1987

False positive rates encountered in the detection of changes in periodontal attachment level

John C. Gunsolley

School of Basic Health Science
Virginia Commonwealth University

This is to certify that the thesis prepared by John C. Gunsolley entitled False positive rates encountered in the detection of attachment level changes has been approved by his committee as satisfactory completion of the thesis requirements for the degree of Masters of Science.

[redacted]
Director of Thesis and Department Chairman

[redacted]
Committee Member

[redacted]
Committee Member

[redacted]
School Dean

5 May 1987
Date

False positive rates encountered in the detection of changes in periodontal attachment level

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science at Virginia Commonwealth University

By

John C. Gunsolley , B.S.E.E., General Motors Institute 1972, D.D.S. Indiana University School of Dentistry 1976

Thesis director: Dr. Walter H. Carter Jr.
Department Chairman
Department of Biostatistics

Virginia Commonwealth University
Richmond, Virginia
May, 1987

Acknowledgements

I would like to thank my wife Pam, my children John, Cindy, and Shawn and pets Mischief, Princess and Dusty for their patience and understanding in the more frustrating phases of this thesis. I would like to thank Dr. W. H. Carter for his direction of the thesis and committee members, Dr. A. M. Best and Dr. J. A. Burmeister for their help.

I would like to thank Drs. J. M. Goodson, A. D. Haffajee, and S. S. Socransky for supplying the data used in this thesis. Also I would like to acknowledge support of N. I. H. Grant # DE00130 for financial support.

Table of Contents

Page

List of tables

List of figures

Figure Page

Chapter 1

Review of the research problem

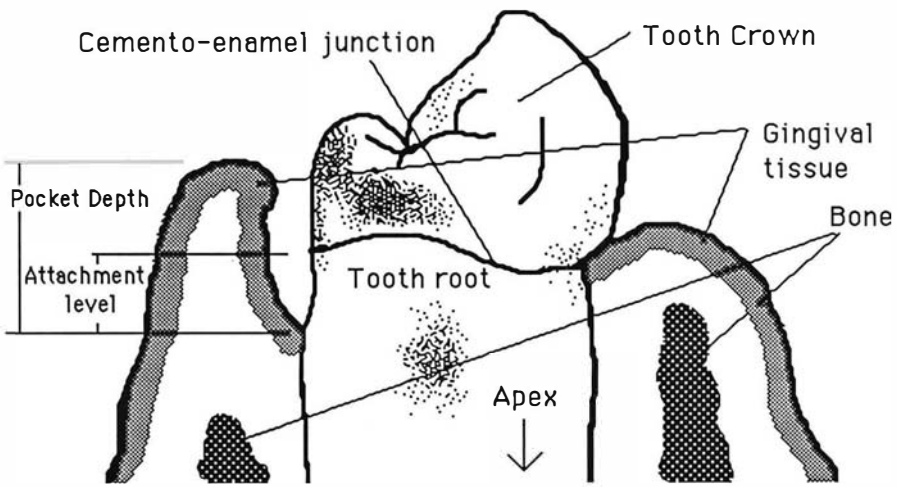
1.1 Description of attachment level

Periodontics is the dental field in which the supporting structures of teeth are studied. In both periodontal research and the clinical practice of periodontics, attachment levels of numerous sites are monitored to determine disease activity. Attachment level is the most coronal position on the tooth where the soft tissue, termed gingiva, attaches to the tooth. This attachment consists on a microscopic level of junctional epithelium and, apical to the junctional epithelium, a connective or fibrous tissue attachment (Listgarten, M. A., Mao, R., Robinson, P. J., 1976). When a tooth erupts into the oral cavity, the most coronal portion of the attachment is at the cemento-enamel junction (Figure 1). So in the absence of periodontal destruction, the attachment is at the cemento-enamel junction. Periodontitis is the destruction of the periodontium which results in the loss of a portion of the supporting structure of a tooth or teeth. Any change of the periodontal attachment in an apical direction from the cemento-enamel junction is an indication of loss of some of the supporting structure of the tooth.

Attachment level is measured from the cemento-enamel junction (Figure 1) to the most coronal position of the tooth where the soft tissue attaches. It is customary to measure attachment level at either 4 or 6 sites around the tooth. Human subjects have from 0 to 32 teeth and if 6 sites are sampled per tooth, from 0 to 192 sites are measured in each subject. Since periodontal destruction can be very localized it is

Figure 1

Cross Sectional View of the Supporting Structure of a Tooth



necessary to measure multiple sites in each subject. Some areas in the mouth may be undergoing severe destruction, resulting in large changes in attachment level, while other areas experience little or no destruction. Due to the localized nature of the disease, it is not sufficient to identify only those subjects with periodontitis, but clinicians must also identify areas of disease activity and inactivity within the one individual subject. One of the unresolved questions in periodontal research is why some areas in the periodontium undergo very rapid destruction, while other areas in the same subject remain stable or may even gain attachment.

1.2 Models of destructive periodontal disease

In the field of periodontics there is a controversy as to whether attachment loss occurs as very rapid destruction over a short period of time or as a slow, gradual change. It is inferred from epidemiological studies (Suomi et al., 1971; Axelsson & Lindhe, 1978; Löe et al., 1986) that periodontal disease is a chronic disease due to its low annual rate of attachment loss. In the study by Löe et al. an average annual rate of .17 mm. of attachment loss per mesial site was reported in Sri Lanka tea workers. This small mean rate of attachment loss was found in both the cross-sectional and longitudinal aspects of the study. However, the subjects were monitored only once every three years. Thus, from these studies it is difficult to determine the course of the disease process. The observed loss of attachment level may be due to either a slow, gradual process or a very rapid process followed by periods of inactivity.

