Forensic Science International 219 (2012) 183–198

Contents lists available at SciVerse ScienceDirect

Forensic Science International

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

Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools

Glenn Langenburg [a,b,*], Christophe Champod [b,1], Thibault Genessay [b,2]

[a] Minnesota Bureau of Criminal Apprehension (BCA), 1430 Maryland Avenue East, Saint Paul, MN, USA
[b] Ecole des Sciences Criminelles, Institut de police scientifique, Batochime, University of Lausanne (UNIL), CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 21 June 2011
Received in revised form 2 November 2011
Accepted 31 December 2011
Available online 24 January 2012

Keywords:
Fingerprints
Error rates
Statistics
Forensic science
ACE-V
Decision making

ABSTRACT

The aim of this research was to evaluate how fingerprint analysts would incorporate information from newly developed tools into their decision making processes. Specifically, we assessed effects using the following: (1) a quality tool to aid in the assessment of the clarity of the friction ridge details, (2) a statistical tool to provide likelihood ratios representing the strength of the corresponding features between compared fingerprints, and (3) consensus information from a group of trained fingerprint experts. The measured variables for the effect on examiner performance were the accuracy and reproducibility of the conclusions against the ground truth (including the impact on error rates) and the analyst accuracy and variation for feature selection and comparison.

The results showed that participants using the consensus information from other fingerprint experts demonstrated more consistency and accuracy in minutiae selection. They also demonstrated higher accuracy, sensitivity, and specificity in the decisions reported. The quality tool also affected minutiae selection (which, in turn, had limited influence on the reported decisions); the statistical tool did not appear to influence the reported decisions.

1. Introduction

Fingerprint analysts employ a generic protocol called ACE-V or Analysis-Comparison-Evaluation-Verification [1,2]. These phases of the protocol can be loosely described as the following: (1) analysis: an information gathering phase, where a fingermark[3] is assessed for its overall discriminating value and the reliability of the available discriminating features, (2) comparison: a side-by-side examination of features selected from the fingermark against corresponding or discordant features in a known print from a fingerprint card (exemplar), (3) evaluation: the weighing stage leading to the decision of the analyst to associate the images to a shared common source (“identification”[4]), the decision that the images do not share the same source (“exclusion”), or the analyst cannot make a determination (“inconclusive”), and (4) verification: the review of the decision by a second analyst.[5]

Generally speaking, these phases are not applied (or documented) consistently from laboratory to laboratory, or even analyst to analyst within the same laboratory [3,4]. This can be an issue of personal preference, case complexity, or simply a lack of enforceable standards[6] requiring any documentation. Furthermore, the quantum of corresponding information necessary to offer a decision of “identification” or “exclusion” is not clearly

* Corresponding author at: Minnesota Bureau of Criminal Apprehension, 1430 Maryland Avenue East, Saint Paul, MN 55106, USA. Tel.: +1 651 793 2967.
E-mail addresses: glenn.langenburg@state.mn.us (G. Langenburg), cchampod@unil.ch (C. Champod), thibault.genessay@unil.ch (T. Genessay).
[1] Tel.: +41 021 692 4629.
[2] Tel.: +41 021 692 4613.

[3] Throughout this paper, the authors have adopted the convention of referring to the friction ridge impressions left under uncontrolled conditions by an unknown source as “fingermark”, as opposed to “latent fingerprint”, which is more commonly used in the United States. When referring to a friction ridge impression from a known source, the term “print” or “exemplar” is used throughout the paper. Furthermore when using the term fingerprint, this refers generically to friction ridge skin impressions from the finger, but will typically apply equally to friction ridge impressions from the palms and feet.

[4] The term “identification” is used throughout this paper interchangeably with “individualization”. Both terms, in the context of this paper, are meant to describe an association between a mark and print of such degree that the chance of observing the corresponding features in another area of friction ridge skin is too remote to accept it as a realistic possibility.

[5] Depending on agency policies, procedures, or other circumstances, this review may include any of the following events: a review of all decisions made in the case, a review of the notes and procedures used, in blinded conditions or in sequentially unmasked conditions, where the verifying analyst is not privy to all of the case details or notes during key stages of the review.

[6] We note the recent efforts of SWGFAST in this context (Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). Standard for Documentation of Analysis, Comparison, Evaluation and Verification (ACE-V) (Latent), ver. 1.0. 2010: http://www.swgfast.org/documents/documentation/100310_Standard_Doumentation_ACE-V_1.0.pdf).

0379-0738/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2011.12.017

defined, but instead it is a fuzzy line of '*sufficient*' agreement or disagreement of ridge details based on the quality and quantity of features in the mark [5]. When there is inadequate documentation and an unclear threshold for conclusions, the process of decision making in fingerprint examinations can be best described as "opaque". This lack of transparency however does not necessarily equate to "unreliable", but rather the processes and steps toward decision making are not clear [6]. However, when an error is discovered, lack of transparency can lead to significant confusion and a host of problems [7]. In these cases, documentation and rationalization is usually done during post hoc inquiries and is unlikely to be insightful or accurate of the view at the time of the examination.

Similarly, studies that focus solely on the decision outcome (i.e. "a black box model" [8]) offer little insight into the decision-making strategies and nuances of the process [9]. Notwithstanding, black box studies do serve an important purpose of establishing decision error rates for a large volume of decisions and analysts, but at the sacrifice of less understanding of the steps taken to achieve the outcome [10,11].

An aspect to achieving transparency in decision making is the use of *reliable* measurements upon which to form an opinion. Two such measurements are suggested: (1) quantifying the quality of the friction ridge image and (2) assessing the significance of corresponding features between a pair of images. Quantifying the quality of the image involves an assessment of the friction ridge details against the background noise. Expressing the strength of the corresponding features (also referred to as "the weight of the evidence") through likelihood ratios (LRs) has been suggested as a means of representing the significance of a "match" [12–15].

The use of these tools can be explored with emerging technology. Several research groups are actively exploring the development of quality metric for images involved for pattern evidence. For assessing the quality of prints, we note the development of quality measures used in fingerprint recognition systems [16–18]. The U.S. National Institute of Standards and Technology (NIST) hosted two conferences (in 2006 and 2007) on quality of biometric data,[7] including fingerprints. For assessing the quality of marks, less published work is available, but we note the work carried out by the FBI in collaboration with Noblis, Inc. [19]. With respect to LR modeling for fingerprints, recent publications describe multiple efforts to develop tools to implement this approach [20–23].

These tools are not currently available to all forensic departments (for various reasons such as validation, acceptance, and commercial availability). Nonetheless, we can still aspire to improve transparency and reliability. Even without the use of quality assessment software or likelihood ratio tools, it is desirable to have identified reliable features to use throughout the examination process. One way to achieve this is to use features that are selected by a consensus of experts. In this manner, by definition, the features are reproducible. In the present research, the use of expert consensus information will also be explored.

The aim of the present study was to explore the effects of introducing tools which provided information about the examination at various stages throughout the examination process. Specifically, we assessed the following:

(1) How information regarding the clarity of friction ridge features (as provided in various formats to the participants) informed the judgments of fingerprint analysts.

(2) How information representing the strength of the corresponding friction ridge features (as provided in a likelihood ratio format) informed the judgments of fingerprint analysts.

(3) How information regarding the strength of the corresponding features when provided by *other* fingerprint experts, informed the judgments of the participating analysts. (This is typically referred to as a "consultation[8]" among fingerprint experts.)

We did this by comparing the participants' accuracy and consistency in feature selection, feature comparison, and the ultimate decision of source attribution.

2. Methods and materials

Approximately 600 fingerprint analysts were invited to participate in this research. Potential participants represented a wide range of experience, agency size, federal, state and local agencies, etc. A CD-ROM containing a software program was mailed to each potential participant. Each CD-ROM had a unique, random user name and password (to maintain anonymity) and the software allowed the user to log in to a secure server at the University of Lausanne (UNIL) in Switzerland. Users logged into a platform, developed by UNIL, called PiAnoS (Picture Annotation System).[9] Over 200 users successfully[10] logged in and started the study. At the close of the data collection period, 176 users completed all of the 12 experimental trials. This resulted in a total of 2112 completed trials. Some basic information (e.g. sex, years of experience, status of expertise) was captured for each participant using a short questionnaire during the initial log-in session.

The PiAnoS platform allowed for a controlled presentation of the fingermark and exemplar images in a manner that follows ACE-V—namely, an analysis is first conducted on the fingermark, absent an exemplar, and then the exemplar is presented for comparison, and finally a decision is reported. Analysts were provided a drop down menu to select from three reportable decisions: "identification", "inconclusive", and "exclusion". A comment box was provided for further explanation when "inconclusive"[11] was chosen as a decision. During Analysis and Comparison phases, tools for selecting and annotating features were provided (Figs. 1 and 2). PiAnoS provides a tool for annotating minutiae during the Analysis phase. The analyst could select the type of minutia (ending ridge, bifurcation, or unknown) during feature selection. The analyst could also use a ridge tracing tool to trace the ridges. Minutiae selected during the Analysis phase appeared on the screen as red annotations. During the Comparison phase, with the exemplar present, any newly selected minutiae appeared as yellow. Thus it was possible to see what minutiae were annotated during the Analysis phase (no exemplar) versus what minutiae were selected (and any changes made to pre-existing minutiae) during the Comparison phase (with the exemplar).

Participants were given twelve different fingermark comparisons to perform. They could save their work and return at their leisure, working at their own pace. Analysts were randomly assigned upon their first log-in to one of six experimental groups. Each experimental group applied combinations of the tools that were to be tested. Table 1 provides a summary of the experimental variables and tools for each group of participants. Prior to any experimental trials, participants were given a set of instructions and background information specific to their assigned experimental group. For instance, Group 3 was instructed regarding what LRs represent, how they were calculated, and the general principles of an LR approach. All groups were presented with a practice trial and detailed instructions for using their tools.

2.1. Quality Map tool

Participants under the Quality Map tool condition were provided information regarding the quality (clarity) of the ridge features during the Analysis phase. Low quality regions, containing potentially unreliable features, were distinguished from higher quality regions, according to the tool's assessment. A beta-version[12] of Noblis' Image Quality Tool was used to initially generate the Quality Maps. However, these images were then cleaned and modified for this experiment. Minutiae selected *by experts* during pilot testing were then overlaid onto the Quality

[7] Presentations and papers can be obtained from http://www.nist.gov/itl/iad/ig/bio_quality_wkshopi_present.cfm and http://www.nist.gov/itl/iad/ig/bio_quality_wkshopii_present.cfm.

[8] SWGFAST defines a consultation as a "significant interaction between examiners" and generally should be documented when it has an impact on decisions or outcome of a case [24].

[9] The initial development of PiAnoS is due to Julien Furrer and Romain Voisard under funding from the innovative pedagogical development fund of the University of Lausanne. PiAnoS has been modified to allow the administration of the scenarios used in this study.

[10] Many potential participants were unfortunately unable to participate due to technical problems stemming from overly restrictive government agency firewalls, protections, IT permissions, hardware/software compatibility issues, etc.

[11] For example, common reasons for an "inconclusive" conclusion were requests for additional, better quality exemplars or that the analyst stated there were insufficient, reliable features in the fingermark to effect a positive identification, etc.

[12] Universal Latent Workstation (ULW) Beta 5.6.0 software was used.

