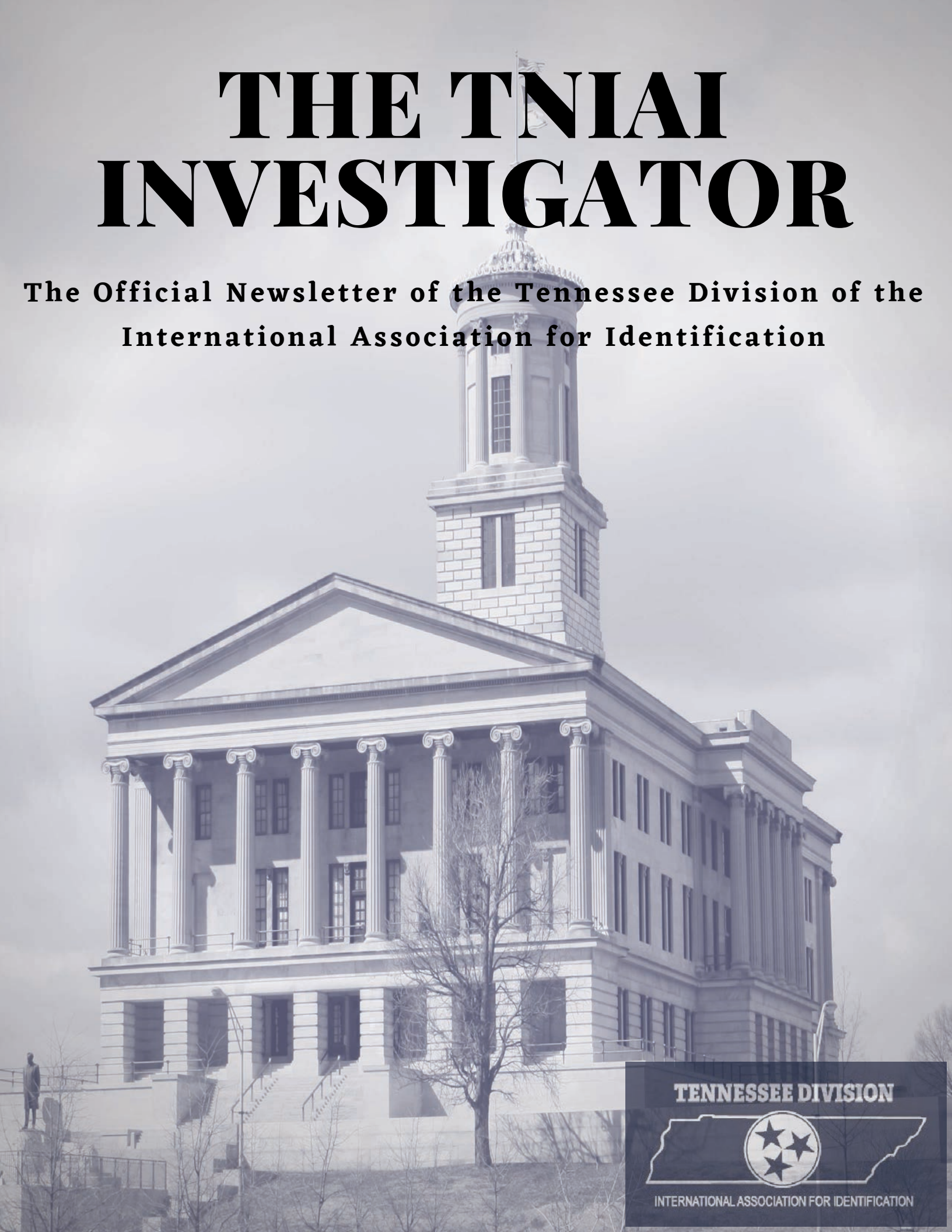
THE TNIAI INVESTIGATOR

The Official Newsletter of the Tennessee Division of the International Association for Identification



WATCH WEBSITE FOR MORE DETAILS

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION'S
105th EDUCATIONAL CONFERENCE | AUGUST 1 - 7, 2021

GAYLORD OPRYLAND RESORT, NASHVILLE, TENNESSEE

TENNESSEE DIVISION

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION

INSIDE THIS ISSUE

CRIME SCENE DO NOT CROSS

CONTENT:

Online Training Courses

Affordable Learning at Your Fingertips!

Not able to attend one of our many in-person training courses? Try one of our webinars taught by expert instructors, all of which have been submitted for approval to the IAI Certification Boards.

