Science & Justice xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Science & Justice

journal homepage: www.elsevier.com/locate/scijus

Validation studies in forensic odontology – Part 1: Accuracy of radiographic matching

Mark Page[a,b], Russell Lain[b], Richard Kemp[c], Jane Taylor[a,b,*]

[a] School of Health Sciences, University of Newcastle, Australia
[b] Department of Forensic Medicine, New South Wales, Australia
[c] School of Psychology, University of New South Wales, Australia

ARTICLE INFO

Keywords:
Forensic science
Odontology
Validation
Radiographic matching
Radiographs

ABSTRACT

As part of a series of studies aimed at validating techniques in forensic odontology, this study aimed to validate the accuracy of ante-mortem (AM)/postmortem (PM) radiographic matching by dentists and forensic odontologists. This study used a web-based interface with 50 pairs of AM and PM radiographs from real casework, at varying degrees of difficulty. Participants were shown both radiographs as a pair and initially asked to decide if they represented the same individual using a yes/no binary choice forced-decision. Participants were asked to assess their level of confidence in their decision, and to make a conclusion using one of the ABFO (American Board of Forensic Odontology), INTERPOL (International Criminal Police Organisation) and DVISys™ (DVI System International, Plass Data Software) identification scale degrees. The mean false-positive rate using the binary choice scale was 12%. Overall accuracy was 89% using this model, however, 13% of participants scored below 80%. Only 25% of participants accurately answered yes or no > 90% of the time, with no individual making the correct yes/no decision for all 50 pairs of radiographs. Non-odontologists (lay participants) scored poorly, with a mean accuracy of only 60%. Use of the graded ABFO, DVISYS and INTERPOL scales resulted in general improvements in performance, with the false-positive and false-negative rates falling to approximately 2% overall. Inter-examiner agreement in assigning scale degrees was good (ICC = 0.64), however there was little correlation between confidence and both accuracy or agreement among practitioners. These results suggest that use of a non-binary scale is supported over a match/non-match call as it reduces the frequency of false positives and negatives. The use of the terms “possible” and “insufficient information” in the same scale appears to create confusion, reducing inter-examiner agreement. The lack of agreement between higher-performing and lower-performing groups suggests that there is an inconsistency in the cognitive processes used to determine similarity between radiographs.

1. Introduction

Some of the basic science underpinning the identification sciences such as fingerprint analysis, bitemark analysis, handwriting analysis, fibre analysis, and footwear analysis has been challenged [1], most notably by the National Academy of Sciences Report of 2009 [2]. This report recommended two approaches to remedy these shortfalls; the use of validation studies and the consultation with forensic psychologists to ensure effective design of these validation studies.

These recommendations have not been taken up with vigour by many disciplines. A number of larger-scale studies have been conducted using judicial casework as validation for particular forensic techniques, such as fingerprints, toolmarks and handwriting analysis, but these types of studies fail to ensure independent verification of the true source [3,4]. Many studies that have attempted to bolster the argument for forensic examiner accuracy, such as those run by the Collaborative Testing Service and the American Society of Crime Laboratory Directors, have been criticised as having many of the features of poor-quality studies, including no control over the conduction or timing of the tests, uncertain sample sizes, and varying difficulty levels [4,5]. DNA evidence, as the relative newcomer to forensic science, demonstrates the most rigorous background when it comes to validation of scientific methodology, although these too are not without their failings, as few look to address the human factors implicit in interpretation of DNA results [6].

A basic activity in forensic odontology casework is the reconciliation of ante-mortem and post-mortem data, principally radiographic data, to facilitate identification in cases where the remains are not

* Corresponding author at: Oral Health, University of Newcastle, PO Box 127, Ourimbah, NSW 2258, Australia.
E-mail address: jane.taylor@newcastle.edu.au (J. Taylor).

http://dx.doi.org/10.1016/j.scijus.2017.11.001
Received 5 June 2017; Received in revised form 9 October 2017; Accepted 4 November 2017

visually identifiable. While this aspect of odontology practice, and the science underpinning it, did not attract the criticism of the National Academy of Sciences Report, the limited research undertaken to reinforce the science behind this aspect of forensic odontology practice appears to have gone largely unnoticed. Many of the studies that have been undertaken suffer from similar flaws to those undertaken in other forensic disciplines, with small sample sizes [7–9], limited participation by trained forensic odontologists [10–12], and use of identification scales that do not reflect what odontologists currently use in actual casework [11,13]. This paper reports a study designed to test the ability of Forensic Odontologists to match ante-mortem and post-mortem radiographs in an experimental setting that attempts to remedy some of these flaws in response to the call for validation and verification of dental identification from radiographic matching.

