Original Article

Validity of the Index of Complexity, Outcome, and Need (ICON) in Determining Orthodontic Treatment Need

Allen R. Firestone, DDS, MS[a]; F. Michael Beck, DDS, MA[b]; Frank M. Beglin, DDS, MS[c]; Katherine W. L. Vig, BDS, FDS, D Orth, MS[d]

Abstract: Occlusal indices are used to determine eligibility for orthodontic treatment in several publicly funded programs. The Index of Complexity, Outcome, and Need (ICON), based on the perception of 97 orthodontists from 9 countries, has been proposed as a multipurpose occlusal index. The aim of this study was to investigate the validity of the ICON as an index of orthodontic treatment need compared with the perception of need as determined by a panel of US orthodontists. One hundred seventy study casts, representing a full spectrum of malocclusion types and severity, were scored for orthodontic treatment need by an examiner calibrated in the ICON. The results were compared with the decisions of an expert panel of 15 orthodontic specialists from the central Ohio area. The simple kappa statistic (0.81) indicated very high agreement of the index with the decisions of the expert panel. The sensitivity (94%), specificity (85%), positive predictive value (92%), negative predictive value (90%), and overall accuracy of the ICON (91%) also confirmed good agreement with the orthodontic specialists. The panel found that 64% of the casts required orthodontic treatment; the ICON scores indicated that 65% of the cases needed treatment. There was agreement between the expert panel and the index in 155 of the 170 cases. These results support the use of the ICON as a validated index of orthodontic treatment need. (*Angle Orthod* 2002;72:15–20.)

Key Words: Orthodontics; Treatment need; Validity

INTRODUCTION

Historically, orthodontic diagnosis has been taught and practiced as a descriptive, qualitative subject. However, in response to an external need for information on the prevalence of malocclusions and for a method to objectively quantify the severity of the various features of malocclusion, several indices have been proposed. These indices purport to measure severity of malocclusion objectively, either as a deviation from normal/ideal occlusion or in terms of perceived treatment need. For indices of treatment need, there is a system of protocols or rules to summarize data about malocclusion and return a numeric value. Within each of these indices, a numeric value exists below which the severity of a malocclusion is considered so minor that there is no need for treatment. All numeric values above that point indicate malocclusions for which treatment is indicated. In effect, an index with a cutoff point functions as a diagnostic test for treatment need, although a definitive ''gold'' or ''truth'' standard does not exist.

The pooled decision of orthodontic specialists is generally considered as the gold standard against which any index should be validated. Recently, studies have shown that several of these indices accurately reflect the decisions of local orthodontic specialists.[1,2] Although in practice occlusal indices and treatment need indices have been used interchangeably, there is no single index that has been developed and validated for both treatment need and deviation from normal/ideal. Recently, Daniels and Richmond[3] have proposed the Index of Complexity, Outcome, and Need (ICON), which the authors claim can be used to assess treatment need as well as to assess treatment outcome. The index is based on the perception of treatment need and outcome by 97 orthodontists from 9 countries who judged 240 dental casts for the assessment of treatment need and 98 paired pre- and posttreatment cases for assessment of treatment outcome.[3–5] The authors described the index as simple to use, requiring only a millimeter ruler and an Aesthetic

[a] Associate Professor, Department of Orthodontics, The Ohio State University, Columbus, Ohio.

[b] Associate Professor and Chair, Department of Health Services Research, The Ohio State University, Columbus, Ohio.

[c] Private Practice, Napa, Calif; Former Orthodontic Resident, The Ohio State University, Columbus, Ohio.

[d] Professor and Chair, Department of Orthodontics, The Ohio State University, Columbus, Ohio.

Corresponding author: Dr Allen Firestone, College of Dentistry, 305 West 12th Ave, PO Box 182357, Columbus, OH 43218-2357 (e-mail: firestone.17@osu.edu).

Accepted: May 2001. Submitted: February 2001.

