Technical Note

Analysis, Comparison and Evaluation of Latent Prints: The Results of the 2021 Collaborative Exercise of the ENFSI Fingerprint Working Group

Ido Hefetz [1 4]
Shimon Kimchi [1 4]
Francesco Zampa [2 4]
Aldo Mattei [3 4]

Abstract: In 2021, the Fingerprint Working Group (FIN-WG) of the European Network of Forensic Science Institutes (ENFSI) conducted a collaborative exercise (CE) on the analysis, comparison, and evaluation (ACE) of friction ridge marks. The test proved to be quite challenging, as demonstrated by the high False Negative Rate (FNR = 13.1%). A new category of error was unexpectedly discovered. Two laboratories, while correctly identifying a donor, marked up non-corresponding feature sets to support the identifications. As a result, two false positive rates were calculated, one for source (FPR-1) and one for feature set (FPR-2). FPR-1 was 1.1%, while FPR-2 was 0.7% in this exercise. Additionally, the impact of laboratory approaches to the concept of simultaneity when applied to the lack of continuity of the ridge flow was shown to impact results. Finally, there is also a discussion of the meaning of exclusion given different laboratory approaches.

[1] Fingerprint Database Lab, Israel Police – Jerusalem – Israel
[2] Reparto Carabinieri Investigazioni Scientifiche (R.I.S.) – Parma – Italy
[3] Reparto Carabinieri Investigazioni Scientifiche (R.I.S.) – Messina - Italy
[4] On behalf of the ENFSI Fingerprint Working Group (FIN-WG)

Introduction

Between 2004 and 2015 five collaborative tests of fingerprint comparison were conducted within the Fingerprint Working Group (FIN-WG) of the European Network of Forensic Science Institutes (ENFSI) (formerly EFP-WG) and a summary of these tests have been published previously [1, 2]. In 2016, a permanent advisory group was established with the task of managing an on-going testing program. The goal of the advisory group is to ensure continuity and plan testing in a way that meets the strategic objectives of ENFSI. Moreover, two main reports, the 2009 NAS, "Strengthening Forensic Science in the United States: A Path Forward", and the 2016 PCAST, "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods", recommended the urgent need for strengthening the scientific basis of forensic disciplines, particularly in the form of performance testing in pattern evidence [3, 4]. In general, proficiency tests are a tool to evaluate and compare the performance of laboratories and examiners. Collaborative exercises further enhance understanding of forensic processes by illuminating differences in decision-making.

The purpose of this article is to provide an overview of the results of a collaborative exercise carried out in 2021 that focused on analysis-comparison-evaluation (ACE) of friction ridge marks. This review will shed light on differences in the establishment of sufficiency for comparison thresholds and to describe the distribution of decisions made by the participant laboratories for the same set of marks. This data may assist in the development of protocols that could increase consistency in decision-making or reduce errors.

Materials and Methods

Fingerprint Database Lab of the Israeli Police volunteered for the selection of the marks. The advisory group supported the activities in the pre-testing phase, in the collection and analysis of the results, and in reviewing the final report. The test was provided at no cost to the participants.

Ten marks (five finger marks and five palm marks) and four sets of reference prints (donors #A through #D, with complete tenprint and palmprint cards) were made available for registered participants to download from a web-based platform. Both marks and reference prints were given in TIFF format at 1000 ppi resolution. The quality of the marks as well as of the reference prints was determined to be representative of the average quality in casework per the experience of the fingerprint examiners involved in the preparation of the test.

The Israeli Police selected all marks and reference prints from a ′True Source Database′ created for the collaborative exercise. All marks in this database were made by Israeli Police employees simulating different types of activity (donors gave permission for use of their friction ridge impressions on an international level). Marks were deposited on several surfaces such as plastic, metal, tape, glass and paper and were visualised using common methods. Like casework circumstances, the marks were provided in variable orientations to the participants. All images included a scale or background that indicated the size of the marks. Figure 1 displays two of the finger marks and two of the palm marks used in the exercise.

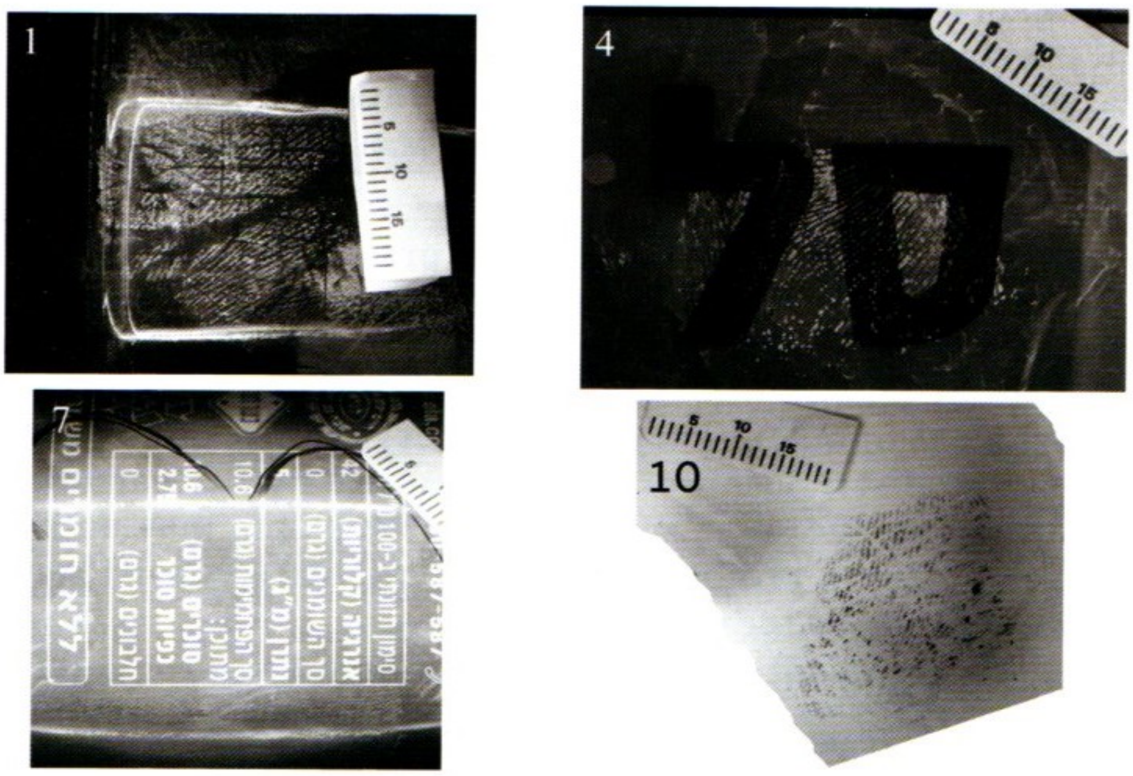

Figure 1

Examples of marks (palm marks #1 and #10; fingermarks #4 and #7) used for the exercise.

The participants were asked to analyze the marks and determine if each mark was suitable for comparison. For those marks determined to be suitable for comparison, the participants compared the reference prints and were asked to provide an opinion using one of the following terms: match, no match, or an inconclusive decision. The authors acknowledge the use of terms "match" and "no match" are debated within the literature and may not have been used in a formally correct way within this study. Given this was the direction provided to the participants for the collaborative exercise, these terms will be used in this paper to discuss participant answers.

Participants were instructed to follow their standard operating procedures while performing the examination of the marks and reference prints. The analysis was conducted without predetermined constraints, allowing for a natural initiation of the documentation habits regarding the marks. In terms of the verification step, which serves as a crucial quality control measure, it can be considered implemented based on current indications provided by the scientific literature [5, 6]. However, the organisers do not know if participants followed through with the verification step, nor the type of verification applied. It is important to note that only laboratory answers were accepted, and not individual practitioners.

The test also allowed for a probabilistic evaluation of the evidential value of the marks, using either a verbal scale, or a numerical value of the likelihood ratio, or both. With this regard, the participants were required to use their standard operating procedure, explaining the method they used to derive the probabilistic assessment.

In case of a match, the participants were requested to attach the side-by-side comparison chart. In case of no match or "not of value for comparison" outcomes, the participants were requested to attach the image of the friction ridge mark with the minutiae mark-up, thus disclosing their analysis. These instructions applied also for the inconclusive decisions.

Finally, as per the indications provided to the participants, a "no match" outcome was only to be used when no significant corresponding features were located in any of the four sets of reference prints provided. It should be noted that there is considerable confusion and variation regarding exclusion opinions in the friction ridge discipline. It is unknown how the various participants in this study approached the exclusion opinion. In general, there are two basic approaches:

- Exclusion Approach #1 – Person who provided reference prints is not the source of the mark. The examiner has determined that all necessary reference prints have been provided to exclude the person from having deposited the mark. If the examiner does not have the needed profile regions of the friction ridge skin from the donor, the result will be inconclusive, and the agency will request the needed reference prints to complete the comparisons and provide a source opinion regarding the person.

- Exclusion Approach #2 – No corresponding features could be found in the submitted reference prints, but the person who provided the reference prints may still be the source of the mark. In practice, it is unknown if agencies request the additional reference prints that would be needed to issue a source opinion regarding the person.

As a result of this variation in practices, the same mark can produce exclusion decisions with two very different meanings. Hopefully, the consumers of a particular laboratory's forensic results clearly understand when a person has been excluded versus when the person may still be the donor of a mark. Lack of understanding the differences in these two approaches to exclusions and inconclusive results could have dire consequences for forensic investigations and trial outcomes.

Results

Sixty-two (62) agencies initially agreed to participate in the collaborative exercise, with fifty-two (52) returning results (84% response rate). Limited data regarding the operational environments of the laboratories was collected and is presented below.