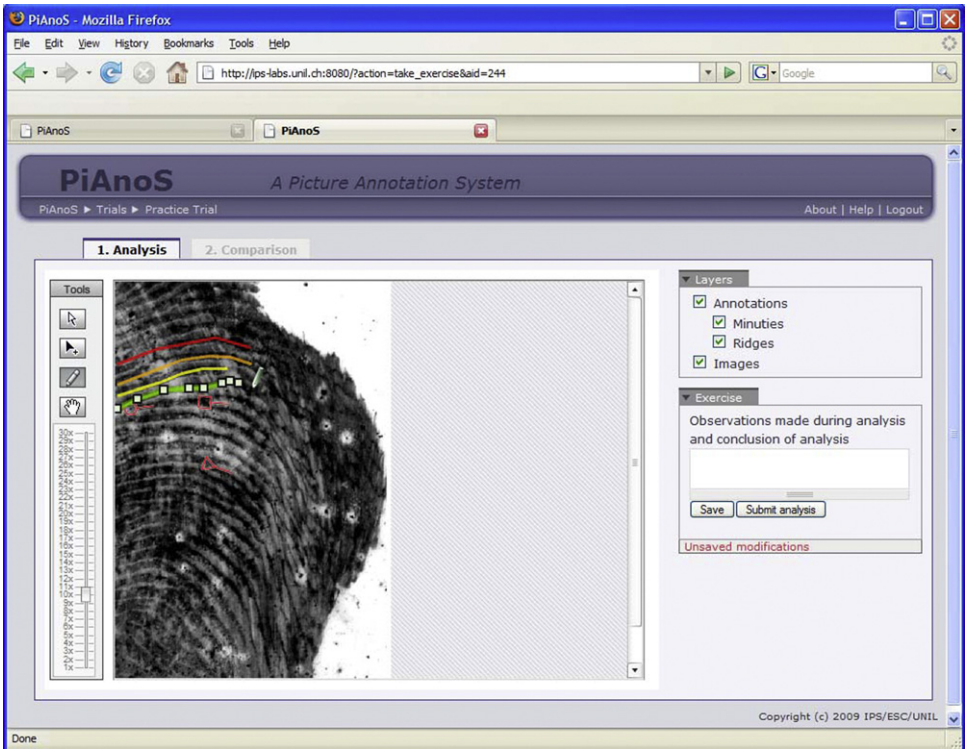

Fig. 1. Screenshot of PiAnoS showing a fingermark available for annotation of minutiae during the Analysis phase. Minutiae annotations and the ridge marking tool can be seen.

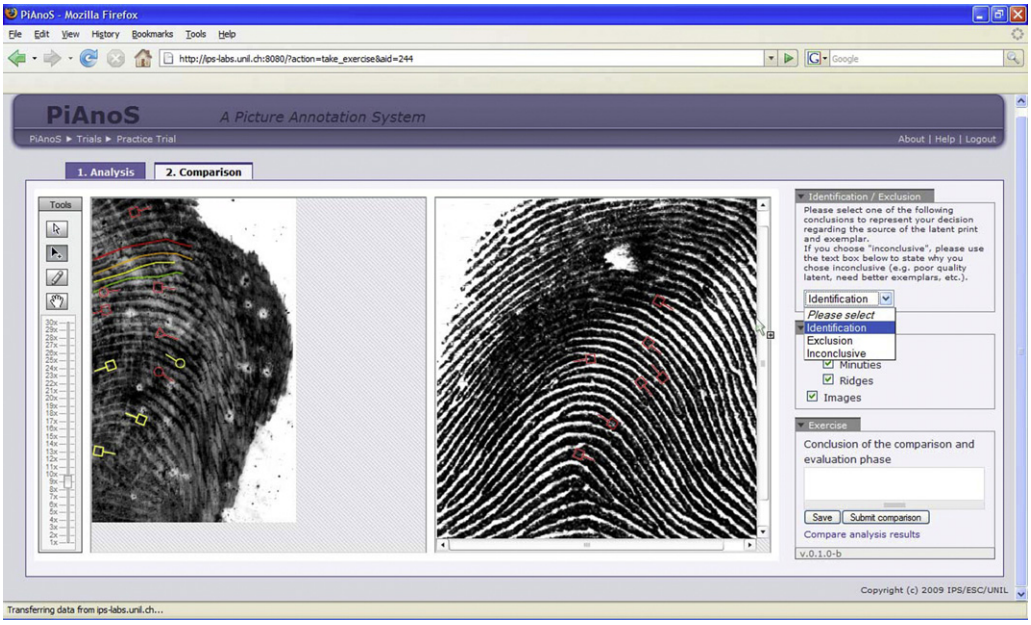

Fig. 2. Screenshot of PiAnoS showing the availability of the exemplar during the Comparison phase. The decision drop-down menu is shown on the right side of the screen.

Map. Thus a realistic proxy quality tool was produced, but participants were not informed that they were viewing the minutiae selected by other experts; rather, they were led to believe that the minutiae were selected by the software. Only the Quality Map colors were used from the Noblis software. See Fig. 3.

2.2. Likelihood Ratio Statistic tool (LR tool)

Participants under the statistical tool condition were provided information during the Comparison phase. This tool provided a statistic that represented the strength associated with the corresponding features. The tool used a likelihood ratio to represent the weight of the corresponding minutiae. The larger is the LR (greater than "1"), the stronger is the evidence in favor of the view that the unknown mark and the known exemplar share a common source. The smaller is the LR (less than "1") the stronger is the evidence in favor of the view that the images do not share a common source. The LRs were calculated using a statistical model for fingerprints developed by UNIL.[13] See Fig. 4.

[13] Research project funded by the Technical Support Working Group (TSWG) and the National Institute of Justice (NIJ): Statistical Assessment of Latent Print Evidence—TSWG task IS-FE-2478.

Table 1
Experimental variables and tools assigned to each group.

Experimental group	Analysis phase tool	Comparison phase tool
Group 1 (control)	No tool presented	No tool presented
Group 2	Expert Consensus Minutiae Map	No tool presented
Group 3	No tool presented	Likelihood ratio (LR)
Group 4	Quality Map	No tool presented
Group 5	No tool presented	Expert Consensus Decision table
Group 6	Quality Map	Likelihood ratio (LR)

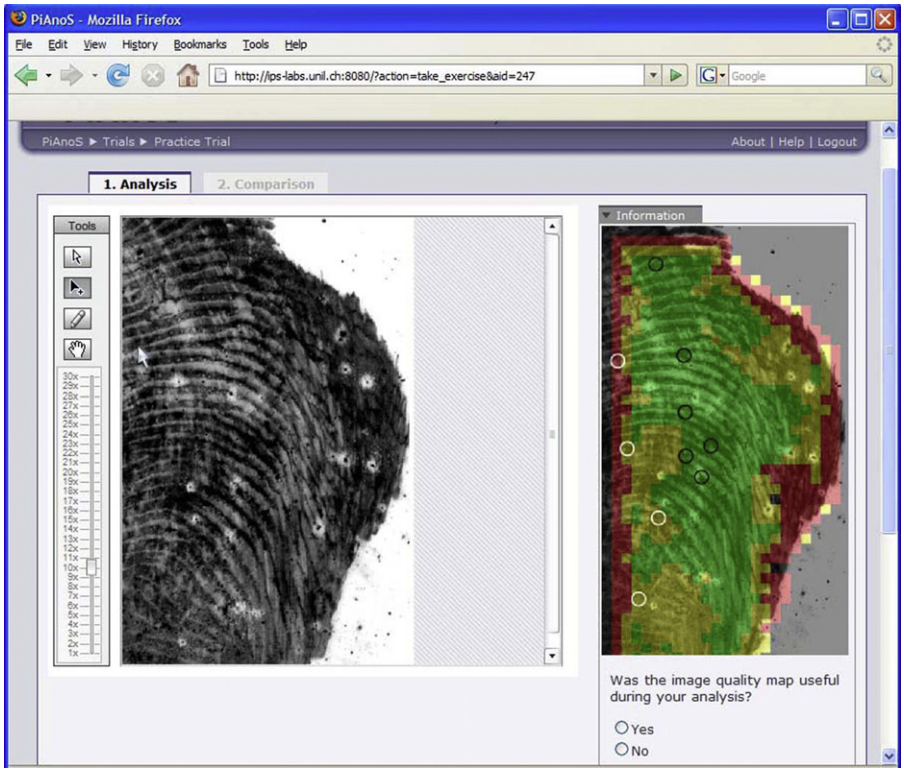

Fig. 3. A screenshot of PiAnoS showing the Quality Map tool that was provided during the Analysis phase for Groups 4 and 6.

2.3. Expert Consensus tools

In current practices, in the absence of more objective measurement tools, analysts commonly consult each other. Participants under the consultation conditions were either provided with the consensus minutiae or the conclusions of other participating analysts. Prior to starting this study, a pilot test of the trials was performed using 25 experts, each having at least 5 years of experience. Figs. 5 and 6 show the Expert Consensus Minutiae Map and the Expert Consensus Decision table that were provided to Group 2 and Group 5, respectively. Compare the minutiae selected in Fig. 3 to those selected in Fig. 5 (they are the same).

2.4. Case selection

Twelve cases were selected where the ground truth was known for all of the fingermarks. The fingermarks were produced "naturally" and not altered in any way. Participants were presented with seven trials where the fingermarks originated from the same source as the exemplar. These cases were determined (during pilot testing) to be fairly challenging and possessed significant distortions. Some of these same source cases exhibited apparent differences, yet were from the same source. The remaining five cases presented fingermarks against exemplars which did not originate from the same source. All five of these cases were close non-matches resulting from searches in a large fingerprint database.[14] Searches were maximized to find the closest candidates that would produce the most challenging trials for experienced analysts. During pilot testing, most of these cases produced errors and disagreement among experts. This was evidence that the cases selected were sufficiently difficult to provide a challenge to experienced experts. It was important to use difficult cases to promote errors for a measurable effect on the accuracy, error rate, and variance.

2.5. Data analysis

Data were tabulated into a spreadsheet and then analyzed using SPSS (v. 15.0) statistical analysis software. A table of the raw data is provided as an Appendix. Standard non-parametric tests for comparing distributions such the Mann–Whitney, Kruskal–Wallis, and Kolmogorov–Smirnov tests were used. These tests are more robust to outlier effects and make no assumptions about the parameters of a distribution. Pearson's Chi-square test was employed when comparing differences in categorical opinions (nominal data). The comparisons of levels of reproducibility were carried out using resampling techniques in R [25].

For the analysis of the minutiae on the marks, these data were extracted manually. Using the ground truth of the clear undistorted exemplar prints from the same area of friction ridge skin, we could determine which minutiae were accurately selected and which minutiae did not exist. When minutiae were annotated that did not correspond to a true minutia in the source exemplar, these minutiae were labeled as "false". There was a "one-ridge" lee-way in determining if the analyst correctly marked it. Analysts were scored false minutiae only when the issue was whether the feature existed. They were not scored an error with their classification of the minutiae type (ending ridge or bifurcation). For the analysis of the conclusions reported by the analysts, the ground truth for each trial was known. Therefore, an "identification" decision on a trial showing two images from different sources was labeled an "erroneous identification" decision. Conversely, an "exclusion" decision on a trial with images from the same source was scored as an "erroneous exclusion" decision. "Inconclusive" decisions were scored, but they were neither scored as correct decisions, nor as errors (see discussion below under "error rates"). All demographic/survey data were coded and entered for possible contribution effects.

[14] At the time of the searches, the size of the database was approximately 600 million fingerprint images.