CURRENTLY SCHEDULED WEBINARS

INTRO TO CRIME SCENE MANAGEMENT - MARCH 24, 2021

INTRO TO CRIME SCENE STAGING DYNAMICS IN HOMICIDE - APRIL 6, 2021

INTRO TO FORENSIC BIOLOGY AND DNA ANALYSIS - APRIL 8, 2021

BLOODSTAIN PATTERN RECOGNITION - APRIL 12, 2021

ADVANCED CSI: SHOOTING SCENES - APRIL 19, 2021

RULES OF FINGERPRINT CLASSIFICATION - APRIL 27, 2021

EXPERT WITNESS TESTIMONY - JUNE 1, 2021

RULES OF FINGERPRINT CLASSIFICATION - JUNE 8, 2021

And here's an in-person course to consider:

COMPREHENSIVE LATENT PRINT COMPARISON TRAINING

HOOVER, AL | JUNE 14 - 18, 2021 | For Basic to Advanced Examiners

MORE COURSES AVAILABLE ONLINE

CHECK TRITECHTRAINING.COM FOR ADDITIONAL INFORMATION

EXCLUSIVE TRAINING PARTNER OF THE International Association for Identification

TRITECH TRAINING FORENSICS

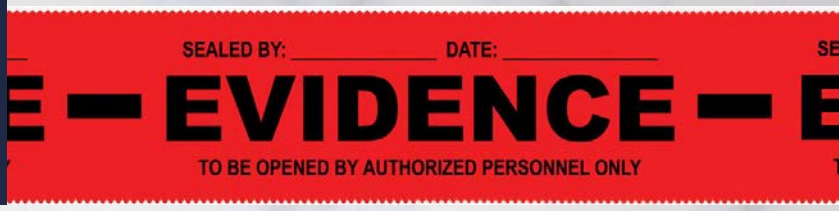
training@tritechusa.com | 800.438.7884 ext. 1025

Home of Tri-Tech Forensics Training Division Headquarters

TENNESSEE DIVISION

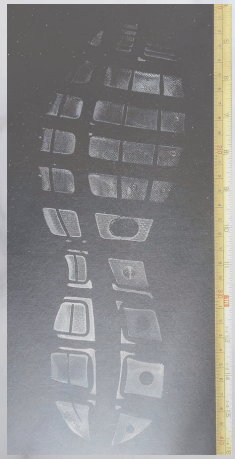
INTERNATIONAL ASSOCIATION FOR IDENTIFICATION

EDITOR'S MESSAGE

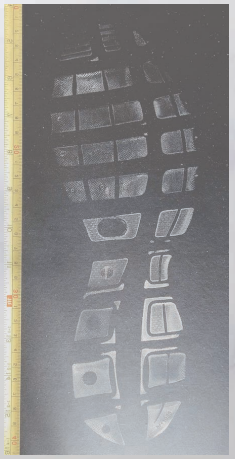


Message from the Editor and Committee

We hope everyone is doing well and staying healthy and safe. As always, we are very proud of our newsletter and the progress we have made. Creating content that is applicable to our members and sponsors is our top priority. The goal of this newsletter is to promote further education and forensic training, encourage research and outreach, and share advancements and developments within the forensic community. We would like to take this opportunity to thank our amazing members for their continued support and participation in this organization. If you or your company have any achievements, promotions, or kudos that you would like to award, please let us know and we would be happy to include them in the next newsletter.



- Isabella Barnett



The Double Loop Podcast is a weekly show featuring Glenn Langenburg and Eric Ray discussing latent print topics, current events in forensic science, the newest research articles, interesting guests, and analysis of notable cases from a forensic scientist perspective.

CHECK IT OUT

Double Loop Podcast

http://doublelooppodcast.com

The DOUBLELOOP podcast

TENNESSEE DIVISION

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION

PRESIDENT'S MESSAGE

Message from the President

Hello and Happy Spring!!

I am pleased to announce that the planning of the 105th Educational Conference of the IAI is well underway! This year our local division is being called upon for assistance in conference volunteer duties throughout the week. If you are interested in helping in any way, please reach out to me ASAP.

As I have stated in previous correspondence, there are a lot of opportunities for you to be involved in the conference, such as poster presentations and a photo contest! For more information visit https://www.theiai.org/conference_poster_photo_contes.php.

I personally hope you consider attending this event, as it is the biggest meeting of our professional association, a fantastic place to network, and offers so many workshops and lectures to gain a wealth of knowledge. Visit https://www.theiai.org/2021_iai_conference_nashville.php for more information on the tentative schedule, upcoming registration information, and hotel accommodations. Be sure to make your reservation ASAP if you plan to stay onsite during the week.

I would also like to encourage our membership one last time to apply to present!! This would be a great "local" opportunity to present your on the job knowledge, skills, research, technology, case experience, etc. to our professional organization at a larger level and bigger platform (great for certification points and to build on resumes/CVs!). For presenter information and applications visit https://www.theiai.org/conference_presenter_informati.php. If you have any additional questions, do not hesitate to contact us at tennesseeiai@gmail.com or you may contact Lesley Hammer, the IAI Educational Program Coordinator, directly at iaiedplanner@gmail.com.

****Please note that although the workshop proposals have closed, you may still inquire if you have a workshop format. Lecture proposals are scheduled to close at the end of April, but will also continue to take inquiries. Consider applying and apply as early as you can!****

Deadlines from the IAI website:
Lecture proposals may be submitted until April 30, 2021.
Posters may be submitted until June 30, 2021.

