Iain A. Pretty,[1] B.D.S. (Hons), M.Sc., and David Sweet,[2] D.M.D., Ph.D.

Digital Bite Mark Overlays—An Analysis of Effectiveness

REFERENCE: Pretty IA, Sweet D. Digital bite mark overlays—an analysis of effectiveness. J Forensic Sci 2001;46(6):1385–1391.

ABSTRACT: U.S. courts have stated that witnesses must be able to identify published works that define operational parameters of any tests or procedures that form the basis of scientific conclusions. Such works do not exist within the field of bite mark analysis. As the most commonly employed analytical technique in bite injury assessment, this study defines quantifiable variables for transparent digital overlays. A series of ten simulated, postmortem bites were created on pigskin and, with accompanying overlays, assembled into cases. Using two separate studies with four examiner groups, the study defined values of intra- and inter-examiner reliability, accuracy, sensitivity, specificity, and error rates for transparent overlays. Methods and statistical treatments from medical decision-making and diagnostic test evaluation were employed. Forced decision models and receiver operating characteristic analyses were utilized. Sensitivity and specificity values are described, and the results are consistent with other dental diagnostic systems. It was concluded that the weak inter-examiner reliability values explain the divergence of odontologists' opinions regarding bite mark identifications often stated in court. The effect of training and experience of the examiners was found to have little effect on the effective use of overlays within this study. The authors conclude that further research is required so that the results of the current study can be placed into context, but this represents a significant first step in establishing the scientific basis for this aspect of forensic dentistry.

KEYWORDS: forensic science, forensic dentistry, reliability, validity, examiner agreement, bite marks

It is not unusual to see dentists testifying in court. Forensic odontologists assist criminal proceedings by identifying the deceased victims of crime and by analyzing bite marks to identify the biter (1). Contemporary legal history is littered with cases where it has been possible to identify a bite on a victim to the person who has caused the bite. In many cases, this type of evidence may be crucial to the successful outcome of the trial (2). Bite mark evidence has been almost universally accepted in the courts, but the fundamental validity and scientific basis for its use is frequently challenged (2,3).

Rapid advances in forensic science have caused concern to the judicial system. Recent rulings, such as *Daubert* and *Kumho* in the United States, have placed a greater emphasis on the validity and reliability of opinion testimony based on supposed scientific principles. Judges have stated that witnesses must be able to identify published works that define the operational parameters of any tests or procedures that form the basis of scientific conclusions (2). Such works do not exist within the field of bite mark analysis (1).

The purpose of this study was to determine values of intra- and inter-examiner reliability, sensitivity, and specificity on both a dichotomous scale and the recommended American Board of Forensic Odontology conclusions scale (4). Methods from medical diagnostic assessments were employed to analyze the data. The impact of the examiners' training and experience was measured.

Materials and Methods

Selection of Examiners

To address the impact of training and experience on bite mark overlay use, the following groups of examiners were selected:

- Diplomates of the American Board of Forensic Odontology (ABFO).
- Members of the American Society of Forensic Odontology (ASFO).
- General Dental Practitioners (GDP).

The ABFO Diplomates were the examiners with the highest level of training and experience. Two separate groups were studied. The first ABFO group provided data for intra-examiner reliability. The second ABFO group was involved in determining the inter-examiner reliability.

Members of the ASFO who were practicing dentists with an interest in forensic dentistry and had been involved in at least one bite mark case or had attended a training course on the subject were recruited. General dental practitioners were recruited from a forensic dental study group concerned with responses to mass disasters. These dentists had no practical bite mark experience other than attending three lectures on the subject.

Ten simulated bite mark cases were presented to each of ten examiners. Each bite mark case included two suspects resulting in a total of 20 decisions for each examiner and 200 decisions for each examiner group. Overall, this represented 40 examiners (two ABFO groups, one ASFO group, and one GDP group) and 800 identification decisions.

Selection of Suspect Dentitions

Twenty-two sets (upper and lower) of high quality dental casts were selected to ensure that the bite marks represented a range of difficulty. This difficulty ranged from straight, even teeth to displaced, crowded teeth. Each of the ten bite mark cases had two sets of casts associated with it. One set of casts was used to produce the bite and the other was used as a foil (nonbiter). The casts that pro-

[1] Doctoral student, Faculty of Medicine, Department of Clinical Dental Sciences, The University of Liverpool, England.

[2] Director, Bureau of Legal Dentistry, Vancouver, British Columbia, Canada.

