Application of statistical decision theory to radiographic diagnosis of endodontically treated teeth

CLAES REIT AND HANS-GÖRAN GRÖNDAHL

Departments of Endodontics and Oral Radiology, Faculty of Odontology, University of Gothenburg, Gothenburg, Sweden


Abstract - The presence or absence of periapical radiolucencies, after a predetermined healing period, have been used as a criteria for therapeutic success or failure in endodontic treatment. However, in radiologic diagnosis it has been shown that variations within and between examiners are substantial. In the present study six endodontists evaluated the periapical tissues in radiographs of 119 endodontically treated roots using a five-graded rating scale. An earlier study on the same radiograph material served as reference and “true” states of the periapical tissues were established. True positive and false positive reports on the presence of periapical lesions could then be calculated. For each observer it was noticed that a higher true positive percentage always was coupled with a higher false positive percentage and vice versa. If pairs of true positive and false positive percentages were plotted on a two-dimensional graph they corresponded well to a Receiver Operating Characteristic (ROC) curve. Variations between the observers could be explained by their adoption of different criteria of periapical disease resulting in different positions on the ROC curve. Influence of observer variation on reported frequencies of periapical lesions was greater, the lower the prevalence of the disease. The best opportunities for revealing relative differences in disease prevalences was created when the examiner defined a strict criterion for disease and reported a positive finding only when absolutely certain.

Key words: decision theory; endodontics; observer variation; periapical diagnosis; periapical osteitis.

O. Reit, Department of Endodontics, Faculty of Odontology, Box 33070, S-400 33, Gothenburg, Sweden. Accepted for publication 27 December 1982.

In radiographic studies of the results of endodontic therapy the presence of a periapical radioluency at the end of a predetermined healing period is generally considered a sign of failure. The relative frequencies of such radiolucencies have therefore been used to analyze the influence of various factors on the healing processes (1–6).

In a recent investigation by Reit & Hollender (7) six experienced observers examined the periapical conditions on radiographs of 119 endodontically treated roots. A total of 37 periapical bone lesions were diagnosed but only 10 were reported in agreement. Similar findings have also been published by other authors (8, 9). Since the observer is an integral part of the radiographic diagnostic procedure, variations among observers may influence the reported results of radiographic studies. This could, at least partly, explain the great discrepancies
among various investigations on the results of endodontic therapy. Since endodontic treatment procedures to a great extent are based on radiographic studies it seems necessary to further elucidate problems associated with interobserver variation.

Interpreting radiographs is a visual perception process and could be regarded as a task of detecting a signal against a background of noise. A view of detection as a statistical process of decision making has been used extensively in psychophysics (10). The use of signal detection theory in medical radiology has been suggested by, for example, Garland (11), Lusted (12) and McNeil et al. (13). In dental radiology Webber et al. (14) and Grøndahl (15) have shown the usefulness of this theory. The periapical bone lesion associated with an infected dental pulp is a gradually increasing, or during healing decreasing, process. In the radiographic image it develops from being impossible or difficult to reveal, to being easily distinguished from the background (16). It therefore seems that radiographic diagnosis of periapical bone lesions has attributes in common with a signal detection task.

The aim of the present investigation was to apply statistical decision theory to periapical radiographic diagnosis to elucidate the consequences of adopting different criteria of disease.

Material and methods

Radiographs of 119 endodontically treated roots were sampled at random from the files of the Department of Endodontics. The examined teeth were equally distributed among incisors, premolars and molars. Each root was visible on two separate radiograms obtained according to the technique described by Eggén (17). Six dentists specialized in endodontics served as observers. The film readings were made with the aid of a viewing box and a magnifying viewer equipped with a masking frame corresponding to the size of a dental film.

According to signal detection theory the stimulus of interest (signal) occurs against a background of stimulation already present (noise). The probability of detecting the signal depends upon its intensity relative to the background. Neither the background nor the signal is constant in its appearance. Frequency distributions of noise and signal plus noise may look as shown in Fig. 1. As long as the two distributions overlap it is not possible for the observer always to be correct in determining whether or not a signal is present. The theory assumes that the observer, along a continuum of observations, adopts a cut-off point, a criterion, above which he reports the presence of a signal, and below which he reports no signal. Two types of observer errors are possible; either a signal may be missed (a false negative response) or a signal will be reported where there is none (a false positive response). By shifting his criterion the examiner may vary the proportion of these two errors. By plotting pairs of true positive (TP) and false positive (FP) reports on a two-dimensional graph the obtained values will form a so called Receiver Operating Characteristics (ROC) curve. In diagnostic radiology the generation of pairs of TP% and FP% could be done by letting the observer quantitate, on a rating scale, the certainty of the response regarding the presence or absence of a signal. In the present investigation the observers evaluated the periapical tissues according to the following classification. No measures were taken to calibrate the examiners.