Recent work, primarily at the Forsyth Periodontal Research Center (Goodson et al., 1982; Socransky et al., 1984), questions the conclusion that the disease process is gradual. These authors measured periodontal attachment level every 2 months for

periods up to 2 years. Two months between examinations is a much shorter time interval than previous studies had used. In addition, instead of focusing on mean attachment loss for a subject, the Forsyth studies focused on individual sites and tried to determine if sites had experienced change. This approach was used because they believe that periodontal disease occurs at only some sites, while the large majority of sites remain unchanged. Conclusions from their reports suggest that individual sites undergo an episodic burst of destruction followed by either remission or a period of repair. Their model is termed the "burst" model to distinguish it from the "chronic" model. At this time the controversy continues over the pattern of attachment level change encountered in periodontitis since there is great difficulty in confidently identifying attachment level changes.

1.3 Problems in the longitudinal monitoring of multiples sites

Numerous problems exist in the evaluation of longitudinal attachment loss measurements. The detection of a slow rate of attachment level change is difficult to determine in both the burst and chronic model of destructive periodontal disease. With a model of slow, gradual change at many sites, the rate shown by Löe et al. (1986) represents an average change of only .028 mm. per site over a two month period of time. Attachment level is measured using a periodontal probe marked in 1 or 3 mm. increments. The measurements are commonly rounded off to the nearest millimeter, making one millimeter the minimum detectable change in attachment level. The rate of change estimated from a model of chronic disease of .028 mm. is considerably less than the minimum detectable change of 1 mm., thus making changes of .028 mm.

impossible to detect. Conversely, under a burst model, the low mean rate of change may represent a very small percentage of the sites changing at a more detectable level.

When detecting change in attachment level, the probability of incorrectly identifying changes at sites is nearly as great as the probability of real change occurring. To demonstrate this point, two calibration studies (Haffajee et al., 1983; Baderstein, A., Nilveus, R., and Egelberg, J., 1984) are compared to two longitudinal studies (Lindhe et al., 1983; Haffajee et al., 1983). Calibration studies are studies in which measurements are replicated at short periods of times. Therefore, it is assumed that replicated measurements are obtained when no real change in attachment level has occurred, and any difference in measurements replicated at the same site must be due to the error of the two measurements. In comparing these studies, change in attachment level is concluded when the difference in consecutive measurements is equal to or greater than 2 mm. In the calibration studies, from 3 to 6% of the sites have a difference in replicate measurements of 2 mm. or more. In the longitudinal studies approximately 9% of the sites over a period of one year and 16% of the sites over three years demonstrate differences in consecutive attachment level measurements of 2 mm. or more. Therefore, as many as 66% of the sites in the one year longitudinal study could be incorrectly identified as having changed. Thus, a significant percentage of the perceived change in attachment level, may be due to measurement error.

1.4 The use and evaluation of diagnostic rules

Many clinicians make decisions on whether or not destructive periodontal disease has occurred based on changes in attachment level measurements. These decisions are usually based on either implicit or explicit rules. Change in attachment

level is concluded when the difference in attachment level measurements taken at consecutive time points is greater than or equal to a given threshold of **k** millimeters. However, measurement of attachment level change includes both measurement error and actual change. If the change in attachment level measurement is greater than or equal to **k** millimeters, but true attachment level has not changed, then a false positive test to the diagnostic rule is obtained. If the change in attachment level measurement is not greater than or equal to **k** millimeters, but true attachment level is, then a false negative test is obtained. In order to evaluate the ability of a diagnostic test to correctly identify change, the impact of false tests must be taken into account.

The impact of false tests of diagnostic rules is evaluated by estimating the specificity, sensitivity, negative predictive value and positive predictive value of a diagnostic rule. To describe these values the following notation and definitions are introduced:

P(T-)	is the proportion of sites that test negative.
P(T+)	is the proportion of sites that test positive.
P(C+)	is the proportion of sites that have real change of attachment level.
P(C-)	is the proportion of sites that have no real change of attachment level.

Specificity is the proportion of sites with no change in attachment level that test negative. Sensitivity is the proportion of sites with change in attachment level that test positive (Yerushalmy, 1947). Positive predictive value is the relative proportion of positive tests that occur in the presence of real change in attachment level. Negative predictive value is the relative proportion of negative tests that occur in the absence of real change in attachment level (Imery, 1986).

Two of these four rates, sensitivity and the negative predictive value of a diagnostic rule, are not evaluated here. Previous reports (Ralls and Cohen, 1986; Aeppli, D. M., Boen, J. R., and Bandt, C. L. 1984; and Imery, 1986) present varied estimates of sensitivity, both within and among the various reports. The estimates are highly dependent on the assumed magnitude of actual change in attachment level and the threshold, **k**, used to detect the change. As the threshold is decreased or the assumed attachment level change increased, sensitivity increases. The broad range of estimates of sensitivity and those estimate's basis on arbitrary assumptions bring to question their value in evaluating the ability of attachment level measurements to detect change in attachment level. For these reasons, sensitivity is not estimated. The negative predictive value is a function of sensitivity and therefore it also is not estimated.