Received 15 Feb. 2001; accepted 12 March 2001.

duced the bite in each case were determined randomly. Case 3 and Case 4 had three sets of dental casts associated with them to create a situation in which neither suspect was the biter. In these cases, the third cast was used to produce the bite. Models were labeled "Suspect A" and "Suspect B" for each of the ten cases; the unseen biters were labeled "Suspect C" (See Table 1).

Production of Overlays

Sweet et al. describe the most accurate form of producing digital overlays that is currently available, and this method was used (5,6). Table 2 illustrates the equipment employed. This technique was used to produce 1:1 (life-sized) overlays of the anterior teeth of Suspect A and Suspect B for each case (See Fig. 1). Note that overlays were not produced for Suspect C in Cases 3 and 4.

Simulation of Bites on Animal Model

The use of animal skin analogues to produce simulated bite marks is well established within forensic dentistry (7). It was decided to create in situ postmortem bites on pigskin since this is widely accepted as an accurate analogue of human skin (8). Previous studies have used postmortem pigskin (7), antemortem dog skin (9), and postmortem sheepskin (10).

Two piglets (7 to 8 weeks old), freshly slaughtered, and weighing approximately 15 kg each, were obtained from a local abattoir. Anatomical locations were selected on each piglet that represented areas of minimal skin curvature and distortion. The lower abdomen and ears were found to be ideal sites. The dental casts from each randomly selected biter were clamped to the skin for 10 m to create a bite mark. Following the release of the clamp the bite mark was subjectively examined to ensure that sufficient detail was recorded.

TABLE 1—*Distribution of biters among the ten simulated cases.*

Case Number	Suspect A	Suspect B	Suspect C
1	Biter	Non Biter	
2	Non Biter	Biter	
3	Non Biter	Non Biter	Biter
4	Non Biter	Non Biter	Biter
5	Non Biter	Biter	
6	Biter	Non Biter	
7	Biter	Non Biter	
8	Non Biter	Biter	
9	Biter	Non Biter	
10	Non Biter	Biter	

TABLE 2—*Equipment for production of digital overlays.*

Item	Model	Manufacturer	Location
Scanner	HP ScanJet 4c	Hewlett Packard Co.	Palo Alto, CA
Scanning software	HP DeskScan	Hewlett Packard Co.	Palo Alto, CA
Scale	ABFO No. 2	Lightning Powder Co., Inc.	Salem, OR
Computer	PowerMac G3	Apple Computer Inc.	Cupertino, CA
Imaging software	Photoshop v5.0.2	Adobe Systems Inc.	Mountain View, CA
Laser printer	LaserWriter 4/600PS	Apple Computer Inc.	Cupertino, CA
Transparency film	Catalogue no. 9055	3M Visual Systems Division	Austin, TX

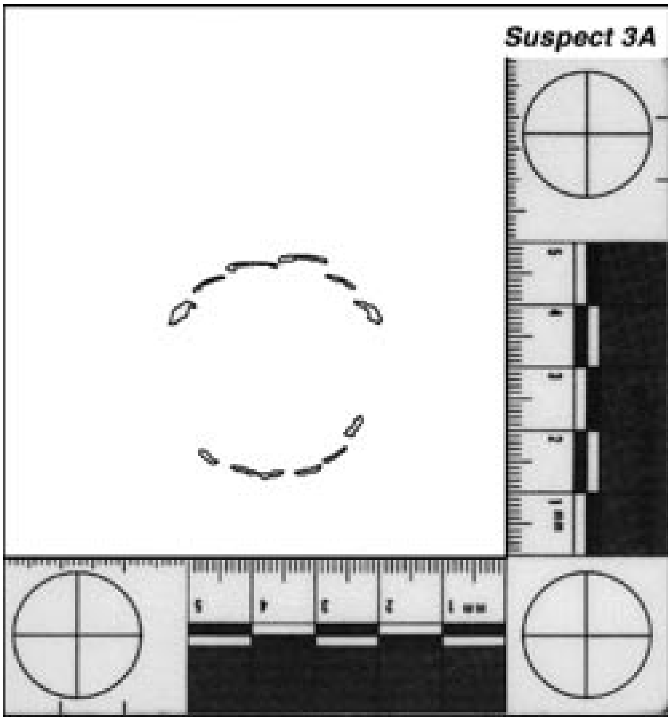

FIG. 1—*Digital overlay for Case 3, Suspect A showing 12 anterior teeth.*

The injury was photographed following the ABFO guidelines for evidence collection (4). Color and black-and-white photographs were exposed with the ABFO No. 2 scale in place. The best reproduction of each bite mark was selected and photographs were printed at 1:1 (life-sized). Subsequently, the photographs were scanned into a computer and stored in JPEG format at 1440 dpi. These images were printed with an inkjet printer at 1440 dpi on photographic paper. Prints were made for each examiner. An example of one of the bitemark photographs is shown in Fig. 2.