I ask that you help the IAI spread the word and be on the lookout for future emails as we assist the parent body in planning this great event. We hope to plan to see you in August!
-Monica Kent

Monica Kent
2019-2021 TNIAI President

Forensic SERIES

Air Science provides forensic laboratory equipment to meet the needs of each step in the evidentiary chain, from field processing, to transport and storage, to analytical procedures in the laboratory.

- Ductless Fume Hoods
- Forensic Evidence Drying Cabinets
- Automatic Cyanoacrylate Fuming Chambers
- Mobile Forensic Evidence Benches
- DFO and Ninhydrin Fingerprint Development Chambers
- Fingerprint Powder Workstations
- Benchtop Decontamination Chambers
- Evidence Storage Cabinets
- Swab Drying Cabinets
- Mobile Evidence Transporters
- Fume Chambers

Review our product offerings at **airscience.com/forensics**

Air Science®

Fort Myers, FL 33907 \ **Toll Free.** 800-306-0656
www.airscience.com \ info@airscience.com

TENNESSEE DIVISION

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION

OFFICERS AND BOARD MEMBERS

2019-2021 TNIAI Officers and Board Members

President - **Monica Kent** - Metro Nashville
1st Vice-President - **Amber Smith** - Metro Nashville
2nd Vice-President - **Angela Christian** - Montgomery Co. Sheriff's Dept.
Secretary - **Kendra Fleenor** - TBI
Treasurer - **Elizabeth Reid** - TBI
Webmaster - **Daniel Anselment** - UT National Forensic Academy
Sgt. At Arms - **LJ Davidson** - Metro Nashville
Historian - **Brooke Duke** - TBI

<u>Board of Directors</u>

Chairperson - **Heather Hammond** - TBI (2019-2021)
David Hoover - TBI (2018-2020)
Monica Kent - Metro Nashville (2019-2021)
Adam Liberatore - Montgomery Co. Sheriff's Dept. (2019-2021)
Kristine Keeves - Lavergne Police Dept. (2018-2020)
Rebecca Hooper - Metro Nashville (2021)
Charles "Chip" Linville - Metro Nashville (2019-2020)

<u>Associate Member Representatives</u>

Philip Sanfilippo - Tri-Tech Forensics

<u>Committee Chairs</u>

Certification Committee Chair - **Elizabeth Reid** - TBI
Conference Committee Chair - **Rebecca Hooper** - Metro Nashville
New Membership Committee Chair - **Easton Haynes** - Metro Nashville
Public Relations Chair - **Jessica Davis** - Metro Nashville
Scholarship Committee Chair - **John Dunn** - TBI
By-laws Committee Chair - **Heather Hammond** - TBI

MEGAfume cyanoacrylate fuming chambers

attestor

innovation in forensics

clear, clean & innovative

ATTESTOR-NORTHAMERICA.COM

TENNESSEE DIVISION

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION

CONFERENCE INFORMATION

CLASSIFIED

Conference Deadlines from IAI Website:

Lecture proposals may be submitted until April 30, 2021

Posters may be submitted until June 30, 2021

SCOPE OUT PARENT BODY WEBSITE FOR IMPORTANT INFORMATION:

For information regarding registration and hotel arrangements --> https://www.theiai.org/2021_iai_conference_nashville.php

For information regarding poster presentations and photo contests --> https://www.theiai.org/conference_poster_photo_contes.php

For information regarding presentation proposals--> https://www.theiai.org/conference_presenter_informati.php

Congratulations to our newest lifetime member

Elizabeth Reid

25 consecutive years of TNIAI membership

~ and ~

Congratulations to our spring student scholarship recipient

Rachael Akins

of MTSU

Students attending an accredited college or university (full or part time, undergraduate or graduate) taking courses in the pursuit of a career in the various phases of the science of identification or the law enforcement field and are a student member of the TNIAI, please visit https://www.tniai.org/students to apply for scholarships

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION'S
105th EDUCATIONAL CONFERENCE | AUGUST 1 - 7, 2021

GAYLORD OPRYLAND RESORT, NASHVILLE, TENNESSEE
FOR UPDATES, VISIT THE IAI WEBSITE - THEIAI.ORG

SPRING into action and have an EGG-TASTIC year! Stay safe and HOP into a fruitful 2021!