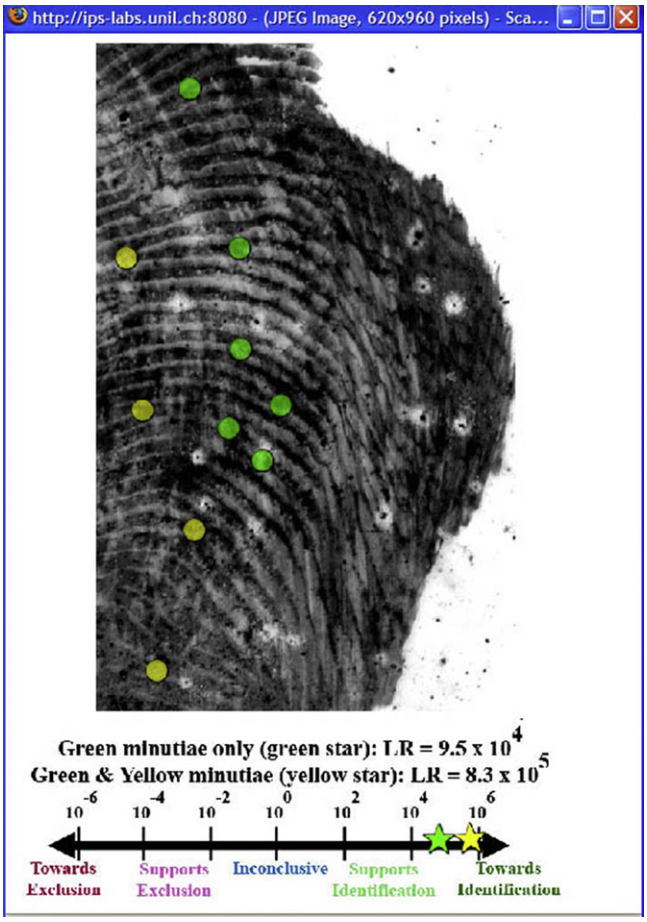

Fig. 4. A screenshot of the LR tool provided during the Comparison phase to Group 3. Participants in Group 6 received a similar image, but with a Quality Map overlaid on the image.

As discussed in Koehler [26], error rates are very appropriate measures for conveying the reliability of a method. However, there can be confusion on how best to handle the "inconclusive" decisions, when calculating false positive and false negative error rates.

To demonstrate the importance of this issue, we can calculate *three different* false positive rates from the data in Table 2. If we include "inconclusive" decisions in the total, but we do not count them as errors, and only count "erroneous identifications" as errors, then the false positive error rate is: 23/880 or 2.6%. If we do not include "inconclusive" decisions at all in the totals or as errors, as suggested by Koehler [26], then we have a false positive error rate of 23/788 or 2.9%. Finally, if we count "inconclusive" decisions as errors (a failure to exclude when the images came from different sources), then we have a false positive error rate of 115/880 or 13.1%.

For purposes of this paper, we have not chosen to report "inconclusive" decisions as errors per se, but still have considered them in the total number of trials for calculations. This is because it is reflective of actual practice, where many fingerprint analysts consider "inconclusive" a perfectly valid decision and in some instances more appropriate than a definitive source attribution decision (i.e. "identification" or "exclusion"). In some instances an "identification" decision, even if it reflects the ground truth, may be inappropriate to analysts, because essentially a decision of "identification" would be too strong a claim in light of the corresponding features at hand and more modest guidance is warranted.

However, this entire issue can be avoided by using *false positive discovery rates*. Indeed a major advantage to reporting false discovery rates is that the rate is not affected by the number of "inconclusive" decisions. The false positive error rate, can be written as "the proportion of cases in which an "identification" decision was reported, *given* that the fingermark and exemplar do not share a common source (see Eq. (1)[15]). The false positive discovery rate however is "the proportion of cases in which the fingermark and exemplar did not originate from the same source, *given* an "identification" decision has been reported (see Eq. (2)). Conversely, false negative error rates and false negative discovery rates are derived in a similar manner (see Eqs. (3) and (4)).

$$\text{False positive error rate} = \frac{\text{of cases of "identification" given } \bar{S} \text{ is true}}{\text{of cases where } \bar{S} \text{ is true}} = \frac{B}{B+D+F} \quad (1)$$

$$\text{False positive discovery rate} = \frac{\text{of cases where } \bar{S} \text{ is true among "identification" conclusions}}{\text{of cases of "identification"}} = \frac{B}{A+B} \quad (2)$$

$$\text{False negative error rate} = \frac{\text{of cases of "exclusion" given S is true}}{\text{of cases where S is true}} = \frac{E}{A+C+E} \quad (3)$$

$$\text{False negative discovery rate} = \frac{\text{of cases where S is true among "exclusion" conclusions}}{\text{of cases of "exclusion"}} = \frac{E}{E+F} \quad (4)$$

3. Results and discussion

The following generic hypothesis was tested in this study: experts would benefit from the implementation of tools that objectively and transparently assess clarity of friction ridge features and/or provide statistics regarding the strength of a match. This benefit would come in the form of reduced error rates, reduced variance during feature selection, and increased reproducibility of decisions among analysts. A table of the raw data can be found in Appendix A.

3.1. Error rates of reported conclusions

Using approaches for calculating error rates described by Koehler [26], a testing matrix was produced from the results of the experiment. Table 2 shows the results for all 176 participants.

In addition, as noted by Koehler [26], false positive discovery rates are more relevant to the task-at-hand in a trial. In the courtroom trial, typically the relevant question is: Given an "identification" has been reported, what is the chance the fingermark is not from this source (i.e. the analyst is wrong)? This is best represented by false discovery rates. The question answered by false positive error rates is a different one: given the images are not from the same source, what is the chance an analyst will provide an "identification" decision? One might argue the question of false positive error rates speaks to the accuracy of the method under an array of scenarios and possible outputs. This statistic may be more appropriate for an admissibility hearing (i.e. under *Daubert* considerations), while false positive discovery rates

[15] S and $\bar{S}$ are used to represent "same source" and "different source", respectively.

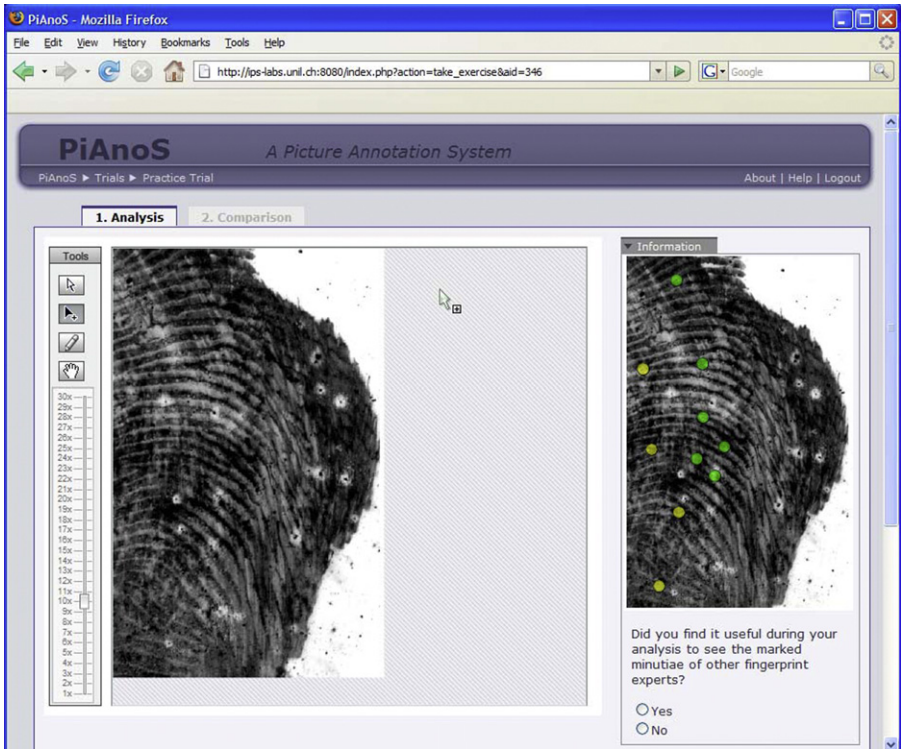

Fig. 5. The Expert Consensus Map that was provided to participants in Group 2 during the Analysis Phase. The yellow dots represent minutiae annotated by 50% or more of the 25 experts during pilot testing. The green dots represent minutiae annotated by 75% or more of the experts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

would be more appropriate to the task-at-hand during a trial where the defendant has been identified as the source of the fingermark. Regardless of the approach taken, at least with respect to the data in the present study, the difference between the false positive error rates and false positive discovery rates are minimal (2.6% compared to 2.7%, see Table 3).

How such rates should be used by the trier of fact is far from clear. Koehler argues that data and rates such as these are informative as "base rates" when the trier of fact is assessing the reliability of the method [26,27]. Others have argued that base rates are misleading, only the probability of a random match is important [28] or that the only relevant error rate is the chance of an error in *this* case, given

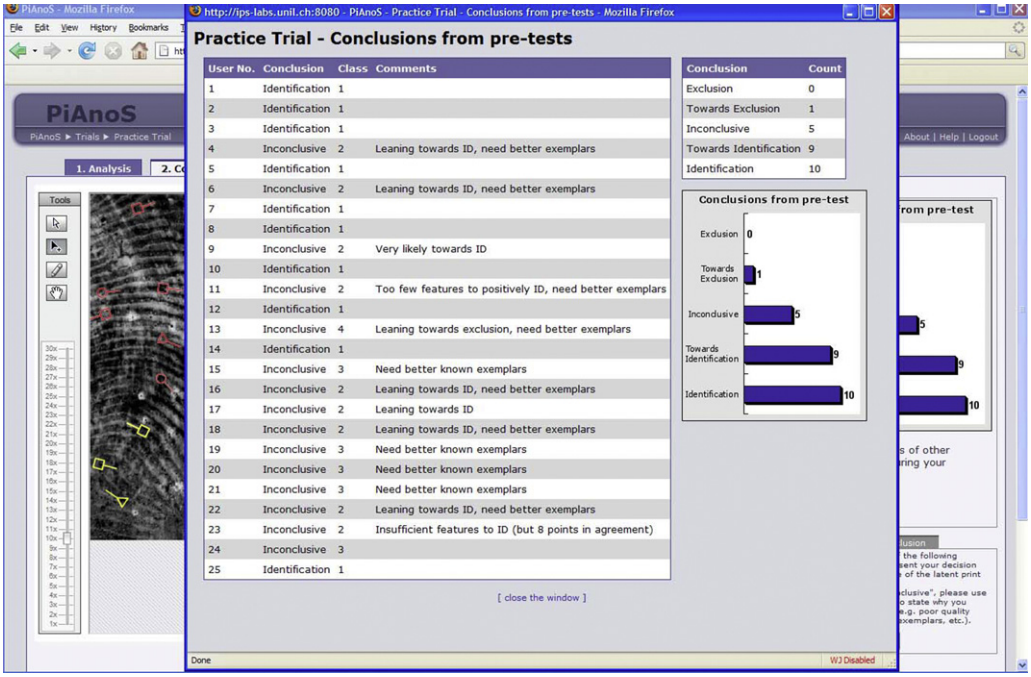

Fig. 6. A screenshot of the Expert Consensus Decision table shown to participants in Group 5, during the Comparison phase. The opinions presented and the comments shown are the actual decisions and results of the 25 experts tested during pilot testing.

Table 2
Test results matrix for all 176 participants who each completed all 12 trials.

	Ground truth of fingermark source		Total
	Same source (S)	Different source ($\bar{S}$)	
Analyst's decision			
Identification	840 (A)	23 (B)	863 ($A+B$)
Inconclusive	322 (C)	92 (D)	414 ($C+D$)
Exclusion	70 (E)	765 (F)	835 ($E+F$)
Total	1232 ($A+C+E$)	880 ($B+D+F$)	2112 ($A+B+C+D+E+F$)

quality assurance methods and the potential to reanalyze [29]. Still others have argued that while random match probabilities are useful, they are overshadowed by methodological and human error which may be larger by orders of magnitude [30]. Nonetheless, courts and commentators are asking for them [4,31,32]. We are hopeful that data in the present study will be useful to triers of fact when assessing the reliability of ACE-V, although the authors are not prescriptive about how they should be used. Furthermore, it cannot be overlooked that error rates are essential for research studies such as this when comparing variables introduced into the methodology. They are important for measuring and comparing effects. As a research tool, they cannot be avoided.

