J Forensic Sci, November 2003, Vol. 48, No. 6
Paper ID JFS2002032_486
Available online at: www.astm.org

Iain A. Pretty,[1] *B.D.S. (Hons), M.Sc.; Robert J. Pretty*[2]*; Bruce R. Rothwell,*[3] *D.M.D.; M.S.D.; and David Sweet,*[4] *D.M.D., Ph.D.*

The Reliability of Digitized Radiographs for Dental Identification: A Web-Based Study

ABSTRACT: In the era of *Daubert* and other judicial rulings pertaining to the acceptability of forensic evidence, it is increasingly important that experts are able to testify that their methods have been scientifically tested and that error rates and other factors relating to reliability have been published. The purpose of this study was to determine the reliability of digitized radiographic comparisons for the purposes of dental identification. Participants with various forensic backgrounds and experience levels were passively recruited to the website. Ten forensic identification cases composed of antemortem and postmortem dental radiographs were supplied to examiners using a bespoke website. Participants responded to the cases on two occasions after a one-month washout interval using the ABFO conclusion levels for forensic identifications. A total of 115 first attempts and 87 matched second attempts were received. Of the total responses, 72% were dentally trained respondents who had completed at least one forensic identification case; of these, 38% were experienced forensic dentists who had completed more than 25 identifications. Data relating to accuracy, intra- and inter-examiner agreement, and the effect of case difficulty are presented. Mean accuracy was 85.5% for all cases, with the experienced forensic dentists obtaining a 91% success rate. The inter-examiner agreement on the negative identification cases was classified as poor. The data suggest that dental identifications resulting from the comparison of postmortem and antemortem radiographs are valid, accurate, and reliable when undertaken by experienced odontologists.

KEYWORDS: forensic science, forensic dentistry, reliability, validity, examiner agreement, identification

The process of employing dental records to identify individuals is well established and will usually represent the majority of a forensic odontologist's caseload (1–6). Dental identification is premised upon the distinctive features of the human dentition in terms of anatomy, pathology, and treatments. The use of postmortem radiographs is key to this process by enabling the examination of the dental evidence and the comparison of these to antemortem radiographs taken by the deceased's dentist (1). Forensic odontologists are employed in cases of severe head and neck trauma, gross decomposition, burning, and other perimortem assaults, and, despite the advent of biomolecular identification techniques, they continue to be called upon by medical examiners, coroners, and investigative agencies to provide this valuable service (1).

It is not uncommon for odontologists to be required to justify their decisions in Court (7). As such, the need for forensic dentistry to satisfy the requirements of *Daubert* is well recognized, especially the need for peer-reviewed data on error rates.

The majority of dental identifications are routine and are not contested. However, in homicide cases, life insurance claims, civil cases involving, for example, motor vehicle accidents, the determination of identity may be questioned. It is essential that the testifying odontologist can quote from the scientific literature studies that have produced reliability and error rate data. In this manner, the trier-of-fact is able to place the evidence in context and afford it appropriate weight (8).

Few studies have undertaken the task of assessing dental identification reliability. Sholl and Moody used groups of forensic dentists, recent dental graduates, and dental hygienists to assess radiographs produced from skulls (9). They determined a mean identification success rate of 93.3% among the odontologists with a range of 63.6 to 100%. Recent dental graduates scored 85.2% and hygienists 89.7%. Unfortunately, the study involved only nine individuals in each group and thus limits the value of the data. The expense of radiographic duplication and delivery to odontologists may be a limiting factor for both the number of studies undertaken and the number of participants in such studies.

MacLean et al. also examined reliability of dental radiographs and altered the antemortem–postmortem time period to examine the effect of this on accuracy (10). Interestingly, the investigators arrived at the same accuracy as the Sholl study, 93%, although with increasing antemortem–postmortem interval this decreased. A further study by Kogon (11) also revealed a high degree of sensitivity and specificity when using bitewing radiographs, although this decreased when the time period between antemortem and postmortem exceeded 20 years.

It is the purpose of this study to determine, via the use of digitized dental radiographs, the rate of error among a large population of forensic odontologists for dental identifications. By assessing both intra- and inter-examiner agreement and using ROC statistical analysis, these data will satisfy the requirements of *Daubert* for published error rates concerning dental identification procedures (8).

[1] Graduate student, The University of Liverpool, Department of Clinical Dental Sciences, Liverpool, United Kingdom.