TENNESSEE DIVISION

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION

INFORMATION

Please join us in welcoming our newest member! Julie McDowell - Middle Tennessee State University

<u>TN Regional Training Opportunities:</u>

University of Tennessee Law Enforcement Innovation Center

PoliceTraining.net

TN Law Enforcement Training Officers Association

TRITECH Forensics Training

<u>National Training Opportunities:</u>

International Association for Identification

TRITECH Forensics Training

University of Tennessee Law Enforcement Innovation Center

Please visit <u>https://www.tniai.org/forum</u> for more information

PRESERVE & PROTECT

Latitude Fentanyl Filtered Hood

Ductless Fume Hoods

FORENSIC MYSTAIRE® SOLUTIONS

Cyanoacrylate Fuming Chamber

Evidence Drying Cabinets

MYSTAIRE®

Phone: +1-919-229-8511 • Toll Free: 1-877-328-3912 • www.mystaire.com

MADE IN THE USA

Forensic Science International 318 (2021) 110457

Contents lists available at ScienceDirect

Forensic Science International

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

Testing the accuracy and reliability of palmar friction ridge comparisons – A black box study

Heidi Eldridge[a,b,*], Marco De Donno[b], Christophe Champod[b]

[a] RTI International, 3040 E. Cornwallis Rd., Research Triangle Park, NC 27709, USA
[b] University of Lausanne, Batochime Quartier Sorge, Lausanne-Dorigny, VD, CH-1009, Switzerland

ARTICLE INFO

Article history:
Received 27 March 2020
Received in revised form 30 July 2020
Accepted 5 August 2020
Available online 8 August 2020

Keywords:
Black box
Palm prints
Error rates
Fingerprint examiners
Expertise

ABSTRACT

Critics and commentators have been calling for some time for black box studies in the forensic science disciplines to establish the foundational validity of those fields—that is, to establish a discipline-wide, base-rate estimate of the error rates that may be expected in each field. While the well-known FBI/Noblis black box study has answered that call for fingerprints, no research to establish similar error rates for palmar impressions has been previously undertaken. We report the results of the first large-scale black box study to establish a discipline-wide error rate estimate for palmar comparisons. The 226 latent print examiner participants returned 12,279 decisions over a dataset of 526 known ground-truth pairings. There were 12 false identification decisions made yielding a false positive error rate of 0.7%. There were also 552 false exclusion decisions made yielding a false negative error rate of 9.5%. Given their larger number, false negative error rates were further stratified by size, comparison difficulty, and area of the palm from which the mark originated. The notion of "questionable conclusions," in which the ground truth response may not be the most appropriate, is introduced and discussed in light of the data obtained in the study. Measures of examiner consistency in analysis and comparison decisions are presented along with statistical analysis of the ability of many variables, such as demographics or image quality, to predict outcomes. Two online apps are introduced that will allow the reader to fully explore the results on their own, or to explore the notions of frequentist confidence intervals and Bayesian credible intervals.

1. Introduction

While critics had for some time been calling for the friction ridge comparison discipline to produce research that supported their claim to accurately associate unknown impressions back to their sole source[1–4], it was only with the release of the 2009 National Research Council's watershed report[5] that the research community began to take notice and answer the call. To date, there have been two large-scale black box studies completed that have attempted to establish a discipline-wide error rate estimate for friction ridge comparisons.

The first of these studies is the well-regarded FBI/Noblis black box study[6], which reported a false positive rate of 0.1% and a false negative rate of 7.5%. However, this study did not address the accuracy of palm comparisons, presenting only correctly-oriented phalanx impressions (fingerprints) to its participants. The second study, as yet unpublished, is NIJ-funded research by Miami-Dade Police Department[7]. This study did include palmar comparisons in its design. However, two limitations prevent its use as an estimate of palm comparison accuracy. The first is that the study authors did not calculate and report a separate error rate for the palm comparisons, but lumped all comparison types together, reporting a single false positive error rate. The second is that the exemplars for the different source trials were constructed by selecting exemplars from study donors who did not create the mark in a given trial, without any attempt to locate a close non-match. Without deliberately sought-out close non-match distractors, it is highly unlikely that the different source trials presented a meaningful challenge.

Due to the limitations of these two studies as regards palmar comparison, there has not been a black box study to date that has measured the accuracy and reliability of friction ridge examiners in performing palmar comparisons. According to the President's Council of Advisors on Science and Technology report[8], these studies are necessary to establish the foundational validity of a pattern comparison method. There is scant, if any, scientific support to suggest that the error rates calculated for fingerprint comparisons can be extrapolated to palmar comparisons, and in fact, it seems likely that the false negative rate for palmar

* Corresponding author.
E-mail addresses: heldridge@rti.org (H. Eldridge), Marco.DeDonno@unil.ch (M. De Donno), Christophe.Champod@unil.ch (C. Champod).

http://dx.doi.org/10.1016/j.forsciint.2020.110457

comparisons should be higher given that there is a much larger area to search, that orientation clues are often ambiguous or missing, and that practitioners often receive less training and practice in this area compared to the comparison of fingerprints.

We suggest that there are three criteria that should be met to properly establish an informative error rate for palmar impressions. These criteria are:

1. Error rates should be constructed for palmar impressions separate from those of distal phalanges;
2. Test impressions at different quality levels should be used and error rates calculated for each so that meaningful comparisons to casework images can be made; and
3. Close non-matches should be incorporated to present a realistic chance of making a false-positive error.

This paper presents the results of a recent black box study that takes as its blueprint the FBI/Noblis study, but specifically tests the accuracy and reliability of friction ridge examiners in making comparisons of palmar impressions, in accordance with the three criteria outlined above.