From Table 2, we can also calculate two other standard diagnostic testing terms: sensitivity and specificity. Sensitivity (in terms of ACE-V) refers to the proportion of reported "identification" decisions when the images originated from the same source (see Eq. (5)).[16] Specificity in this instance, refers to the proportion of reported "exclusion" decisions when the images did not originate from the same source (see Eq. (6)).

$$\text{Sensitivity} = \frac{\text{of cases of "identification" given S is true}}{\text{of cases where S is true}} = \frac{A}{A+C+E} \quad (5)$$

$$\text{Specificity} = \frac{\text{of cases of "exclusion" given } \bar{S} \text{ is true}}{\text{of cases where } \bar{S} \text{ is true}} = \frac{F}{B+D+F} \quad (6)$$

From the data in Table 2, sensitivity was calculated as 840/1232 or 68%; specificity was calculated as 765/880 or 87%. These statistics demonstrate an imbalance toward "exclusion" decisions in this study. Participants were more prone to report an "inconclusive" decision in a 'same source' trial where the available information was at or near the analyst's decision threshold than report an "inconclusive" decision in a 'different source' trial with a similar profile of available ridge details. We offer two (not mutually exclusive) reasons for this. The first is that the threshold for an "exclusion" decision is most likely much lower. Concepts such as using class characteristics to exclude when they are not discriminating enough to individualize, and the "one-dissimilarity doctrine" [33] would seem to exemplify this. The second reason is based on the risk-benefit utility function of the process [34]. While both erroneous "identifications" and erroneous "exclusions" are errors, the community historically has treated these two errors very differently, especially in terms of punitive repercussions for the erring analyst [35].

[16] Please note that the authors are being very specific in the language here (to the point of excessive wordiness) by avoiding phrases such as "correct identifications" or "correct exclusions". We wish to stress that although an "identification" or "exclusion" decision may be in agreement with the ground truth, some of these decisions may not be qualified as "correct" because such a claim may not be justified in the light of the observed features. An example could be for an analyst to identify the correct source using only three matching minutiae with very low discriminating power.

To assess the impact of the tools, error rates for each group were calculated. These are displayed in Table 3. Group 2 and Group 5 committed fewer errors, but a Chi-square test did not show a statistically significant decrease (due potentially to the relatively low error rates to begin with). Both Groups 2 and 5 were the groups exposed to Expert Consensus tools.

Examining how experience may affect error rates and performance, the experts (those analysts independently working cases) and trainees were separated into different categories and the experts were stratified by years of experience analyzing and comparing fingermarks on a routine basis (see Table 4). A clear trend was observed. False positive error rates started quite high for trainees, but then significantly dropped (to almost zero) during the first couple of years of casework. Then as experience levels rose, false positive rates increased to the highest levels for the most experienced analysts. This trend was mirrored in the sensitivity (i.e. they were attempting more "identification" decisions after gaining more years of experience). A possible explanation for this trend is that experts become overly confident with time and their self-confidence induces them to "push the envelope". Similar trends have been reported in the medical community and other expert domains [36–38].

Conversely, false negative error rates are higher in the least experienced group of experts (2 years of experience or less) and are reduced in the most experienced group. Simultaneously, the specificity is increasing. Thus we can see that experts are becoming more efficient at excluding with more experience (i.e. they are attempting more "exclusion" decisions while simultaneously making fewer erroneous "exclusions"). Comparing the false negative rates for trainees, we see they are actually the lowest of all the experience groups. However, the specificity is also the lowest, which means that they are attempting fewer "exclusion" decisions, in general. The trainees also have the highest sensitivity, which means they are attempting more "identification" decisions (but at the price of the most erroneous "identification" decisions). These statistics level out though once they are working independent casework. These data show a transition toward a conservative "identification" threshold, but a less conservative "exclusion" threshold.

With respect to the "inconclusive" decisions, readers may have concerns about the number of "inconclusive" decisions in this study (i.e. 20% of the decisions were "inconclusive"). At first glance, this may appear to be a surprisingly high number of "inconclusive" decisions, compared to what may be found reported in casework. Recall however, that considerable effort was put into selecting twelve challenging and difficult trials that would push decision making to the thresholds. A number of trials had been pilot tested with other experts prior to data collection. Based on the results of the pilot testing, these twelve were chosen for the challenges they would present. Therefore, given the experimental design, it is not surprising that many decisions resulted in "inconclusive". The images are included in Appendix A and the appropriateness of this view is left to reader.

Other factors were explored for the effect on error rates. Some of these factors included the sex of the participant, whether the participant had taken any courses in statistics applied to forensic science, whether the individual was certified by a recognized professional organization, whether the participant traced ridges with the ridge tool, and whether the participant personally found the tools useful during trials. None of these factors appeared to significantly affect error rates and performance. One interesting factor that did impact error rates was whether the participant documented and annotated at all. A handful of participants ($N = 4$)[17] chose not to annotate their features or document the ACE process at

[17] We must use extreme caution in interpreting these results given such a small sample size.

Table 3
Error rates and other performance statistics calculated for each experimental group. As a reminder: Group 1 (control group, no tool), Group 2 (Expert Consensus Minutiae Map), Group 3 (LR tool), Group 4 (Quality Map tool), Group 5 (Expert Consensus Decisions table), and Group 6 (Quality Map and LR tool).

	N	Error rates		Discovery rates		Sensitivity	Specificity
		False +	False −	False +	False −		
Group 1	27	3.7%	6.3%	4.0%	10.1%	63%	79%
Group 2	33	0.6%	5.2%	0.6%	7.3%	67%	92%
Group 3	28	2.1%	5.1%	2.1%	7.6%	71%	86%
Group 4	31	4.5%	6.5%	4.6%	9.9%	67%	82%
Group 5	32	0.6%	5.8%	0.7%	7.8%	67%	96%
Group 6	25	4.8%	5.1%	4.3%	7.9%	75%	84%
All groups	176	2.6%	5.7%	2.7%	8.4%	68%	87%

Table 4
Error rates and performance statistics separated by years of experience. Note: five participants were not included above because they were not actively performing casework or in a training program.

Expertise	*N*	Error rates		Discovery rates		Sensitivity	Specificity
		False +	False −	False +	False −		
Trainees	13	9.2%	3.3%	8.0%	6.0%	76%	72%
Experts (all)	158	2.2%	6.0%	2.2%	8.7%	67%	88%
Experts with							
≤2 years exp	26	0.8%	8.8%	0.8%	12.6%	65%	85%
3–7 years exp	48	1.7%	5.7%	1.8%	7.9%	66%	92%
8–15 years exp	49	2.9%	5.2%	3.0%	8.2%	66%	82%
16–35 years exp	35	2.9%	5.3%	2.8%	7.5%	71%	92%
Total	171	2.7%	5.8%	2.7%	8.4%	68%	87%

all. The false positive and false negative error rates in this group were 15.0% and 7.1%, respectively, compared to 2.7% and 5.8% reported by the participants that documented the process. This is a significant increase in error and may indicate the value of careful documentation, exercise of caution, and working slowly when dealing with complex cases. Incidentally, those that did not document were all case working experts, ranging from 10 to 30 years of experience.

3.2. Reproducibility of decisions

Reproducibility is defined as the ability of a test or method to provide consistent results when the same sample is provided to different instruments (here fingerprint examiners). If we look at the reproducibility of the conclusions for all groups, as shown in Fig. 7, it can be seen that some trials were more reproducible than others.

When the reproducibility of the decisions was compared by experimental group, Group 2 and Group 5 (the Expert Consensus tools) had the highest reproducibility. The probability of any randomly paired analysts in Group 1 (the control group) having the same conclusion (regardless of accuracy) was 63%. Groups 2 and 5 had the highest reproducibility probability—both were at 78%. Groups 3 and 6 had reproducibility probabilities of 72% and 76%, respectively. Group 4 (the Quality Map) produced the lowest reproducibility (69%).

Using a resampling technique for comparing proportions, the significance of these reproducibility values was tested [39]. Fig. 8 shows the boxplots resulting after the computation of the

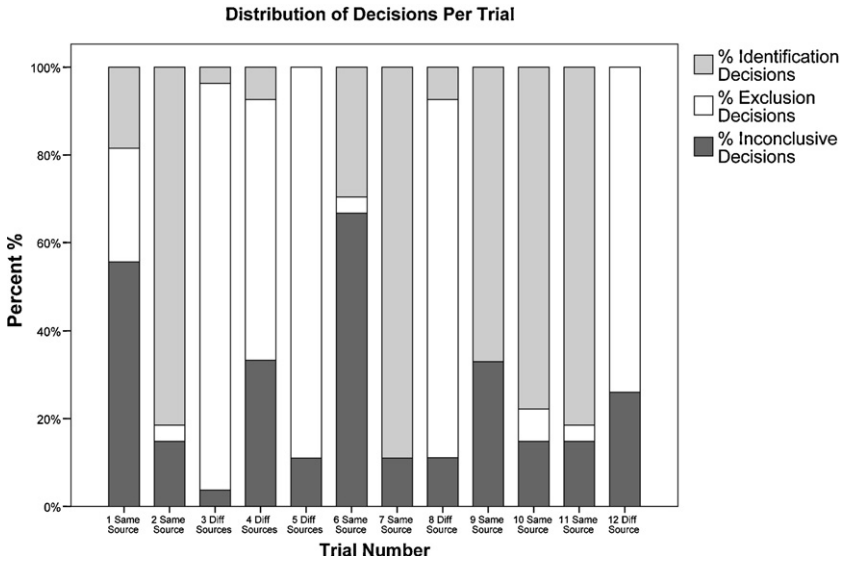

Fig. 7. The graph shows the variation of reported decisions for each trial. All groups have been pooled.

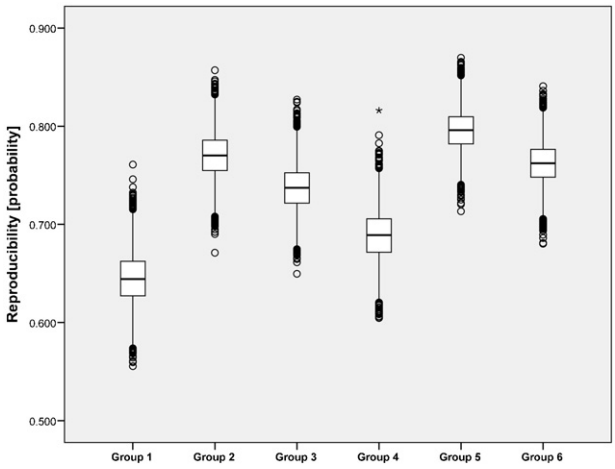

Fig. 8. Boxplots of bootstrap samples (N = 10,000) for the reproducibility values of each experimental group.

reproducibility on 10,000 bootstrap samples taken with replacement from the obtained empirical samples from each group. The distributions of these estimated reproducibility probabilities show increased reproducibility in all test groups when compared to the control group, and Groups 2 and 5 are demonstrating the highest and most similar level of reproducibility.

Comparisons between Group 1 and Groups 2 through 6 can easily be carried out based on these bootstrap samples by computing the difference between each sample and assessing how the proportion of simulations that led to a difference of reproducibility is below zero. Results are presented graphically in Fig. 9. The only group that is above 1% is Group 4 (11%), which is the closest to the control Group 1.

3.3. Ridge and minutiae annotations

Recall that users could annotate minutiae in the fingermarks during the Analysis phase without the presence of an exemplar, adjust their annotations in the fingermark once they have viewed the exemplar, and lastly, annotate minutiae in the exemplar print that corresponded to features in the fingermark. These data were labeled, respectively, A_{min}, O_{min}, and C_{min}. Additionally, participants could use a ridge marking tool and trace ridges. Ashbaugh states the importance of analyzing and comparing ridge systems, while avoiding focusing solely on minutiae configurations [2]. While several papers discuss the highly discriminating value of minutiae configurations (e.g. [20,21,23,40,41]), there is additional discriminative value with the inclusion of ridge counts and ridge tracings.

