终实验间隔 2 ~ 4 个月,利用专家在这两次实验中对重复样本进行主观判断的结果分析专家自身可靠性。DeGuzman 等[9]曾从 200 例总体样本中随机抽取 50 例样本用于首次实验结束后的 4 周再次评价,以检测专家自身可靠性,结果显示专家自身可靠性高($r=0.98$);此外,Pae 等[3]和 Arruda 等[15]的研究也得出类似结果,与本研究一致。但也有研究得出相反结果,如 Richmond 等[16-17]请 74 位专家对 272 例病例构成的分层随机抽样样本进行错𬌗严重程度主观评价以检验 PAR 指数(peer assesment rating index)测量准确性,该研究使用了直接包含在评价样本中的 40 例重复样本进行检测,结果显示专家自身可靠性并不理想(Kappa 值为 0.39 ~ 0.81)。

Richmond 等[16]认为专家自身可靠性并不理想的原因可能包括:一定程度上的记录错误、专家评价疲劳及评价顺序偏倚。结果记录错误并不能通过不同的实验设计完全避免。从专家自身可靠性评价方法来看,抽取一定量小样本待一定时间间隔后请专家再次评价这种方法比将重复样本放于总样本中请专家一并评价更能有效地降低专家评价疲劳的程度;本研究即请专家于预实验结束 2 ~ 4 个月后在最终实验中评价与预实验部分重复的样本以检测专家自身可靠性,有效地降低了专家评价疲劳。此外,本研究在预实验结束后对样本进行了重新抽样分组,有效地避免了专家评价重复样本时可能存在的评价顺序偏倚。

3.2　专家之间评价一致性

本研究结果显示,在预实验和最终实验两次评价中专家之间均高度一致(r 值为 0.919 ~ 0.989),与 Pae 等[3]的研究结果近似。但 Richmond 等[17]的研究却显示专家之间一致性仅达到中度水平(Kappa 值为 0.44 ~ 0.59),这是因为参与该研究的专家由正畸咨询医师、正畸专业医师和不具备正畸医师资质的全科医师或社区医师组成,其研究结果显示社区医师之间和全科医师之间的判断一致性最差。Pae 等[3]的研究及本研究均请正畸专业医师参与评价,依照本研究的专家资格标准,专家之间具有近似的教育背景和一定年限的专科临床经验积累,其正畸专业知识水平和对错𬌗特征的认知水平更为相近。

3.3　错𬌗畸形严重程度评价和正畸治疗难度评价之间的联系与区别

本研究请正畸专家仅仅对错𬌗畸形严重程度进行主观评价,并未请正畸专家同时对错𬌗畸形严重程度和正畸治疗难度进行主观评价。有研究请 30 位正畸医师依据治疗前临床资料评价 6 例病例的错𬌗畸形严重程度和正畸治疗难度,结果表明,对正畸医师而言,错𬌗畸形严重程度和正畸治疗难度不同,仅凭错𬌗特征并不能有效地预测正畸治疗难度[18]。DeGuzman 等[9]的研究也证实了这一点,该研究通过检验 PAR 指数测量指标是否能有效地反映错𬌗畸形严重程度和正畸治疗难度,结果显示 PAR 指数的测量分值与错𬌗畸形严重程度的相关性($r=0.83$)高于其与正畸治疗难度的相关性($r=0.68$)。这可能与两方面原因有关:首先,对治疗前的病例资料进行正畸治疗难度的评价难以体现患者个体在生物学或者社会心理方面对正畸治疗反应的差异;其次,与治疗相关的因素可能影响正畸医师对治疗难度的判断,比如正畸医师可能会认为通过拔牙矫治即能改善的拥挤病例的治疗难度低于需要纠正覆盖

或者中线异常的病例,这些患者相关因素或者治疗相关因素均会影响正畸治疗难度评价。因此,除了错𬌗特征以外,Cassinelli 等[2]将患者相关因素和治疗相关因素也纳入到正畸治疗难度评价的研究中,使用 PAR 指数和 IOTN 指数(index of orthodontic treatment need)对治疗前咬𬌗状况进行测量以分析是否能够依据治疗前咬𬌗状况评价病例治疗难度,并且设计了包含上述三方面因素的调查问卷,以了解当正畸医师对完成的病例进行难易评价时可能影响正畸治疗难度评价的因素,其结果显示,与治疗难度低病例组相比,在治疗难度高病例组中,不仅其治疗前的模型指数测量分值更高,反映咬𬌗状况偏离正常程度更严重,而且其在治疗中出现口腔卫生维护不佳的次数和患者依从性问题也明显增加。除此以外,治疗方案变化、拔牙数量、复诊次数和疗程长短等治疗相关因素在治疗难度高的病例组中出现频次均明显增加[2]。