2. Method

Wherever possible, the design of this study mimicked that of the FBI/Noblis study so that we could get as close as possible to an apples-to-apples comparison between the error rates for palmar comparisons and finger mark comparisons. However, there were instances where the aims of this research or the logistics of building the study samples and participant population required some deviation.

Participants for the study were recruited from among working friction ridge examiners, examiner trainees, and retirees. A demographic survey administered to all participants allowed separation of the trainees from the fully qualified analysts. Participants were provided with anonymized usernames and maintained contact with the researchers through a confidential liaison, such that their identities were never known to the research team. Informed consent was provided to all donors and participants and was reviewed and approved by RTI International's Institutional Review Board.

A total of 328 participants enrolled in the study; however, only 133 completed all 75 trials that were requested of them. An additional 93 participants completed between one and 74 comparison trials. There were 226 total participants designated as "active" in that they completed at least one analysis. The data from all completed conclusions (analysis or comparison) were used in data analysis. Only demographic information from the 226 active participants is reported here. Most examiners were between 30 and 50 in age, female, had between zero and 20 years of experience, and worked for accredited US state or local laboratories. Approximately 44% reported that they were certified as latent print examiners by the International Association for Identification (IAI). More detailed demographic information and information on the impressions used in the study can be found in Appendix A. Supplementary Material.

Fifty individuals at 6 partner laboratories each donated palm marks of known source and multiple sets of exemplars. Marks were made on a variety of substrates, in a variety of matrices, and using a variety of development techniques to mimic the range of samples seen in casework. Marks also varied in the amount of distortion present. Distractor (different source) exemplars were selected from an AFIS database containing approximately 25,000 palm records. By combining marks and exemplars, a study pool of 526 paired cases were drawn. These included 400 (76%) same source trials and 126 (24%) different sources trials. Each participant was assigned a group of 75 cases pseudo-randomly drawn from this pool. The distribution of same source to different sources trials and trial difficulty levels assigned to each participant is described in Appendix A. Supplementary Material. The assigned cases could be worked in any order the participant chose.

Marks of low quality were included to test where participants' thresholds for declaring a mark to be suitable for comparison lay. In some cases, exemplars of low quality were deliberately selected to increase the difficulty of a comparison to a relatively clear mark. All same source trials contained overlapping areas of corresponding features to make the comparisons fair, although for some, the expectation was that an "inconclusive" decision would be reached due to the low quantity or degraded state of the data.

Participants took part in the study using a custom online Picture Annotation System (PiAnoS) software interface, which was developed at University of Lausanne. This interface presented users with the unknown mark first, as depicted in the workflow depicted in Fig. 1. If the mark was declared not to be suitable for comparison, the trial ended. However, if it was declared suitable for comparison, it was then presented side-by-side with a candidate exemplar.

Participants were provided with tools for zooming in and out, rotating the mark, dragging to different areas of the exemplar, annotating minutiae, pairing minutiae between the mark and the print, tracing ridges and other features, and demarcating areas of different qualities. These tools were made available for the convenience of the participants and their use was not required for the study. Participants were also provided with an optional text box where they could produce written notes to allow participants to express their thought processes when reaching decisions. These boxes often allowed insight into how errors occurred and are the subject of a separate publication [9].

The custom PiAnoS interface that was developed for this research was limited in the tools that were provided for the participants' use. No image processing tools (such as brightness or contrast adjustments or the ability to invert dark and light pixels) were provided. This was a source of frustration to multiple participants, who commented that they were accustomed to being able to digitally "enhance" images prior to comparing them. However, it would have introduced an unwanted variable into the analysis. We were not concerned so much with examiners' skills at digital processing as we were in the conclusions multiple examiners would reach when presented with the same stimuli.

Participants were initially presented with an image of the unknown mark on its own and were asked to render a decision about its suitability for comparison. Three response options were provided: no value, suitable only for exclusion, and suitable for identification. If "no value" was selected, the trial terminated and the participant was free to select a new trial to begin. If either of the other two options was selected, the participant proceeded to a side-by-side comparison.

At the end of their comparison, participants were required to select a conclusion from the following three options: Identification, Inconclusive, or Exclusion. If the participant selected "Inconclusive", they were prompted to provide a reason for the decision from among the following options, which were taken from the FBI/Noblis study:

- Inconclusive due to no overlapping area
- Inconclusive due to insufficient information
- Inconclusive, but with corresponding features noted

If the participant selected "Exclusion", they were prompted to provide a reason for the decision from among the following options, which were taken from the FBI/Noblis study:

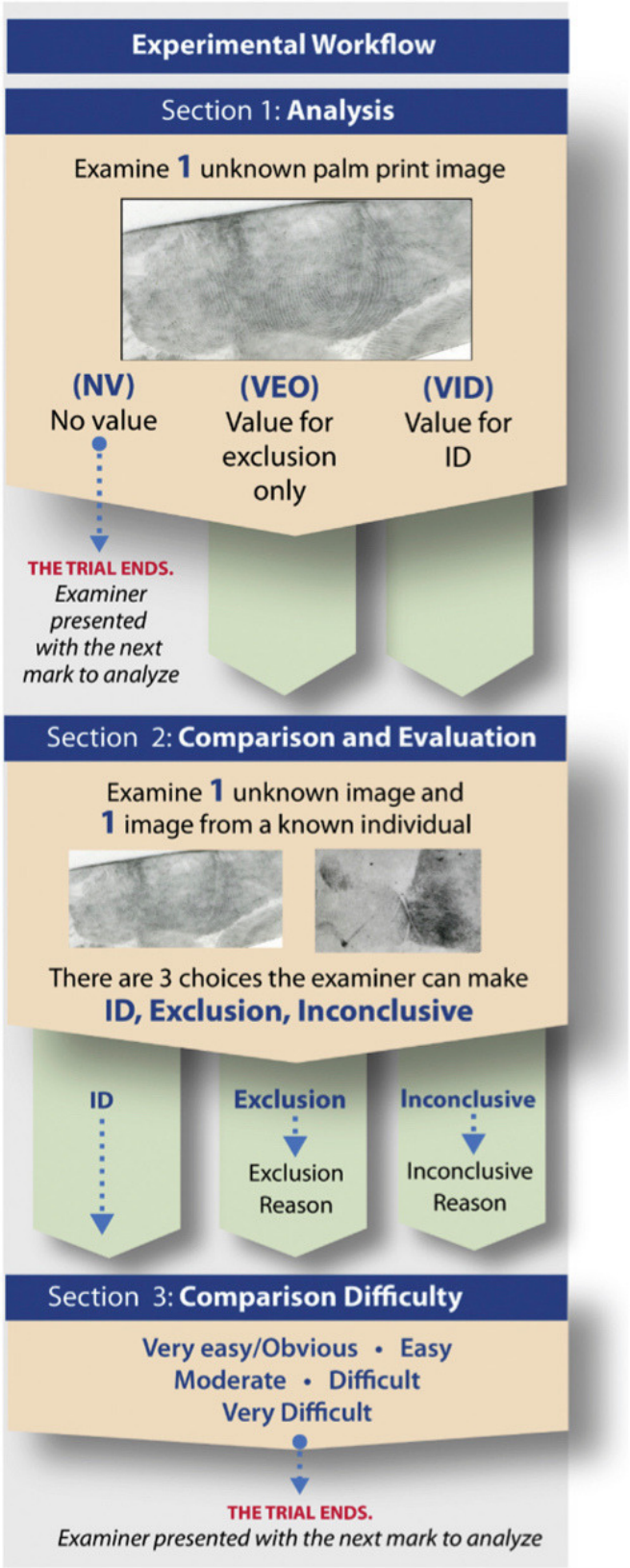


Fig. 1. Experimental workflow in PiAnoS.

- Pattern class/ridge flow alone
- Minutiae and/or level 3

Finally, at the end of each comparison trial, participants were asked to indicate the difficulty of reaching the comparison decision, from among the following options, which were taken from the FBI/Noblis study:

- Very easy/Obvious
- Easy
- Moderate
- Difficult
- Very Difficult

We took advantage of three algorithms to automatically measure the quality of the marks:

- LFIQ1: a quality metric developed by [10].
- LFIQ2: a second quality metric developed by [11].
- LQmetric: A set of quality indicators currently implemented in ULW developed by FBI/Noblis [12].

LQmetric can process palm marks directly as it contains an auto-encoder for the detection of minutiae. LFIQ1 and LFIQ2 require the set of detected minutiae as input in addition to a greyscale image. We used the auto-encoder (version 11) of an Idemia MorphoBis AFIS system acquired in 2015 to obtain the necessary minutiae for LFIQ1 and LFIQ2 analysis. Only minutiae meeting Quality Level 11 (a quality metric associated with auto-encoded minutiae) were retained.

Statistical analysis of the results was carried out in R version 3.6.3 RC (2020-02-21 r77847)[13] coupled with RStudio Version 1.2.5033[14] using the following packages: *tidyverse*[15] for data wrangling, *caret* for machine learning and computing confusion matrices and associated error statistics [16],*vip* [17] for computing variable importance, and *proportion* [18,19] for computing confidence and credible intervals. The production of tables of results has been done using *knitr*[20] and *kableExtra* [21]. All datasets and R code associated with this research and paper (Rmarkdown) can be foundin https://doi.org/10.5281/zenodo.3726896.[1] The results were also prepared in an interactive web-based user interface developed in RStudio using the following packages: *shiny*[22], *shinydashboard* [23], *shinyjs* [24], *shinyBS* [25], *rintrojs* [26], and *rhandsontable*[27]. These results interfaces have been deployed on shinyapps.io.