The two remaining rates, specificity and the positive predictive value of a diagnostic test, are complements to false positive rates. The type I error rate, **P(type I)**, is the proportion of sites with no change in attachment level that test positive. **P(type I)** is equal to one minus the specificity. Using the notation developed by Fleiss (1981, p. 4), $\mathbf{P_{f+}}$ is the relative proportion of positive tests that occur in the absence of real change and is equal to one minus the positive predictive value.

P(type I) provides the proportion of false positive decisions for the population of sites that have not changed. Of greater interest to the clinician is the false positive error rate for a diagnostic test in a specific clinical situation. Examples of clinical situations are the monitoring of change in attachment level before treatment, during treatment or after treatment. For these clinical situations, varying rates of attachment level change are obtained. $\mathbf{P_{f+}}$ provides an estimate of the proportion of tests that are

false in a specific clinical situation. P_{f+}, however, may be different for each specific clinical situation.

The proportion of positive tests that are false, P_{f+}, can be expressed in terms of the type I error rate **P(type I)**, the proportion of a positive tests to a decision rule **P(T+)**, and the probability of a site changing **P(C+)**:

$$P_{f+} = \frac{\text{P(type I)} * (1-\text{P(C+)})}{\text{P(T+)}} \qquad 1.1$$

To evaluate equation 1.1, the three quantities involved must be known; however, only two of the three can be obtained. **P(T+)** can be estimated by applying a diagnostic rule to longitudinal data and finding the proportion of sites that test positive to the rule. **P(type I)** must be determined from data where no real change in attachment level has occurred. An example of data where no real change has occurred is a data set in which measurements of attachment level are replicated at the same time, such as in calibration studies. The third quantity, **P(C+),** the proportion of sites that actually changed, cannot be readily determined. Since attachment level can not be measured without error, real change and the probability of it occurring can not be determined.

The common practice in evaluating diagnostic tests is to estimate **P(C+)** from another already established, highly accurate diagnostic test. Such a diagnostic test is referred to as the "gold standard." In finding change in attachment level, no "gold standard" exists. Since **P(C+)** cannot be determined, P_{f+} cannot be calculated. However, examination of equation 1.1 reveals that an upper bound to P_{f+} can be obtained. An upper bound for a given ratio of **P(type I)** to **P(T+)** is obtained when

$P(C+)$ approaches zero. Therefore, the upper bound of P_{f+} will be presented with the assumption that $P(C+)$ is zero. This is not an unreasonable assumption in a model of destructive periodontal disease where infrequent bursts of attachment level changes are assumed to occur.

1.5 Evaluation of the Type I error rate of the tolerance method

The tolerance method has been proposed by Haffajee, A. D., Socransky, S. S. and Goodson, J. M. (1983) as a method to find change in attachment level. Goodson (1986) estimates the type one error rate of the tolerance method by computer simulations. The method used by Goodson is a modification of the original method proposed by Haffajee. The tolerance method, as originally described, consists of comparing the difference in the means of paired measurements taken at consecutive time points to the maximum of three thresholds. The three thresholds are: 2 times the population standard deviation of the difference in replicated measurements; 3 times the subject standard deviation of the difference in replicated measurements; and 3 times the pooled standard deviation of the difference in replicated measurements of the site. If the mean difference is greater than or equal to the maximum, then the site is considered to have undergone change. However, in Goodson's simulation the tolerance statistic is compared to a single value of 2.5 mm. Goodson makes the additional assumption that the errors in the attachment level measurement are normally distributed. The normal distribution has a standard deviation estimated from the difference in replicate measurements on multiple sites within 56 subjects. A type I error rate of .00012 is estimated from the simulation. Goodson concludes from this low type I error rate that a false positive is an extremely rare event.

Kent and Goodson (1986), using the same data set, describe the distribution of the difference in replicate measurements. The distribution has a standard deviation of .77 mm. The distribution is symmetrical with skewness of -.099, but exhibits positive kurtosis of 9.7. Positive kurtosis suggests that the tails of the distribution are heavier than a normal distribution. If the tails of the normal distribution are used to estimate the tails of a distribution with positive kurtosis, then the resultant probabilities will be underestimated. Due to the discrete nature of attachment level measurements (all values are rounded to the nearest mm.), the distribution of replicate differences may not be properly estimated by a normal distribution. This suggests a possible problem in the estimation of type I error by the simulation method described by Goodson (1986). If the normal distribution is not appropriate, then other methods should be used to estimate the type I error rate.

1.6 Purpose

The purpose of this thesis is to estimate two false positive rates for two sets of diagnostic decision rules used in the detection of change in attachment level. The two rates to be estimated are: **P(type I),** the relative proportion of unchanged sites that test positive; and $\mathbf{P_{f+}}$, the relative proportion of positive tests that occur in the absence of change in attachment level. In the case of $\mathbf{P_{f+}}$ an upper bound will be estimated. To estimate **P(type I)** for a given decision rule, a resampling technique similar to bootstrapping will be used.

Two sets of decision rules will be evaluated. The first set of decision rules is based on single attachment level measurements at each time point. Change in attachment level is concluded when the absolute value of the difference in consecutive

attachment level measurements is greater than or equal to a given threshold **k.** This set of decision rules simulates the clinical practice of periodontics. The second set of decision rules is based on a pair of attachment level measurements at each time point. Change in attachment level is concluded when the absolute value of the difference in the mean of paired measurements is greater than or equal to a given threshold **k.** This second set of decision rules simulates some recent clinical research, where pairs of measurements of attachment level are taken (Goodson, 1986).