TABLE 1. Distribution of Index of Orthodontic Treatment Need Dental Health Component Grades in Sample

Grade	Treatment Need	n	Percentage
1	None	3	2
2	Little	47	28
3	Moderate	37	22
4	Great	42	36
5	Very great	21	12

Component Scale.[6] The index is intended for use in the late mixed dentition and permanent dentition. Further, the index may be applied clinically to patients and to casts without any modification. The ICON is unique in incorporating an aesthetic score as an integral part of the evaluation of treatment need. Because it is both an index of treatment need and an occlusal index of malocclusion severity, the ICON offers significant advantages over other indices of treatment need.

There is evidence that geographic location may affect the specialist's determination of treatment need and outcome.[4,5] There is ample evidence in medicine that treatment delivery varies by geographic region, even though prevalence of the underlying disease does not vary.[7–9] Thus it would seem prudent to validate an index purporting to correctly identify treatment need with the opinion of orthodontic specialists practicing within a limited geographic region.[1,2] Therefore, the aim of the present investigation was to validate the ICON as an index of treatment need based on the perceptions of orthodontic specialists practicing in central Ohio.

MATERIALS AND METHODS

A set of 156 pairs of orthodontic study casts consisting of treated and untreated study cases from the University of Pittsburgh Orthodontic Department were duplicated and delivered to The Ohio State University Orthodontic Department. These casts had previously been used for validation of other indices in western Pennsylvania and central Ohio.[1,2] The duplicate casts had been evaluated for accuracy by comparing measurements (overjet, overbite, midline deviations, and anterior tooth contact point displacements) taken from a sampling of the original casts with those of the duplicated casts. These study casts represented a full spectrum of malocclusion types and severity. After the sample was reviewed, 14 pairs of casts with elective treatment need were added, increasing the final sample to 170 pairs. The distribution of the cases within the 5 categories of the Dental Health Component (DHC) of the Index of Orthodontic Treatment Need (IOTN) and a description of the degree of treatment need for cases within that category is presented in Table 1.[10] The same panel of orthodontists from central Ohio had previously validated the DHC of the IOTN.[1]

Volunteers were solicited from among 90 orthodontists who were members of the American Association of Orthodontists and whose practice addresses were within the central Ohio area; 15 volunteers were selected. To be selected, the volunteer had to have been a practicing orthodontist with at least 5 years of experience. In addition, they had to agree to attend 1 of 3 dates for the initial rating and 1 of 2 dates for the repeat rating. The Human Subjects Institutional Review Board of The Ohio State University[1] approved the design of the study.

The 15 orthodontist-raters scored the 170 casts and recorded the need for treatment of each as a score from 1–7 on an adjectival scale where 1 equals ''none/minimal'' need and 7 equals ''very great'' need. At a second session, approximately 30 days after the first session, each rater again assigned a score to a random subset of 40 study casts, stratified by occlusal severity, to test intrarater reliability. For both sessions, the casts were displayed in numerical order on tables in a large room. The raters were asked to start at staggered points throughout the sample. The raters were instructed to work at their own pace with no time limit.

At the beginning of both rater sessions, the following verbal and written instructions were given to the raters:

> *You are the orthodontic consultant for a private corporation for which a fund has been established to provide orthodontic treatment for personnel. You are to evaluate these study casts of personnel and answer the following question: In your opinion, to what extent does this occlusion need orthodontic treatment? Please circle the corresponding number:*

None/ Minimal						Very great
1	2	3	4	5	6	7

At the end of the second session, each rater was asked to answer the following question:

> *On the 7-point scale that you have used throughout this rating session, indicate the score at or above which you feel orthodontic treatment is indicated.*

This score was termed the ''indicated treatment point'' (ITP) and was recorded for each of the 15 raters.

One examiner (Dr Firestone) who was calibrated in the ICON scored the 170 study casts using the ICON index. One month later, a random subset of 40 study casts was chosen and scored again by the calibrated examiner to test intrarater reliability.