[2] Undergraduate student, The University of Northumbria at Newcastle, Department of Computing Sciences, Newcastle-Upon-Tyne, United Kingdom.

[3] Formerly of the University of Washington, Seattle, WA.

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

Received 8 Feb. 2003; and in revised form 18 May 2003; accepted 7 June 2003; published XXXX.

Method and Materials

Real forensic identification cases were selected to ensure that a wide range of difficulty was present. In some instances, radiographs were removed from the available series to make cases more difficult. Many radiographic views were used, including bitewing, periapical, occlusal, and panoramic films. Ten cases were compiled with three negative identifications and seven positive identifications. In each case, the identity was confirmed by odontology and at least one alternate method, i.e., DNA, visual.

For each of the cases in this study, a comparison between a single antemortem series and postmortem series is conducted. This is a common situation in forensic odontological casework where dental methods are frequently used for confirmatory identification based upon strong presumptive evidence. It is, however, important to note that, for example, in multiple-victim road traffic incidents and mass disasters, that it may be necessary to compare a single postmortem record to a number of antemortem records. The advent of computer-assisted comparison systems for such multiple fatality identifications has fundamentally altered the methods employed for such identifications, and therefore the procedures and techniques differ from those under investigation in this study (12–14).

The radiographs were digitized using a high-resolution scanner. The radiographic images were then uploaded to a website that was custom-designed by one of the authors (RJP) using the PERL programming language. Participants logged on to the website and

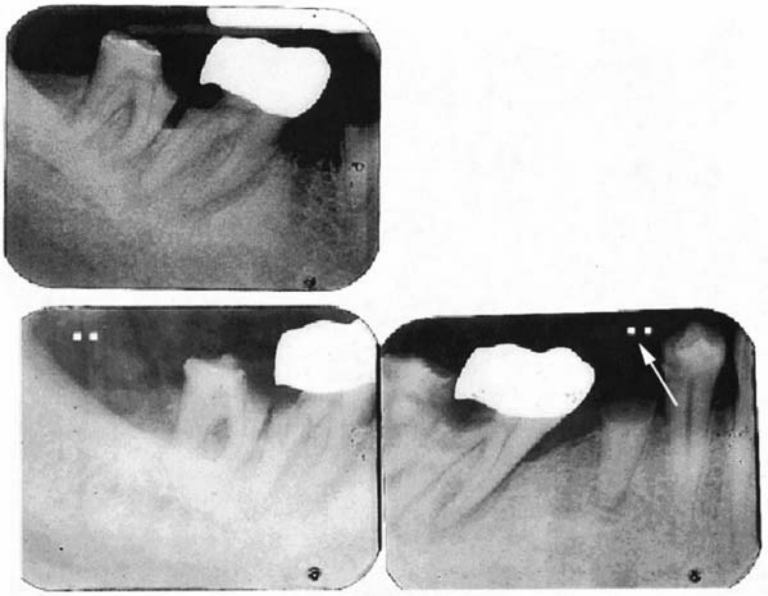

FIG. 1—*Example of a positive identification result. Arrow shows two perforations of the radiograph indicating it is from the postmortem record.*

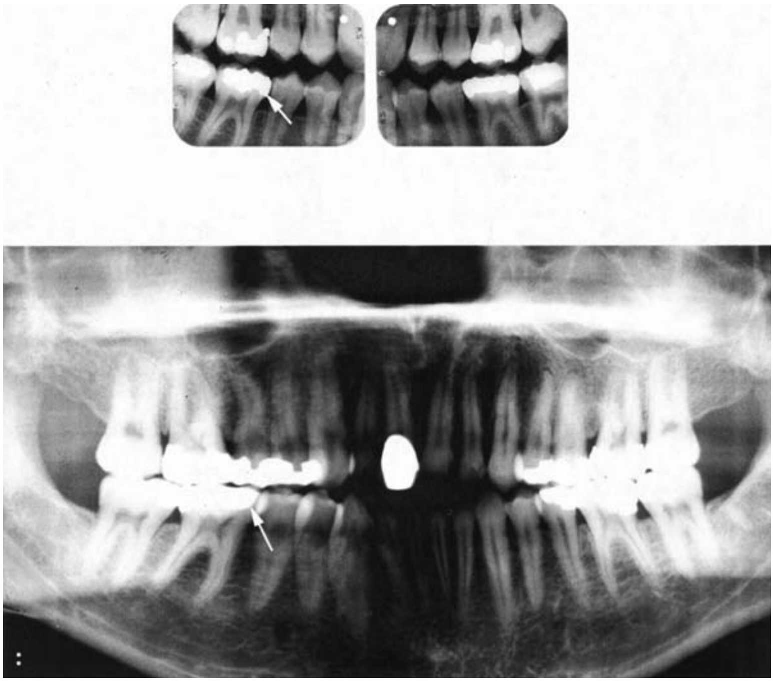

FIG. 2—*Example of a negative identification result. Arrow shows a single unexplainable discrepancy.*