Study 1: Intra-Examiner Reliability

An anonymous group consisting of ten diplomates of the ABFO was selected. Each participant received ten simulated bite mark cases, which contained one color and one black-and-white photograph of the bite, two computer-generated overlays labeled Suspect A and Suspect B, occlusal views of the suspects' dentition, instructions, and an answer sheet. The examiners were asked to determine whether each suspect was the biter or not for the appropriate case. The examiners were asked to indicate "Positive" for the biter and "Excluded" for the nonbiter. No other option was available.

Ten diplomates returned answer sheets for the first assessment (100%). However, only seven returned the study materials. Since three Diplomates retained the materials, the second assessment to study intra-examiner reliability, which was carried out three months later involved only seven of the Diplomates. These diplomates were sent the same materials again and asked to repeat the exercise.

The results were entered into tables and treated statistically. Each of the examiners' responses was compared between the two different assessments and kappa was applied to correct for chance. PEPI statistical software was used to analyze the raw data (11).

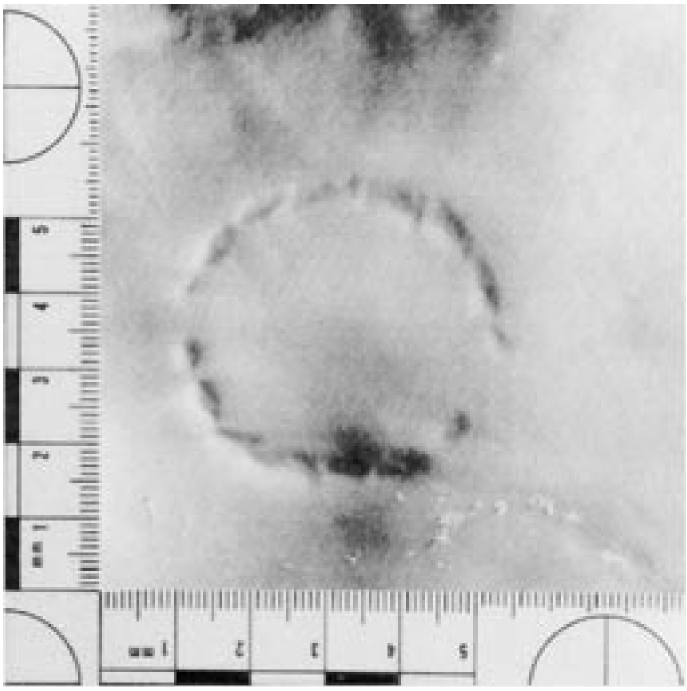

FIG. 2—*Example of case photograph from a simulated bite mark on pigskin.*

Study 2: Inter-Examiner Reliability

Three groups consisting of ten diplomates of the ABFO, ten members of the ASFO, and ten general dental practitioners were selected. Each participant received ten bite mark cases, which contained one color and one black-and-white photograph of a simulated bite mark, two computer-generated overlays labeled Suspect A and Suspect B, occlusal views of each suspect's dentition, instructions, and an answer sheet. The instructions and answer sheet were revised from Study 1 to make available the five levels of certainty described by the American Board of Forensic Odontology (4) and a "Don't Know" option within the forced decision model (FDM). Thirty examiners (100%) returned responses. Receiver-operating characteristics (ROC) were used to analyze the multiple-threshold data. Results were entered into tables and analyzed using the PEPI statistical application (11).

Results

Intra-Examiner Reliability

Seven examiners returned completed answer sheets on both occasions (70%) and the intra-examiner reliability was calculated for each (See Table 3). Kappa values were calculated to measure agreement between each of the examinations and to control for chance agreement (12). The kappa values varied from 0.30 to 1.00, or from fair to almost perfect agreement (13). Mean kappa was 0.72, indicating substantial agreement. Percent agreement (non-chance corrected) ranged from 65 to 100% with a mean value of 87.2%.

The mean accuracy for the seven examiners' first and second attempts was 85.7% and 83.5% respectively, with no statistically significant difference between the attempts ($p = 0.6286$). When examining kappa values for comparisons with the gold standard, a mean of 0.70 resulted from the first examination. This decreased slightly to 0.65 for the second examination. Both scores rate as substantial agreement and no significant differences were detected between the attempts ($p = 0.5568$).