Statistical analysis

The simple kappa statistic was used to assess agreement of the index with the expert panel. Weighted (Fleiss-Cohen)

TABLE 2. Comparison of the Diagnostic Performance Characteristics of the Index of Complexity, Outcome, and Need (ICON) at the Standard and Optimized ICON Score Cutoff Point for Determining Orthodontic Treatment Need When Applied to the 170 Test Casts

ICON Score Cutoff	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Accuracy, %	Kappa
>42 (Standard)	94.4	85.5	91.9	89.8	91.2	0.81
>52 (Optimized)	91.7	93.5	96.1	86.6	92.4	0.84

kappa statistics were used to assess both intra- and interrater reliability.[11] The kappa statistic is a measure of agreement that has been corrected for chance agreement.[12] A kappa value of 0 indicates no agreement beyond chance, whereas a kappa value of 1 indicates perfect agreement.

Interrater reliability was calculated by comparing all raters on the entire sample of 170 sets of casts during the first session. Intrarater reliability was based on a comparison of the scores assigned by the raters to the subset of 40 casts at the second session to the scores assigned by the raters to those same casts at the first session.

The ''truth'' or ''gold standard'' was determined in the following manner. First, the mean ITP for the 15 raters was calculated. Second, the mean rater score for the 15 raters for each cast was calculated. Finally, the mean score for each cast was compared to the mean ITP value, and if the cast score was below the mean ITP score, the case was assigned to the ''no treatment'' category. If the mean rater score for a cast was equal to or greater than the mean ITP value, the case was assigned to the ''treatment'' category.

The developers of the ICON have proposed the cutoff point for treatment as a score >42.[3] Each of the 170 study casts was assigned to a ''treatment'' or ''no treatment'' category by comparing the calibrated examiner's score for each cast with the recommended cutoff point for the index. For each of the casts, the (mean) decision of the raters, the gold standard, was compared to the decision assigned by the calibrated examiner using the index.

From these comparisons, the following values were calculated for the index: sensitivity, specificity, positive and negative predictive values, accuracy (percentage agreement), and kappa statistic. *Sensitivity* is the percentage of all cases needing treatment that the index identified as needing treatment. *Specificity* is the percentage of all cases not needing treatment that the index identified as not needing treatment. *Positive* and *negative predictive values* are the percentage of cases that the index identified as needing (positive) or not needing (negative) treatment that in fact need or do not need treatment. *Accuracy* in this study was defined as the percentage agreement with the decisions of the expert panel. This measure does not take into account agreement due to chance. An optimized cutoff point for the index was determined by plotting a receiver operating characteristic (ROC) curve.[13–15] The significance of the area under a ROC curve has been described as representing the probability that a randomly chosen subject in need of orthodontic treatment will be correctly rated or ranked with greater need than a randomly chosen subject with no need for orthodontic treatment.[16] It has been proposed that a ROC curve is a more meaningful measure of the value of a diagnostic test than ''accuracy,'' or the percentage of cases in which the dentist is correct. This is because, unlike accuracy, the ROC curve is not dependent on the prevalence of a disease in the population. Neither will 2 tests with the same accuracy, but different sensitivity and specificity, give the same ROC curves.[14] A ROC curve plots the sensitivity vs 1-specificity at different decision thresholds.[14]

RESULTS

Both the calibrated examiner and the panel of orthodontic experts gave evidence of high levels of reliability. The calibrated examiner demonstrated high intra-examiner reliability for the 40 casts that were evaluated twice. The weighted kappa value (95% confidence limit, with lower confidence boundary and upper confidence boundary in parentheses) was 0.89 (0.74, 1.00). The 15 raters exhibited a high level of interrater reliability; the overall weighted kappa value was 0.81 (0.81, 0.82).[1] For intrarater reliability, the 15 raters also achieved high levels of reliability. The overall weighted kappa value was 0.92 (0.90, 0.93) for the 40 casts that were evaluated twice by each rater.[1]

The mean ITP for the 15 raters (mean ± SD) was 3.53 ± 0.74. Those casts with mean scores equal to or greater than 3.53 were assigned to the ''treatment'' category, and the remaining casts, with scores below 3.53, were placed into the ''no treatment'' category. There were 108 (64%) casts in the ''treatment'' category and 62 (36%) casts in the ''no treatment'' category.