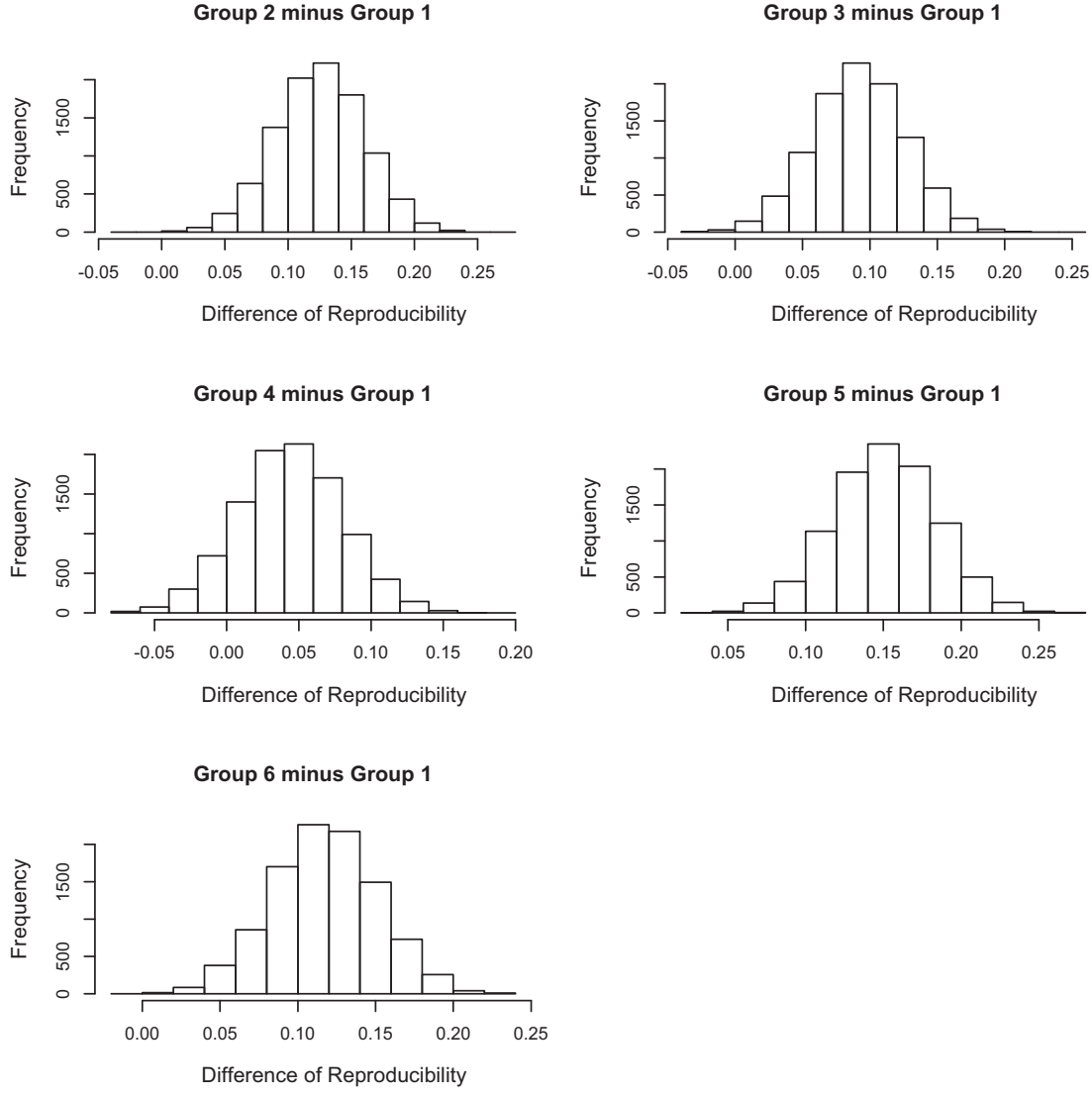

Fig. 9. Histograms representing the proportion of differences in comparisons of reproducibility between Group 1 (control group) and the experimental groups.

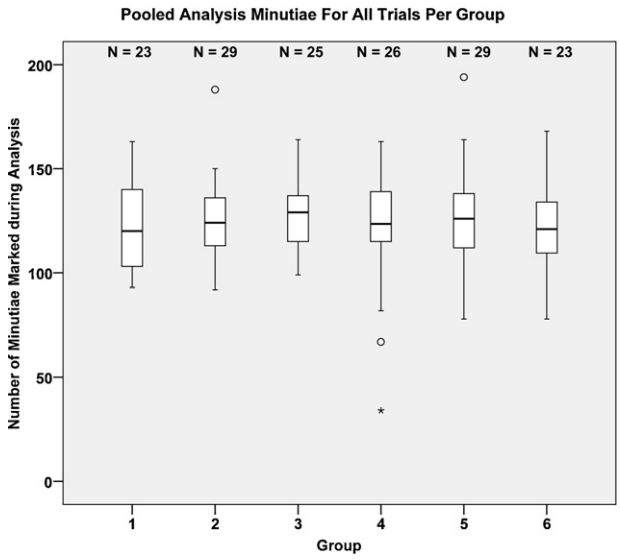

Fig. 10. Boxplots showing the total number of Analysis minutiae (A_{min}) pooled for all trials, across groups for experts only. There were no statistically significant differences for means, variance, and relative standard deviation (RSD) across groups. Note that Groups 2, 4, and 6 used minutiae maps to focus the analysis of the participants toward consensus minutiae.

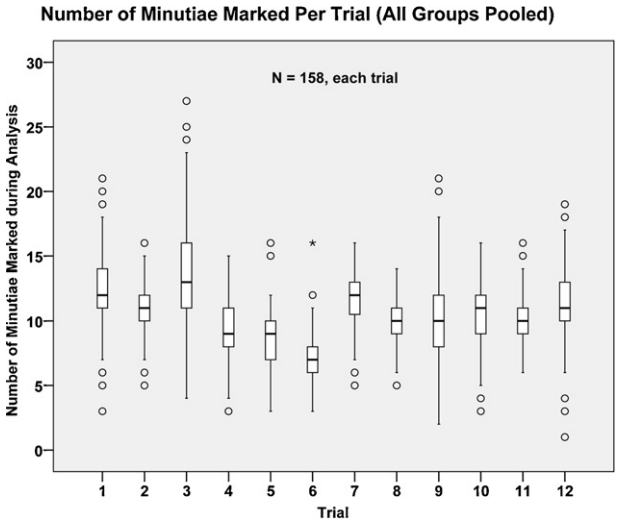

Fig. 11. Boxplots showing the number of Analysis minutiae (A_{min}) marked by experts (all groups combined) for the twelve trials. It is clear that some of the fingermark images produced a larger variance of annotated minutiae than other images.

In the trials where experts employed the ridge tool during the Analysis phase (i.e. traced some of the ridges), a higher mean number of A_{min} were annotated by these users compared to the mean number of A_{min} for users that did not use the ridge tool [Mann–Whitney test, $p < 0.001$; mean (did not use tool) = 10.1 (3.4 SD), compared to mean (used tool) = 11.2 (3.6 SD)]. However, the use of this tool did not reduce the total number of reported erroneous decisions regarding source attribution. Those using the ridge tool had a 5% erroneous decision rate and those not using the ridge tool had a 4% erroneous decision rate. However, it must be clarified, that just because an analyst did not use the ridge tracing tool, it does not mean the analyst did not analyze and compare ridges, it simply means the analyst did not formally document it in PiAnoS.

As for the Analysis minutiae (A_{min}), there were no statistically significant differences for means, variance, or relative standard deviations (RSD) across experimental groups (Group 1–6) (see Fig. 10) reported by expert participants. *Thus we conclude that the tools did not affect the number of minutiae selected during the Analysis phase*. For Analysis minutiae changed or added during the Comparison phase (O_{min}) and for minutiae marked in the exemplar during the Comparison phase (C_{min}), no statistically significant differences for means, variance, or relative standard deviations (RSD) across experimental groups were observed either. Kruskal–Wallis tests were used to test the means and RSDs for significant differences between groups. Levene's test was used to test the variances for significant difference between groups. All of these statistical tests produced p values typically much greater than 0.500, and thus it was concluded that there were no statistically significant differences for these metrics when comparing the total numbers of A_{min}, O_{min}, and C_{min} pooled for all trials.

It should be noted that given the instructions of the experiment, which did not instruct participants to necessarily focus on creating corresponding pairs of minutiae during the comparison phase, caution must be exercised when interpreting the C_{min}. It is quite likely, that if participants found few features in agreement, they may not have marked many minutiae in correspondence or few at all if the impressions were well below their level of sufficient ridge detail to effect a decision. Conversely, participants may have stopped annotating corresponding minutiae pairs once they achieved a threshold decision level. If they had been instructed to mark all pairs, regardless of whether the corresponding pairs fell well below or well above their decision threshold, the data may have been quite different at the extreme low and high ends.

Factors such as sex, expert status (i.e. trainee versus certified expert versus expert, etc.), having had a statistics course, and years of experience were considered as possible contributors to differences in minutiae selection. Only "years of latent print experience" showed some statistically significant difference in the distribution of minutiae (Kruskal–Wallis, $p = 0.073$): the "less than 2 years of experience" group reported the lowest RSDs for the A_{min} (all trials pooled).

Not surprisingly, the largest driving factor that showed differences in mean number of minutiae (for A_{min}, C_{min}, and O_{min}), variance, and RSD was the fingermark itself. In other words, as observed in other studies [42–45], the quantity and quality of ridge detail available in the fingermark drove the minutiae selection process (Kruskal–Wallis, $p \ll 0.001$). In Fig. 11, the boxplots showing the distribution of A_{min} across all trials, clearly demonstrate these significant differences.

3.4. Number of minutiae and the expert's decision

There was a statistically significant difference in the number of minutiae reported in each trial with respect to the decisions reported by the expert participants. Experts reporting a definitive conclusion ("identification" or "exclusion" decision) were more likely to have annotated a higher number of minutiae in the Analysis and Comparison phases (Kruskal–Wallis, $p < 0.001$).

Table 5
The mean and standard deviation for the number of minutiae (A_{min}) annotated during the Analysis phase (without an exemplar present) and during the Comparison phase (C_{min}) (annotated in the exemplar print and presumably corresponding to minutiae selected in the fingermark). The data have been stratified according to the decision reported by the expert. The data above only represent the seven same source trials.

Number of minutiae	Mean	Std. dev.	N
A_{min} | "Identification"	10.9	2.3	727
A_{min} | "Inconclusive"	9.2	3.6	287
C_{min} | "Identification"	11.5	3.1	744
C_{min} | "Inconclusive"	9.2	4.0	296

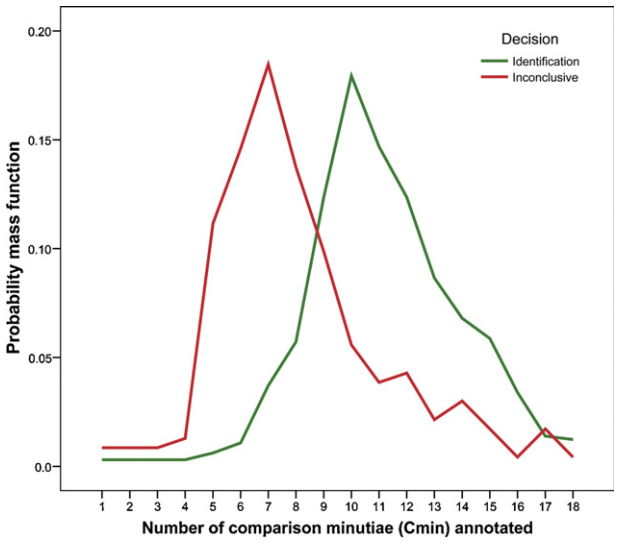

Fig. 12. Probability mass function for the number of minutiae annotated in the exemplar during the Comparison phase (C_{min}) in the seven same source trials. The distributions are conditioned on the decision reported by the expert ("inconclusive" versus "identification").

