Science & Justice 59 (2019) 93–101

Contents lists available at ScienceDirect

Science & Justice

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

Review

Validity of forensic odontology identification by comparison of conventional dental radiographs: A scoping review

Sher-Lin Chiam[a,*], Mark Page[a], Denice Higgins[b], Jane Taylor[a]

[a] School of Health Sciences, University of Newcastle, Australia
[b] Forensic Odontology Unit, The University of Adelaide, Australia

1. Introduction

Identification of the deceased by forensic dental comparison is well accepted as valid and efficient and is one of the primary methods relied upon in disaster victim identification. A vital component of this method of identification is image comparison with dental radiographs, which continues to provide the most valuable source of evidence [1,2]. Radiographs as a graphic record of dental status contain more verifiable information and detail than written descriptions or charts. Human error can lead to inaccuracies in written records but an image provides an irrefutable source of information [2]. Recently, there has been increased focus on quantification of accuracy, reliability and the objectivity of comparative forensic science disciplines [3]. The 2011 President's Council of Advisors on Science and Technology (PCAST) report [3], which focused specifically on the issues pertaining to comparative forensic sciences, identified the need for empirical investigations into validation of methods. While identification via dental comparison, in contrast to bitemark analysis, did not come under direct scrutiny in either the 2009 National Academy of Sciences (NAS) report [4] or the PCAST report [3], it is important that validation studies are conducted on all aspects of forensic odontology work.

To date, only one review of the literature from the period of 1990 to 1994 regarding the validity of radiographs for forensic identification has been undertaken [5]. This review looked specifically at the use of bitewing radiographs and discussed only four studies. Empirical research since then has included additional types of radiographs and variations in method design, making it a heterogeneous group of studies. In this study we undertake a scoping review to provide an overview of existing empirical research with regards to the validity of using dental radiographs for identification. Information from identified relevant studies has been extracted, collated and summarised to present a landscape of the research. The main issues pertaining to research into this method of forensic identification are also discussed.

2. Method

This scoping review employs the method described in Arksen and O'Malley 2005 [6], which requires identification of specific research questions, systematic search and selection of studies, charting of the data and summarisation of the outcomes. No evaluation of the quality of studies is included in scoping studies [6]. The objective of this review accordingly, was to map existing empirical validity research on the use of dental radiographs for identification and to extract, collate, tabulate and summarise the findings to provide a landscape of the research to date.

A search strategy using the Boolean search terms; 'forensic dentistry' or 'forensic odontology' combined with 'dental radiology' and 'identification' was employed to search "Web of Science", "Science Direct" and "Medline" databases, restricted to publications in English and excluding duplicates. Based on the title and abstract relevant articles were identified. After full text reading, a second exclusion was performed based on pre-determined inclusion and exclusion criteria. A second search was performed by hand-searching through the references cited in these articles for suitable studies not found in the first search.

Included were all primary research publications on validity or reliability using conventional (analogue and digital) extra and intra oral radiographs for comparison in forensic dental identification by human observers. The exclusion criteria were studies that used advanced imaging techniques i.e. computerised tomography (CT) or magnetic resonance imaging (MRI). Studies that compared other maxillofacial structures for identification such as sinuses, trabecular bone structure in edentulous subjects or included biometric or automated systems for comparison of dental radiographic structures were also excluded.

The main research question identified was:

What existing research on the validity of using conventional dental radiographs for forensic identification is available?

The supporting and specific research questions were:

1. What type and sample size of dental radiographs were used?
2. How many participants were involved and what was their experience/skill level?
3. What methods were employed for comparison of the radiographs?
4. Was case information provided to aid the comparison of the radiographs?

* Corresponding author.
E-mail address: SherLin.Chiam@uon.edu.au (S.-L. Chiam).

https://doi.org/10.1016/j.scijus.2018.08.008
Received 12 March 2018; Received in revised form 6 August 2018; Accepted 27 August 2018

Table 1
A list of the articles in this review.