Based on the index score as determined by the calibrated examiner and the cutoff point for the ICON (>42), each of the 170 casts was assigned into the ''treatment'' or the ''no treatment'' category. The diagnostic performance of the index at its recommended cutoff points was satisfactory (Table 2). The overall agreement (simple kappa coefficient) with the gold standard (the decisions of the orthodontists) was 0.81 (0.72, 0.90). The results of the comparison between the decisions of the orthodontists and the index are summarized in a 2-by-2 contingency table (Table 3).

The area under the ROC curve (Figure 1), 97%, indicates the high validity of the index, ie, the degree to which the index reflects the decisions of the gold standard panel of orthodontists. The ROC curve can be used to locate an

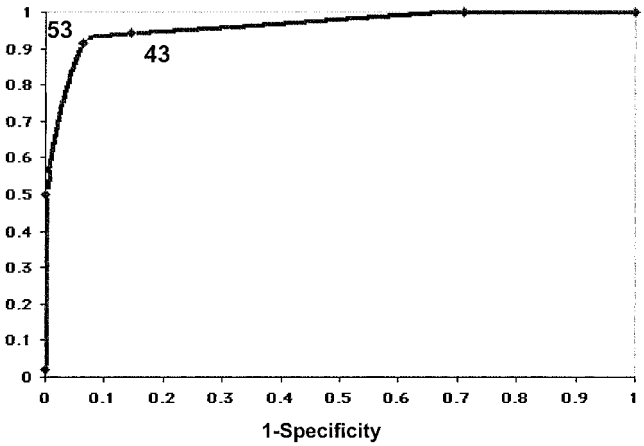

FIGURE 1. Index of Complexity, Outcome, and Need (ICON) receiver operating characteristic curve with the standard (>42) and optimized (>52) ICON score cutoff points identified.

optimized cutoff point: the point most superior and most to the left on the curve (Figure 1). This may be understood as the point where both the sensitivity and specificity are maximized. A comparison of the recommended cutoff point with the optimized cutoff point is presented in Table 2.

Under the optimized, more stringent cutoff, the number of false positives, ie, a case recommended for treatment by the index but not by the expert panel, was reduced from 9 to 4 (Table 3). The number of false negatives, ie, cases recommended for no treatment by the index but classified as needing treatment by the panel, rose from 6 to 9. The net change was a gain of 2 additional cases correctly classified by the index at the higher, more stringent cutoff; ie, 8 cases were moved from the ''treatment'' to the ''no treatment'' group, 5 cases correctly and 3 cases incorrectly.

DISCUSSION

The results of this investigation are similar to those of other investigators who evaluated the validity of other indices of treatment need.[2,17] The 15 orthodontic raters achieved high levels of intra- (0.81) and interrater agreement (0.92), which are comparable to levels of 0.84 and 0.91, respectively, reported by Younis et al.[2] Agreement between the ICON, as applied by a calibrated examiner, and the gold standard panel of orthodontic experts was also comparable to results with other indices. Agreement (simple kappa) in the present investigation was 0.81. Landis and Koch[18] have proposed that a kappa statistic for agreement in the range of 0.81–1.00 be considered as ''almost perfect.'' In an earlier study with the same expert panel and a different calibrated examiner, agreement for other indices was as follows: Dental Aesthetic Index, 0.83; Handicapping Labio-lingual Deviations Index (California modification), 0.62; IOTN DHC, 0.84; and IOTN Aesthetic Component, 0.67.[1] Other investigators, using a subset of the sample of casts employed in the present study, have reported areas under the ROC curve, a measure of the utility or diagnostic value of an index, ranging from 0.96 to 0.99.[2] These results are similar to the results in the present investigation: an area under the ROC curve of 0.97. During the development of the ICON, for treatment need, the sensitivity, specificity, and accuracy of the index were reported as 85.2%, 86.4%, and 85.5%, respectively, when compared with the decisions of the international panel of orthodontists.[3] In the present study, these values were 94.4%, 85.5%, and 91.2%, respectively, when compared with the decisions of the local panel of orthodontists (Table 2).