TABLE 1—*Conclusion levels available to the participants.*

Level	Description
5. Inconclusive	Insufficient evidence to make any statement
4. Excluded	Negative identification
3. Possible	Could be, may or may not be
2. Probably	More likely than not
1. Reasonable medical certainty	No reasonable or practical possibility that it is someone other than that represented in the antemortem records

provided demographic information, including the number of previous forensic identifications they had completed. Participants were then shown sequentially each of the cases with the antemortem and postmortem radiographs viewable on the screen simultaneously. The participants were asked to record their conclusions using the four ABFO levels for body identification (see Table 1, www.abfo.org). The desired conclusion was selected using a drop-down menu that was available for each case. Once the conclusion was selected, the participants were automatically advanced to the next case.

Following a one-month washout interval, a message was sent by e-mail inviting the participant to complete the study a second time. Each participant repeated the exercise, although the order in which the cases were presented was changed.

The conclusions from each participant were automatically forwarded to the investigators, and the data were entered into both SPSS and PEPI for statistical analysis. The following variables were assessed: overall accuracy, intra-examiner agreement, inter-examiner agreement, sensitivity, and specificity. Since the case conclusions included a degree of certainty, there are several methods by which the data could be handled. The simplest method was to include a threshold, or cutoff. In this case, exclusion was considered as negative and all other responses as positive, with the exception of insufficient evidence. This conclusion was disregarded within the analysis. This method is simple to analyze and understand, but it ignores the level of certainty expressed by the participants in their conclusions. Therefore, the data were also analyzed using receiver operating characteristics (ROC) (15). This statistical treatment permits a measure of accuracy (area under the curve) without requiring an artificial cut-off point (15). Kappa statistics were used to measure agreement (16,17).

Results

Demographics

A total of 155 participants attempted the exercise on one occasion, and 87 individuals repeated the exercise. For the purposes of statistical analysis, all 155 responses were used to determine accuracy, and the results from the additional 87 responses were used to determine intra-examiner agreement. Seventy-two percent of the respondents were dentists; of these, 38% were experienced forensic odontologists that had completed more than 25 forensic identification cases and 41% had completed some identifications. For simplicity of analysis, the respondents were split into the following four groups: (1) experienced forensic odontologists (EXPFD); (2) dentists with more than one but less than five identification cases (MODFD); (3) general dentists with no forensic experience (DENT); and (4) nondentists (NODNT). Eighty-seven percent of the respondents were from the United States, 8% were from Europe, and 5% were from Australia and New Zealand.

Accuracy

Accuracy can be defined, in its simplest form, as percentage correct. This is an easily understood term but ignores the possibility of correct answers that have been arrived at by chance. Kappa is a statistical treatment that corrects for this and produces a scale that has been interpreted by Landis and Koch (Table 2). The accuracy results are presented in Table 3. Table 3 also contains the ROC area under the curve (AUC) data. Tables 4, 5, 6, and 7 contain the data from the ROC analysis with sensitivity and specificity presented

TABLE 2—*Guide to the intepretation of Kappa scores.*

Kappa	Strength of Agreement
0.00–0.01	Poor
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

From Landis and Koch, 1977.