1 = Periapical destruction of bone definitely not present.
2 = Periapical destruction of bone probably not present.
3 = Unsure.
4 = Periapical destruction of bone probably present.
5 = Periapical destruction of bone definitely present.

Fig. 1. Probability density distributions of two populations with regard to observation strength. On the x-axis the observer chooses a cut-off point, criterion (C), at which he wishes to operate. When the operating points is moved along the x-axis, the relative proportion of false negative and false positive reports will change.
In an earlier study (7) the same radiographs were examined by another group of observers (three oral radiologists and three endodontists). These observers were to distinguish between “normal periapical conditions”, “increased width of the periodontal space” and “periapical bone destruction”. In 22 of the 119 cases four of the six examiners independently agreed on the presence of a periapical lesion. In the present study these 22 roots were regarded as having “true” periapical lesions. Thus it was possible to calculate, for each observer, the TP% and FP% using the formulas below.

\[
TP = \frac{\text{True positive reports}}{\text{True positive reports} + \text{False negative reports}} \times 100
\]

\[
FP = \frac{\text{False positive reports}}{\text{False positive reports} + \text{True negative reports}} \times 100
\]

**Results**

The reports of the examiners are presented in Table 1. Interobserver variations were recorded with positive reports of disease ranging between nine and 30 cases when a strict criterion (score 5) was used. Forty of the 119 examined roots were assigned score 5 by at least one observer. In six cases (15%) the opinions of all observers coincided (Fig. 2). If score 4 also was regarded as a positive response the interexaminer range increased to encompass 13–51 cases. Fifty-nine roots were given either score 4 or score 5 by at least one observer and in nine cases (15%) consensus was reached.

Variation was also disclosed between observers concerning their use of the end parts of the rating scale (i.e. scores 1 and 5). Observer E2, for example, placed 103 (87%) of his reports in the extreme positions, while the corresponding values for observer E6 were 53 (45%).

When values of TP% and FP% were calculated considering score 5 as a positive response the observers diagnosed “true lesions” with an interindividual range from 41% to 95% of the cases (Table 2). “Healthy” periapical tissues were diagnosed as diseased in up to 9% of the roots. Intra- and interobserver comparisons showed that an increase of the TP% values as a rule was accompanied by an increase of the FP% values and vice versa. When the TP% and FP% pairs were plotted on a graph they corresponded well to an ROC curve (Fig. 3).

In order to elucidate the consequences of different positions on the ROC curve, the obtained TP% and FP% values were used on four hypothetical populations with different assumed prevalences of periapical disease (Table 3).

![Fig. 2. Interobserver agreement on radiographic evidence of periapical disease. On the x-axis least number of agreeing observers are tabled. Filled parts of bars represents number of root with positive reports when a strict criterion (score 5) for disease is used. Unfilled parts denotes additional positive reports when a liberal criterion (score 4 or 5) is used.](image-url)
Table 2

True positive (TP) and false positive (FP) percentages calculated for each observer on two criteria for periapical disease. Criterion A = score 5, criterion B = score 4 or 5

<table>
<thead>
<tr>
<th>Examiners</th>
<th>Criterion A</th>
<th></th>
<th>Criterion B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP%</td>
<td>FP%</td>
<td>TP%</td>
<td>FP%</td>
</tr>
<tr>
<td>E1</td>
<td>41</td>
<td>0</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>E2</td>
<td>41</td>
<td>2</td>
<td>45</td>
<td>3</td>
</tr>
<tr>
<td>E3</td>
<td>82</td>
<td>1</td>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td>E4</td>
<td>68</td>
<td>5</td>
<td>95</td>
<td>15</td>
</tr>
<tr>
<td>E5</td>
<td>91</td>
<td>6</td>
<td>100</td>
<td>16</td>
</tr>
<tr>
<td>E6</td>
<td>95</td>
<td>9</td>
<td>100</td>
<td>29</td>
</tr>
</tbody>
</table>

It was then found that at 1% prevalence of periapical lesions the reported frequencies would vary among the observers from 0.4% to 9.9%. At 50% prevalence the range would enclose 20.5%-52.0%. The values shown in Table 3 were used to calculate relations between observer reported prevalences (Table 4). It was found that the reported relations approached the true values as the FP% decreased. True relations were only reported by observer E1 when using score 5 and giving no false positive reports.