Chapter 2

Estimating false positive rates

2.1 Description of the bootstrap resampling technique

The previous chapter presented some of the problems encountered in estimating the type I error rate of various decision rules. A need was demonstrated for a technique to estimate the type I error rate that does not rely on a normal approximation to the distribution of replicate differences. An appropriate alternative technique is that of obtaining estimates of type I error rates by resampling the data. Resampling techniques estimate the distribution of the data by repeatedly and randomly sampling the data. This avoids making an assumption about the form of the underlying distribution of the data. The obvious advantage of resampling procedures is that theoretical calculations are not necessary to determine the distribution of a function of the data. The disadvantage of the method is the large amount of computer resources required to carry it out.

The resampling algorithm to be used is similar to the bootstrap resampling technique described by Efron (1982). The algorithm suggested by Efron has four steps:

1) Assume that the data consists of m independent and identically distributed observations from a unknown probability distribution F with parameter φ. The data are denoted by $x_1, x_2....x_m$.

2) Draw with replacement a sample $x_1^*, x_2^*....x_m^*$ from the data x_1,

$x_2 \ldots x_m$.

3) Calculate $\hat{A}$ based on $x_1^*, x_2^* \ldots x_m^*$, where $\hat{A} = f(x_1^*, x_2^* \ldots x_m^*)$ and $\hat{A}$ is an estimate of φ. Denote this as $\hat{A}_i$, where i=1, 2,...B. B is the number of bootstrap samples.

4) Repeat step 2 and 3 until i=B.

From the collection of B estimates of $\hat{A}$, an estimate of the distribution of $\hat{A}$ is obtained.

The resampling procedure used here is a modification of the bootstrap method. In the classical bootstrap procedure, if the sample size is **m**, **m** observations are used to estimate the statistic of interest. In the case of the decision rules used in finding change in attachment level, only **p** observations (1 in the case of single measurements and 2 for paired measurements) are necessary for the calculation of the test statistic.

Additionally, the goal of the resampling is to estimate the probability of a function of attachment level measurements meeting or exceeding a threshold **k**. This is demonstrated below.

$$\text{Type I error} = \frac{\#(f[Xs_1, Xs_2, \ldots Xs_p] \geq \mathbf{k})}{B} \qquad 2.1$$

where

\# is the number of times the function in the brackets meets or exceeds **k.**

k is the threshold to be evaluated.

$f(Xs_1^*, Xs_2^* \ldots Xs_p^*)$ is a function of a bootstrap sample.

B is the number of bootstrap samples.

This is in contrast to the more common applications described by Efron (1982). The usual application consists of either an estimate of the standard error of $\hat{A}$ or a confidence interval around it. However, the resampling procedure provides an estimate of the entire distribution of $\hat{A}$, so the probability of exceeding any value **k** can be estimated.

2.2 Data used for the estimation of false positive rates

In order to estimate the type I error rate, the distribution of functions of attachment level measurements, such as simple differences in consecutive attachment level measurements, must be determined under conditions of no change in attachment level. The replicate measurements that Goodson uses (1986) provide data under conditions of no real change. The data set consists of two measurements of attachment level taken at each time point **i**, $\mathbf{Al_{i,1}}$ and $\mathbf{Al_{i,2}}$ respectively. The time points are separated by two month intervals. The attachment level measurements are taken at 6 sites per tooth from 56 untreated periodontal subjects. Table 1 shows the distribution of the difference in measurements replicated at the same time point **i**, $\mathbf{Al_{i,1} - Al_{i,2}}$. This difference is obtained when no real change in attachment level could have occurred. Therefore, this data set can be used to estimate the type I error rate of various decision rules.

To estimate an upper bound to $\mathbf{P_{f+}}$, more information than is given in Table 1 is needed. In addition to the data presented in the table, the actual measurements taken at each time are needed. So that this analysis can be done, Goodson provides the entire data set. Table 2 presents the mean and range of mean clinical indices for the subjects in this study. The table demonstrates that the patient population has severe periodontal

Table 1

Distribution of differences between replicate measurements

Difference (mm.)	N	%
-8	2	0.00
-7	4	0.01
-6	7	0.01
-5	13	0.03
-4	36	0.07
-3	152	0.32
-2	946	1.97
-1	7,723	16.07
0	30,464	63.38
1	7,733	16.09
2	843	1.75
3	95	0.20
4	19	0.04
5	14	0.03
6	7	0.01
7	1	0.00
8	5	0.01

Note: Data are from 56 subjects representing 48,064 measurement pairs Goodson (1986).

Table 2

Summary of Clinical Indices

Index	Mean	Range
Attachment loss[1]	3.17	1.39 - 9.00
Pocket Depth[1]	3.25	2.30 - 5.96
Redness[2]	.46	.08 - 1.00
Bleeding on probing[2]	.22	.01 - 1.00
Suppuration[2]	.02	.00 - .23
Teeth affected ≤ 2 mm.[3]	.98	.96 - 1.00
Teeth affected ≤ 5 mm.[3]	.59	.04 - 1.00

1 In millimeters

2 Dichotomous index (0,1 values)

3 Proportion of teeth with at least one site of attachment loss greater than or equal to the value shown

disease with a wide range of mean clinical indices, representing a varied but severely involved periodontitis patient population.