The mean values for sensitivity (79.8%) and specificity (90.0%) for the first examination were calculated and compared with the mean sensitivity (73.2%) and specificity (89.3%) values for the second examination. No statistically significant difference was detected between these values (sensitivity $p = 0.5218$, specificity $p = 0.5792$).

Inter-Examiner Reliability

Thirty examiners (ten ABFO, ten ASFO, and ten GDP) returned completed answer sheets. The multiple-threshold ROC data reveal two main results: (a) the individual sensitivity and specificity of each conclusion threshold, and (b) the area under the curve (AUC) as a measure of overall effectiveness (See Table 4). Youden's Index, a measure of agreement using sensitivity and specificity, was also calculated for each of the five possible conclusion levels (See Table 5). The closer Youden's Index is to 1.0 the greater the level of agreement.

ABFO Diplomates—Forced Decision Model

Ten diplomates of the ABFO returned completed answer sheets (100%). Out of 200 decisions, 28 (14%) were "Don't Knows." However, 24 (12%) of these "Don't Knows" were attributable to two examiners (Examiner 2 = 16, Examiner 10 = 8). Excluding these examiners, the uncertain decisions are reduced to only 4 (2%). Sensitivity was calculated for each examiner and ranged

TABLE 3—*Study 1 summary data illustrating percentage agreement between examinations conducted three months apart.*

Examiner	Kappa	S.E.	Percent Agreement
1	0.30	0.222	65
2	0.38	0.219	70
3	1.00	0.224	100
4	1.00	0.224	100
5	0.52	0.224	80
6	0.88	0.222	95
7	1.00	0.224	100
Mean	0.72		87.2%

TABLE 4—*Mean values from ROC analyses.*

Mean Values	ABFO (%)	ASFO (%)	GDP (%)
Area Under the Curve	80.5 ± 11.8	81.0 ± 8.8	80.8 ± 8.0
Sensitivity			
Reasonable Medical Certainty	27.5	23.8	12.5
Probable	57.5	53.8	60.0
Possible	81.3	72.5	76.3
Exclusion	88.8	77.5	60.0
Inconclusive	100.0	100.0	100.0
Specificity			
Reasonable Medical Certainty	98.3	98.5	99.2
Probable	94.9	94.3	93.4
Possible	55.3	74.4	64.2
Exclusion	47.7	68.7	55.9
Inconclusive	0.0	0.0	0.0

TABLE 5—*Summary of ROC results for the three groups studied.*

Level of Conclusion	Sensitivity (%)	Specificity (%)	Youden's Index
ABFO diplomates			
Reasonable Medical Certainty	27.5 ± 24.1	98.3 ± 5.2	0.26
Probable	57.5 ± 26.5	94.9 ± 11.0	0.52
Possible	81.3 ± 22.2	55.3 ± 30.0	0.40
Exclusion	88.8 ± 19.1	47.7 ± 24.0	0.36
Inconclusive	100.0 ± 0.0	0.0 ± 0.0	0.00
ASFO Members			
Reasonable Medical Certainty	23.8 ± 17.1	98.5 ± 4.9	0.24
Probable	53.8 ± 17.7	94.3 ± 8.4	0.48
Possible	72.5 ± 12.9	74.4 ± 11.2	0.47
Exclusion	77.5 ± 14.1	68.7 ± 14.7	0.46
Inconclusive	100.0 ± 0.0	0.0 ± 0.0	0.00
GDP Novices			
Reasonable Medical Certainty	12.5 ± 11.8	99.2 ± 2.3	0.13
Probable	60.0 ± 18.4	93.4 ± 5.3	0.55
Possible	76.3 ± 10.9	64.2 ± 11.9	0.37
Exclusion	83.6 ± 10.3	55.9 ± 11.3	0.37
Inconclusive	100.0 ± 0.0	0.0 ± 0.0	0.00

from 28.6 to 100% with a mean sensitivity of 73.7 ± 22.0%. Specificity for this group ranged from 54.5 to 100% with a mean specificity of 84.1 ± 14.9%. There was no significant difference between the sensitivity and specificity scores ($p = 0.2721$).