2. Method

2.1. Test design

The test was designed to be implemented as an online survey using the *Qualtrics*™ software platform [14]. The images used in the test were selected from actual Disaster Victim Identification (DVI) casework undertaken by two of the authors, and all positive identifications had been confirmed by DNA analysis. 50 cases were assembled for the study and for each case the participants were presented with a pair of radiographs with the ante-mortem (AM) image appearing on the left of the screen, and the post-mortem (PM) image on the right.

In half of the cases the two presented images were from the same individual, and in the other half they were from different individuals. If the images were from different individuals care was taken to ensure the non-match AM image was a plausible match for the PM image. The authors also attempted to construct cases with a range of clinical difficulty, including some easy and some more difficult cases in both the matching and not matching cases. None of this information was revealed to the participants.

Participants were recruited via emails sent to targeted professional groups in Australia, New Zealand, the United Kingdom, the United States and Europe.

Psychology students at the University of New South Wales (UNSW) were also recruited as a control group with no dental training. Ethical approval was obtained from the UNSW Human Research Ethics Committee.

On opening the link to the study participants read a Participant Information Statement and after consenting to participate was asked a number of demographic questions relating to age, sex, experience, professional training and qualifications. They were then provided with instructions on how to complete the test. Participants were instructed that for each of the 50 cases they would be required to determine if the AM and PM images were from the same individual. Enlargement of the radiograph was possible by clicking on the image. Initially they would be required to indicate their decision as a binary same/different decision, but that later they would be able to express their decision using one or more of three standardized forensic odontology scales for identification in common use: ABFO (American Board of Forensic Odontology), INTERPOL (International Criminal Police Organisation) and DVISys™ (DVI System International, Plass Data Software) (Table 1). Participants were then asked to indicate their level of confidence in the decision they had made and record this on a 100 point scale. After confirming that they understood these instructions participants were provided with a practice case, which could be repeated as many times as wished.

The order of the cases was randomised for each participant. Once the test was started participants could stop if they chose and re-enter the test where they had previously left off. On completion of the last case participants were asked if they wished to receive feedback on their individual performance. Those who indicated in the affirmative were taken though each of the cases again with their response and the correct response indicated. Finally, participants were asked about the perceived difficulty of the test and the extent to which they felt the test reflected real casework.

Table 1
ABFO, INTERPOL and DVISys scale choices – levels of identification.

ABFO	Interpol	DVISys
Positive	Positive	Positive
Possible	Probable	Probable
Exclude	Possible	Possible
Insufficient Evidence	Exclude	Exclude
		Insufficient Evidence

2.2. Data analysis

The data were analysed with regard to accuracy using the binary response; accuracy using the forensic odontology scales for identification; and the relationship between confidence and accuracy, and the relationship between accuracy and training, experience and qualifications. Outliers in confidence data were defined as values determined by the interquartile range $[Q_1 - k\ (Q_3 - Q_1),\ Q_3 + k\ (Q_3 - Q_1)]$, using Tukey's definitions where $k > 1.5$ indicates an outlier and $k > 3$ indicates an extreme value. These data were excluded from further analysis between confidence and accuracy, but denoted in the box plots by ♦ and X respectively.

3. Results

3.1. Participants

29 participants with dental training completed all 50 examples. Mean years of experience in dentistry was 28.5 years (range 5–48 years). 26 (87%) stated they had specific post-graduate training in forensic odontology, with 27 of these 29 participants (93%) having forensic casework experience.

3.2. Accuracy

The mean accuracy rate (defined as those radiographs correctly identified as either a match or non-match) for this group was 87.5% (range 70–98%). The median accuracy was 90%. 4 (13%) individuals got < 80% of cases correct, with the majority of the cohort achieving between 80 and 90% accuracy (Fig. 1). Correct identification of a match occurred 89.3% of the time, with 6 participants correctly identifying all matches. The least accurate match rate was for one individual who identified only 60% of matches correctly. Correct identification of a non-match (i.e. correct rejection rate) had a mean accuracy of 85.6%. 5 participants correctly identified all non-matches, with the worst performer identifying only 44% of non-matches correctly. Kuder-Richardson analysis (KR-20) indicated that internal consistency among the group was considered at best fair ($r = 0.58$). No participant simultaneously scored 100% accuracy for both matches and non-matches. The mean non-match accuracy for the 6 participants who correctly identified all of the matches was 75%. The mean match accuracy for those 5 participants who identified all of the non-matches was 84%.

3.3. Experience and qualifications

69 non-dentally trained individuals (psychology students) also completed the online tool and served as a control group with no dental training. The mean accuracy for this group was 59.8% (range 42%–80%, with a match-accuracy of 57.9% and a non-match accuracy of 61.7%. A comparison of overall accuracy between dentists and non-