可见,与错𬌗畸形严重程度相比,对正畸治疗难度的评价受到错𬌗特征、患者相关因素和治疗相关因素等多种因素共同影响,各因素相互影响作用,导致专家对正畸治疗难度的评价难以达成一致,这为探究影响正畸治疗难度的因素带来困难[6]。本研究请专家对病例进行错𬌗畸形严重程度判断,除错𬌗特征本身以外,避免了其他因素的混杂影响,这可能是本研究专家自身可靠性及专家之间一致性均很好的原因之一。

中国正畸专家依据病例治疗前的整体临床资料,包括模型、头颅侧位片、面像、曲面断层片和病例基本信息,对错𬌗畸形严重程度进行主观判断,专家自身可靠性好,并且专家之间评价一致性高。本研究所得到的主观判断结果能够有效地反映中国正畸专家对错𬌗畸形特征的认识,并且中国正畸专家对错𬌗畸形严重程度主观判断的高度一致为建立错𬌗畸形严重程度的客观分级系统奠定了良好的基础。

参考文献

[1] 林久祥. 现代口腔正畸学[M]. 北京:中国医药科技出版社, 1999: 105 - 106.
[2] Cassinelli AG, Firestone AR, Beck FM, et al. Factors associated with orthodontists' assessment of difficulty [J]. Am J Orthod Dentofac Orthop, 2003, 123(5): 497 - 502.
[3] Pae EK, McKenna GA, Sheehan TJ, et al. Role of lateral cephalograms in assessing severity and difficulty of orthodontic cases [J]. Am J Orthod Dentofacial Orthop, 2001, 120(3): 254 - 262.
[4] Bergstrom K, Halling A. Comparison of three indices in evaluation of orthodontic treatment outcome [J]. Acta Odontol Scand, 1997, 55(1): 36 - 43.
[5] Daniels C, Richmond S. The development of the Index of Complexity, Outcome, and Need (ICON) [J]. Br J Orthod, 2000, 27(2): 149 - 162.
[6] Richmond S, Aylott NA, Panahei ME, et al. A 2-center comparison of orthodontist's perceptions of orthodontic treatment difficulty [J]. Angle Orthod, 2001, 71(5): 404 - 410.
[7] Bergström K, Halling A, Huggare J, et al. Treatment difficulty and treatment outcome in orthodontic care [J]. Euro J Orthod, 1998, 20(2): 145 - 157.
[8] Summers CJ. The occlusal index: a system for identifying scoring occlusal disorders [J]. Am J Orthod, 1971, 59(6): 552 - 567.
[9] DeGuzman L, Bahiraei D, Vig KW, et al. The validation of the PAR Index for malocclusion severity and treatment difficulty [J]. Am J Orthod Dentofac Orthop, 1995, 107(2): 172 - 176.
[10] Gramling JF. The Probability Index [J]. Am J Orthod Dentofac Orthop, 1995, 107(2): 165 - 171.
[11] Landis JR, Koch GG. The measurement of observer agreement for categorical data. [J] Biometrics, 1977, 33(1): 159 - 174.
[12] 李春波. 一致性检验方法的合理应用[J]. 上海精神医学, 2000, 12(4): 228 - 232.
[13] Rosner B. Fundamentals of Biostatistics[M]. 4th ed. Belmont: Duxbury Press, 1995: 518 - 519.
[14] Armitage P, Berry G, Matthews JNS. Statistical methods in medical research [M]. 3rd ed. Cornwall: Blackwell Publishing, 1994.
[15] Arruda AO. Occlusal indexs as judged by subjective opinions [J]. Am J Orthod Dentofacial Orthop, 2006, 134(5): 671 - 675.
[16] Richmond S, Daniels C. International comparisons of professional perceptions in Orthodontics. Part 1: treatment need [J]. Am J Orthod and Dentofacial Orthop, 1998, 113(2): 180 - 185.
[17] Richmond S, Shaw WC, O'Brien KD, et al. The development of the PAR Index (Peer Assessment Rating): reliability and validity [J]. Eur J Orthod, 1992, 14(2): 125 - 139.
[18] Rowe KGT. The concordance of pre-treatment malocclusion assessment among orthodontic specialty practitioners [D]. Ann Arbor: University of Michigan, 1989.

(2011-09-13 收稿)
(本文编辑:赵 波)