In this study we are considering three different types of decisions:

1. The decisions reached by the participants at the end of the *Analysis* phase (NV, VEO and VID)
2. The decisions reached by the participants at the end of the *Comparison* phase (ID, EXC and INC) compared against the ground truth state of the trials submitted.
3. The decision reached by the participants at the end of the *Comparison* phase (ID, EXC and INC) but now compared with the majority voted conclusion by the participants.

For comparison decisions against ground truth (2 in the above list), we will compute the following efficiency indicators:

- False positive rate (FPR) is obtained by the proportion of the number of same source trials where the response was "exclusion" to the total number of same source trials. One minus FPR gives the value known as the *sensitivity* or *true positive rate*. For this study, the FPR measures false identifications.
- False negative rate (FNR) is a similar proportion but considers the proportion of different sources trials where the response was

[1] The images of the marks and corresponding prints can be made available by the corresponding author upon request

"identification" to the total number of different sources trials. One minus FNR is the value known as the *specificity* or *true negative rate*. For this study, the FNR measures false exclusions.
- Positive predictive value (PPV), also known as *precision*, is the proportion of trials declared to be an "identification" that are truly from the same source. One minus PPV gives the value known as the *false discovery rate*.
- Negative predictive value (NPV) is the proportion of trials declared to be an "exclusion" that are truly from different sources. One minus NPV gives the value known as the *false omission rate*.

Readers who are more accustomed to interpret results in terms of *sensitivity*, *specificity* or *false discovery rate* can obtain these values using the above 1 minus relationships.

We noted in the literature that the computation of these rates and values may differ with regards to what is counted to build up the total number of trials. For example in the FBI/Noblis black box study[6], although all data were included in the appendices and many data analyses were reported, when reporting the main false positive rate of 0.1%, only the cases declared VID in analysis were taken into account (excluding the cases declared VEO). In addition the authors kept the inconclusive (INC) responses in their total when calculating the same FPR for VID comparisons. Our position is that INC decisions should not be accounted for as it was suggested by PCAST[8]. We also treat VID and VEO cases equally. Hence our total number of cases will have VEO cases but ignore the INC decisions, whereas Ulery et al.[6] would, for the rates presented in the main paper, remove VEO cases and keep INC decisions. In order to facilitate comparison with previous studies, we present both options in the results associated with the comparison decisions against ground truth.

For analysis and comparison decisions against the response voted by the majority (1 and 3 in the above list), we will compute the following efficiency indicators:

- Examiner Response Disagreement Rate column (ERD) represents the proportion of the time that the considered Response (given by the examiner) did not match the majority vote.
- Majority Response Disagreement Rate column (MRD) represents the proportion of the time that the considered Response (given by the majority votes) was not made by the individual examiner.

In both the above cases, the "Response" refers to a decision that is being considered, i.e.identification, inconclusive, exclusion, no value, VID, or VEO.

ERD and MRD will receive a subscript corresponding to each context. For example, ERD_A will give the rate at which the response of the individual examiner disagrees with the majority vote for the Analysis decision being considered while ERD_C will give the same rate for the Comparison decision being considered. For the Comparison decision, we have considered the VEO cases but excluded the INC decisions.

In the shiny app `https://cchampod.shinyapps.io/app_CI/`, the reader will find illustrations (using the button "show the cells used to compute that proportion") of the counts used to compute each error/disagreement rate and predictive values.

3. Results and discussion

This paper provides an overview of the main results obtained from this study.It focuses mainly on results that are analogous to those provided in the FBI/Noblis study and on additional trends and observations the authors found to be of note. While we set up the