Using the local panel of orthodontic experts as the gold standard and the ICON scores of the calibrated examiner, an ''optimized'' cutoff point was calculated. A comparison of this optimized cutoff score, 53, applied to the sample of 170 casts with the results of the standard cutoff point is presented in Table 2. The net result was that 2 additional cases were correctly assigned by the index to the ''no treatment'' group. Thus, the specificity of the index increased, but the sensitivity decreased. It is a property of diagnostic tests that any increase in sensitivity or specificity achieved by changing the cutoff point will result in a decrease in the other parameter.

In effect, a diagnostic test is ''described'' by its ROC curve. By changing the cutoff point, one can ''move'' along the curve, but not change the curve. Decisions about where to place cutoff points are subject to discussion and disagreement. Decisions depend on the costs, risks, and benefits incurred by increasing the number of true-positive diagnoses and the false-positive vs those incurred by increasing the number of true-negative diagnoses and the number of false-negatives.[19]

In the present study, changing the cutoff point to a stricter value would bring the benefits of a correct decision to assign 5 additional patients to a ''no treatment'' group. These benefits would consist of savings in money, time, inconvenience, and suffering by avoiding unnecessary orthodontic treatment. The costs involved in changing the cutoff point would be incurred as a result of the decision to incorrectly assign 3 patients who needed treatment to the ''no treatment'' group. These costs might include the psychosocial costs of having a malocclusion and missing an opportunity to use growth modification and the costs associated with orthodontics and possibly orthognathic surgery later. The benefits of leaving the cutoff point at the more lenient value are correctly assigning 3 additional patients to the ''treatment'' group. The costs are a consequence of incorrectly assigning 5 patients to the ''treatment'' group. These costs might include the costs associated with an orthodontic consult/case workup in which one expects the orthodontic expert to reverse the decision of the index. At the extreme, the costs may include inappropriate treatment.

The limitations of this investigation include the use of a local panel of experts to establish a gold standard. This may

TABLE 3. A 2-by-2 Contingency Table of Decisions to Treat or Not Treat for Index of Complexity, Outcome, and Need (ICON) vs Gold Standard

	ICON	
	Don't Treat	Treat
Gold standard		
Don't treat	53	9
Treat	6	102

limit the ability to generalize the results. There is evidence that the country where orthodontic specialists practice has an effect on their evaluation of treatment need.[4] Thus, the validity of an index may depend on the origin of the panel of experts serving as the gold standard. Previous investigators have shown that there is excellent agreement between the panel of central Ohio orthodontists and a panel from western Pennsylvania on the need for treatment when examining the same sample of casts.[1] We conclude that the ICON has local, regional validity and hypothesize that it may be generally valid.

Using one calibrated examiner allows determination of intra-examiner reliability using the index. It does not, however, allow interexaminer reliability to be examined. Even the calibration process is not a guarantee against differences due to experience, personal biases, or individual aptitude.[20] Thus, a study design such as the present one must only be considered to demonstrate the efficacy of the use of the ICON in determining orthodontic treatment need in a controlled study environment, but the effectiveness of the ICON as a tool in practice remains undetermined.

A further limitation of the study lies in the nature of the test sample of 170 casts (Table 1). The accuracy of a test is influenced by, among other things, the number of easy (extreme and thus easy to diagnose) and difficult (borderline) cases in the sample.[21] The sample of casts in the present investigation included only 24 cases (14%) in the 2 extreme categories of the IOTN Dental Health Component.[10] There were 47 cases (28%) in the second category of ''little'' orthodontic treatment need and 37 cases (22%) in the third, ''borderline need'' category. It remains to be seen how a sample consisting of, eg, only cases in category 3, ''borderline'' need, would have affected the results of this study.