Authors	Year	Title	Journal
Borrman et al. [16]	1990	Accuracy in establishing identity by means of intraoral radiographs.	The Journal of Forensic Odonlo-Stomatology
Ekstrom et al. [10]	1993	Accuracy among dentists experienced in forensic odontology in establishing identity.	The Journal of Forensic Odonlo-Stomatology
MacLean et al. [17]	1994	Validation of dental radiographs for human identification. *Journal of forensic sciences*.	Journal of Forensic Sciences
Kogan et al. [18]	1996	Long-term validation study of bitewing dental radiographs for forensic identification.	Journal of Forensic Sciences
Sholl et al. [15]	2001	Evaluation of dental radiographic identication: an experimental study.	Forensic science international
Pretty et al. [12]	2003	The reliability of digitized radiographs for dental identification: a Web-based study.	Journal of forensic sciences
Soomer et al. [11]	2003	Dentists' qualifications affect the accuracy of radiographic identification.	Journal of forensic sciences
Fridell et al. [7]	2006	The use of dental radiographs for identification of children with unrestored dentitions.	The Journal of Forensic Odonlo-Stomatology
Wenzel et al. [14]	2010	Matching simulated antemortem and postmortem dental radiographs from human skulls by dental students and experts: testing skills for pattern recognition.	The Journal of Forensic Odonlo-Stomatology
Pinchi et al. [9]	2012	Dental identification by comparison of antemortem and postmortem dental radiographs: Influence of operator qualifications and cognitive bias.	Forensic science international
Kaur Bhullar et al. [19]	2014	Evaluation of dental expertise with intraoral periapical view radiographs for forensic identification.	Journal of forensic dental sciences
Balla et al.[8]	2017	Identification by comparison of caries free bitewing radiographs: Impact of observer qualifications and their clinical experience	Forensic science and Criminology
Page et al. [13]	2017	Validation studies in forensic odontology – Part 1: Accuracy of radiographic matching	Science and Justice

5. What type of scale was used for decision making?
6. Do these studies all have the same focus?
7. What and how were the results derived and presented?

3. Results

The first search produced a total of 336 articles; 154 from Medline, 90 from Science Direct and 42 from Web of Science. A search of relevant titles and review of the abstracts reduced this to 21 articles. Full text reading narrowed this to 11 articles and two other relevant articles were located through hand searching. A total of 13 publications were identified for this review. These publications are summarised in Table 1. The frequency and types of terms in the title that are semantically and conceptually associated with validity or reliability studies are summarised in Table 2. Three studies [7–9] did not use any specific term in the title that indicated the main research interest was evaluating the use of dental radiographs.

The main theme of all 13 articles was validity or reliability of the method of comparing ante-mortem and post-mortem dental radiographs for identification of deceased persons. The articles were summarised and analysed according to the following parameters:

1. The types and number of radiographs used.
2. Number of participants and their skill level.
3. The method employed for comparison of the radiographs.
4. Presence or absence of case information to aid the comparison of the radiographs.
5. The scale used for decision making.
6. Specific research questions within the general theme of validation pertaining to the use of dental radiographs for forensic identification.
7. Analysis of the results.

Table 2
A summary of terms used in the title that suggest the concept of validity and reliability.

Term used in title	Study
Validity	MacLean et al.; Kogan et al.; Page et al.
Accuracy	Borrman et al.; Ekstrom et al.; Soomer et al.
Evaluate	Sholl et al.; Kaur Bhullar et al.
Reliability	Pretty et al.
Test	Wenzel et al.
No specific terms used	Fridell et al.; Pinchi et al.; Balla et al.

3.1. The type and number of radiographs used in the studies

The type and number of dental radiographs used in the different studies is summarised in Table 3. The sample sizes of radiographs ranged from six to 280 pairs. Four studies [10 − 13] used radiographs from actual forensic cases. Page et al. [13] and Pretty et al. [12] used radiographs from cases that had also had the identifications confirmed by DNA. Radiographs taken on dry skulls were utilised by two studies [14,15], while the remaining studies [7–9,16–19] used radiographs from living patients seen in clinical practice.

Three studies [9] scanned analogue radiographs for conversion to digital format. One study [14] used digital intra oral radiographs for comparison to analogue radiographs.

Six studies [7,8,10,16–18] used only bitewing radiographs. One study [19] used only periapical radiographs, while two studies [14,15] used a mixture of bitewing and periapical radiographs. Four studies [9,11–13] used a mixture of extra-oral and intra-oral radiographs.

Two studies [8,16] used only radiographs which were from the same individuals (true positive) for comparison. The other studies utilised radiographs from different individuals (true negative) as well as true positive radiographs for comparison, however only 4 studies [13,14,17,18] used equal numbers of positive and negative radiographs.

3.2. The method employed for comparison of the radiographs

Two main comparison strategies were employed in the studies, paired presentation and free matching. Paired presentations are where only one ante- mortem and one post-mortem radiograph are presented and made available for comparison for each case. Free matching allowed participants to identify matched pairs from all the radiographs presented.

Six studies [7–10,15,16] used the free matching method, and seven studies used the paired presentation method as summarised in Table 3. Two studies [8,16] did not include extra non-matching radiographs in their free matching method.

3.3. Types and numbers of participants in each of these studies

Information relating to the participants in the studies is summarised in Tables 4 and 5. Participants had varying levels of expertise in radiographic interpretation and comparison ranging from lay persons to forensic odontologists. Forensic odontologists in some studies were referred to as dentists with forensic experience rather than specialist

Table 3
Summary of the radiographs used in all the studies.