Standards for Sufficiency & Accreditation

Participants were asked if the laboratory followed a numeric, holistic, mixed, or probabilistic approach when providing an opinion of same source (or support for same source). As shown in Figure 2, most participating laboratories reported using a numerical standard (requirement for a minimum number of minutiae). A "mixed" approach was indicated when a laboratory deployed different standards based on the specific circumstances of the case.

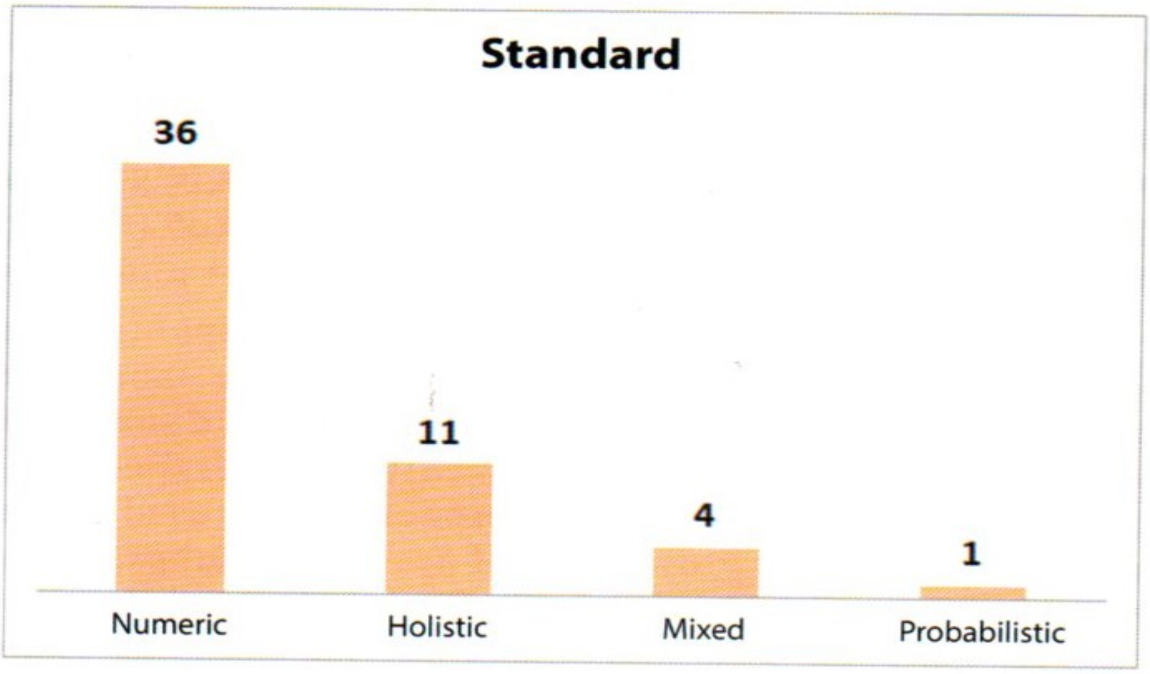

Figure 2

Overview of the standard used for identification (number of laboratories).

For the thirty-six laboratories operating with a numerical standard for providing an opinion of same source (or support for same source), there were three main thresholds (10, 12, and 16 points) reported. Figure 3 shows the number of laboratories using each reported threshold.

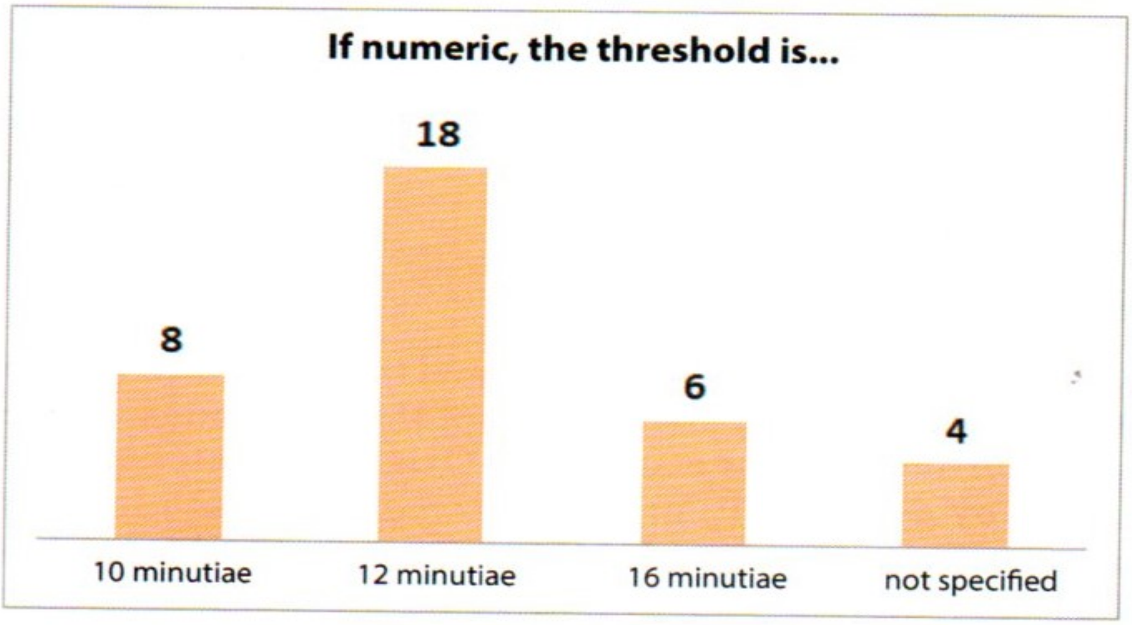

Figure 3

Thresholds in the numerical standard (number of laboratories).

It is important to note that some laboratories reporting the use of a numerical standard permitted lowering the specific threshold under certain circumstances. The main factors influencing this decision was the presence of other features (e.g., creases or scars), or certain attributes of the ridges (e.g., reliable edge shapes), or certain attributes of minutiae (e.g., combinations of minutiae exhibiting high specificity).

Figure 4 provides an overview of the participants regarding ISO/IEC 17025 accreditation. Thirty-three laboratories were accredited, while eighteen were not (one did not provide accreditation status). From the accreditation perspective, no preference was noted regarding approaches to providing same source opinions, with a prevalence of accredited laboratories in this exercise reporting a numerical standard.

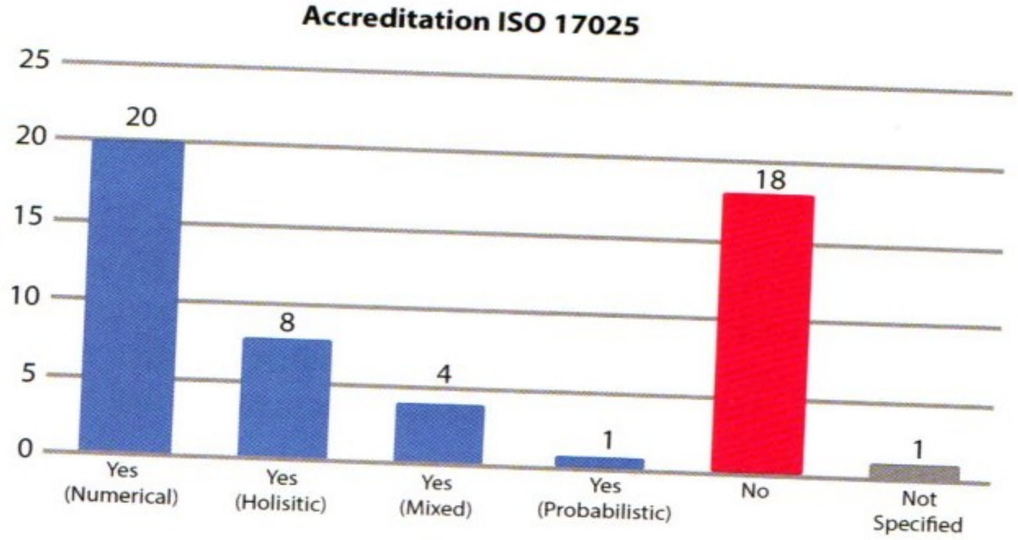

Figure 4

Accreditation ISO/IEC 17025 for fingerprint comparison (number of laboratories) based on the standard used for fingerprint comparison.

Overview of the Responses

The exercise consisted of eight mated trials and two non-mated trials. Table 1 shows the distribution of the 52 responses received from the participants for each mark. Origin indicates if the mark was made by a finger (F) or palm (P). Ground truth indicates if the comparison was a mated (M) or non-mated (NM) trial. Mark #4 and Mark #5 were the non-mated trials. As highlighted in the blue text in Table 1, the majority comparison opinion for each mark coincided with ground truth. The "match" data for Mark #6 and Mark #8 in Table 1 are noted with an asterisk (*) because each of these marks resulted in one instance where the wrong feature set within correct donor was "matched". Further discussion of this issue is provided below. Analysis and comparison results based on the ground truth status of the trial are summarized in Table 2. The "match" data in Table 2 is noted with an asterisk (*) because it includes the two erroneous matches of the correct donor (as previously indicated for Mark #6 and Mark #8).