Accuracy, determined as percent agreement with the gold standard, ranged from 65.0 to 100% with a mean value of 83.2%. Agreement determined by Cohen's Kappa ranged from 0.22 (fair agreement) to 1.00 (almost perfect agreement). Mean kappa was 0.58 (moderate agreement). Mean false positive rate (FPR) was 15.9%, ranging from 0 to 45.5%. Mean false negative rate (FNR) was 25.0%, ranging from 0 to 71.4%. Positive predictive value (PPV) ranged from 55.5 to 100% with a group mean of 77.7%. Negative predictive value (NPV) ranged from 66.6 to 100% with a group mean of 83.2%.

ROC Analysis—The mean sensitivity, specificity, and Youden's Index for each of the conclusion levels is shown in Table 5. The AUC for the ABFOs ranged from 62.0 to 97.7% (mean 80.5 ± 11.8%).

Reliability—Using Cohen's Kappa, each of the examiners was paired and compared using a crosswise system based on their FDM decisions. From these data it was determined that ten pairs (22%) had slight agreement, 11 pairs (24%) had fair agreement, 13 pairs (29%) had moderate agreement, three pairs (7%) had substantial agreement and eight pairs (18%) had almost perfect agreement. Mean kappa from the crosswise analysis was 0.47 ± 0.31 (moderate agreement).

ASFO Members—Forced Decision Model

Ten members of the ASFO returned completed answer sheets (100%). Out of 200 decisions, 18 (9%) were "Don't Knows." Sensitivity was calculated for each examiner and ranged from 28.6 to 85.7% with a mean sensitivity of 60.9 ± 22.9%. Specificity for this group ranged from 34.6 to 100% with a mean specificity of 82.4 ± 19.7%. There was no significant difference between the sensitivity and specificity scores ($p = 0.378$). Accuracy, determined as percent agreement with the gold standard, ranged from 55.0 to 94.1% with a mean value of 75.8%. Agreement, determined by Cohen's Kappa, ranged from 0.16 (slight agreement) to 0.88 (almost perfect agreement). Mean kappa was 0.50 (moderate agreement).

Mean FPR was 11.9%, ranging from 0 to 27.3%. Mean FNR was 39.3%, ranging from 14.3 to 74.4%. PPV ranged from 59.9 to 100% with a group mean of 79.7%. NPV ranged from 58.4 to 91% with a group mean of 78.1%.

ROC Analysis—The mean sensitivity and specificity for each of the conclusion levels is shown in Table 5. The AUC for the ASFO members ranged from 62.5 to 89.6% (mean 81.0 ± 8.8%).

Reliability—Using Cohen's Kappa, it was determined that three pairs (7%) had poor agreement, five pairs (11%) had slight agreement, nine pairs (20%) had fair agreement, 16 pairs (36%) had moderate agreement, 11 (24%) pairs had substantial agreement and one pair (2%) had almost perfect agreement. Mean kappa from the crosswise analysis was 0.44 ± 0.22 (moderate agreement).

General Dental Practitioners (GDP)—Forced Decision Model

Ten GDPs returned completed answer sheets (100%). Out of 200 decisions, 15 (7.5%) were "Don't Knows." Sensitivity was calculated for each examiner and ranged from 62.5 to 100% with a mean sensitivity of 80.7 ± 13.5%. Specificity for this group ranged from 50 to 100% with a mean specificity of 77.9 ± 15.0%. There was no significant difference between the sensitivity and specificity scores ($p = 0.6001$). Accuracy, determined as percent agreement with the gold standard, ranged from 55.6 to 84.2% with a mean value of 74.7%. Agreement, determined by Cohen's Kappa, ranged from 0.14 (slight agreement) to 0.89 (almost perfect agreement). Mean kappa was 0.56 (moderate agreement).

Mean FPR was 22.0%, ranging from 0 to 50.0%. Mean FNR was 19.3%, ranging from 0 to 37.5%. PPV ranged from 46.0 to 100% with a group mean of 72.7%. NPV ranged from 70.1 to 100% with a group mean of 85.7%.

ROC Analysis—The mean sensitivity, specificity, and Youden's Index for each of the conclusion levels is shown in Table 5. The AUC for the GDPs ranged from 64.1 to 90.6% (mean 80.8 ± 8.0%).

Reliability—It was determined that three pairs (7%) had poor agreement, six pairs (13%) had slight agreement, eight pairs (18%) had fair agreement, 17 pairs (38%) had moderate agreement, ten pairs (22%) had substantial agreement and one pair (2%) had almost perfect agreement. Mean kappa from the crosswise analysis was 0.45 ± 0.23 (moderate agreement).