CONCLUSION

The ICON, applied as an index of treatment need by one calibrated examiner, is a valid and reliable instrument. The cutoff point closely matches the collective treatment/no treatment decision threshold of a panel of orthodontic experts from central Ohio. These results and those of other investigators have shown indices of treatment need to be valid and reliable when applied by calibrated examiners.[1,2,17,22] The use of the ICON as an index of complexity or treatment outcome remains to be validated using an independent sample of cases and/or raters. At this time there still does not exist an index validated for use to measure both treatment need and treatment outcome.

ACKNOWLEDGMENTS

The authors wish to recognize all the orthodontists who agreed to participate in the study and to especially thank the 15 specialists who served as raters.

REFERENCES

1. Beglin FM, Firestone AR, Vig KWL, Beck FM, Kuthy RA, Wade D. A comparison of the reliability and validity of three indices of orthodontic treatment need. *Am J Orthod Dentofac Orthop.* In press.
2. Younis JW, Vig KWL, Rinchuse DJ, Weyant RJ. A validation study of three indices of orthodontic treatment need in the United States. *Community Dent Oral Epidemiol.* 1997;25:358–362.
3. Daniels C, Richmond S. The development of the Index of Complexity, Outcome and Need (ICON). *Br J Orthod.* 2000;27:149–162.
4. Richmond S, Daniels CP. International comparisons of professional assessments in orthodontics, I: treatment need. *Am J Orthod Dentofac Orthop.* 1998;113:180–185.
5. Richmond S, Daniels CP. International comparisons of professional assessments in orthodontics, II: treatment outcome. *Am J Orthod Dentofac Orthop.* 1998;113:324–328.
6. Shaw WC, Richmond S, O'Brien KD, Brook P, Stephens CD. Quality control in orthodontics: indices of treatment need and treatment standards. *Br Dent J.* 1991;170:107–112.
7. Carlisle DM, Valdez RB, Shapiro MF, Brook RH. Geographic variation in rates of selected surgical procedures within Los Angeles County. *Health Serv Res.* 1995;30:27–42.
8. Polednak AP. Geographic variation in postmastectomy breast reconstruction rates. *Plast Reconstr Surg.* 2000;106:298–301.
9. Wilt TJ, Cowper DC, Gammack JK, Going DR, Nugent S, Borowsky SJ. An evaluation of radical prostatectomy at Veterans Affairs Medical Centers: time trends and geographic variation in utilization and outcomes. *Med Care.* 1999;37:1046–1056.
10. Brook FH, Shaw WC. The development of an index of orthodontic treatment priority. *Eur J Orthod.* 1989;11:309–320.
11. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability. *Educ Psychol Meas.* 1973;33:613–619.
12. Cohen AJ. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
13. Beck JR, Shultz EK. The use of receiver operating characteristic (ROC) curves in test performance evaluation. *Arch Pathol Lab Med.* 1986;110:13–20.
14. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med.* 1978;8:283–298.
15. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* 1988;240:1285–1293.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982; 143:29–36.
17. Richmond S, Shaw WC, O'Brien KD, et al. The development of the PAR index (Peer Assessment Rating): reliability and validity. *Eur J Orthod.* 1992;14:125–139.
18. Landis JR, Koch CG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
19. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. *N Engl J Med.* 1975;293:211–215.
20. Roberts CT, Richmond S. The design and analysis of reliability

studies for the use of epidemiological and audit indices in orthodontics. *Br J Orthod.* 1997;24:139–147.
21. Swets JA, Getty DJ, Pickett RM, Seltzer SE, McNeil BJ. Enhancing and evaluating diagnostic accuracy. *Med Decis Making.* 1991;11:9–18.
22. DeGuzman L, Bahiraei D, Vig KWL, Vig PS, Weyant MS, O'Brian K. The validation of the peer assessment rating index for malocclusion severity and treatment difficulty. *Am J Orthod Dentofac Orthop.* 1995;107:172–176.