Mark	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Origin	P	F	F	F	P	P	F	F	P	F
Ground Truth	M	M	M	NM	NM	M	M	M	M	M
Opinions Provided										
No Value	0	6	10	0	5	0	0	1	15	5
Match	34	40	37	0	1	49*	51	44*	26	30
No Match	16	2	1	50	38	3	1	7	1	16
Inconclusive	2	4	4	2	8	0	0	0	10	1

Table 1

*Distribution of responses for each mark. Origin indicates if the mark was a made by a finger (F) or palm (P). Ground truth indicates if the comparison was a mated (M) or non-mated (NM) trial. Blue text indicates majority opinion. * indicates "match" data that includes an erroneous match of the correct donor.*

Opinion	Mated	Non-Mated	Total
No Value	37	5	42
Match	311*	1	312
No Match	47	88	135
Inconclusive	21	10	31
Total	416	104	520

Table 2

*Counts of opinions reached for each mark given the ground truth. * indicates "match" data that includes two erroneous matches of the correct donor.*

The percentage of inconclusive responses for different source trials (9.6%) was nearly double the percentage of inconclusive responses for same source trials (5%). These results are similar to data reported in a black-box study on palmar friction ridge comparisons [7] but differ from studies evaluating examiner performance on fingermarks (distal phalange only) by Ulery *et al.* [8] and Langenburg *et al.* [9]. It should be noted that half of the ”inconclusive” results reported for same source trials in this collaborative exercise were related to palm marks. It is unknown if the participating laboratories spend equal time training with impressions from all regions of the friction ridge skin (distal phalanges, lower phalanges, palms, and feet).

Analysis of the Responses

Comparison responses were provided on 358 mated trials and 89 non-mated trials for a total of 447 comparison opinions (no value and inconclusive responses omitted). The calculations of a false positive rate and positive predictive value from this collaborative exercise are complicated due to the design of the exercise. These complications also occurred in a similarly designed performance study by researchers from the Miami-Dade Police Department which was funded by the U.S. Department of Justice and summarized in a 2014 Technical Report [10, 11, 12, 13].

In performance studies previous mentioned by Eldridge *et al.* [7], Ulery *et al.* [8], and Langenburg *et al.* [9], each mated or non-mated trial was a one-to-one comparison of a mark to a single reference print. In each of these performance studies, the false positive rate represents the percentage of non-mated pairs that resulted in an identification opinion. The positive predictive value in each study represents the percentage of identifications that were correct.

More broadly, however, the false positive rate can be viewed as the number of times an error is made divided by the number of opportunities to commit the error. Because four subjects were compared in each comparison trial in this collaborative exercise, each trial provided an opportunity to commit a false positive. Like the Miami Dade Study published a decade ago, erroneous "matches" occurred in both the mated and non-mated trials in this collaborative exercise. As a result, the false positive rate will be presented in two ways. The first false positive rate (FPR-1) is the percentage of ***erroneous person matches*** that occurred in the 89 non-mated trials. To contribute to this false positive rate, a person had to be erroneously matched to a mark. The second false positive rate (FPR-2) is the percentage of ***erroneous feature set matches*** that occurred in the 447 trials. Similarly, the positive predictive value will also be presented with two different calculations.

Tables 3 and 4 provide the raw data and associated performance metrics given the following calculations and considerations:

- False Positive Rate 1 (FPR-1) – The percentage of non-mated trials that resulted in an incorrect person match opinion. FPR-1 essentially answers the question: Given the mark is not from one of the four donors provided, what is the chance of a "match" response in this exercise?
- False Positive Rate 2 (FPR-2) – The percentage of all trials that resulted in an incorrect feature set being "matched". FPR-2 essentially answers the question: Given a comparison of four donors, what is the chance that an incorrect feature set in the reference prints will be matched?
- False Negative Rate (FNR) – The percentage of mated comparisons that resulted in a "no match" response. FNR essentially answers the question: Given the mark is from one of the sources provided, what is the chance of a "no match" response in this exercise?
- Positive Predictive Value 1 (PPV-1) – The percentage of match opinions where the correct ***person*** was matched to a mark. PPV-1 essentially answers the question: Given a "match" was reported in the exercise, what is the chance the impressions were from the same person?

- Positive predictive Value 2 (PPV-2) – The percentage of match opinions where the correct ***feature set*** was matched to a mark. PPV-2 essentially answers the question: Given a "match" was reported in the exercise, what is the chance the impressions are recordings of the same feature set in the skin?
- Negative Predictive Value (NPV) – The percentage of "no match" opinions that were correctly reported as no match. NPV essentially answers the question: Given a "no match" was reported in the exercise, what is the chance the mark was not made by the donors of the four reference prints?

Person Opinion	Mated Person	Non-Mated Person
Match	311	1
No Match	47	88
Feature Set Opinion	**Mated Features**	**Non-Mated Features**
Match	309	3
No Match	47	88

Table 3

Confusion matrix with raw data for person opinions and feature set opinions. Opinions are the responses by the participants (match or no match), while mated and non-mated are the ground truth answers for the trials.

Performance Metric	Formula	Value
FPR-1 (person)	1 / 89 * 100	1.1%
FPR-2 (feature set)	3 / 447 * 100	0.7%
FNR	47 / (311+47) * 100	13.1%
PPV-1 (person)	311 / (311+1) * 100	99.7%
PPV-2 (feature set)	309 / (309+3) * 100	99.0%
NPV	88 / (88+47) * 100	65.2%

Table 4

False positive rates (person and feature set), false negative rate, positive predictive values (person and feature set), and negative predictive value for the collaborative exercise.

The choice to omit the "not of value" and inconclusive decisions is somewhat contested but is not uncommon in performance testing research [7]. Additionally, inconclusive results constituted a small percentage (14%) of the total number of comparisons in this exercise. Given the debate surrounding inconclusive decisions [14 - 18], an analysis of method conformance [18] with respect to the expected results [19] was conducted using the feature documentation provided by the participants. Two main assumptions were made during this review process:

Assumption #1: Based on the consensus reached by the fingerprint examiners involved in the preparation of the test (four examiners from two different institutes with varying levels of experience), all the marks should have been considered of value for a comparison. Therefore, the "not of value" outcomes were deemed inappropriate. The authors acknowledge that this decision relied on human judgment regarding the perceivable features of the evidence. Although this approach is criticized by Biedermann *et al.* [16], it provides a practical way to evaluate method conformance based on the value of the feature set observed within the marks.

Assumption #2: The appropriateness of each "inconclusive" result was evaluated in relation to the ground truth, the consensus among the participants, the reported identification standard, and participant feature documentation. In some cases, an inconclusive was considered to conform to the expected outcome. For instance, an appropriate outcome can be inconclusive on a non-mated trial if the participant is reasonably concerned about the completeness of the reference prints. In other cases, inconclusive was considered non-conforming (e.g., in the case of a mated comparison where the participant failed to locate the match).

Given these premises, greatest conformance was obtained with fingermark #7 (98% in conformance) and with fingermark #4 (96% in conformance). Worst conformance occurred with palm mark #9 (50% of the answers were marked as non-conforming), palm mark #10 (42% non-conforming), and palm mark #1 (35% non-conforming). Figure 5 shows the conformance ratings for each mark across all 52 participants.

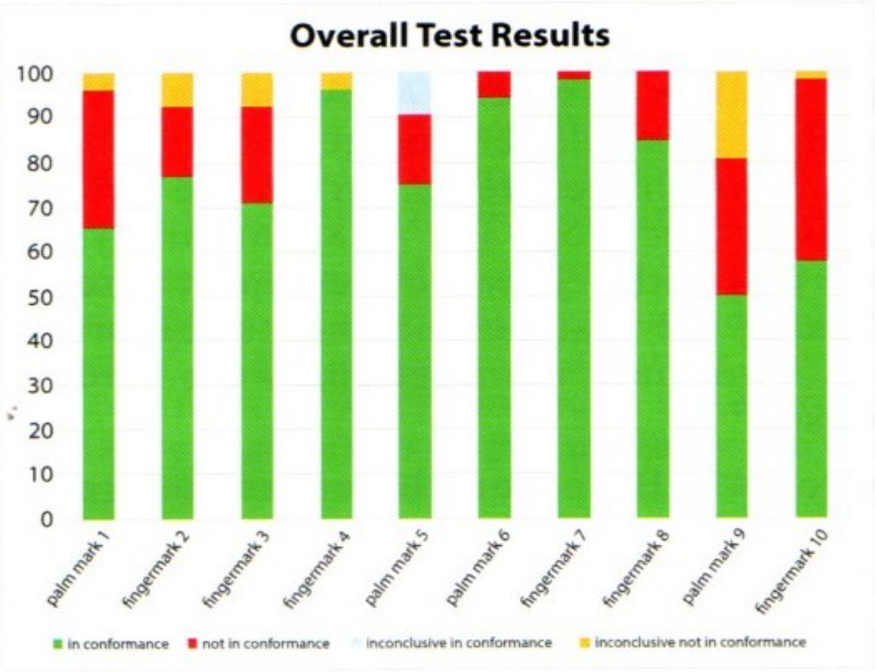

Figure 5

Overview of the conformance of the answers for each mark based on the expected results.

As for Mark #6 and Mark #8, erroneous feature set matches were included with the non-conforming responses. As previously discussed, for each of these marks, there was one instance where the correct donor was listed as a match, but the wrong region of the donor's friction ridge profile was matched to the mark (as demonstrated in the feature mark-up of the comparison by the participants).

The analysis of the data demonstrates that the comparison of some of the marks posed more difficulties than others and on average the test has been proven to be quite challenging. As previously reported by Wertheim *et al.*, determining the level of difficulty of the marks was subjective and can be related to the clarity of the mark [20]. Performance also seemed to be impacted by challenges related to the ability of the participants to recognize the search diagnosticity of the feature set and assess the completeness of the reference prints [7, 21, 22]. Similar to Eldridge *et al.'s* results [7], the data suggest that there are differences between finger and palm comparisons, both in examiners' risk tolerance for reaching definitive conclusions and their accuracy in reaching exclusion decisions. Of note, the performance study by Eldridge *et al.* [7] demonstrated higher error rates for palm marks than the performance studies by Ulery *et al.* [8]. To date, no performance studies have addressed impressions of the lower phalanges or feet.