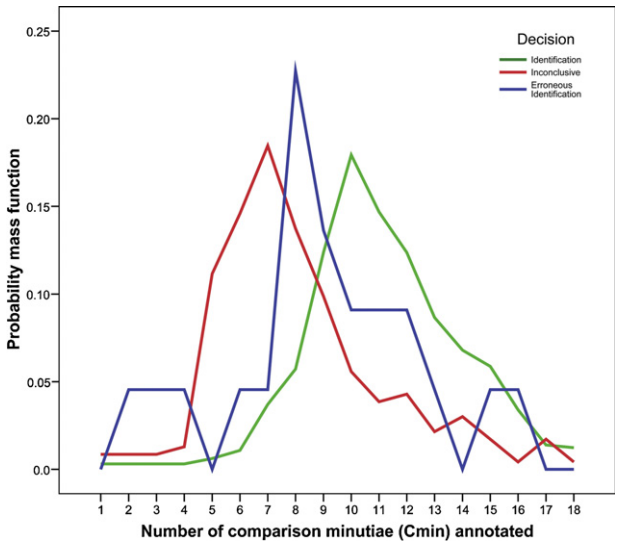

Fig. 13. Added in blue, the distribution of minutiae marked in Comparison phase (C_{min}) when an erroneous "identification" decision was made. Note that these data represent only 22 data points (including 6 data points from trainees) compared to hundreds of data points for the distributions of C_{min} for "identification" and "inconclusive" decisions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 5 provides the means and standard deviations for the minutiae annotated when an "identification" decision was provided (given an "Identification" decision or | "Identification") versus when an "inconclusive" decision was provided (| "Inconclusive"). The data in Table 5 reflect only the trials where the images were in fact coming from the same source.

The probability densities for the number of marked minutiae in the Comparison phase for these conditions are plotted in Fig. 12. From Fig. 12, it can be observed that the point at which the analyst's decision changes from an "inconclusive" decision to an "identification"[18] decision is between 8 and 9 minutiae. Thus we can infer that, at least under the conditions of this study, experts had an *operational decision threshold* around 8 or 9 minutiae. While experts are trained to not solely base their decision on minutiae alone, we can still predict that above 8 or 9 minutiae in correspondence, experts will be more likely to report a positive attribution.

We did not examine the correlation of decision and number of minutiae reported in cases where the images came from different sources. The main reason for not doing so was that participants normally did not completely annotate or marked few minutiae in comparison after reaching an "exclusion" decision. However, we did examine the 23 false positive cases and overlaid the results in Fig. 12 distributions. Remember, in these 23 cases, the participants believed the images were coming from the same source and therefore annotated what they believed to be valid correspondences.[19] Because of the relatively few data points (compared to the correct "identification" decisions and "inconclusive" decisions), the 6 trainee false positive errors were included in these data. The distribution of minutiae annotated in the Comparison phase for these erroneous decisions is shown in Fig. 13. It is interesting to observe that the mean number of Comparison phase minutiae (mean = 9.0, SD = 3.5)[20] for the erroneous "identification" decisions is found at the previously discussed *operational decision threshold*.

[18] In some agencies, analysts would report "No value" or "not of value for identification purposes" instead.

[19] However, one of the 23 false positives was not annotated at all.

[20] Removing the six false identifications made by six (of the thirteen) trainees, who generally had higher numbers of minutiae annotated, the mean and standard deviation change from 9.0 (SD = 3.5) to 8.4 (SD = 3.6).

3.5. Inter-observer variation

If two analysts each annotate a total of 11 minutiae in the same trial, it may be tempting to think these two analysts are in agreement. However, Fig. 14 illustrates just how incorrect this notion is. The images in Fig. 14 depict the Analysis phase annotations of two analysts in the same group (both were classified as experts). While both have 11 minutiae annotated, they only share 6 minutiae between them. The participant on the left has annotated 5 different minutiae than the participant on the right and the participant on the right has annotated 5 different minutiae than the participant on the left. Therefore, it can be said there is a difference of 10 minutiae annotated between these two participants. This type of measurement is called a *Euclidean Squared Distance* (hereafter ESD). ESDs offer deeper insight into the inter-observer differences between analysts.

Pairwise comparisons of the annotated minutiae can be performed for all pairs of participants in each experimental group and across all groups. Fig. 15 is a histogram of the ESDs across all groups for all 158 expert participants (although not all 158 experts annotated minutiae in each trial). This generated $1/2N \times (N-1)$ pairwise comparisons. For each group, this resulted in approximately 300–400 expert pairwise comparisons for each trial. Across all groups, for each trial, this resulted in approximately 11,000–12,000 pairwise comparisons.

Recall from Fig. 11 that the mean number of A_{min} annotated by experts in Trial 1 was 12.2. If we divide the mean ESD (5.24) by the mean A_{min} (12.2), we see that the relative ratio of ESD difference to the average total number of minutiae marked is 0.43. In other words, on average, 43% of the experts' annotations differed between randomly selected pairs of experts. A summary of this ESD ratio for each trial, and other relevant statistics can be found in Table 6.

It can be seen from Table 6, that for this study, the average difference between analysts, across trials was approximately 44% (SD = 9%). Therefore, on average, selecting any 2 expert analysts, one would expect approximately between 37% and 51% (99% confidence interval, $\alpha = 0.01$, SD = 0.09, $n = 12$) of the minutiae annotated in the Analysis phase to differ between the analysts for the fingerprint comparisons in this experiment.

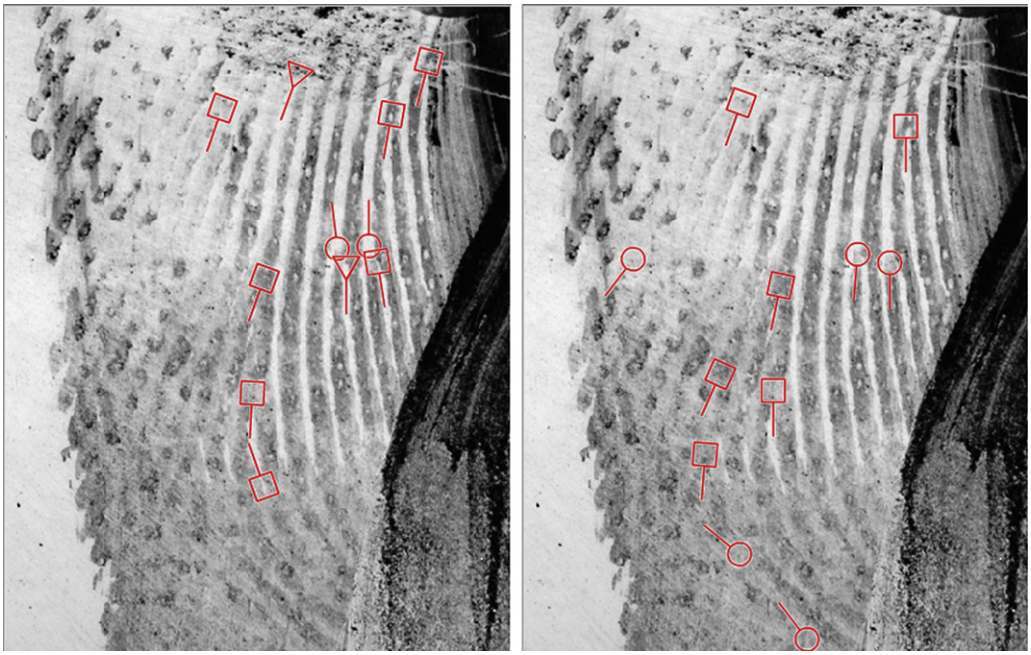

Fig. 14. A screenshot of the Analysis minutiae annotations between two experts from the same group. While both participants have 11 minutiae, they share 6 minutiae in common and have 10 different minutiae between them.

3.6. Minutiae and Quality Maps effect

Participants in Groups 2, 4, and 6 were exposed to "minutiae suggestion" maps during the Analysis phase. In contrast, Groups 1, 3, and 5 did not receive any "minutiae suggestion" maps. Of the two types of "minutiae suggestion" maps, Groups 4 and 6 received the Quality Map and Group 2 had the Expert Consensus Map—although all three of these groups were suggested to look at the exact same minutiae. These groups received two different colored minutiae prompts: one color represented minutiae marked by 75% or more of the pre-tested experts and the other color represented minutiae marked by 50% or more of the pre-tested experts.[21] We investigated whether these groups had an increase in the number of analysts marking the minutiae that corresponded to the "suggested minutiae". No statistically significant increase was noted in any group for the 75+% minutiae. Of course, this is somewhat understandable since these were minutiae that most experts observed naturally, without prompting, during the pre-testing. However, for the 50+% minutiae, a statistically significant increase was observed in all three of the "suggested minutiae" map groups (Chi-square test for all 12 trials compared against control group, 11 d.f., $\chi^2 = 98, 43, 54$; p values = < 0.001). Group 3 did not show a significant increase in the number of 50+% minutiae. Group 5, unexplainably, did.[22] It is unknown why this group had such a significant increase in the number of 50+% minutiae for one of the twelve trials. In summary, the minutiae maps appeared to help direct the attention of the analyst to minutiae that they might not have normally noticed or annotated. *This resulted in greater consistency and conformity in their annotations in those groups.*

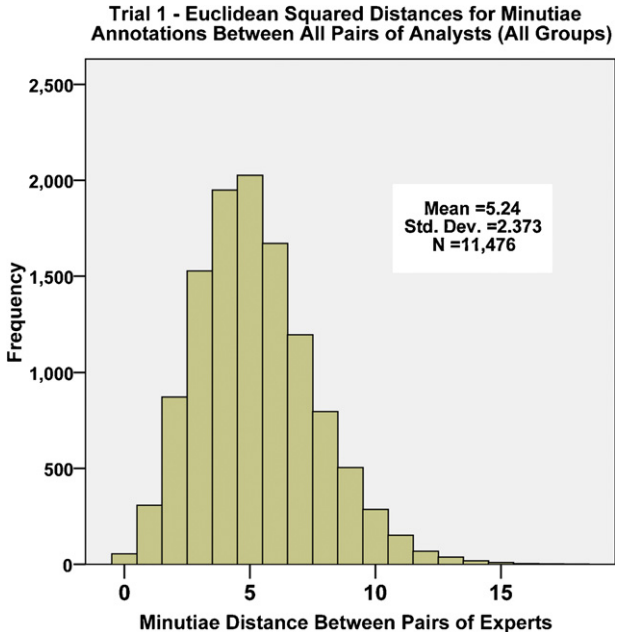

Fig. 15. Euclidean Squared Distances (ESDs) for Analysis phase minutiae for all pairwise comparisons across all experimental groups for Trial 1.

Furthermore this effect can be observed directly in the ESD ratio values when the data in Table 6 are stratified according to experimental group. Table 7 shows the ESD ratios for each group per trial. The ESD differences are significantly lower for all three groups which had a minutiae suggestion map (Groups 2, 4, and 6). A Mann–Whitney test found the distributions of ESD ratios for Groups 1, 3, and 5 to be statistically significantly different from Groups 2, 4, and 6 ($p = 0.003$). The ESD differences are slightly lower for Groups 3 and 5 compared to the control group. It is unknown why these groups would have been slightly lower, since they received the fingermark images exactly as Group 1 (control) during the Analysis phase. Perhaps because they were told they

[21] In Group 2, these colors corresponded to green (75+%) and yellow (50+%). In Groups 4 and 6, these colors corresponded to black (75+%) and white (50+%).

[22] A possible explanation comes from an outlier value. One of the Group 5 trials had one outlier value that made the p-value significant. (Chi-square stat = 27, $p = 0.004$). Over half of this statistic came from the contribution of Trial 1 where inexplicably a large number of 50+% minutiae were marked. Similar values were not seen in the other trials for this group.

Table 6
The left half of the table shows statistics for the number of minutiae marked during Analysis phase (A_{min}) for each trial (including the percentage of incorrectly marked minutiae) by 158 experts. The right half of the table shows statistics for the Euclidean Squared Distances measuring the differences in A_{min} annotations between all expert pairwise comparisons.