Studies	Type of radiographs for comparison	Radiographs			Matching strategy	Sourced from patients	Sourced from forensic cases.	Dry skull	Digital	Analogue	Analogue to digital
		True positive	True negative	Extra - Not paired							
Borrman et al.	BW to BW	60 pairs			Free Matching	✓				✓	
Ekstrom et al.	BW to BW	30 pairs	1 pairs		Free Matching	✓	✓			✓	
MacLean et al.	BW to BW	140 pairs	140 pairs		Paired one on one	✓				✓	
Kogan et al.	BW to BW	100 pairs	100 pairs		Paired one on one	✓				✓	
Sholl et al.	PA to PA & BW to PA	15 pairs		5 extra BW or PA am	Free Matching			✓		✓	
Pretty et al.	Pm PA or BW to Am OPG	7	3		Paired one on one		✓				✓
Soomer et al.	Pm PA or BW to Am OPG	9 cases			Paired one on one		✓				✓
Fridell et al.	BW to BW	30 pairs		Extra 20 BW for am	Free Matching	✓				✓	
Wenzel et al.	BW to BW & Bw to PA	51 pairs	51 pairs		Paired one on one			✓	✓(Am BW)	✓ (Pm BW + PA)	
Pinchi et al.	OPG to multiple PA & BW	16 OPG to 37 Intraoral		5 BW or PA	Free Matching	✓					✓
Kaur Bhullar et al.	PA to PA	6	4		Paired one on one	✓				✓	
Balla et al.	BW to BW	7 pairs			Free Matching	✓				✓	
Page et al.	Pm PA or BW to Am OPG	25 Pairs	25 pairs		Paired one on one		✓				✓

BW- Bitewing radiograph.
PA- Periapical radiograph.
OPG- Orthopanthogram.

Table 4
Summary of the types and number of participants in the studies.

Studies, (Identified by the index in Table 1.)	Forensic	General Dentist	Maxfac radiologists	Layperson	Dental student	Others	Note	Total
Borrman et al.	1		6				Lay person included dental assistants.	7
Ekstrom et al.	17[a]							17
MacLean et al.	1[a]	1			1		One of the dentists had considerable experience	3
Kogan et al.	1		1		1			3
Sholl et al.	9	9			9	9 Dental hygienists		36
Pretty et al.	42[a]	68		44		45 Dentist with less forensic experience	2 experiments. 1st had 155 responses, 2nd had 87 responses.	155 / 87
Soomer et al.	40						Participants from 19 countries.	40
Fridell et al.		5	5					10
Wenzel et al.	2		1		10			13
Pinchi et al.	6	12			20	20		78
Kaur Bhullar et al.		20			20	20		60
Balla et al.		11		5		6	Others were dental postgraduate studies including 2 forensic odontology trainees.	22

(*continued on next page*)

Table 5
Characteristics of the forensic odontologists used in the studies.

Forensic odontologists with some form of formal recognition, member of society or association.	Sholl et al.; Soomer et al.; Wenzel et al.; Page et al.
Stated by author as forensic odontologists	Borrman et al.; Kogan et al.; Pretty et al.; Wenzel et al.
General dentists with experience in forensic odontology	Ekstrom et al.; MacLean et al.; Page et al.

practitioners.

Table 5 shows the main criteria used to define forensic odontologists in the 10 studies [10–18,20]. In four studies [12,14,16,18], the authors stated that the experimental group was forensic odontologists, no further information was given as to how they came to be thus classified. Three studies [10,13,17] used general dentists with forensic odontology experience, while four studies [9,11,13,15] used forensic odontologists who were formally recognised specialists by relevant boards. It should be noted that the study by Page et al. 2017 [13] included both dentists with odontology experience and recognised specialists as the experts.

3.4. The scale used for assessment

Five types of scale were used as shown in Table 6. The most common decision required of the participants was the binary match or non-match two choice decision scale used in nine studies [8–10,13,15–19]. The other scales were three to five levels of choice. A study by Wenzel et al. [14] used three choices but required the participants to reduce this to match or non- match. Three studies [11 − 13] used multiple decision choices. Page et al. [13] used the binary forced choice in addition to three types of multi- level choice scale. It is interesting to note that Pretty et al. [12] described the five-level choice scale as an ABFO scale, however, the ABFO scale has only four levels of choice [21].

3.5. Presence or absence of case information to aid the comparison of radiographs

As shown in Table 7, four studies [7,11,17,18] provided background information, Soomer et al. [11] stated case treatment information was provided, Maclean et al. [17] presented the dates of the ante-mortem and radiographs while Kogan et al. [18] supplied only dates of the ante-mortem radiographs in the investigation of the effect of time interval

Table 6
Scale used for judgment of match or non-match and for calculating accuracy.