Discussion

Details regarding the exercise data were summarized in the previous section. In the following section, key points will be highlighted to illustrate complex issues within the friction ridge community and, hopefully, trigger additional discussion and debate. While the latent marks and donor reference prints were chosen to be representative of casework, the samples were also selected to provide challenging comparisons to the participants. In general, no significant correlation was observed among the performance of the participating laboratories regarding their accreditation status as shown in Figure 6. It is interesting to note that laboratories using a numerical standard for identification (regardless of accreditation status) made half the average number of errors compared to laboratories using a holistic or mixed approach.

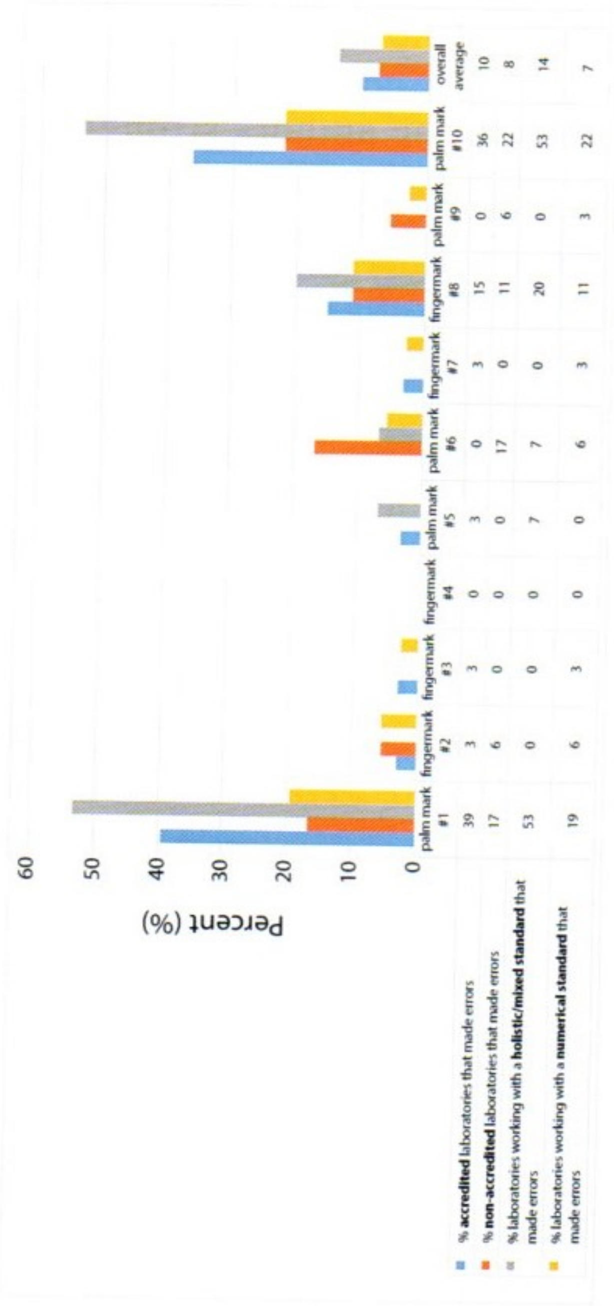

Figure 6

Distribution of the errors in terms of accreditation, and the varying SOP constraints (no numerical standard versus numerical standard).

The calculated error rates of this collaborative test show a significantly higher false negative rate compared to the false positive rate. It is important to note that this exercise only included two non-mated trials, providing very few opportunities to make false positive errors of the first type (FPR-1). Despite the lack of opportunities to commit false positives, the lopsidedness in the false positive and false negative error rates is within expectations given previously published performance studies [7, 8, 9, 10].

At first glance, the reasons of the errors might be derived from low clarity of the marks or reference prints provided, from the digital nature of the testing material supplied to the participants (no hard copies were made available), or from the challenging nature of the samples chosen by the providers. To evaluate the possible impact of these factors and shed light on the decision-making process of the participants, the test providers reviewed the annotated images provided by the participants. As stated in previous publications [23, 24], challenging tests that induce errors can benefit the learner. The authors strongly encourage the participants in testing programmes to embrace the concept of "learning from errors", coping with the feeling of failure, bringing to awareness the risk of errors and its meanings, and moving forward to the next test with the passion to succeed.

Anatomical region and Orientation Issues

Fingermark #3 posed some issues to the participants to assess its correct anatomical region and orientation (search diagnosticity of the feature set). The core area is not reproduced, and the position of the delta can be only inferred. Additionally, the edges of the mark are not clear, and they give indication that the ridge-flow from left to right could end but indeed, it does not, and goes further to the right side, as can be seen in the comparison chart of one participant in Figure 7.

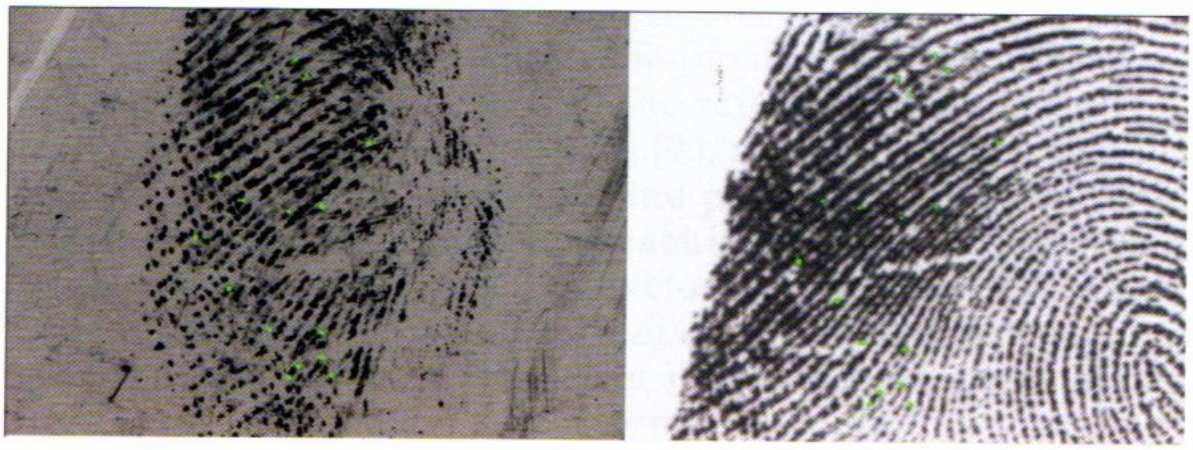

Figure 7 (previous page)

Fingermark #3 – comparison chart of one participant.

One of the laboratories explained their "inconclusive" decision as follows: "*palm: inconclusive: difficult to orient the marks and not enough points; no element to orient it*".

Figure 8 shows examples of mark-up analysis by laboratories that failed to identify the mark. It seems clear that the mark was not correctly oriented, thus being considered as a palm mark whereas the mark originated from a finger. Once again, when an orientation error occurs in the analysis phase, it is almost impossible to complete the comparison successfully. For example, one participant explained the "no value" decision as it follows: "*Overall observations – looks as an edge of palmprint. Ridge flow and distance looks like a palmprint, some areas are double*".

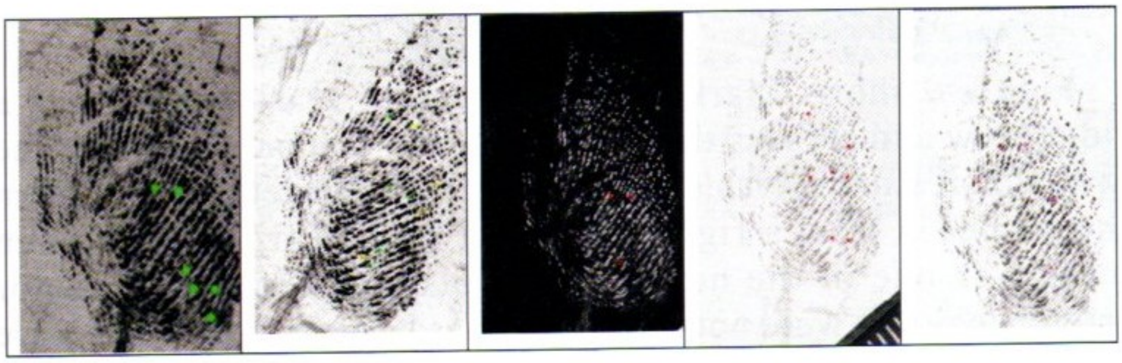

Figure 8

Fingermark #3 – examples of laboratories that wrongly orientated the mark.

Errors in anatomical region and orientation can obviously cause failures to find a match in a set of reference prints when the corresponding features are present. Assigning anatomical region and orientation are also directly related to appropriate exclusion and inconclusive opinions [21, 25] and how errors are tracked within an agency (based on Exclusion Approach #1 versus Exclusion Approach #2). Imagine an examiner assigns a mark in a case as a palm and excludes Donor X, but the mark was made by an area of Donor X's palm that is not recorded in the reference prints. In a lab following Exclusion Approach #1 (exclusion of a person), excluding Donor X would be considered an error (the response should have been inconclusive). Even if the palm mark was not made by Donor X, Donor X's reference prints would not be considered adequately recorded to support the exclusion of Donor X. Under Exclusion Approach #1, the only appropriate outcome for the comparison would be "inconclusive" and the examiner is expected to request the additional reference prints needed to complete the comparisons to Donor X.

If a laboratory follows Exclusion Approach #2 (exclusion of the friction ridge profile present in the reference prints), then the inability to locate corresponding features within the provided reference prints of Donor X would support an exclusion opinion of Donor X, even if the palm mark was made by Donor X. Do consumers of friction ridge reports understand what a given laboratory means by exclusion? Do consumers of laboratory reports understand that an exclusion reported by the DNA section may mean something different than an exclusion reported from the friction ridge section? Performance from different laboratories can be difficult to compare when errors regarding exclusion and inconclusive opinions are handled differently.