	A_{min}			ESD statistics						
	Mean	SD	% false marked	Mean	SD	*N*	Min	Median	Max	ESD ratio
Trial 01	12.2	2.7	3%	5.2	2.4	11,476	0	5	18	0.43
Trial 02	10.9	2.0	3%	3.6	1.8	11,628	0	3	11	0.33
Trial 03	14.1	3.8	3%	7.9	3.6	11,781	0	7	27	0.56
Trial 04	9.2	2.2	4%	3.9	2.1	11,476	0	4	12	0.43
Trial 05	8.6	2.1	2%	4.5	2.2	11,781	0	4	18	0.52
Trial 06	7.1	2.0	12%	3.8	2.1	11,935	0	4	15	0.53
Trial 07	11.6	2.1	5%	4.0	2.0	11,781	0	4	12	0.34
Trial 08	10.0	1.6	3%	3.3	2.0	11,781	0	3	17	0.33
Trial 09	10.3	3.3	7%	5.7	2.7	11,781	0	5	19	0.55
Trial 10	10.6	2.5	11%	5.3	2.3	11,935	0	5	17	0.50
Trial 11	10.4	1.8	4%	3.3	2.1	11,781	0	3	18	0.32
Trial 12	11.1	2.8	2%	4.6	2.7	11,935	0	4	18	0.42
				Mean						0.44
				SD						0.09

Table 7
Euclidean Squared Distance (ESD) ratios (ESD/mean A_{min}) for each group per trial.

	Group 1 ratio	Group 2 ratio	Group 3 ratio	Group 4 ratio	Group 5 ratio	Group 6 ratio
Trial 01	0.51	0.36	0.47	0.40	0.43	0.46
Trial 02	0.37	0.30	0.32	0.31	0.34	0.29
Trial 03	0.65	0.50	0.53	0.49	0.65	0.53
Trial 04	0.42	0.36	0.44	0.40	0.47	0.49
Trial 05	0.58	0.41	0.54	0.49	0.56	0.50
Trial 06	0.60	0.46	0.54	0.44	0.66	0.47
Trial 07	0.39	0.30	0.37	0.31	0.35	0.31
Trial 08	0.41	0.30	0.32	0.29	0.34	0.26
Trial 09	0.63	0.56	0.55	0.50	0.57	0.46
Trial 10	0.56	0.44	0.53	0.44	0.52	0.45
Trial 11	0.40	0.32	0.30	0.28	0.31	0.29
Trial 12	0.52	0.36	0.42	0.40	0.41	0.36
Mean	0.50	0.39	0.44	0.40	0.47	0.41
SD	0.10	0.09	0.10	0.08	0.12	0.10

would be viewing a new tool in the Comparison phase, this created some "conservative minutiae selection" experimental bias.

From Table 7 we can conclude that the Expert Minutiae Consensus map in Group 2 had the strongest effect of reducing analyst variation in minutiae selection/annotation. Groups 4 and 6 had comparable reductions as well. Fig. 16 provides a snapshot of the effect. The decrease in ESD ratio is shown in Fig. 16, by setting Groups 2 through 6 relative to Group 1. The groups have been reordered by increasing effect.

When the false minutiae[23] annotation error rates were stratified by group, a similar trend as that in Fig. 16 was observed. Groups 6, 4, and 2 all showed a reduction in the error rate for false minutiae annotations (see Fig. 17). The groups have been ordered from highest error rate to lowest. The order is identical to the order of the ESD ratios.

Finally, when we attempted to tie accuracy and consistency of A_{min} to decision error rates, we did not see this trend. We could not see convincing evidence that accuracy and consistency of minutiae selection in the Analysis phase necessarily translated to accuracy and consistency in the final decision reported in the Evaluation phase. This does not mean that there is no relationship, but rather, there are likely other factors besides the Analysis phase results that influence examiner judgments. In Fig. 18, the groups were ordered from highest to lowest false positive discovery rates (for reference and completeness, false negative discovery rates were superimposed on the graph). Compared to Figs. 16 and 17, the order of the groups has changed, although Group 2 still has the lowest error rate.

While we were unable to clearly illuminate the connection between the Analysis phase and the Evaluation phase, we are not discouraged. One problem connecting the Analysis phase feature selection process to the Evaluation phase decision may lie in the intermediary Comparison phase. The major function of the

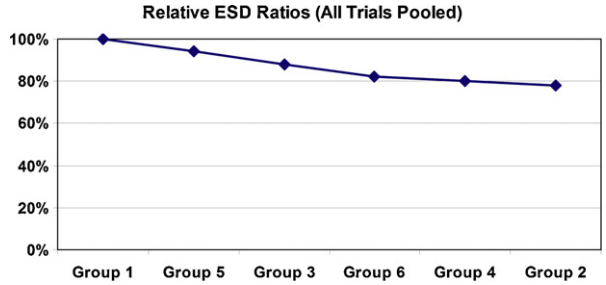

Fig. 16. Relative (to Group 1, control) ESD ratios, showing a decreasing effect for groups with a minutiae suggestion map. Here, a decreasing effect represents less differences (more consistency) between sets of Analysis phase minutiae annotations from paired experts.

[23] False minutiae annotations were defined in this study as the presence of an annotated feature where none existed according to the ground truth exemplar. These were counted manually and we used the following guideline for determining a false minutia: we gave a "one ridge leeway" in their markings (if they were within a ridge of a true feature we counted it as accurate and incipient ridges and dots were not scored as false minutiae). Accuracy of minutiae type was not considered in these determinations.

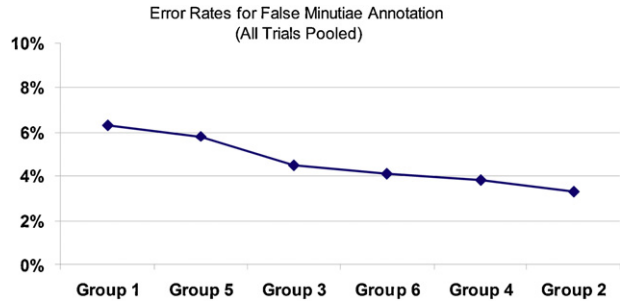

Fig. 17. The percentage of Analysis phase minutiae annotations (for all trials pooled) that reflected false minutiae (i.e. they did not exist in the ground truth exemplar).

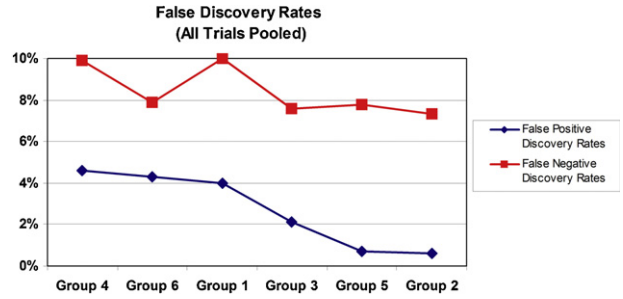

Fig. 18. False discovery error rates ordered according to descending rates and groups.

Comparison phase is to compare all of the reliable features found during the analysis of the fingermark, to the corresponding features in the print. Each set of corresponding features must be determined to be in "agreement" if they are from the same source. An analyst decides if features are "not in agreement" when the observable differences between compared features are so great, that the only reasonable conclusion is that they are not from the same source. Suppose that two analysts both selected the same eight features during the analysis of a mark. Upon comparing to the exemplar print, the first analyst is concerned because one of the eight features looks significantly different and is located in a slightly different position in the exemplar print. This analyst places significant weight on these differences and renders an "exclusion" decision. The second analyst, using the same eight features as the first analyst, places little weight on these apparent differences, citing that distortion effects in the mark can reasonably cause such apparent differences. This analyst is satisfied there are no "discrepancies" and renders an "identification" decision. The erroneous decision by one of the analysts in this scenario had little to do with the reliability of the features selected during the Analysis phase. The issue was the weight that was assigned to the perceived differences and tolerances (the range of maximum acceptable differences) for those features. Generally, tolerances are derived during the feature selection process since the clarity of the feature tends to set the tolerance. But after viewing the exemplar, analysts can be fluid in their assignment of tolerance. This has led to concerns of circular reasoning and bias and is the subject of some debate [3,46,47].

A second problem connecting the Analysis phase to the Evaluation phase may lie in the decision criteria and not the feature selection task. Again, let us assume that two analysts selected the same reliable eight minutiae, and they have the same level of confidence in these eight features. The first analyst reports an "identification". The second analyst declares this trial "inconclusive", stating that there are many features in agreement, but insufficient to satisfy his personal threshold. This could be a possible outcome when different personal weights are assigned to the evidence. In other words, they are placing different discriminating weights on the same features. A second possibility, not mutually exclusive of the first, is that they could have different decision thresholds. The first analyst needs 8 generic minutiae in agreement to reach the lower thresholds of his "identification" decision and the second analyst needs 10 generic minutiae in agreement to satisfy his thresholds. Currently, there are no studies that test the experts' calibration of their personal weights against some externally verifiable measure (such as a likelihood ratio).

In the present study, we attempted to calibrate the expert judgment using the LR tool in Group 3 and Group 6. The results did not show convincingly that analysts were incorporating this information into the current ACE-V practice (namely one that requires an "all-or-nothing" decision process, i.e. "identification" or "exclusion"). While Fig. 8 showed that there was higher reproducibility of decisions in Groups 3 and 6, compared to Groups 1 and 4 (the difference between these respectively paired groups being the addition of the LR tool), error rates for Groups 3 and 6 did not appear to be significantly impacted. Comments by participants using the tool ranged drastically from finding the tool useful to finding the tool completely unhelpful. It may be that presently, many fingerprint examiners are unable at this time to process the significance of statistical measurements. Additional training or familiarity with such tools may lead to better assimilation of LR information.

4. Limitations

A few limitations of the experimental design should be considered when interpreting the reported data. As with any study, we must be cautious when taking a specifically designed and constructed situation and making inferences and generalizations to actual laboratory case-working practices. The experimental design in this study focused only on one type of limited fingerprint examination, namely an on-screen, 1:1 comparison (as opposed to 1:*N* comparison). Participants were using tools that many may have never seen before or were uncomfortable using. Participants were free to use their own equipment, work in their most comfortable comparison environment, and work at their own pace; however, equipment quality (e.g. screen resolution, internet speed, etc.), environment, and frequency of interruptions could vary from user to user. Participants accustomed to viewing fingermarks only in "to scale" photographs or lifts and through a 2× to 5× magnifier, may have been at a serious disadvantage.

A larger experimental concern is the awareness that the participant was being tested. This is unavoidable of course since we tested tools that were not a part of normal case procedure. To this point, participants (as evidenced in many of their comments) attempted to avoid using some of the tools out of concerns that it would "bias" their decisions. Of course, this was the point of the study: to see not how the tools would "bias" them, but rather, as the title implies, how the tools would "inform" their judgments (for better or worse). We took efforts to design the platform so that participants were forcibly exposed to the tools, but it is still possible that some experts avoided significant interaction with the tools.

A second limitation of the study is to recognize that this was entirely an ACE process and no verification step was included. Many, but not all, of the errors would have been caught had a verification step been employed for each decision. Many of the erroneous decisions were repeated independently by multiple examiners in the same trial. Even if a blind verification strategy had been employed, there was still a realistic chance in some trials of drawing a verifier that would have made the same error as the initial analyst. Given the variance in Verification procedures that an agency may choose to employ, we did not estimate error rates if

a Verification step was employed. However, we believe that the error rates presented here could generally be reduced with the inclusion of a rigorous quality Verification step.