Comparison of the Three Examiner Groups—Table 5 shows data from the ROC results of the three groups. Table 6 shows a comparison of mean values obtained from the FDM study. There was no statistically significant difference between the distributions of "Don't Knows," kappa values, AUC, accuracy, sensitivity, or specificity between the three groups of examiners when tested with ANOVA.

Discussion

A key feature of modern forensic science is that scientific principles are no longer accepted based on opinion or anecdotal beliefs. This new doctrine has been enforced by legal judgments, such as

TABLE 6—*Mean values for the FDM and crosswise kappa analyses.*

Mean Values	ABFO	ASFO	GDP
Don't Knows	14.0%	9.0%	7.5%
Sensitivity	73.7 ± 22.0%	60.9 ± 22.9%	80.7 ± 13.5%
Specificity	84.1 ± 14.9%	82.4 ± 19.7%	77.9 ± 15.0%
Accuracy	83.2%	75.8%	74.7%
Kappa (Gold standard)	0.58	0.50	0.56
Kappa (Crosswise)*	0.47	0.44	0.45
False Positive Rate	15.9%	11.9%	22.0%
False negative Rate	25.0%	39.3%	19.3%

* Inter-examiner crosswise kappa comparisons.

those described in *Daubert* and *Kumho.* Claims are now to be checked against empirical evidence. The value of this evidence is based on the way it has been collected and presented (14). The purpose of this study was to establish empirical justification for the use of digital overlays in bite mark analysis.

The increased interest in evidence-based medicine and dentistry has revitalized techniques for the assessment of diagnostic effectiveness. The discipline of medical-decision making has employed these techniques in increasingly novel ways to challenge the basis upon which clinical practice is built. Using these techniques, this study has determined quantitative values for the analysis of overlay effectiveness.

During the initial planning stages of this project, considerable thought was given to the use of cases employing either real or simulated bites. The use of real forensic cases as study material has advantages. First, authenticity is assured. Materials used are the same as those handled by forensic dentists during routine casework. Second, many examples of bite marks exist both at the author's laboratory and in other centers. Therefore, the collation and duplication of such materials would be straightforward.

But, several disadvantages are also associated with the use of real cases. The most important of these is that of the gold standard. One of the criteria for assessing the effectiveness of a particular test is to ensure that it is compared against a suitable gold standard. The use of real case materials requires that the conclusions of the original examining odontologist are regarded as such a standard. Due to the lack of published studies, it is impossible to determine how accurate these original conclusions are likely to be. Indeed, it is the purpose of the current study to provide such data.

The use of simulated bite marks enabled greater control over the injury. Variables such as anatomical location, the teeth used to create the bite, the number of teeth in the bite, and the collection of the evidence were easily controlled and standardized. The use of simulations also permitted a consistent quality of materials to be produced, allowing parity between each of the study cases, and removing any potential biases introduced by case variability. However, simulations do have limitations. Significantly, the simulated bites were not on human skin.

Postmortem bites, as used in this study, do not display any of the ecchymosis or bruising patterns that are seen in antemortem or perimortem bite injuries and this could be considered a limitation. However, postmortem injuries do record the details of teeth well. The use of postmortem simulated bites is well accepted within forensic dental research (6,15).

Before discussing the effectiveness of the overlays, it is important to discuss the issue of examiner and test separation so that the results from the FDM and ROC analyses can be placed in the correct context. The performance of individual examiners and their decision-making processes were thought to be separate entities. Originally, it was decided to assess the use of overlays in the identification of biters. To this end, materials supplied to the examiners were limited to those that permitted the use of overlays only. But it was discovered that examiner performance and decision making are not separate. The use of bite mark overlays has been shown to be both examiner and case sensitive. And despite the objectivity of the overlay production technique, the subsequent application of that technique is highly subjective (16). In tests where subjectivity is high, there is always interplay between the operator and the test (17). The separation of operator and test in assessment of performance is impossible. With this caveat in mind, the discussion of the examiners' performance follows.

FDM Performance

The forced decision model allowed the use of simple statistical analysis. The use of terms such as false positive and true negative are easily understood. Hence, the power of this model is in its ease of use and explanation of results. However, there are drawbacks to the model. First, the American Board of Forensic Odontology recommends the use of particular levels of conclusion that are not replicated in the dichotomous decisions offered by the FDM. (There is a speculative argument, however, that the recommended levels of conclusion are simply extrapolated by courts and jurors to a positive or a negative judgment.) Second, the FDM is especially prone to influence by the personal threshold of the examiner.