TABLE 3—*Accuracy scores, agreement with truth, ROC AUC.*

Group	Kappa	S.E.	% Agreement	AUC
EXPFD	0.83	0.256	91	92
MODFD	0.74	0.234	81	86
DENT	0.79	0.222	86	87
NODNT	0.76	0.223	84	84
Average (Mean)	0.78	. . .	85.5	87.25

TABLE 4—*EXPFD ROC results.*

Conclusion Level	Sensitivity, % (SD)	Specificity,% (SD)
5	100.0 (±0.0)	0.0 (±0.0)
4	89.6 (±19.1)	64.3 (±21.0)
3	74.5 (±12.2)	83.4 (±18.5)
2	63.5 (±11.5)	92.9 (±10.0)
1	33.5 (±8.1)	99.3 (±20.2)

TABLE 5—*MODFD ROC results.*

Conclusion Level	Sensitivity, % (SD)	Specificity, % (SD)
5	100.0 (±0.0)	0.0 (±0.0)
4	75.6 (±20.3)	61.2 (±18.2)
3	68.9 (±18.2)	74.2 (±12.7)
2	54.2 (±17.5)	85.6 (±19.0)
1	29.3 (±3.1)	92.4 (±19.5)

TABLE 6—*DENT ROC results.*

Conclusion Level	Sensitivity, % (SD)	Specificity, % (SD)
5	100.0 (±0.0)	0.0 (±0.0)
4	82.5 (±23.1)	62.5 (±9.2)
3	71.9 (±20.2)	81.6 (±16.1)
2	59.8 (±16.2)	90.0 (±17.7)
1	31.6 (±9.8)	94.9 (±22.3)

TABLE 7—*NODENT ROC results.*

Conclusion Level	Sensitivity, % (SD)	Specificity, % (SD)
5	100.0 (±0.0)	0.0 (±0.0)
4	79.9 (±30.2)	60.2 (±5.3)
3	67.5 (±25.3)	75.9 (±18.9)
2	54.3 (±19.8)	87.3 (±27.3)
1	31.2 (±14.7)	93.4 (±33.3)

TABLE 8—*Mean inter-examiner agreement results.*

Group	Kappa	SD
EXPFD	0.89	5.6
MODFD	0.85	8.6
DENT	0.87	12.3
NONDENT	0.78	22.3

TABLE 9—*Mean intra-examiner agreement results.*

Group	Kappa	SD
EXPFD	0.95	5.6
MODFD	0.86	11.3
DENT	0.88	14.7
NONDENT	0.76	19.8

for each group at each level of conclusion. ANOVA was applied to the group data and statistically significant differences ($p > 0.05$) were found between the EXPFD group and all other groups, but no differences existed between any of the remaining groups.

Agreement

Table 8 contains the inter-examiner data, and Table 9 contains the intra-examiner data. ANOVA was applied, and there were no statistically significant differences between the groups for the inter-examiner scores. However, the EXPFD group was significantly different from all other groups when the intra-examiner scores were compared ($p > 0.05$). Analysis of the participants' decisions showed that the exclusions were most often changed at the time of the second attempt. This accounted for 93% of all non-agreements between each examiner. When analyzed separately, agreement between each of the groups for the exclusion decisions was rated only moderate on the kappa scale.

Discussion

As described in the authors' previous work on the reliability of transparent bitemark overlays (18), a key feature of modern science is that of scepticism; no longer are scientific principles accepted based solely on authority or common-sense anecdotal beliefs. This is certainly the case within forensic science, and the courts are requiring that evidence presented be based on peer-reviewed research establishing the error rate among other requirements. In the *Daubert* era, this is taken further, with individual experts being required to describe their personal performance against those of their peers. Personal identification evidence is presented in court less frequently than bitemark identifications (8). However, the agencies requesting the services of odontologists should be made aware of the accuracy of the techniques employed in the arrival of identification decisions. The purpose of this work is to establish empirical justification for the use of postmortem-to-antemortem radiographic comparisons in the identification of individuals.

Validity of the Design

There are a number of features of the study design that must be discussed. The use of authentic forensic cases lends validity to the study. Each of the matching cases was identified positively through additional methods at the time of the original dental comparisons, so external validity of the conclusions is assured. The cases fairly represent the quantity and quality of the radiographs used in such identifications and represent a good range of radiographic type and time interval between antemortem and postmortem images.

The cases were selected from the case files of the authors and therefore do reflect the range of materials available to odontologists. Two of the authors (IAP, DS) rated each case on a difficulty scale of 1 to 5. The mean difficulty, as rated by both authors, was 2.5, demonstrating that an excellent spread of case complexity was present within the samples. Complexity can be affected by a number of variables, such as quantity and quality of the radiographs, antemortem and postmortem interval, number of restorations, and presence of more distinctive treatments.