One Trial Resulting in the Erroneous Identification of Person

Figure 9 shows Mark #5, a palm mark. Mark #5 displays a ridge flow and creases that are diagnostic of a palm thenar. The available feature set is not highly diagnostic of left or right palm thenar but is from a right thenar per ground truth. This palm mark was one of the non-mated trials in the exercise (donor reference prints were not provided). It is believed that the feature set (ridge flow, creases, and minutiae) are sufficient for comparison purposes. Among the 52 laboratories that submitted their results, 38 correctly provided a "no match" result.

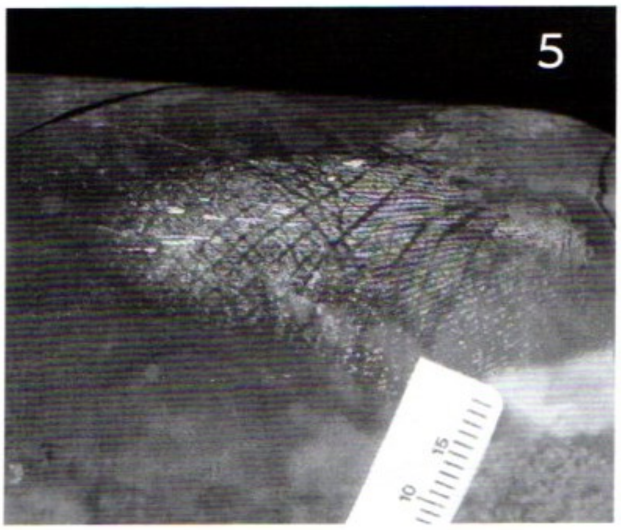

Figure 9

Palm mark #5.

Five laboratories surprisingly found some similarities between the mark #5 and two of the potential donors (donor D for four laboratories and donor C for one laboratory). Feature set mark-ups are shown in Figures 10 through 14 (the comments given by the laboratories are reported in the related captions). It is important to recall that the true donor was not provided to the participants.

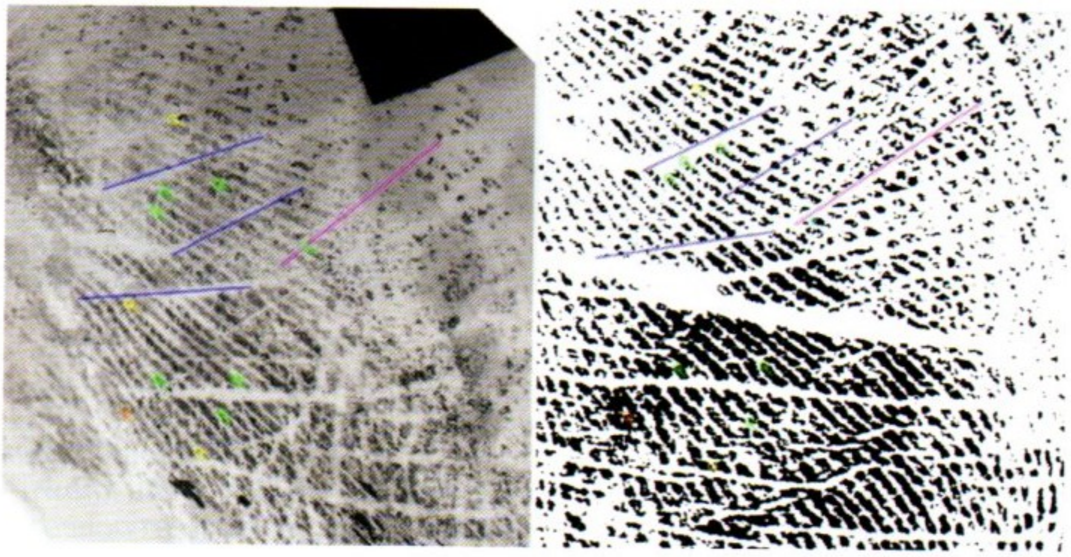

Figure 10

Mark #5 – Lab X1: side-by-side comparison - Conclusion of the lab: "very strong support (estimated LR: 100.000)" for a match.

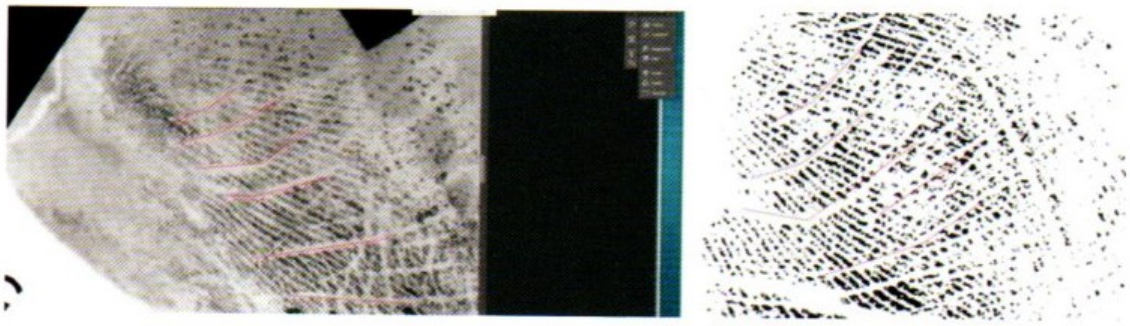

Figure 11

Mark #5 – Lab X2: side-by-side comparison. Comment of the lab "Many fingerprints and palms from donors are not complete in design and rotation: i.e. donor sample "D" has RR and RL incomplete, thus we decided to classify Mark #4 and #5 as INCONCLUSIVE."

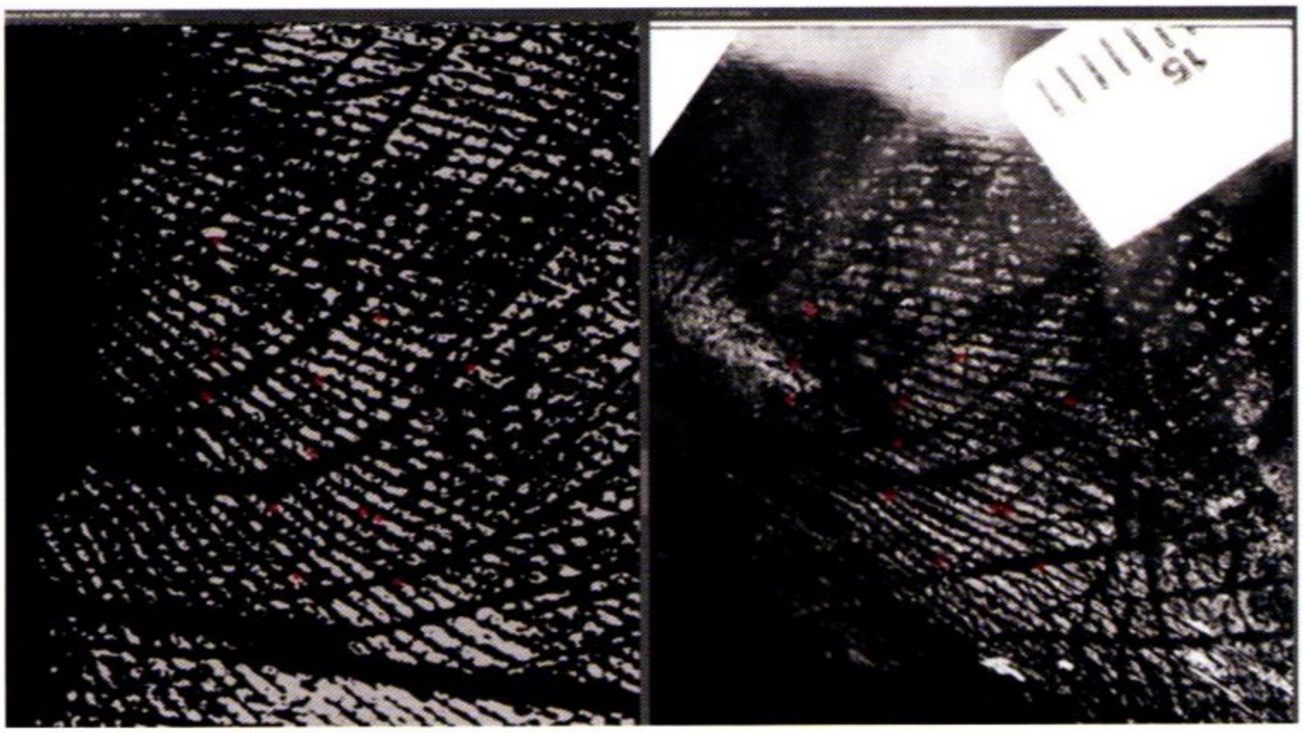

Figure 12

Mark #5 – Lab X3: side-by-side comparison. Comment of the lab "Although the latent print is of value for comparison, and we have found some common minutiae with Right Palm of Donor D, the poor quality of the known palm print induces a low confidence in this match. So we've reached an inconclusive result to the comparison activity. In a real case we will ask a new acquisition of the right palm of the donor D in order to have a better palmprint."