Scale used	Decisions	Studies	Note
Two alternative forced choice	Match or non-match	Borrman et al.; Ekstrom et al.; MacLean et al.; Kogan et al.; Sholl et al.; Pinchi et al.; Kaur Bhullar et al.; Balla et al.; Page et al.	
3 Levels	Certain match, certain non-match, uncertain	Wenzel et al.	3 rounds reduced until certain or uncertain. Binary forced choice used in analysis.
3 Levels	Without doubt, possible probable	Fridell et al.	This is not forced choice as under possible and probable, more than one choice was possible.
4 levels	Positive identification. Possible identification, Insufficient evidence, exclusion.	Soomer et al.[a] Page et al.[b]	ABFO scale used[a] Interpol scale and ABFO scale used[b]
5 levels	(Reasonable medical certainty, probable, possible, exclude, inconclusive)[c] (Positive, Probable, Possible, Exclude and Insufficient Evidence)**c**	Pretty et al.[d] Page et al.[c]	Pretty et al.[d]- Stated as ABFO scale used. Only ABFO used to plot specificity and sensitivity at different threshold. Page et al.[c-] Used DVISys™ scale.

[a] ABFO- American Board of Forensic Odontology. (www. abfo.org).
[b] Interpol -International Criminal Police Organisation: Positive, Probable, Possible and Exclude.
[c] DVISys™ (DVI System International, Plass Data Software).

Table 7
Studies with information provided and types of information.

Study	Type of information provided
MacLean et al.	Dates for the antemortem and post-mortem radiographs provided.
Kogan et al.	Antemortem radiograph exposure dates, no post-mortem dates provided.
Fridell et al.	Sex and antemortem exposure dates of the antemortem radiographs and age of disappearance provided.
Soomer et al.	Case treatment notes for antemortem and short case history for post-mortem radiographs. No details about what the information consisted of.

between ante-mortem and post-mortem radiograph and Fridell et al. [7] provided date of exposure of ante-mortem radiographs and dates of disappearance in the study of ability to identify children without restorations from dental radiographs.

3.6. Specific research questions and specific area of focus in these studies

All studies had an overarching aim of evaluation of the method of using dental radiographs for identification, however, sub themes and specific areas of interest were found within this body of research. An extraction of these sub-themes is presented in Table 8. Four studies [9,11,17,19] queried the effect of the skill sets, qualifications and experience of the participants, while three studies [10,16,17] explored restorative status of the dentition on the accuracy of identification. Wenzel et al. [14] queried the validity of comparing digital to analogue intra oral radiograph, Fridell et al. [7] investigated if children without restorations could be correctly identified by using dental radiographs. Kogan et al. [18] investigated the effect of temporal interval on identification.

3.7. Validity: accuracy and reliability results

The results of the studies were complex to summarise due to the different approaches to analysis and calculations employed. The use of different types of participants, radiographs and different areas of emphasis of research in this group of studies added further complexity.

Table 9 presents the results of the "expert" group of participants in studies that fulfilled at least one criteria for validation studies. All these studies have confirmed the identities of the radiographs used for matching. The first five studies [13,14,17–19] are studies which used the paired strategy affording measurement of accuracy as sensitivity and specificity. The sensitivity ranged from 0.71 to 0.96 while the specificity ranged from 0.85 to 1. It is important to note that Maclean et al. [17] and Kogan et al. [18] only had three participants.

While the fifth study [12] used paired matching, an area under a ROC (Receiver operating characteristic) curve was presented for accuracy, hence, no mean sensitivity or specificity rate was presented.

The last four studies [7,9,10,15] used a free matching strategy and therefore only accuracy rates were available, the accuracy rates ranged from 88 to 94%. Only two studies Pretty et al. [12] and Ekstrom et al. [10] repeated the trials on the same practitioners. Pretty et al. [12] found that forensic odontologists had the highest repeatability and reproducibility while Ekstrom et al. [10] found that the participants did not make the same mistakes in the repeat trial.

Other significant results included the expert group was found to have higher accuracy with higher specificity in seven studies [7,9,12,14,15,17,19]. In addition, three studies [12,14,15] found practicing experience was correlated to performance. Kogan et al. [5] found that sensitivity decreased after 25 years.

4. Discussion

The PCAST report [3] advocated and provided a framework for establishing foundational validation research in the comparative forensic sciences. It was acknowledged that human decision and judgment was integral to these methods, yet little consideration was given to the human factors in the recommendations for the design of research. The social-behavioural aspect of such research is important as commented by Martire and Kemp 2016 [22]. The following discussion when approached from the perspective of requirements for basic foundational validity and the psychosocial aspect of evaluating human performance, provided some points of interest with regards to the material, method and research design.