This study resulted in 539 decisions from the FDM (excluding "Don't Know" decisions). The data that were most useful were the values of sensitivity, specificity, accuracy, and kappa agreement with the gold standard. It should be noted that no forensic dental study, either on the subject of bite marks or on other topics, describing these values was found in the literature. This makes it difficult to compare the values obtained for overlay effectiveness in the current study to other tests in forensic dentistry.

Sensitivity values for the three groups of examiners were not significantly different. The mean sensitivity from the three groups was 71.8%. The GDP novices had the smallest standard deviation among the groups (GDP>ABFO>ASFO) and achieved the highest sensitivity. Specificity values were not significantly different for the three groups. The mean specificity was 81.5%. The ABFO expert group achieved the highest score. In no group was there a significant difference between the sensitivity and specificity scores. These mean values are similar to values for sensitivity and specificity from other dental diagnostic tests.

The use of percentage agreement (accuracy) and kappa allowed a different perspective on the data obtained. In simple terms, how often were the examiners correct? Percentage agreement is a simple measure of this, and the mean across all three groups was 77.9%. The ABFO diplomates were the most accurate examiners scoring a group mean of 83.2%. However, the differences between the groups were small and not statistically significant.

It is interesting to note that two of the diplomates chose "Don't Know" responses for more than half of the cases, resulting in over 85% of the "Don't Know" decisions for this group. Significantly, both of these participants obtained 100% accuracy. This could indicate that they had very high personal thresholds to identify or exclude biters. Mathematically, their responses resulted in increasing the diplomates' mean accuracy. When these participants are removed from consideration, the mean accuracy of the diplomates group dropped to 78.5%. The results indicate that these two examiners are unlikely to render opinions in bite mark cases that are presented to them. However, if they were prepared to reach a conclusion, then it would most likely be highly accurate.

A more powerful technique for quantifying agreement with a gold standard is the chance-corrected kappa value. The mean for all three groups with this value was 0.54; the diplomates scored the highest kappa at 0.58. When Examiners 2 and 10 were removed, the mean kappa for diplomates dropped to 0.54, which placed the GDP kappa (0.56) as the highest. Regardless if these outlying examiners are included or excluded, the mean kappa score for all three groups falls into the "moderate agreement" category of the Landis rating scale (12,13).

No significant difference was detected between the three groups of examiners using any of the measured values. This indicates that training and experience have little effect on the application of overlays to bite mark identifications. However, caution must be applied in this conclusion since more detailed questionnaires would be required to identify correctly all of the variables surrounding experience and training.

ROC Analysis

The use of ROC enabled a range of conclusions, including "Don't Knows," to be incorporated into the analysis. Because this technique allowed the examiners to express their certainty within the established levels of conclusions, the operator sensitivity issues found in the FDM were minimized. ROC analysis provides a means by which the identification of biters using transparent overlays can be distinguished from the judgment of the operator. This separation is achieved by using a rating scale that is equivalent to varying the examiner's personal threshold while holding the properties of the bite mark constant. The area under the curve provides an objective parameter of the diagnostic accuracy of the test (the ability to determine biters) that is far superior to comparing single combinations of specificity and sensitivity because the influence of threshold is eliminated (18–20) (See Table 4). The AUC is a combination and generalization of the concepts of sensitivity and specificity into a single measure of accuracy (21). In this study, the AUC values from the three groups were very similar, with the ASFO members having the value closest to 100% (perfect diagnostic test). Six hundred decisions made up the AUC analysis. The mean AUC for the combined groups was 80.7%, which means that the biter was correctly determined approximately eight out of ten times.

It is difficult to place this result into context. A value of 50% assumes that a test is nondiagnostic. Thus, bite mark overlays are closer to the perfect diagnostic test than a purely random allocation of biters and nonbiters. Whittaker's study determined a mean AUC of 63% for the determination of whether bites were caused by adults or children (22). Comparison of these results with those of the current study indicate that the use of overlays in determining biters is more effective than the subjective determination of biter age group. But, this is not a particularly useful comparison and serves only to allow a point of reference. Further research into bite mark identification techniques is required to produce a range of AUC values from other methods and contexts. These data will then enable a comparison of techniques and move the discipline to a more evidence-based approach. The ease by which AUC can be calculated and compared promises to allow exciting additional research possibilities in the future. Studies could be carried out using the same base materials as in this study (i.e., bite mark photographs) but adding other items of dental evidence from suspects, including wax test bites or dental casts. Following calculation of the area under the curve, it would be possible to determine the relative impact of each item on the identification of biters from bite marks.