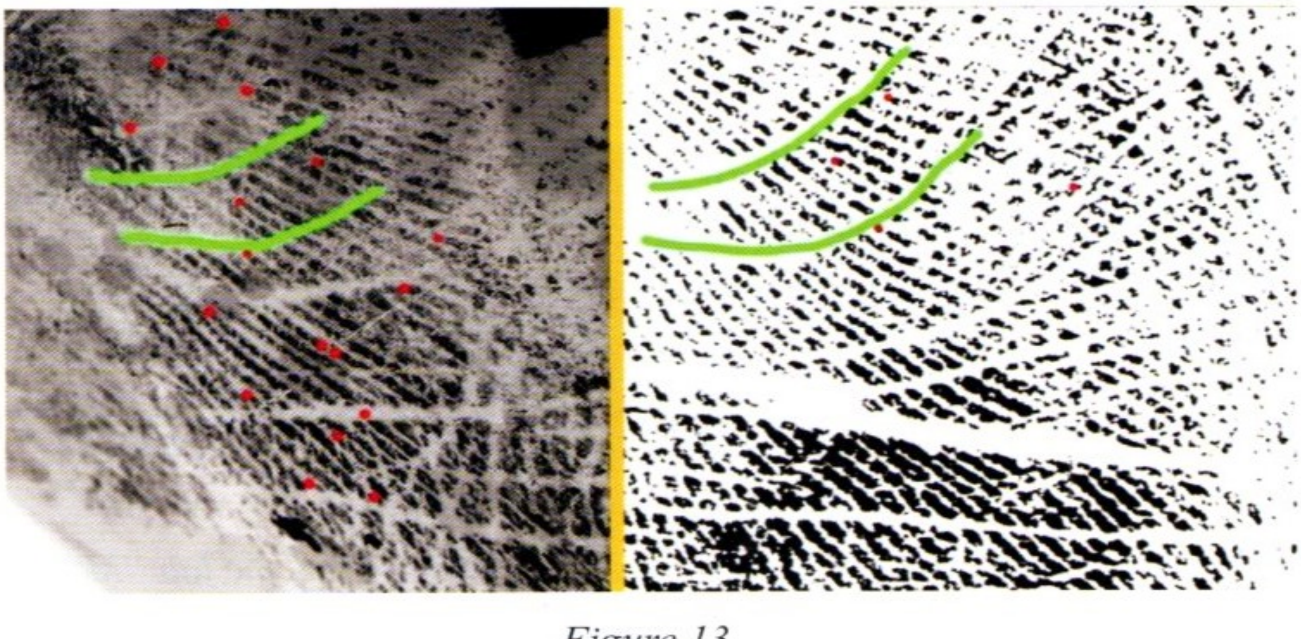

Figure 13

Mark #5 – Lab X4: side-by-side comparison. Comment of the lab "The area for a possible match (right palm, donor D – thenar area) in the known exemplar for comparison was of low quality (not sufficient for an unequivocal match)."

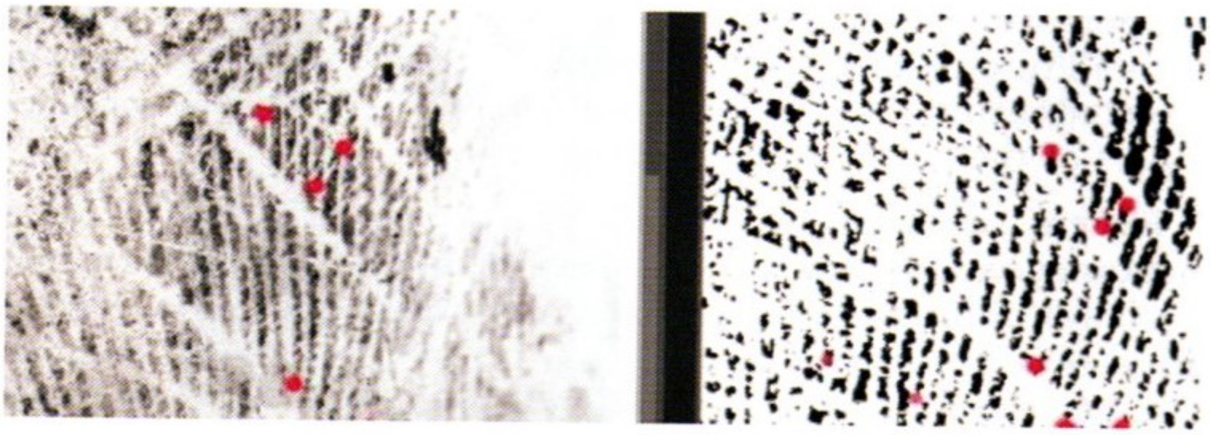

Figure 14

Mark #5 – Lab X5: side-by-side comparison. Comment of the lab "Visible Galton details, or other details with uniqueness – Is not considered sufficient. Despite the above, similarities are still noted. Some similarities with Right Palm Print Candidate D– not sufficient and not Inconclusive for ID."

It is important to note that the test designers did not consider providing any close non-match reference prints in the exercise. When evaluating Figures 10 through 14, it is possible to note that the minutiae mark-up shows some inconsistencies. Moreover, where creases were considered, it seems evident that only the apparent similarities were documented, possibly ignoring differences (or at least not documenting differences) and seemingly stretching tolerances (acceptable variations in appearance of features or relationships between features). Therefore, these laboratories were recommended to carefully review their results, which convey misleading information to their potential customer, even if not properly being considered as false identifications.

For Mark #5, the "*very strong support for a match – estimated L.R. 1:100000*" has been considered as a false positive, as well as "inconclusive" answers of those laboratories that marked the minutiae on the palm mark were considered as not in conformance with the ground truth.

The test designers recognized that not all the records of the donors were complete and easily readable. It is the opinion of the test designers, however, that the reference prints should not have prevented correct "no match" responses in the non-mated trials. The choice of the quality of the reference prints was intentional to increase the level of complexity of the test, and this does align with recent attitudes to challenge the level of expertise, rather than submitting a proficiency test for the sake of lab's accreditation [26]. It should be noted that in casework, the ability to recognize the sufficiency of reference prints to support an opinion is critical and current proficiency testing and performance testing regiments do not include "inconclusive - additional reference prints needed" as the appropriate and expected result of a comparison. Perhaps this should be included as part of the performance testing process.

Recent research on the low prevalence effect [27, 28], a phenomenon whereby targets prevalence impacts performance in visual comparison tasks, showed that people more often 'miss' infrequent target stimuli. In terms of fingerprint comparison, participants could more often misjudge non-mated pairs as matches when non-matches were rare. While mark #5 is a piece of a whole test carried out by the participants, there are previous knowledge and experience that most of the marks in this kind of test will mate one of the donors. This bias may explain the above-mentioned feature set similarities noted for Mark #5.

Two Trials Resulting in Erroneous Identification (Incorrect Feature Set) of the Correct Donor

Palm mark #6 and fingermark #8 each resulting in one match to the wrong portion of the friction ridge profile of the correct donor. These two erroneous matches were committed by two different laboratories. In the case of palm mark #6, the laboratory provided a screenshot showing a different area of the palm than expected (triradius in the hypothenar rather than a triradius in the interdigital region), with *de facto* non-corresponding minutiae as shown in Figure 15A. Figure 15B shows the correct corresponding region of the donor.

As for fingermark #8, another laboratory, while indicating the correct donor, provided a screenshot showing a wrong orientation of the mark #8 than expected (should be rotated 90° clockwise), with de facto non-corresponding minutiae as shown in Figure 16A. Figure 16B shows the correct orientation and corresponding minutiae.

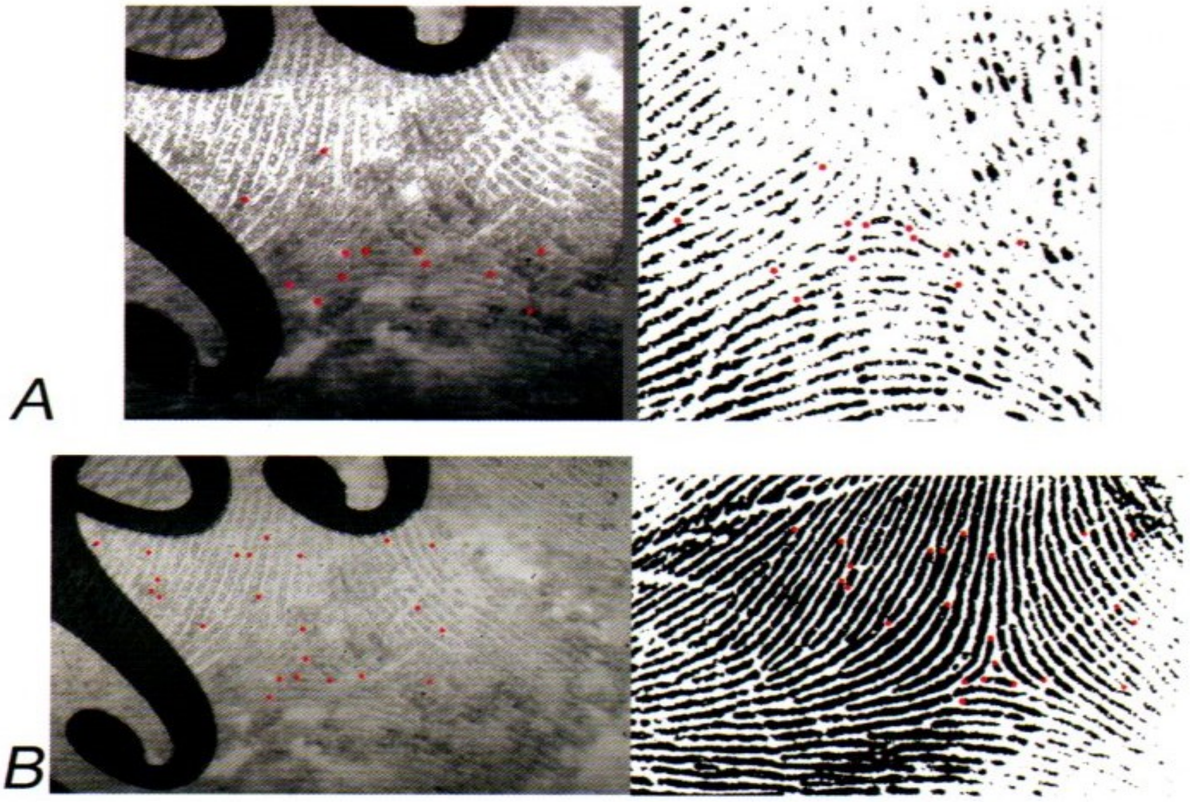

Figure 15

Mark #6 – A) Incorrect mark-up in the side-by-side comparison of one laboratory. B) Annotation for the correct area of the same donor.