The appropriate sample for validity testing according to the PCAST report [3] needs to be representative of those encountered in case work. Earlier studies [5,10,15–17] concentrated on comparison of the most common radiographs available at that time; analogue bitewings. Later studies [9,11–14] investigated the effect of the use of digital dental radiographs because technological advancement has impelled changes in methods. This demonstrates that the forensic odontology community was aware of the need for scientific validation of the method before the queries by the advisory committee. More importantly, it indicates that the forensic odontology community was also aware of the need to re-evaluate techniques as new technologies arise.

A number of the studies examined in this review fall short of certain considerations in the design and method for validation studies. Two studies [10,11] used forensic cases where the identities of radiographs were not independently verified. While the use of forensic cases may add authenticity and relevance, confirmation of identification should be counter checked by another modality i.e. DNA. This is significant because "ground truth" is required for computation of the true positive and negative rates. These two complementary rates are composites in an accuracy rate. As recognised by many authors in this group of studies the inclusion of specificity rates is important because the cost of false positive identification is higher than that of false negative [5,10,14,17,19]. For this reason, inclusion of non-matching rather than the exclusive use of matched radiographs is important as is the use of equal numbers of non-matching and matching radiographs for a balanced study design. Only four studies [5,13,14,17] had this balanced design. Apart from a balanced design, the total number of radiographs in a trial is also an important consideration. Inclusion of too many radiographs lead to failure of participants to complete the trials, likely due to boredom [13] and fatigue [5], which could confound the accuracy rate. This is an example of the reason for the need to take human factors into account. Ultimately it is human decision and judgment that are being validated and calibrated.

Obtaining large sample sizes of participants is an inherent problem

Table 8
Specific area of focus and research questions found in the articles.

Specific research questions	Studies
Comparison of analogue intra oral radiographs with digital intra oral radiographs.	Wenzel et al.
Identification of children without restorations via the use of dental radiographs.	Fridell et al.
The effect of use of digital dental radiographs using a web interface.	Pretty et al.
The effect of long time span between antemortem and post-mortem radiographs.	Kogan et al.
The effect of presence and absence of restorations between antemortem and post-mortem radiographs.	Borrman et al.; Ekstrom et al.; MacLean et al.
Quality of the observers and the effects on the accuracy when using this method.	Soomer et al.; Pinchi et al.; Kaur Bhullar et al.; Balla et al.;

Table 9
Results from the nine studies that partially fulfill the Pcast criteria for a validation study.

Study	Type of radiograph for comparison	Accuracy			Other methods used for calculating accuracy	Accuracy rate calculated by other method
		Overall accuracy = $\frac{TP+TN}{Match+non\ match}$	Sensitivity- TP rate = $\frac{TP}{Total\ match}$	Specificity- TN rate = $\frac{TN}{Total\ non-match}$		
MacLean et al.	BW to BW	mean = 0.93 as stated	0.71–0.99	0.97–0.99		
Kogan et al.	BW to BW	mean = 0.93	ave = 0.85	ave = 0.99		
Kaur Bhullar et al.	PA to PA	90.5%	89.3% SD ± 13.	92.3% SD ± 12.6		
Page et al.	Mixed OPG, PA and BW	mean accuracy = 87.5%. (mean accuracy = 96.7% when calculated from multi-level decision scale) [a]	TP = 89.3%,	TN85.6%		
Wenzel et al.	BW to BW Bw to PA	No accuracy rate calculated	0.96 TP	TN = 1	Scores presented as table.	
Pretty et al.	Mixed OPG, PA and BW	0.93 – for forensic odontologists			Receiver operator curve (ROC)- Area under Curve	
Fridell et al.	BW to BW				Total correct, no specificity	91% for 6–7 and 94 for 12–13
Ekstrom et al.	BW to BW				Not calculated presented results of 2 trials	0.88 first trial, 0.89 s trial
Sholl et al.	PA to PA BW to PA				Not calculated. Stated in abstract	93.3%- (Stated only in abstract)
Pinchi et al.	OPG to multiple PA and BW	96% = termed correct attribute	0.96	1	Unspecified for accuracy as described in article.	0.97

Study	Statistical method used	Decision scale	Matching design	Case information	Significant notes or conclusion	Repeat trials and notes of those with repeat trials.
MacLean et al.	Overall accuracy, sensitivity and specificity.	Binary forced choice	Paired matching (140 matched and 140 non-matched)	Dates for ante mortem radiographs and post-mortem radiographs	Forensic odontologists higher specificity rate	No
Kogan et al.	Overall accuracy, sensitivity and specificity.	Binary forced choice	Paired matching (100matched and 100 non -matched)	Partial - patient name and antemortem radiograph date	Sensitivity decreased significantly after 25 year interval between radiographs. Most of these older radiographs had significant restorations and changes from tooth lost.	No
Kaur Bhullar et al.	χ test and Mann Whitney *U* test to compare sensitivity and accuracy	Binary forced choice	Paired matching (6 matched and 4 non-matched)	No	Specialist performed better than general dentists. Higher specificity	No
Page et al.	Overall accuracy, sensitivity and specificity.[a]	Binary forced choice	Paired matching (25 matched and 25 unmatched) (verified)	No	Dental knowledge needed. Dental professionals accurate than layman	No
Wenzel et al.	NA	Binary choice after initial 3 choices	Paired matching (51 pairs of match and non-match)	No	Forensic dentists were more accurate and needed less number of trials to reach final decision.	No