Discussion

Substantial inter- and intraindividual variation has been shown in radiographic diagnosis of periapical inflammatory lesions (5-9). In a study by Reit & Hollender (7), for example, six uncalibrated observers agreed in only 27% of the reported lesions. In the present investigation different criteria for the presence of periapical bone destruction were used. A liberal criterion (score 4 or 5) gave an interobserver agreement of 15%. The same result was obtained with a more strict criterion (score 5). Thus an increase in the certainty with which radiographic assessments were made did not give an increase in observer agreement.

The establishment of an ROC curve is a technique suggested for detectability studies and

Table 3

Disease prevalences as reported by different observers in populations with assumed true prevalences. Only score 5 is regarded as positive report of periapical disease

<table>
<thead>
<tr>
<th>True prevalence</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>20.5</td>
<td>21.5</td>
<td>41.5</td>
<td>36.5</td>
<td>48.5</td>
<td>52.0</td>
</tr>
<tr>
<td>10</td>
<td>4.1</td>
<td>5.9</td>
<td>9.1</td>
<td>11.3</td>
<td>14.5</td>
<td>17.6</td>
</tr>
<tr>
<td>5</td>
<td>2.1</td>
<td>4.0</td>
<td>5.1</td>
<td>8.2</td>
<td>10.3</td>
<td>13.3</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>2.4</td>
<td>1.8</td>
<td>5.6</td>
<td>6.8</td>
<td>9.9</td>
</tr>
</tbody>
</table>
DECISION THEORY IN PERIAPICAL DIAGNOSIS

Table 4
Relation of reported prevalences of periapical disease compared to true relation. Two criteria for periapical disease are used. 
(A = score 5, B = score 4 or 5)

<table>
<thead>
<tr>
<th>True relation</th>
<th>Criterion</th>
<th>Reported relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50/10 = 5.0</td>
<td>A</td>
<td>5.0 3.6 4.6 3.2 3.3 3.0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>3.4 3.3 2.7 2.4 2.4 1.8</td>
</tr>
<tr>
<td>5/1 = 5.0</td>
<td>A</td>
<td>5.3 1.7 2.8 1.5 1.5 1.3</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>1.5 1.5 1.3 1.2 1.2 1.1</td>
</tr>
</tbody>
</table>

has been used as a tool in clinical decision making (12, 13, 15, 18). To generate the TP and FP pairs in such a curve it is necessary to know the actual presence or absence of disease in the population under study. This knowledge is difficult to achieve in clinical studies. In this context it would be preferred to obtain biopsy specimens from the periapical tissues. In the absence of such data examiners are limited to use radiologic evidence of periapical disease. In the present investigation a "true" periapical state therefore was defined by using an earlier examination of the same radiographs as a reference (7). Where at least four out of six observers independently had reached the conclusion that a periapical destruction of bone was present, this was regarded as indicating the presence of a "true" lesion. This way of establishing a reference system has been used by Garland (11). It must be noted, however, that by using this method the calculated values have no definite meaning, but must be viewed as relative only.

In the present study it was observed that by changing the criterion from a strict to a more liberal one, all examiners increased both their TP and FP percentages. In order to detect more of the cases with periapical bone lesions they had to pay a "cost" in increased overreading, i.e. reporting healthy periapical conditions as diseased. The obtained TP% and FP% pairs corresponded well to an ROC curve indicating equal ability among the observers to distinguish signal from noise. Thus, the interindividual variation is above all a result of the adoption of different cut-off points on the stimulus continuum. Radiographic evaluation of the results of endodontic therapy has been carried out by numerous investigators. However, different studies show considerable discrepancies in the obtained results, reports on development or persistence of periapical radiolucencies varying from 5% to 35% of the treated cases (1–3, 6, 19). The present study indicates that one explanation to this variation could be due to the adoption of different criteria by the examiners resulting in different positions on the ROC curve.

Since it is difficult to reveal the actual prevalence of periapical pathology by radiographic means, different decision strategies must be used. In a clinical setting diagnosis has to be accompanied by treatment decisions. When diagnosing a disease which can lead to serious complications if undetected, we should be likely to accept a large proportion of false positive diagnoses to ensure that we discover more patients with disease. For less serious conditions, or when treatment is associated with higher risks, we are willing to overlook more lesions to reduce the number of false positive diagnoses. What attitude the clinician should adopt when treating periapical disease still has to be evaluated. When diagnostic radiology is used as a tool in determining the influence of various factors on periapical healing, other considerations must be made. Between different factors or populations
true relation is revealed only by observers giving no false positive reports. Thus, in order to create the best opportunities for disclosing actual relative differences the observer has to define a strict criterion for disease and report a positive finding only when absolutely certain. He will then decrease his ratio of false positive diagnoses. This attitude is more important, the lower the prevalence of the disease under study.

References

This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.