2.3 Decision rules and their simulation by resampling

The first set of decision rules is based on single measurements at each time point. When the absolute difference in single measurements taken at consecutive time points is greater than or equal to a threshold **k**, it is concluded that a site has changed:

$$| Al_{i-1} - Al_i | \geq k \qquad 2.2$$

Goodson shows that both the difference in measurements taken at consecutive time points under a hypothesis of no change in attachment level, and the difference in measurements replicated at the same time, are equal to the difference in the errors of the measurements. If the errors of the measurements are independent, then both statistics should be distributed in an identical manner. Therefore, the type I error rate for the difference in measurements from consecutive time points can be evaluated from the distribution of replicate differences.

There are two purposes for evaluating decision rules based on the differences in single measurements. First, in the clinical practice of periodontics only single measurements of attachment level are made at each site. Thus, estimating false positive rates of decision rules based on single measurements is applicable to routine procedures used in clinical practice. The second purpose is to evaluate the accuracy of simulations based on both the resampling procedure and the normal distribution. This is done by evaluating the ability of each method to simulate the distribution of replicate differences. The second set of decision rules is based on paired attachment level measurements for a given site at each time. The statistic used in this set of decision rules is the absolute value of the difference in the mean of paired measurements, **D-pair**.

D-pair is also the statistic that Haffajee et al. (1983) used in their "tolerance method" :

$$\text{D-pair} = \left| \frac{Al_{i-1,1} + Al_{i-1,2}}{2} - \frac{Al_{i,1} + Al_{i,2}}{2} \right| \qquad 2.3$$

where $Al_{i,j}$ is the attachment level measurement at time **i** and examination **j.**

This equation can be rewritten to demonstrate that it is the mean of two differences in attachment level measurements:

$$\text{D-pair} = \left| \frac{(Al_{i-1,1} - Al_{i,1}) + (Al_{i-1,2} - Al_{i,2})}{2} \right| \qquad 2.4$$

This set of decision rules concludes change in attachment level when the absolute difference in the mean of paired measurements, **D-pair**, is equal to or greater than a given threshold, **k**:

$$\left| \frac{(Al_{i-1,1} - Al_{i,1}) + (Al_{i-1,2} - Al_{i,2})}{2} \right| \geq k \qquad 2.5$$

Under a null hypothesis of no change in attachment level, this function can be simulated by taking a random sample of two observations from the distribution of differences between replicate measurements, Table 1, and then by taking the mean of the two observations. Results will be shown in .5 mm increments for thresholds ranging from .5 mm. to 3.5 mm. This will include the threshold used by Goodson (1986) of 2.5 mm.

In order to estimate the proportion of positive tests that are false positives, $\mathbf{P_{f+}}$, the proportion of positive tests **P(T+)** needs to be estimated. **P(T+)** is estimated for periodontitis subjects with untreated periodontal disease. These patients are monitored every two months for up to two years. As discussed in section 1.4, by obtaining the ratio of the type I error rate, **P(type I),** to the proportion of positive tests, **P(T+)**, an upper bound to $\mathbf{P_{f+}}$ can be calculated.

Chapter 3

Results

Both simulation methods, resampling and using a normal distribution, are compared in their ability to reproduce the distribution of replicate differences. In the resampling method, observations are created by randomly sampling with replacement from the distribution in Table 1. In the normal distribution method, observations are created from a normal distribution with a zero mean and a variance estimated from the distribution of replicate differences. These observations are then rounded to the nearest millimeter. Table 3 demonstrates that the resampling method reproduces the distribution of replicate differences to a greater degree of accuracy than the normal approximation. The method based on the normal distribution is not able to reproduce the distribution of replicate differences because it overestimates the frequency of differences of 1 mm. and 2 mm. and underestimates the frequency of more severe differences (3 mm. or greater). The simulation based on a normal with 200,000 repetitions is not able to produce any differences of 5 mm. or greater. As a result of the failure of the normal distribution to adequately reproduce the distribution of replicate differences, the resampling method is used for the remainder of this thesis.

Table 3

Comparison of resampling and simulating a normal distribution in reproducing the distribution of replicate differences

	Goodson's actual data			
			Resampling*	Normal Distribution*
Difference (mm.)	Frequency	%	%	%
-8	2	0.004	0.006	
-7	4	0.008	0.008	
-6	7	0.015	0.015	
-5	13	0.027	0.033	
-4	36	0.075	0.062	
-3	152	0.316	0.331	0.105
-2	946	1.968	2.011	3.471
-1	7,723	16.068	16.094	22.670
0	30,464	63.382	63.386	47.329
1	7,733	16.088	16.024	22.763
2	843	1.754	1.751	3.560
3	95	0.200	0.184	0.102
4	19	0.040	0.043	0.002
5	14	0.029	0.027	
6	7	0.015	0.015	
7	1	0.002	0.002	
8	5	0.010	0.009	

* Estimates created by simulation, n=200,000

When decision rules are based on single measurements at each time point, type I error rates can be low, but the proportion of positive tests that are false remains high (Table 4). For thresholds of 3 mm. or greater, type I error rates are less than .01, but more than 3 out of 10 positive tests are false. Therefore, taking differences in consecutive single attachment level measurements results in a large proportion of false positive tests, even for a threshold as large as 3 mm.