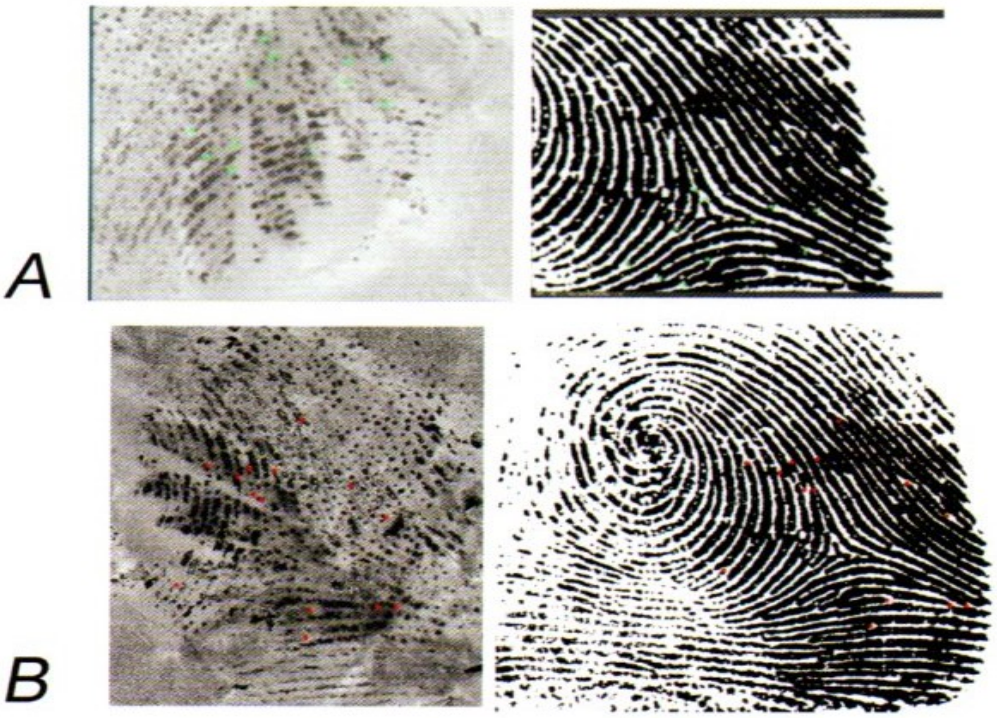

Figure 16

Mark #8 – A) Incorrect mark-up in the side-by-side comparison of one laboratory. B) Annotation for the correct area of the same donor.

The participants responsible for these errors were recommended to carefully review their side-by-side comparison and to conduct a thorough root-cause analysis of the mistakes. The misrepresentation of evidence, caused by incorrect orientation and misidentification of minutiae, highlights the critical importance of rigorous and accurate fingerprint analysis in criminal investigations.

Even if in the context of common testing settings, where only the correctness of the answers is matched against the ground truth, these results might have been considered as accurate conclusions, and not considered as false identifications. The review of the side-by-side charting provided by participants allowed the authors to detect and correctly classify this type of error. While other performance studies have documented false identifications to the correct donor [10, 20, 29], the researchers often faced considerable difficulty determining if the error was clerical in nature (correctly finding the corresponding feature set but writing down a wrong donor or wrong region of the correct donor) or technical in nature (incorrectly associated feature sets). In these performance tests, however, the researchers were alerted to the mistakes because the wrong section of the friction ridge skin was noted on the answer sheet. In other words, the examiner noted the mark was made by a left middle finger, but it was actually made by the right middle finger. The researcher then reviewed mark and exemplar (not charted images by the examiner) and determined if the error was clerical in nature (e.g., if the pattern is obviously wrong) or if it is possible the examiner matched the wrong region of skin. In previous performance testing and current known proficiency testing, if the examiner writes down the correct region of skin (e.g., right palm), it would be unknown to the researchers or test designers if the examiner actually identified the correct feature set within the right palm.

The documentation required by this collaborative exercise permitted unambiguous assessment of the cause of both clerical errors and feature set errors, which is the first of its kind. Any researchers or agencies completing performance testing should require sufficient documentation to permit a full investigation into false or unsupportable opinions. Naturally, case documentation should also permit this type of investigation.

Isolation Techniques for Marks – Inferring Contact Events and Managing Discontinuities

Figure 17 shows palm mark #9. Palm mark #9 resulted in nearly 30% of the participants judging the print as "not of value for comparison" and almost 20% providing an "inconclusive" opinion with the right palm of donor C. The inconclusive opinions were mostly due to minutiae in the mark being located in sub-regions of the mark that were not connected to each other by continuous ridge paths and an inability to locate a sufficient number of corresponding minutiae within any given sub-region of the mark.

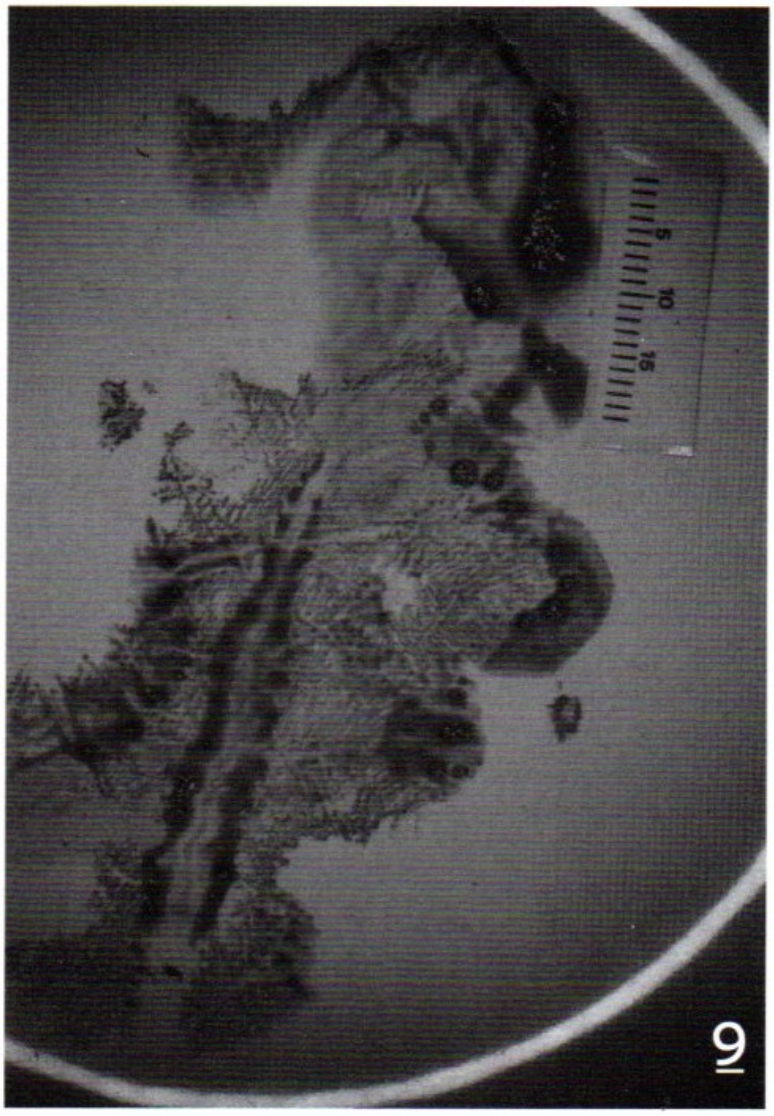

Figure 17

Palm mark #9.

One laboratory, whose side by side comparison is provided in Figure 18, described their results as, *"Although we have found local correspondences of minutiae in different disconnected areas between the mark and the Right Palm of Donor C, the low number of correspondences for each area (under our threshold)* *and the overall poor quality of the mark induce us to judge the latent print of no value for next comparison activities".*

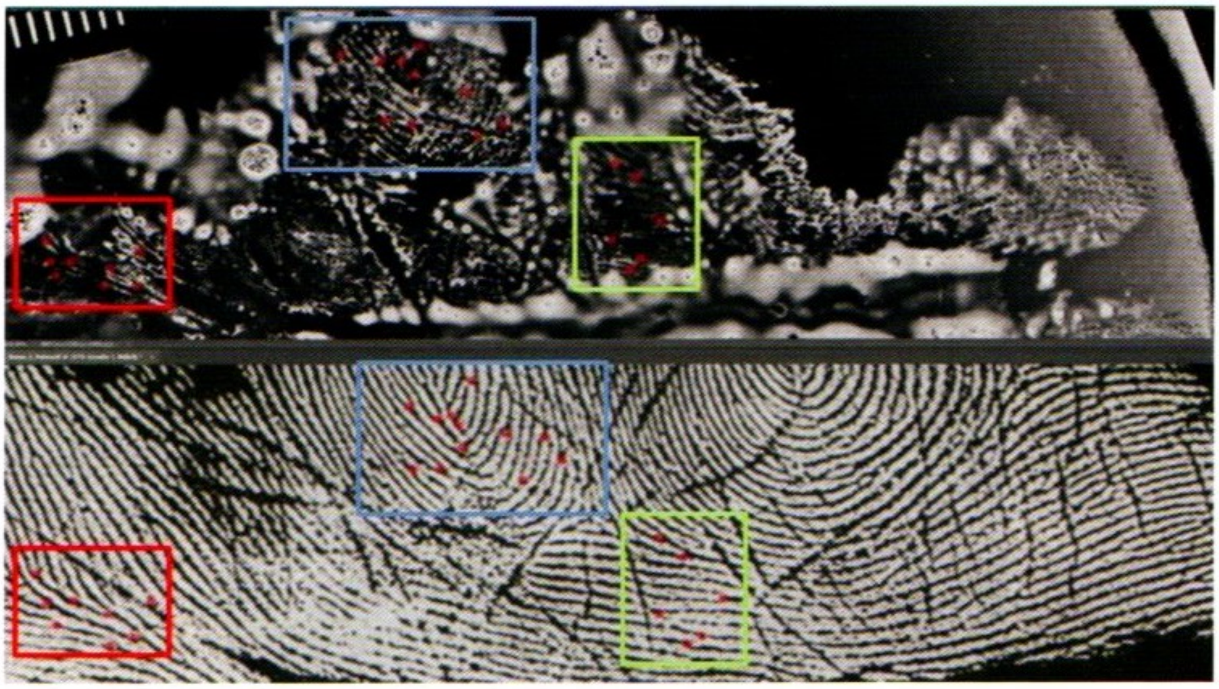

Figure 18

Side-by-side comparison of mark #9 provided by one laboratory.