Table 9 (*continued*)

Study	Statistical method used	Decision scale	Matching design	Case information	Significant notes or conclusion	Repeat trials and notes of those with repeat trials.
Pretty et al.	Are under ROC curve and Kappa for agreement over 2 trials	5 level ABFO	Paired matching (7 matched and 3 non-matched) (verified)	No	Practical experience of the forensic odontologists correlated with accuracy, and highest in repeatability and reproducibility	Yes. 2 trials. Dentists highly experienced had higher inter and intra- rater agreement; higher specificity scores. Kappa inter 0.89, intra 0.95
Fridell et al.	Fishers exact test and Pearson chi-square test	3 levels – 1. without doubt 2. possible 3. probable	Free matching- (extra 20 for unmatched antemortem)	No	Oral maxillofacial radiologists with dentists for each age group of children	No
Ekstrom et al.		Binary forced choice	Free match- one extra non-matched	No	Greatest number of mistakes in cases without restorations.	Yes. 2 trials. Mistakes not consistent inter and intra-rater. Judgments of difficult is different both times and to that of author.
Sholl et al.	NA	Binary forced choice	Free match- extra non-matched (antemortem)	No information but participants told of the extra “antemortem”	Forensic Odontologists with practical experience did better than those qualifications. Some Oral hygienist did better than dentists.	No
Pinchi et al.	Overall accuracy, sensitivity and specificity.	Binary forced choice	Free matching - with extra non-match. “No match” had to be specified for the non -matching radiographs by participants	No	Experienced forensic odontologists performed better then less experienced or higher qualified forensic odontologists. Better sensitivity rate of students attributed to Hawthorn effect.	No

OPG – panoramic radiograph.
PA- periapical radiograph.
BW- bitewing radiograph.
TP- True positive.
TN- True negative.
[a] Page et al. used both a binary decision and multi- level decision scale for the experiment.

in such research because forensic odontology is a subspecialty in dentistry and therefore the number of qualified or active practitioners is limited and performance of practitioners is the crux of such validation research.

The experimental group considered as the "expert" group of practitioners performed better consistently through all studies, however it is important to define who this "expert" should be as the definition of this group varied between studies. Determination of who is an expert is complex because specialist recognition and registration is both societal and evolving [23]. Dentists who practice forensic odontology identification but are not forensic odontology experts by training or education or recognition should rightly be included in validation studies. Comparison of the performance of "expert" with the "non-expert" is a "white box" study. "White box" studies attempt to provide insights into factors that affect decision making and to unpack the implicit cognitive processes [3]. These studies may provide the foundation for refining the factors that affect the quality of decision and cognitive aspects of judgment and decision in this discipline. Foundational validation studies can incorporate designs to allow "white box" observations. For example, Wenzel et al. [14] employed a strategy of reduction of three decision level choices to binary choices. Inference from the choice behaviour revealed that experts needed less number of trials than novices, reflecting their higher degree of confidence together with higher accuracy in the decisions.

A number of studies within this group included information relevant for decision making when comparing the radiographs [5,7,11,17], while the remaining studies [8–10,12–16,19] adopted the model similar to the design for a primary blackbox study referenced in the PCAST report [24], whereby fingerprint samples were presented for matching without any case history. Similarly, these studies isolated the ability to discriminate match and non-match without additional cognitive input apart from the information from the radiographs. The addition of case history or similar information may confound the results if the purpose is primarily validity as defined by PCAST. The inclusion of additional case information may introduce cognitive bias in the interpretation of the radiographs. It may also confound the enhanced ability of certain groups of subjects to integrate the information when comparing the radiographs with the isolated ability to discriminate matched pairs from non-matched pairs of radiographs. In fact, the effect of contextual information on the decision making process in forensic odontology identification has never been explored.

The contention with such primary validation research is the external validity as expressed by a few authors [10,12,17], mainly in actual cases of identification, other dental information is required and is integrated into the final decision. In addition to the inclusion of case information, some studies [7,11,12,14] used multi-level decision scales and the free matching strategy [7–10,15,16] in an attempt to replicate actual working conditions.