In order to employ a web-based strategy, it was necessary to scan the radiographs and view them on a computer monitor. The clarity and resolution of computer monitors is variable and this may have an impact on accuracy. Similarly, odontologists view radiographs on light boxes, and the illumination can be adjusted according to the density of the film image. In this regard, the methodology is limited. In order for the radiographic images to be loaded on slow Internet connections, a degree of compression was employed. This might also impact on the accuracy of comparisons since some data from the original image were lost.

Despite these limitations, the web-based implementation of this study has several advantages. The most significant of these is cost reduction. The traditional method of such investigations would be to duplicate the radiographs and distribute them to participants. This is an expensive process and requires the individual recruitment of participants.

The web-based system permits passive recruitment of participants. This facilitates a large number of respondents across a large geographical area. It does, however, limit participation to those individuals who have access not only to a computer system but also to an Internet connection. The authors cannot be assured of veracity of the participants supplied demographic data as well due to the higher than normal level of anonymity provided by electronic mail and the Internet. Despite this, 155 respondents represent a significant subject pool, and use of the Internet was praised by many taking part in the study. One respondent claimed that the quality of the images was such that they were unable to participate. It should be noted that the quality of antemortem radiographs is highly variable because of factors such as time since exposure, development technique, and storage conditions. As such, high-quality images in forensic casework are not always available. The radiographic images can be viewed at www.forensicdentistryonline.com.

A final aspect of the validity of the design is that the study was limited to radiographic images only. The odontologist will use all of the elements of the dental record available to them when completing a dental identification comparison. This can include written notes, charting symbols, odontograms, study casts, photographs, laboratory prescriptions, and other data. Indeed, some dental records contain no radiographs (especially in young, caries-free patients), and the identification is established using other elements of the record.

Therefore, it is essential that the results of this study are not taken as an indication of the accuracy of dental identifications *per se.* Rather, the results are an indication of the accuracy of that portion of the dental identification process that involves radiographic comparisons.

Accuracy

The use of cutoffs to combine the conclusion levels and extrapolate a positive or negative decision allowed the use of simple statistical analysis. However, this does not replicate the ABFO-recommended methodology for body identifications. Nonetheless, the data are easily understood. All groups performed well, with the most experienced forensic dentists achieving a 91% success rate (substantial agreement). This group was significantly better than all others. Interestingly, there were no significant differences detected between the other groups, including the nondentists. This result is similar to those from our assessment of bitemark comparison overlays where experience was not an indicator of accuracy (18). There could be many explanations for this. One possible explanation is that the comparison of radiographs is essentially a pattern association, which is the overarching principle of bitemark comparisons. The ability to match patterns may well be independent of dental knowledge. However, dental experience is required to understand the natural history of teeth, for example, that a small restoration might be replaced by a larger one and that a large restoration might be replaced by a root canal filling and a crown.

ROC analysis is a statistical technique in which the levels of conclusion afforded to each decision are considered. Results of this more sophisticated statistical treatment of the data are encouraging. Measured as an area under the ROC curve (AUC), this is a combination and generalization of the concepts of sensitivity and specificity into a single measure of accuracy (see Table 3). All of the groups performed well. Each obtained an AUC over 0.75, and this is regarded as a satisfactory level of performance (15). Again, the more experienced forensic dentists were significantly better than all of the other groups.

Sensitivity and specificity data for digitized radiographic identifications from the ROC analysis follow the general trend of data from diagnostic tests: that as specificity increases, sensitivity will decrease. However, it is useful to note that the data suggest the participants in this study were slightly (although significantly) better at establishing a positive identification than a negative identification. This could be related to the proportion of negative identifications encountered. A recent self-audit by one of the authors (IAP) showed that out of 48 identification cases conducted within one 12-month period, only one was a negative identification. If this is true for most forensic dentists, then this may explain the reluctance to establish a negative identification. There is simply insufficient experience in such situations.

There are other reasons for the poorer performance with respect to negative identifications. Studies have demonstrated that anticipatory bias can lead participants in obvious tests (such as this one) to anticipate a certain number of question types, etc., and some examiners may have felt that all the IDs would be positive and that it was the conclusion level assigned to them that was under investigation. Interestingly, each of the negative identifications was allocated a difficulty score of either 1 or 2 by both authors (IAP, DS), indicating either easy or very easy.