When decision rules are based on paired measurements at each time point, false positive rates are lower than those found for single measurements (Table 5). Type I error rates of less than .01 are obtained with thresholds of 2 mm. or greater. The proportion of positive tests being false for the same thresholds ranges from .11 - .17, this compared to a range of .27 - .43 for single measurements. Thus, taking an additional measurement at each time point helps in the detection of change in attachment level by reducing $\mathbf{P_{f+}}$ by a factor of more than 2.

Also, note in Table 5 the published value of Goodson (1986). The value obtained by the resampling method is much larger than the value Goodson found using a normal approximation to the distribution of replicate differences.

Table 4

Type I error and false positive predictive rates of decision rules based on single measurements

A positive test to the decision rule occurs when

$$|Al_{i-1} - Al_i| \geq k$$

where Al_i is the attachment level measurement at time **i** and

k is the threshold.

k (mm.)	Type I error rate P(type I)	Proportion of positive tests P(T+)	P_{f+}
1	.37	.49	.74
2	.045	.10	.43
3	.0074	.0023	.32
4	.0023	.008	.27

Table 5

Type I error and false positive predictive rates of decision rules based on paired measurements at successive time points

A positive test to a decision rule occurs when

$$\left| \frac{(Al_{i-1,1} - Al_{i,1}) + (Al_{i-1,2} - Al_{i,2})}{2} \right| \geq k$$

where $Al_{i,j}$ is the attachment level measurement at time **i** and examination **j** and **k** is the threshold.

k (mm.)	Type I error rate P(type I)	Proportion of positive tests P(T+)	P_{f+}
.5	.55	.64	.86
1.0	.13	.27	.46
1.5	.026	.10	.26
2.0	.0067	.04	.17
2.5	.0027*	.02	.14
3.0	.0014	.01	.11
3.5	.0007	.006	.12

* This can be compared to the published result of Goodson (1986), where a normal distribution is used in his simulation. His estimate of the type I error rate is .00012.

Chapter 4

Conclusions

This thesis demonstrates that the assumption of normality used by Goodson results in the underestimation of the type I error rate of the tolerance method by a factor of 10. This underestimation is due to the positive kurtosis demonstrated in the distribution of replicate differences. Therefore, the assumption of normality does not seem warranted. It is shown here that a resampling technique more accurately estimates the type I error rate.

The estimates of false positive rates have important implications in the field of periodontics. When diagnostic decisions are based on single measurements, false positive rates are high. Even when thresholds as high as 3 mm. are used, over 3 out of 10 sites identified as "changed" have not changed. Unfortunately, in the clinical practice of periodontics, single measurements are commonly used. Therefore, clinicians who make treatment decisions based on attachment level measurements, may be treating a large percentage of sites that have not undergone destructive periodontal disease. Clinical periodontists generally regard a loss of attachment of 3 mm. or more as evidence of progressively worsening disease requiring additional therapy. The consequences of treating areas that are erroneously concluded as having progressed have to be compared to the consequences of not treating areas that are progressing. If a clinician treats sites when a change of 3 mm. in attachment level is detected, it is likely that as many as 32% of the sites may not have progressed. However, if the change in

attachment level is real and the site is not treated, a significant proportion of the attachment may be lost. Changes of 3 mm. are large compared to the length of the root of the tooth. Weine (1982, p. 208-209), using Black's (1902) description of tooth anatomy, presents average root length of 13 categories of teeth. Average root lengths range from 12 to 16.5 mm. for the 13 categories. If a tooth with a root of 14 mm. (near the middle of the range of average tooth length) has a change in attachment level measurements of 3 mm., the clinician is faced with a dilemma as to whether the site should be treated. The dilemma is increased if prior to the change of 3 mm., the site had already lost 50% of its attachment. In this situation the 3 mm. change represents nearly half of the remaining attachment. For these reasons, better measurement techniques would be beneficial in the clinical practice of periodontics.

A controversy exists in the periodontal literature on the ability of single attachment level measurements to find actual change in attachment level. Two recent reports are in general agreement with this study. Imrey (1986) evaluates the ability of single measurements of attachment level to find change in attachment level. He concludes: "If true disease is uncommon and sensitivity to it is not high, these false positives may exceed in number the true positives detected" (p. 521). Ralls and Cohen (1986) reach similar conclusions: "the major issue is that 'bursts' of change can be explained by chance events which arise from measurement error and which occur at low but theoretically expected levels" (p. 751). The results of the present research demonstrate that a large percentage of the perceived change in attachment level is due to measurement error, but not to the degree that Imrey (1986) and Ralls and Cohen (1986) suggest. These researchers attribute almost all the attachment level changes to measurement error. In contrast, Aeppli, D. M., Boen, J. R., and Bandt, C. L. (1984)

reach a different conclusion: "using an observed increase of greater than 1 mm. as a diagnostic rule leads to high sensitivity and yet satisfactorily high specificity" (p. 264).