Categorical scales may appear representative of actual practice where opinions are expressed as levels of confidence; however, it requires a resolution to a binary decision for validity studies. If this resolution is not done by the participants, the thresholds between correct and incorrect responses have to be determined by the experimenter for statistical analysis [11–13]. The accuracy rate can also differ from such threshold determination as seen in Page et al. [13] where the accuracy rate was higher when the multi-level scales were used. The use of an experimental threshold also introduces an additional layer of interpretation [25] and may not reflect the actual decisions, hence, information about the process and behaviour is lost. Pretty et al. [12] used the categorical choices as threshold and cut off points summarising and reporting using a ROC curve. While this method removes the bias in the observers and provides an overview of accuracy, ROC curve does not allow obvious access to the sensitivity and specificity rates, this is because the area under a ROC curve provides a summed measurement of the ability to discriminate correct and incorrect decisions [26].

Free matching, instead of paired comparison does not allow calculation of specificity because active exclusions cannot be made unless non-matching radiographs were included and actively excluded as employed by Pinchi et al. [9]. Using free matching without inclusion of non-matching radiographs also potentially increases the number of incorrect decisions, as one wrong decision would mean another wrong decision especially where only true matching radiographs are used. This was recognised by Balla et al. [8] and pointed out in the PCAST report [3].

It is apparent from this review that the research questions, design and methodology of existing studies are diverse. The terms used in the titles provided a sense of the diverse approach to the concept to validating a method. Apart from the diversity in approach, it was noted from Table 8, that a number of studies [5,7,8,10,12,14,16,17,19] focused on specific types of radiographs and conditions for example, the comparison of only bitewings [5,10,16,17], restoration free bitewings [8,17], restoration free radiographs of children [7].

It is interesting to query if particular types of radiographs, population or dental status and conditions warrant specific investigations. Should this be analogous to separate error rates required for partial latent fingerprints differing from high quality fingerprint matching recommended in the PCAST report [3]? This has implications for appropriate inference of the external validity of the research. Laboratory based research with good internal validity is the most viable way of establishing foundational validity, however the external validity and application of these error rates in practice may require some deliberation. In the application of an error rate, it is the posterior probability that is paramount for the end user. To illustrate, the main interest for the decision maker is the probability that the remains are the presumed identity given that the opinion is a match. The error rate from research however is the error rate given a match and non-match; sensitivity and specificity. This is analogous to the probability of having a disease given a positive diagnostic test. This probability depends on the prior probability or the prevalence of the disease, however, in the forensic situation, the "prevalence" is not readily available.

The confidence in the opinion of an identification from comparison of radiographs results from personal beliefs in the probability of similarity in natural and iatrogenic dental characteristics. For example, it would appear that confidence in identification is inherently lower in the absence of restorations and outstanding features. As to whether discrimination rate for restoration free radiographs requires specific validation, as was done by Borrman et al., Ekstrom et al. and Macklean et al. [10,16,17] or if such conditions can be treated as a prior probability for modification of the foundation error rate generated from samples of all types of radiographs may require some deliberation.

Further examples will be dental identification in disaster victim identification. In such situations, a common practice is to compare multiple radiographs to find the right match. This was the reason for free matching designs in studies. There is a possibility that the rate will be significantly different due to cognitive bias and choice behaviour of such matching tasks.

Evett et al. [27], who advocates a logic inference approach to evidence, emphasised that an expert opinion is a personal belief, however, the opinion should be evidence based, and calibrated among peers. Perhaps as a discipline, forensic odontologists would benefit from consensus to the approach, model and framework of validation research.

5. Conclusion

The majority of studies in this scoping review were conducted prior to the NAS and PCAST reports. However, the existence of studies of this nature attests to the previous awareness of the forensic odontology community for the need to establish the scientific foundation of this method of identification. The heterogeneity of methodologies employed with regards to approach and research design in previous studies does

not enable the collation of results to allow for meaningful comparison of the conclusions drawn. More homogeneous studies with a degree of agreement by the fraternity as to the model and framework suitable would allow for determination of reliable foundational error rates. It remains uncertain if the types of radiographs used for comparison or circumstances of identification i.e. mass disaster scenario versus single identification warrants specific considerations. If further research establishes that these require specific considerations of error rates, possible benefit might also be gained from some consensus on what specific processes would warrant specific error rates separate from the general error rate. This is significant because continual re-validation of this method will be required as technology advances. For example, the use of computed tomography instead of analogue intra oral radiographs for comparison. This review highlights the need for an agreed framework and model within the discipline to serve as a foundation for studies, which can then be integrated and compared to provide more meaningful results.

Declaration of interest

None.