5. Conclusions

The major conclusions drawn from the data presented in this paper are the following:

1. Diagnostic testing statistics such as error rates, false discovery rates, sensitivity, and specificity were calculated for all participants ($N = 176$) and stratified according to group, experience, etc. Groups 2 and 5 (exposed to Expert Consensus tools) had the lowest error rates compared to the control (Group 1) and the other groups. Trained experts ($N = 158$) had lower false positive error rates, but higher false negative error rates, than novice trainees ($N = 13$). Overall, and keeping in mind the difficulty of the cases, the false positive and false negative error rates were 2.6% and 5.7%, respectively.
2. The quantity and quality of the available ridge detail in the fingermark was clearly driving the mean number and variance of minutiae selected. The Analysis phase tools (Expert Consensus Minutiae Map or a Quality Map) did not impact the overall number of minutiae selected during the Analysis phase.
3. While the tools did not appear to affect the *number of minutiae* selected during the Analysis phase, they did appear to affect *which minutiae* were selected by experts and the *consistency* with which they selected certain minutiae. When experts used either the Expert Consensus Minutiae Map, a tool that displayed minutiae selected by a consensus of experts (Group 2), or the Quality Map, a tool which highlighted minutiae in reliable, clear regions of the fingermark (Groups 4 and 6), these experts had a higher *accuracy* of correct minutiae selection and a higher *consistency* among experts with respect to which particular minutiae they selected compared to experts that did not use these tools. Therefore, a major benefit to these tools is that there is greater consensus and accuracy during the Analysis phase when selecting features.
4. Experts that analyzed and compared a higher number of minutiae (approximately 9 or more minutiae) were more likely to report a positive source attribution (when the images were truly coming from the same source). An operational decision threshold for "identification" decisions was observed at approximately 8–9 minutiae, even if experts were not trained to a specific numerical standard to effect an identification.
5. We cannot say if higher consistency and accuracy of feature selection during the Analysis phase directly contributes to higher accuracy and consistency in reported Evaluation phase decisions. Group 2 did exhibit higher consistency and accuracy of feature selection *and also* exhibited the lowest attribution decision error rates; however Groups 4 and 6 which did have higher consistency and accuracy during the Analysis phase did not have lower decision error rates (relative to the control group). Similarly, Group 5 exhibited low error rates and high efficiency in decision making (high sensitivity and specificity) but did not exhibit statistically significant improvement in feature selection. These data do not show that improving feature selection processes will necessarily lead to improved decision making. However, this does not mean that accurate and consistent feature selection has no value in the examination process; it does. Accurate and consistent feature selection speaks to the reliability of the features that are being used.

The major advantage of the tools tested in this experiment was the demonstration that the Quality Map and Expert Consensus Map increased the accuracy and consistency in feature selection. While not leading directly to increased accuracy and consistency in the reported decisions, this effect could lead to reliable feature sets as a product of the Analysis phase.

A set of reliable features selected during the Analysis phase has usefulness. Features selected by a consensus of experts will, by definition, show a reduction of variability and noise in feature selection, when only the features that are selected by a majority of experts are used in the feature set. This subset of "consensus features" would then form the basis of the decision. These features could be used in conjunction with a statistical model, or in the absence of such a model, experts would make a subjective assessment of the weight of the evidence, based solely on these consensus features. A process very similar to this, called the "Questionable Identification procedure" is described by analysts in the Netherlands [48]. The procedures outlined by this "Questionable Identification procedure" (without the use of any statistical tools) echoes of similar clinical diagnostic approaches used in complex medical diagnosis or treatment cases [49,50].

By identifying tools that will increase accuracy and consistency in feature selection, we have identified methods for selecting reliable feature sets. Once methods for feature weighting and decision thresholds can be implemented, a more reliable and transparent process can be recommended.

Acknowledgments

The authors would like to acknowledge the following people: the anonymous reviewers who gave their time to review this paper and thankfully caught several errors; Jordan Jones, University of Minnesota, the student intern that had the unpleasant and daunting task of data entry and running statistical test macros *ad nauseum*; the pilot test group of experts. Their views, time, and patience were necessary for a successful project; BCA management and my fellow co-workers and the knowledgeable and helpful staff at UNIL; the MFRC for their financial support; and most importantly, the hundreds of participants (including those who tried so hard to make it work!) for their commitment and participation. In particular, we would like to thank those supervisors that gave their examiners the time and support to participate.

Funding: This study was funded by a grant from the Midwest Forensic Resource Center (MFRC), Iowa State, Ames, Iowa (SC-10-339).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.forsciint.2011.12.017.

References

[1] Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST), Friction Ridge Examination Methodology for Latent Print Examiners, Ver. 1.01. http://www.swgfast.org/Friction_Ridge_Examination_Methodology_for_Latent_Print_Examiners_1.01.pdf (accessed 29.05.11).
[2] D.R. Ashbaugh, Qualitative-Quantitative Friction Ridge Analysis—An Introduction to Basic and Advanced Ridgeology, CRC Press, Boca Raton, 1999.
[3] L. Haber, R.N. Haber, Challenges to Fingerprints, Lawyers & Judges Publishing Company, Inc., Tucson, Arizona, 2009.
[4] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, DC, 2009.
[5] J.R. Vanderkolk, Forensic individualization of images using quality and quantity of information, Journal of Forensic Identification 49 (1999) 246–256.
[6] C. Champod, Fingerprint examination: towards more transparency, Law Probability and Risk 7 (2008) 111–118.
[7] The Fingerprint Inquiry Scotland. www.thefingerprintinquiryscotland.org.uk, available 20 July 2009.
[8] M.J. Saks, Forensic identification: from a faith-based "science" to a scientific science, Forensic Science International 201 (2010) 14–17.

[9] J.L. Mnookin, The courts, the NAS, and the future of forensic science, Brooklyn Law Review 75 (2010) 1–67.
[10] G. Langenburg, A method performance pilot study: testing the accuracy, precision, repeatability, reproducibility, and biasability of the ACE-V process, Journal of Forensic Identification 59 (2009) 219–257.
[11] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proceedings of the National Academy of Sciences 108 (2011) 7733–7738.
[12] C. Champod, Friction ridge examination (fingerprints): interpretation of, in: A. Moenssens, A. Jamieson (Eds.), Wiley Encyclopedia of Forensic Sciences, vol. 3, John Wiley & Sons, Chichester, UK, 2009, pp. 1277–1282.
[13] I.W. Evett, G. Jackson, J.A. Lambert, S. McCrossan, The impact of the principles of evidence interpretation on the structure and content of statements, Science and Justice 40 (2000) 233–239.
[14] L.M. Al-Haddad, C. Neumann, Benefits and challenges of the use of fingerprint statistical models in casework, Science and Justice 50 (2010) 32–33.
[15] N.M. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems—modelling within finger variability, Forensic Science International 167 (2007) 189–195.
[16] E. Tabassi, C.L. Wilson, C.I. Watson, Fingerprint Image Quality, NIST NISTIR 7151, August 2004.
[17] E. Tabassi, C.L. Wilson, A novel approach to fingerprint image quality, in: IEEE International Conference on Image Processing (ICIP-05), 2005, 37–40.
[18] N.B. Nill, Image Quality of Fingerprint (IQF) Software Application. http://www.mitre.org/tech/mtf (accessed 28.04.11).
[19] A. Hicklin, et al., Latent fingerprint quality: a survey of examiners, Journal of Forensic Identification 61 (2011) 385–418.
[20] C. Neumann, C. Champod, R. Puch-Solis, D. Meuwly, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of three minutiae, Journal of Forensic Sciences 51 (2006) 1255–1266.
[21] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, Journal of Forensic Sciences 52 (2007) 54–64.
[22] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, Journal of the Royal Statistical Society 175 (2012) 1–26, Part 2.
[23] N. Egli, Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System, Université de Lausanne, Ecole des Sciences Criminelles/Institut de Police Scientifique Lausanne, Suisse, 2009.
[24] Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST), Standard Terminology of Friction Ridge Examination. http://www.swgfast.org/Documents.html (accessed 15.10.11).
[25] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.
[26] J.J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter, Hasting Law Journal 59 (2008) 1077–1098.
[27] J.J. Koehler, When do courts think base rate statistics are relevant, Jurimetrics Journal 42 (2002) 373–402.
[28] National Research Council—Committee on DNA Technology in Forensic Science, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996.
[29] B. Budowle, M.C. Bottrell, S.G. Bunch, R. Fram, D. Harrison, S. Meagher, C.T. Oien, P.E. Peterson, D.P. Seiger, M.B. Smith, M.A. Smrz, G.L. Soltis, R.B. Stacey, A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement, Journal of Forensic Sciences 54 (2009) 798–809.
[30] W.C Thompson, F. Taroni, C.G. Aitken, How the probability of a false positive affects the value of DNA evidence, Journal of Forensic Sciences 48 (2003) 47–54.
[31] State of Maryland V. Bryan Keith Rose, The Circuit Court for the Baltimore County, Memorandum decision of Judge Susan Souder, K06-0545.
[32] J.L. Mnookin, Of black boxes, instruments, and experts: testing the validity of forensic science, Episteme: A Journal of Social Epistemology 5 (2008) 343–358.
[33] J.I Thornton, The one-dissimilarity doctrine in fingerprint identification, International Criminal Police Review 32 (1977) 89–95.
[34] A. Biedermann, S. Bozza, F. Taroni, Decision theoretic properties of forensic identification: underlying logic and argumentative implications, Forensic Science International 177 (2008) 120–132.
[35] S.A. Cole, More than zero: accounting for error in latent fingerprint identification, The Journal of Criminal Law and Criminology 95 (2005) 985–1078.
[36] G. Norman, K. Eva, L. Brooks, S. Hamstra, Expertise in medicine and surgery, in: K.A. Ericsson, N. Charness, P. Feltovich, R. Hoffman (Eds.), The Cambridge Handbook of Expertise and Expert Performance, Cambridge University Press, New York, 2006, pp. 339–353.
[37] K.A. Ericsson, The influence of experience and deliberate practice on the development of superior expert performance, in: K.A. Ericsson, N. Charness, P. Feltovich, R. Hoffman (Eds.), The Cambridge Handbook of Expertise and Expert Performance, Cambridge University Press, New York, 2006.
[38] K.A. Ericsson, Development of Professional Expertise, Cambridge University Press, New York, 2009.
[39] W.R. Stephenson, A.G. Froelich, W.M. Duckworth, Using resampling to compare two proportions, Teaching Statistics 32 (2010) 66–71.
[40] C. Champod, P.A. Margot, Computer assisted analysis of minutiae occurrences on fingerprints, in: J. Almog, E. Springer (Eds.), Proceedings of the International Symposium on Fingerprint Detection and Identification, Ne'urim, Israel, June 26–30, 1995, Israel National Police, 1996, pp. 305–318.
[41] S. Pankanti, S. Prabhakar, A.K. Jain, On the individuality of fingerprints, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 1010–1025.
[42] G. Langenburg, Pilot study: a statistical analysis of the ACE-V methodology—analysis stage, Journal of Forensic Identification 54 (2004) 64–79.
[43] I.W. Evett, R.L. Williams, A review of the sixteen points fingerprint standard in England and Wales, in: Proceedings of the International Symposium on Fingerprint Detection and Identification, Ne'urim, Israel, (1996), pp. 287–304.
[44] I.E. Dror, C. Champod, G. Langenburg, D. Charlton, H. Hunt, R. Rosenthal, Cognitive issues in fingerprint analysis: inter- and intra-expert consistency and the effect of a 'target' comparison, Forensic Science International 208 (2011) 10–17.
[45] B. Schiffer, C. Champod, The potential (negative) influence of observational biases at the analysis stage of fingermark individualisation, Forensic Science International 167 (2007) 116–120.
[46] J.R. Vanderkolk, ACE+V: a model, Journal of Forensic Identification 54 (2004) 45–51.
[47] R.B. Stacey, A report on the erroneous fingerprint individualization in the Madrid train bombing case, Journal of Forensic Identification 54 (2004) 706–718.
[48] Interpol, Interpol European Expert Group on Fingerprint Identification II—IEEGFI II, Interpol, Lyon, 2004.
[49] M. Banning, A review of clinical decision making: models and current research, Journal of Clinical Nursing 17 (2008) 187–195.
[50] C. Thompson, D. McCaughan, N. Cullum, T. Sheldon, A. Mulhall, D. Thompson, Research information in nurses' clinical decision-making: what is useful? Journal of Advanced Nursing 36 (2001) 376–388.