Conclusions

The continued use of computer-generated overlays in bite mark analysis appears to be justified, although further work is required to investigate the effect of examiner factors. In this study, no statistically significant differences were detected between the three examiner groups. This suggests that training and experience in forensic casework does not affect the success of overlays in correctly determining the biter. This work has satisfied the requirements of *Daubert* in relation to determining error rates and other quantifiable values.

This study has examined the scientific basis for bite mark comparisons. The significance of the results will be realized in courts of law. While the overall effectiveness of overlays has been established, the variation in individual performance of odontologists is of concern. This variation is of particular importance to those odontologists testifying in court who must be aware of their own values of accuracy and reliability. Poor performance as an expert witness during testimony can have very serious implications for the accused, the discipline, and society.

Acknowledgments

The authors wish to sincerely thank all of the examiners that took part in this study, and Drs. Douglas Waterfield, and Gerardo Maupome. This work was supported by research grants from the American Board of Forensic Odontology and the American Society of Forensic Odontology. Dr. Pretty is supported by the Forensic Science Society (UK). This work is dedicated to the memory of Dr. Bruce R. Rothwell, D.M.D., M.S.D.

References

1. Sweet D, Pretty IA. A look at forensic dentistry—Part 2: Teeth as weapons of violence—identification of bitemark perpetrators. Br Dent J 2001;190(8):415–8.
2. Pretty IA, Sweet D. A Comprehensive examination of bitemark evidence in the American legal system. In: Proceedings of the American Academy of Forensic Science; 2000; Reno, NV; 2000:146.
3. Hale A. The admissibility of bitemark evidence. So Calif Law Rev 1978;51(3):309–34.
4. American Board of Forensic Odontology, Inc. Guidelines for bite mark analysis. JADA 1986;112(3):383–6.
5. Sweet D, Parhar M, Wood RE. Computer-based production of bite mark comparison overlays. J Forensic Sci 1998;43(5):1050–5.
6. Sweet D, Bowers CM. Accuracy of bite mark overlays: a comparison of five common methods to produce exemplars from a suspect's dentition. J Forensic Sci 1998;43(2):362–7.
7. Whittaker DK. Some laboratory studies on the accuracy of bite mark comparison. Int Dent J 1975;25(3):166–71.
8. Zhang Z, Monteiro-Riviere NA. Comparison of integrins in human skin, pigskin, and perfused skin: an in vitro skin toxicology model. J Appl Toxicol 1997;17(4):247–53.
9. Rawson RD, Vale GL, Sperber ND, Herschaft EE, Yfantis A. Reliability of the scoring system of the American Board of Forensic Odontology for human bite marks. J Forensic Sci 1986;31(4):1235–60.
10. Ligthelm AJ, DeWet FA. Registration of bite marks: a preliminary report. J Forensic Odontostomatol 1983;1(1):19–26.
11. Gahlinger PM, Abramson J. Computer programs for epidemiologists (PEPI). 2nd ed. London: Brixton Books and Software; 1999.

12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.
13. Koch GG, Landis JR, Freeman JL, Freeman DH, Lehnen RC. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 1977;33(1):133–58.
14. Dunn G, Everitt B. Clinical bio-statistics—an introduction to evidence-based medicine. London: Edward Arnold; 1995.
15. Rawson RD, Bell A, Kinard BS, Kinard JG. Radiographic interpretation of contrast-media-enhanced bite marks. J Forensic Sci 1979;24(4): 898–901.
16. Strom F. Investigations of bitemarks. J Dent Res 1963;42(1):312.
17. Brunette D. Critical thinking. London: Quintessence Books; 1998.
18. Van Erkel AR, Pattynama PM. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. Eur J Radiol 1998;27(2):88–94.
19. Moise A, Clement B, Ducimetiere P, Bourassa MG. Comparison of receiver operating curves derived from the same population: a bootstrapping approach. Comput Biomed Res 1985;18(2):125–31.
20. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiol 1983;148(3):839–43.
21. Swets JA, Pickerr RM. Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press.
22. Whittaker DK, Brickley MR, Evans L. A comparison of the ability of experts and non-experts to differentiate between adult and child human bite marks using receiver operating characteristic (ROC) analysis. Forensic Sci Int 1998;92(1):11–20.

Additional information and reprint requests:
Iain A. Pretty, B.D.S., M.Sc.
The University of Liverpool
Edwards Building, Daulby Street
Liverpool, L69 3GN
England
Phone: 0151 706 5288
E-mail: ipretty@liv.ac.uk