References

[1] D. Sweet, Forensic dental identification, Forensic Sci. Int. 201 (2010) 3–4.
[2] A.S. Forrest, Collection and recording of radiological information for forensic purposes, Aust. Dent. J. 57 (Suppl. 1) (2012) 24–32.
[3] President's Council of Advisors on Science and Technology, Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, https://obamawhite-house.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_foren-sic_science_report_final.pdf, (2016) (Washington DC).
[4] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, Strengthening Forensic Science in the United States: A Path Forward, National Academy of Sciences, Washington, DC, 2009.
[5] S.L. Kogon, A review of validation studies of dental bitewing radiographs for forensic identification, J. Can. Soc. Forensic Sci. (1996) 113–117.
[6] H. Arksey, L. O'Malley, Scoping studies: Towards a methodological framework, Int. J. Soc. Res. Methodol. Theory Pract. 8 (2005) 19–32, https://doi.org/10.1080/1364557032000119616.
[7] S. Fridell, J. Ahlqvist, The use of dental radiographs for identification of children with unrestored dentitions, J Forensic Odontostomatol. 24 (2006) 42–46.
[8] S.B. Balla, A. Forgie, Identification by comparison of caries free bitewing radiographs: Impact of observer qualifications and their clinical experience, Forensic Sci. Criminol. 2 (2017) 1–5, https://doi.org/10.15761/FSC.1000108.
[9] V. Pinchi, G.A. Norelli, F. Caputi, Dental identification by comparison of antemortem and postmortem dental radiographs: Influence of operator qualifications and cognitive bias, Forensic Sci. Int. 222 (2012) 252–255.
[10] G. Eksrom, Accuracy among dentists experienced in forensic odontology in establishing identity, J Forensic Odontostomatol. 11 (1993) 45–52.
[11] H. Soomer, M.J. Lincoln, H. Ranta, A. Penttilä, E. Leibur, Dentists' qualifications affect the accuracy of radiographic identification, J. Forensic Sci. 48 (2003) 1121–1126.
[12] I.A. Pretty, R.J. Pretty, B.R. Rothwell, D. Sweet, The reliability of digitized radiographs for dental identification: a Web-based study, J. Forensic Sci. 48 (2003) 1–6.
[13] M. Page, R. Lain, R. Kemp, J. Taylor, Validation studies in forensic odontology – Part 1: Accuracy of radiographic matching, Sci. Justice. (2017), https://doi.org/10.1016/j.scijus.2017.11.001.
[14] A. Wenzel, A. Richards, J. Heidmann, Matching simulated antemortem and postmortem radiographs from human skulls by dental studens and experts: Testing skills for pattern recognition, J Forensic Odontostomatol. 28 (2010) 5–12.
[15] S.A. Sholl, G.H. Moody, Evaluation of dental radiographic identication: an experimental study, Forensic Sci. Int. 115 (2001) 165–169.
[16] H. Borrman, H.-G. Grondahl, Accuracy in establishing identity by means of intraoral radiographs, J. Forensic Odonlo-Stomatology. 8 (1990) 31–35.
[17] D.F. MacLean, S.L. Kogon, L.W. Stitt, Validation of dental radiographs for human identification, J. Forensic Sci. 39 (1994) 1195–1200.
[18] S.L. Kogan, D.F. Maclean, S.L. Kogan, D.F. Maclean, Long-term validation study of bitewing dental radiographs for forensic identification validation study of bitewing dental radiographs for forensic, J. Forensic Sci. 41 (1996) 230–232.
[19] K. Kaur Bhullar, R.S. Bhullar, S. Balagopal, A. Ganesh, M. Raian, Evaluation of dental expertise with intraoral periapical view radiographs for forensic identification, J. Forensic Dent. Sci. 6 (2014) 171–176.
[20] G. Ekstrom, T. Johnsson, H. Borman, Accuracy among dentists experienced in forensic odontology in establishing identity, J. Forensic Odonlo-Stomatology. 11 (1993) 45–52.
[21] ABFO, ABFO Body Identification Guidelines, http://gbforensicservices.com/identification1.html, (2000) , Accessed date: 18 June 2017.
[22] K.A. Martire, R.I. Kemp, Considerations when designing human performance tests in the forensic sciences, Aust. J. Forensic Sci. (2016) 1–17.
[23] I.E. Dror, D. Charlton, The paradox of human expertise: why experts get it wrong, in: N. Kapur (Ed.), The Paradoxical Brain, Cambridge University Press, 2011.
[24] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. 108 (2011) 7733–7738.
[25] S. Halligan, D.G. Altman, S. Mallett, Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach, Eur Radiol. 25 (2015) 932–939.
[26] A.-M. Šimundić, Measures of diagnostic accuracy: Basic definitions, Med. Biol. Sci. 19 (2008) 1–9.
[27] I.W. Evett, C.E.H. Berger, J.S. Buckleton, C. Champod, G. Jackson, Finding the way forward for forensic science in the US—A commentary on the PCAST report, Forensic Sci. Int. 278 (2017) 16–23.