All three of the above referenced studies base their conclusions on estimates of sensitivity and specificity. The methods of obtaining estimates of sensitivity and specificity vary between the studies. Aeppli, D. M., Boen, J. R., and Bandt, C. L. base their estimates of specificity and sensitivity on a calibration study involving 34 patients and 3 examiners. Their distribution of differences in replicated measurements is similar to the distribution that Goodson (1986) reports. Imrey (1986) and Ralls and Cohen (1986), instead of using actual data, simulate the distribution of differences by using a normal approximation with standard deviations of 1.125 mm. and 1 mm. respectively. Even though the methods of obtaining data vary, all the reports obtain high values of specificity (Table 6). However, estimates of sensitivity vary both within and among the three studies. Table 6 demonstrates that for similar thresholds the studies obtain a wide range of estimates of sensitivity. Within each study estimates of sensitivity are shown to be highly dependent on the assumed magnitude of actual change and the threshold used to detect the change. As the threshold decreases or the assumed attachment level change increases, sensitivity increases. The possible wide range of estimates that can be obtained within a study is demonstrated by Ralls and Cohen (1986). Their estimates of sensitivity range from .0668 to .9772. As discussed in chapter 1, the broad range of estimates of sensitivity and those estimates' basis on arbitrary assumptions brings to question their value.

Table 6

Comparison of sensitivity and specificity for various studies

For $Al_{i-1} - Al_i \geq 2.0$ mm.

and assuming that real attachment level change is 2 mm.

Study	Specificity	Sensitivity
Ralls and Cohen (1986)	.977	.50
Aeppli et al. (1984)	.979	.82
Imrey (1986)	.975*	.50
Goodson (1986)	.976**	---

* For $Al_{i-1} - Al_i \geq 2.5$ mm.
** calculated from frequency distribution

Table 7

Comparison of distribution of replicate differences for Aeppli et al.(1985) and Goodson (1986)

Proportion of differences ≥ k

k (mm.)	Aeppli et al. (1985)	Goodson (1986)
1	.178	.185
2	.021	.024
3	.002	.004

Ralls and Cohen (1986) attempt to explain the difference in conclusions between their study and Aeppli's. In their attempts they perpetuate some misconceptions. First, they state that Aeppli's estimate of the standard deviation of measurement error for a single measurement of attachment level is much lower than other published studies. However, Aeppli's standard deviation of .46 is very close to the value reported by Goodson (1986) of .55. In fact, the distribution of replicate differences from both studies are very similar (Table 7). Secondly, Ralls and Cohen report that the standard deviation of a single measurement is equal to the standard deviation of the difference times $\sqrt{2}$. However, the standard deviation of a single measurement is equal to the standard deviation of the difference divided by $\sqrt{2}$. Ralls and Cohen have another misconception. They misunderstand the 1 mm. rule of Aeppli's. Ralls and Cohen incorrectly believe that a positive response to the diagnostic rule is a difference in attachment level measurements greater than or equal to 1 mm. According to Aeppli, the rule is only greater than 1 mm. The later definition means that a positive response is obtained when the difference in attachment level measurements is greater than or equal to 2 mm. Table 6 demonstrates that Ralls and Cohen using the correct rule, obtain a specificity very close to the value that Aeppli obtains.

The difference in the conclusions of the studies is that diagnostic rules are not evaluated in the clinical situation where they are going to be used. However, conclusions and inferences about the use of diagnostic tests in clinical situations are made in the reports. Aeppli, D. M., Boen, J. R., and Bandt, C. L. feel that a high specificity and sensitivity are sufficient to conclude that a diagnostic test is adequate. Fleiss (1981, p. 7), however, shows that a diagnostic test with high specificity and sensitivity can result in a high proportion of incorrect diagnostic tests when detecting rare events. Aeppli, D. M., Boen, J. R., and Bandt, C. L. appropriately acknowledge

that if the frequency of change is low, a large proportion of positive tests may be false. However, their report is based on calibration data, so they can not evaluate $\mathbf{P_{f+}}$. In the absence of data to estimate $\mathbf{P_{f+}}$, Imrey (1986) and Ralls and Cohen (1986) speculate on its value. Therefore, their conclusions are heavily based on assumptions with no supporting data. This thesis estimates an upper bound to $\mathbf{P_{f+}}$. While this is a "worse case" estimate of $\mathbf{P_{f+}}$, it does provide a conservative evaluation of a diagnostic test. This evaluation is for the clinical situation of monitoring patients with untreated periodontal disease.

Untreated patients are not usually monitored in clinical practice. An analogous situation, however, is the monitoring of maintenance patients, patients previously treated for periodontal disease. Maintenance patients are brought in every three months for routine cleaning and scaling. The clinician must monitor the patient and make decisions on whether more aggressive therapy is necessary. The rate of change in these patients is shown to be lower than untreated patients (Pihlstrom et al., 1983; Knowles et al. 1979). Treated patients also have lower measurement error (Cerek et al., 1984). Therefore, it appears that the monitoring of these patients could have a problem similar to monitoring untreated patients. However, to determine if a problem exists, a similar analysis would have to be done on this patient population.

There may be situations in the clinical practice of periodontics in which single attachment level measurements would be adequate to monitor change in attachment level. The proportion of positive tests that are false goes down as the frequency of sites that are changing increases. An example of a high frequency of changes may be the comparison of measurements before and after periodontal therapy. A number of studies (Ramfjord et al., 1975; Pihlstrom et al., 1981; Isidor et al., 1984) show that there is considerable change in attachment level during these phases of treatment. A

much smaller proportion of positive tests that are false would be expected in the monitoring of patients during treatment.

One solution to the measurement error problem is to repeat measurements. As shown here, replicating measurements of attachment level reduces the rate of false positives encountered. If the difference in the mean of replicated measurements is greater than or equal to 2 mm., only 15% of the changes can be attributed to error. This is about a third of the value one obtains when single measurements are used. Basing decisions on the difference in the mean of paired measurements is similar to the tolerance method. The results of this study support the positions presented by Haffajee, A. D., Socransky, S. S. and Goodson, J. M.(1983) and Goodson (1986) that the tolerance method can properly identify change in attachment level.