Agreement

The agreement data from all participants rated substantial and those from the experienced forensic dentists as almost perfect. As reflected in the specificity scores, the main area of disagreement concerned the negative identifications. This was the case for both intra-examiner and inter-examiner differences. More negative decisions were changed between the two attempts than any other decision, e.g., probable, possible. This could also be related to the experiences of the odontologist with negative identification cases as described previously or their expectations of the test itself. The Kappa data is presented in Tables 8 and 9.

Conclusions

Dental identifications resulting from a comparison of postmortem and antemortem radiographs are valid, accurate, and reliable when examined under the conditions described within this study. Accuracy and reliability are lower when negative identifications are assessed. Training exercises should include a sample of negative identification cases to provide experience with such comparisons, and advice on arriving at a negative identification conclusion should be provided along with explanations of the common areas of disagreements within such cases.

Previous experience of at least six identification cases significantly ($p < 0.01$) improves individual examiner performance. This finding could be used to establish a baseline for mentors to insist that trainees observe the more experienced odontologist for at least eleven cases before accepting their own casework. Obviously, the impact of field casework complexity would influence the absolute number. The impact of experience should also reflect on the use of less-experienced odontologists within a mass disaster response scenario. In agreement with other studies, these data support the contention that such events should not be used for gaining experience, rather for consolidating it within an environment that offers many additional, extreme challenges to the odontologist.

The use of the World Wide Web to deliver and assess forensic exercises is promising. The use of the methodology in proficiency testing should be further researched, particularly the opinions of those undergoing the tests. The methodology enables the cost of such research to be reduced, and a wider range of participants can be enrolled. This is of importance within the forensic science, field where opportunities for large research grants may be restricted.

References

1. Pretty IA, Sweet D. A look at forensic dentistry-Part 1: The role of teeth in the determination of human identity. Br Dent J 2001;190(7):359–66.
2. Delattre VF. Burned beyond recognition: systematic approach to the dental identification of charred human remains. J Forensic Sci 2000;45(3):589–96.
3. Brkic H, Strinovic D, Kubat M, Petrovecki V. Odontological identification of human remains from mass graves in Croatia. Int J Legal Med 2000;114(1–2):19–22.
4. Brkic H, Strinovic D, Slaus M, Skavic J, Zecevic D, Milicevic M. Dental identification of war victims from Petrinja in Croatia. Int J Legal Med 1997;110(2):47–51.
5. Titsas A, Kieser JA. Odontological identification in two high-impact, high-temperature accidents. J Forensic Odontostomatol 1999;17(2): 44–6.
6. Andersen L, Juhl M, Solheim T, Borrman H. Odontological identification of fire victims-potentialities and limitations. Int J Legal Med 1995;107(5):229–34.
7. Atkinson SA. A qualitative and quantitative survey of forensic odontologists in England and Wales, 1994. Med Sci Law 1998;38(1):34–41.
8. Pretty IA, Sweet D. A comprehensive examination of bitemark evidence in the American legal system. Reno, NV: American Academy of Forensic Science, 2000.
9. Sholl SA, Moody GH. Evaluation of dental radiographic identification: an experimental study. Forensic Sci Int 2001;115(3):165–9.

10. MacLean DF, Kogon SL, Stitt LW. Validation of dental radiographs for human identification. J Forensic Sci 1994;39(5):1195–200.
11. Kogon SL, MacLean DF. Long-term validation study of bitewing dental radiographs for forensic identification. J Forensic Sci 1996;41(2):230–2.
12. Lewis C. WinID2 versus CAPMI4: two computer-assisted dental identification systems. J Forensic Sci 2002;47(3):536–8.
13. McGivney J, Fixott RH. Computer-assisted dental identification. Dent Clin North Am 2001;45(2):309–25.
14. Dailey JC. Computer-assisted identification of Vietnam War dental remains. Mil Med 1987;152(4):179–92.
15. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons, 1986.
16. Koch GG, Landis JR, Freeman JL, Freeman DH, Jr., Lehnen RC. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 1977;33(1):133–58.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.
18. Pretty IA, Sweet D. Digital bitemark overlays-an analysis of effectiveness. J Forensic Sci 2001;46(6):1385–91.

Additional information and reprint requests:
Dr. Iain A. Pretty
The University of Liverpool
Department of Clinical Dental Sciences
Edwards Building, Daulby Street
Liverpool, L69 3GN
Telephone: 0151 706 5288
Fax: 0151 706 5809
E-mail: ipretty@liv.ac.uk
Website: www.forensicdentistryonline.org