The analysis provided by this laboratory is complete. However, the conclusion driven by the available data is disputable, especially the assessment of the mark as "not of value for comparison". Mechanics of friction ridge impressions are quite complicated. This is particularly true when a contaminating substance is involved (as blood in this case). It is not a novelty that friction ridge examiners often face marks that appear different from the related print because changes related to the skin [21, 22, 25], natural breaks due to the existence of regular creases and concave portions of the hands and feet [21, 22, 25], distortion related to contact events with a surface [21, 22, 25, 30, 31, 32].

Regarding mark #9, given the shape, ridge flow, and creases, the feature set is highly diagnostic of a right palm interdigital in a specific orientation. At the same time, considering the quantity of blood involved and the possible dynamics of deposition, some differences could be expected in the final appearance. In short, it could be reasonably inferred that the hand that deposited the mark is the donor's and that some differences may be present due to the transfer event (especially given blood transfer dynamics).

The three areas in Figure 18 can be compared taken together to reach a conclusion referring the mark was deposited during a single contact event with the surface. The process of inferring contact events with a surface has many complicating factors [33], even beyond the issues related to Mark #9. As demonstrated by Maceo in 2009 [30], a single contact event with a surface can create multiple areas of dislocated ridges and minutiae due to movement of the skin across the surface. These were not separate touches (the fingers did not contact twice). As demonstrated by White in 2022 [25] and Gibb and White [21], regular creases, scars, jewellery, and general hand or foot anatomy create natural breaks in ridge paths in impressions that are single contact events with surfaces. It is generally well understood that surfaces can also create breaks in ridge paths [33].

Laboratory policy or habits appear to dictate whether "continuous ridges" are required; these policies/habits were likely made (or evolved) without evaluating the merits of the practice and under what conditions the most probably inference should be followed. Meaning, how often do examiners indicate a mark was created by a single contact event, when in fact it was multiple contact events? How often do examiners indicate a mark represents multiple contact events, when in fact it was a single contact event? If a single palm mark is carved up as three separate impressions, and each is reported individually as a separate identification, does the investigator or the trier of fact naturally assume the person touched the surface three times with that palm?

Another aspect to be considered here is the differences among laboratories in the minutiae mark-up. As an example, the cited laboratory marked ten minutiae in agreement with the exemplar (shown in the blue box in Figure 18) and could not conclude with an identification according to their numerical standard of 12 minutiae. As shown in Figure 19, another laboratory marked nine minutiae at same area of the print and further 20 minutiae (in different areas in groups of four, eight, two and six). Despite the lack of continuity in the ridge flow, the latter laboratory reached the conclusion of identification.

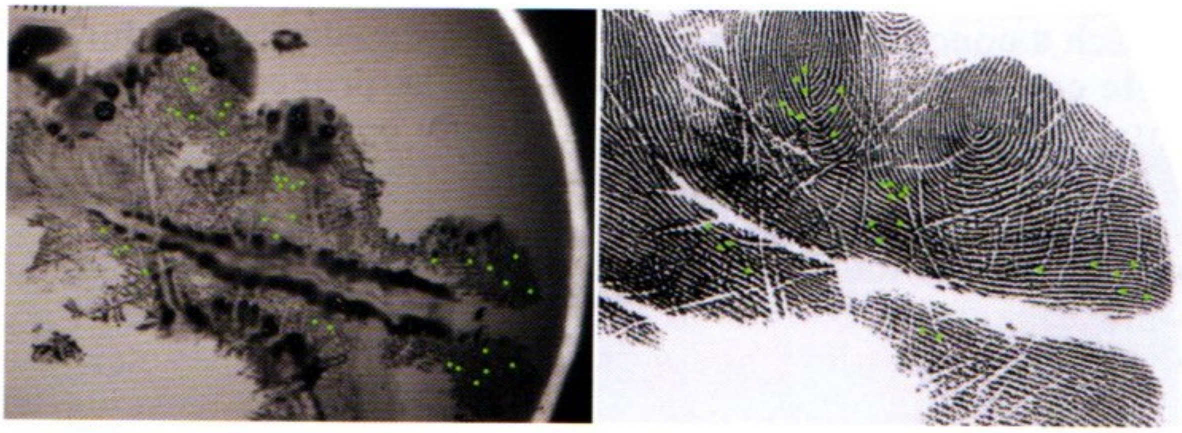

Figure 19
Side-by-side comparison of mark #9 provided by another laboratory.

Here, the question is if the "inconclusive" is a decision not to decide, or whether it could be claimed as a decision. In the first case, as outlined by Dror et al. [14] in an attempt to understand human cognitive processes involved in decision-making, "the inconclusive decision is a broad and imprecise decision category for fingerprint examiners, encompassing the range of "almost an exclusion" all the way to "almost an identification", and because inconclusive decisions are not regarded as error (they do not have the possibility of being false-positive or false-negative errors), an inconclusive decision may be a safe and easy decision choice". On the other hand, participants expressed their findings and conclusions with specific comments on the quality and visibility of mark #9 that led them to conclude "Inconclusive to donor C Right Palm". This focusing on the examination of the correct corresponding donor may be straightforward with the claim that "inconclusive" decisions are no less "definite" than identification or exclusion decisions, as discussed in Biedermann et al. [15].

The "inconclusive" decision as a conclusion deserves further in-depth analysis. As the collected data demonstrates, a different scale of conclusions that allow the examiners to better express similarities or dissimilarities without categorical conclusions may contribute to bring a more scientifically grounded approach to friction ridge comparison. With this regard, it is worth mentioning that the Organization of Scientific Area Committees (OSAC) proposed a 5-conclusion scale, which includes the "true inconclusive" which is a Bayesian LR of one [34, 35]. In any case, whether side-by-side charts are available, it is possible to evaluate the conformance of "inconclusive" results.

Conclusions

The 2021 EFP-WG ACE exercise was designed to test the performance of Forensic Service Providers (FSPs) in fingerprint comparison in realistic conditions. A set of palm and finger marks of various range of difficulty was selected as testing material, providing at the same time reference material also ranging in quality.

According to the feedback from participants, the test resulted to be quite challenging, as demonstrated by the high False Negative Rate (FNR = 13.1%). The False Positive Rates were both considerably lower: FPR-1 was 1.1% and FPR-2 was 0.7%. It is important to note that close non-matches were not included in the test (and would have likely increased the false positive rate) and other regions of the friction ridge skin (lower phalanges and feet) were not included.

This exercise illustrated that managing inconclusive decisions requires more discussion within the friction ridge discipline. When is it appropriate to be inconclusive, given feature set similarities and differences can occur when impressions are from the same source or different sources? What methods should be used by examiners to document non-corresponding features? How should exclusion versus inconclusive results be managed, given examiners rarely receive the full friction ridge profile of a person for comparison in casework and must infer they have the areas of the profile needed to support an opinion?

In this collaborative exercise, charting of the identifications and inclusions (inconclusive decisions that leaned toward same source) permitted unambiguous assessment of errors as being clerical or technical in nature. In order to detect and evaluate errors, all proficiency test and collaborative test designers are encouraged to collect side-by-side comparisons from the participants and review of the images do determine appropriateness of the source opinions and results.

While it was noted that laboratories using a numerical standard for identification made half the average number of errors compared to laboratories using a holistic or mixed approach, ISO 17025 accreditation was not found to influence the accuracy rate of the participants. This suggests that accreditation as implemented in the fingerprint domain does not necessarily result in improved performance. Nonetheless, accreditation does enable a thorough analysis of errors and the implementation of corrective actions.

Finally, the friction ridge discipline needs to continue to research, discuss, and train and test performance on methods used by examiners to determine contact events with a surface and techniques for isolating individual marks. These decisions affect the number of suitable marks reported in a case, the number of opinions provided in a case, and potential inferences made by readers of a report regarding how many times a person may have touched a surface.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors would like to thank the ENFSI Fingerprint Working Group (FIN-WG) for its constant support and all the participants in the test discussed in this paper.

For further information, please contact:

Lt.Col. Francesco Zampa
Parma, Italy
zampa.francesco@gmail.com, francesco.zampa@
carabinieri.it

References

1. Mattei, A.; Fish, J.; Hilgert, M.; Lövby, T.; Svensson M.; Vaughan J.; Zampa F. ENFSI Collaborative Testing Programme for Fingermarks: Past Experiences and Future Perspectives. For. Sci. Int. 2017, 275, 282-301. https://doi.org/10.1016/j.forsciint.2017.03.010
2. Mattei, A.; Fish, J.; Hilgert, M.; Lövby, T.; Svensson, M.; Vaughan, J.; Zampa, F. The 2015 ENFSI Fingerprint Working Group Testing Programme. For. Sci. Int. 2017, 280, 55-63, https://doi.org/10.1016/j.forsciint.2017.09.002
3. National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press; 2009. Available online https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf. Accessed 21 February 2024.