It must be noted that this analysis pools all sites from all patients. The estimates of false positive rates are overall rates. They do not take into account variation due to individual patients or characteristics of individual sites. Baderstein, A., Nilveus, R., and Egelberg J. (1984) suggest that numerous site specific factors influence the error in attachment level measurements. The factors they suggest are the depth of the periodontal pocket and the type of tooth. Further investigation is needed to evaluate these factors.

References

Aeppli, D. M., Boen, J. R., and Bandt, C. L. (1984). Measuring and interpreting increases in probing depth and attachment loss. *Journal of Periodontology*, **56**: 262-264.

Axelsson, P. and Lindhe, J. (1978). Effect of controlled oral hygiene procedures on caries and periodontal disease in adults. *Journal of clinical Periodontology*, **5**: 133-151.

Baderstein, A., Nilveus, R., and Egelberg, J. (1984). Reproducibility of probing attachment level measurements. *Journal of Clinical Periodontology*, **11**: 475-485.

Black, G. V. (1902). *Descriptive anatomy of the human teeth.*. 4th ed., Philadelphia: The S. S. White Dental Manufacturing Co.

Cercek, J. F., Kiger, R. D., Garrett, S., and Egelberg, J. (1983). Relative effects of plaque control and instrumentation on the clinical parameters of human periodontal disease. *Journal of Clinical Periodontology*, **10**: 46-56.

Efron, B., (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions.* 2th ed., New York: John Wiley and Sons.

Goodson, J. M., Tanner, A. C., Haffajee, A. D., Sornberger, G. C., and Socransky, S. S. (1982). Patterns of progression and regression of advanced periodontal disease. *Journal of Clinical Periodontology,* **9**: 472-481.

Goodson, J. M. (1986). Clinical measurements of periodontitis. *Journal of Clinical Periodontology*, **13**: 446-455.

Haffajee, A. D., Socransky, S. S., and Goodson, J. M. (1983). Comparison of different data analysis for detecting changes in attachment level. *Journal of Clinical Periodontology*, **10**: 298-310.

Imrey, P. B. (1986). Considerations in the statistical analysis of clinical trials in periodontitis. *Journal of Clinical Periodontology*, **13**: 517-528.

Isidor, F., Karring, T., and Attström, R. (1984). The effect of root planing as compared to that of surgical treatment. *Journal of Clinical Periodontology,* **11**: 669-681.

Kent, R. L., and Goodsen, J. M. (1986). Statistical analysis of probeable attachment level measurements: Distribution characteristics. *Journal of Periodontal Research*, **65**: 523.

Knowles, J. W., Burgett, F. G. Nissle, R. R., Shick, R. A., Morrison, E. C., and Ramfjord, S. P. (1979). Results of periodontal treatment related to pocket depth and attachment level: Eight years. *Journal of Periodontology*, **50**: 225-233.

Lindhe, J., Haffajee, A. D., and Socransky, S. S. (1983). Progression of periodontal disease in adult subjects in the absence of periodontal therapy. *Journal of Clinical Periodontology*, **10**: 433-442.

Listgarten, M. A., Mao, R., Robinson, P. J., (1976). Periodontal Probing and the relationship of the probe tip to periodontal tissues. *J. Periodontol*, **47**: 511-513.

Löe, H., Anerud, A., Boysen, H. and Smith, M. (1978). The natural history of periodontal disease in man. *Journal of Periodontology*, **7**: 165-176.

Löe, H., Anerud, A., Boysen, H., and Morrison, E. (1986). Natural history of periodontal disease in man : rapid, moderate and no loss of attachment in Sri Lankan laborers 14 to 46 years of age. *Journal of Clinical Periodontology*, **13**: 431-440.

Pihlstrom, B. L., Ortiz-Campos, C., and McHugh, R. B. (1981). Randomized four-year study of periodontal therapy. *Journal of Periodontology*, **52**: 227-242.

Pihlstrom, B. L., McHugh, R. B., Oliphant, T. H., and Ortiz-Campos, C. (1983). Comparison of surgical and non-surgical treatment of periodontal disease. A review of current studies and additional results after 6 1/2 years. *Journal of Clinical Periodontology*, **10**: 524-542.

Ralls, S. A., and Cohen, M. E. (1986). Problems in identifying "bursts" of periodontal attachment loss. *Journal of Periodontology*, **57**: 746-752.

Ramfjord, S. P., Knowles, J. W., Nissle, R. R., Burgett, F. G., and Shick, R. A. (1975). Results following three modalities of periodontal therapy. *Journal of Periodontology*, **46**: 522-526.

Socransky, S. S., Haffajee, A. D., Goodsen, J. M. and Lindhe, J. (1984). New concepts of destructive periodontal disease. *Journal of Clinical Periodontology*, **11**: 21-32.

Suomi, J. D., Greene, J. C., Vermillion, J. R., Doyle, J., Chang, J. J., and Leatherwood, E. C. (1971). The effect of controlled oral hygiene procedures on the progression of periodontal disease in adults. Results after third and fourth year. *Journal of Periodontology*, **9**: 152-160.

Weine, F. S. (1982). *Endodontic Therapy*. 3rd ed.,St. Louis, C. V. Mosby Company.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep.*, **62:** 1432-1449.

Vita