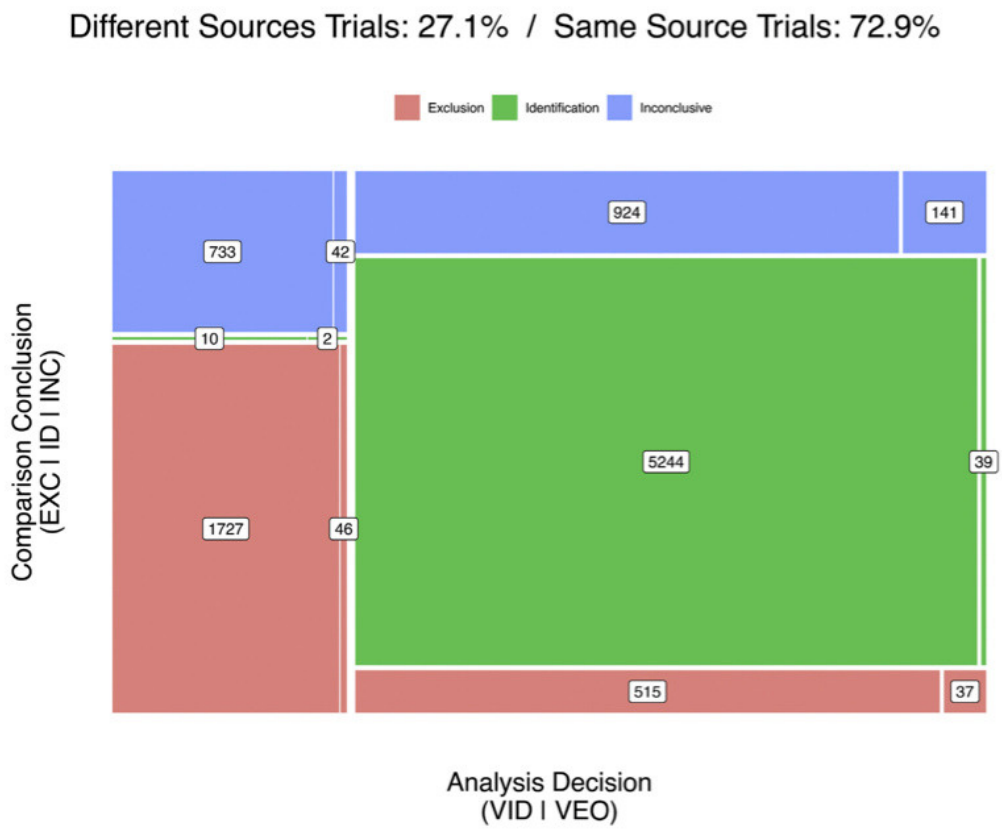


Fig. 2. Summary of the analysis and comparison decisions made in the study (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

design of this research to follow that of the FBI/Noblis study to answer the question, "Generally speaking, are latent print examiners as accurate at comparing palm impressions as fingers?", the results of the two studies are not directly comparable. If a difference in results is noted, we cannot be sure whether that difference is due to a difference in skill comparing palms versus fingers, or to some other variable such as the experimental interface, the difficulty of the selected marks, the identities of the participants, or numerous other possible factors. Interested parties who would like to examine the results in more depth and draw their own conclusions may look at all the gathered data by visiting the following url: `https://cchampod.shinyapps.io/Results_BBStudy/`. Additionally, frequentist confidence intervals and Bayesian credible intervals for all collected data have been constructed and may be explored at the following url: `https://cchampod.shinyapps.io/app_CI/`. Finally, participants in the study who have retained their usernames and passwords may review the images from their own trials at the following url: `https://ips-labs.unil.ch/apps/pianos4-palmbb-nocnm`.

Fig. 2 presents an overview of the analysis and comparison decisions made in the study. Of the 12,279 marks for which an analysis decision was entered, 19.6% (2,406) were declared to be of no value (NV) and are not included in Fig. 2. The left block (representing 27.1% of the data) shows the different sources trials, the right block shows the same source trials (the remaining 72.9%). The divisions on the x-axis of each block are a function of the conclusions reached in analysis (VID or VEO). On the y-axis, the split is made in each block according to the conclusions reached in comparison (EXC, ID, INC). The labels give the numbers of conclusions in each category. For example, if we take the different sources trials, we have 12 erroneous identifications in total, 10 of them concluded VID in analysis and 2 concluded VEO. Or, for the same source trials, we have a total of 552 erroneous exclusions (515 concluded VID in analysis and 37 concluded VEO).

3.1. Analysis

In all, 12,279 analysis decisions were rendered. Because there is no objective "ground truth" for the suitability decision, the majority vote for each mark was used as a ground truth by proxy. This produced the confusion matrix shown in Fig. 3 with the disagreement rates in Table 1.

Of particular note are the 599 instances in which individual examiners determined a mark to be suitable for identification

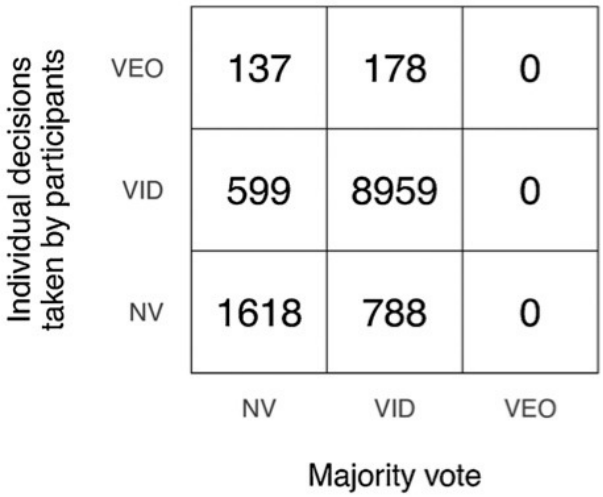


Fig. 3. Confusion matrix obtained in Analysis. The decisions of individual examiners are compared against the majority vote as the expected outcome.

Table 1
Disagreement rates against majority voted conclusions obtained following analysis.

Response	ERD_A	MRD_A
VID	25.4%	9.7%
VEO	2.6%	NA
NV	7.9%	31.3%

when the majority declared it to be of no value, and the 788 instances in which individual examiners determined a mark to be of no value when the majority of examiners declared it suitable for identification. These discrepancies highlight two points. The first is that the threshold for value is not well-defined (and thus, whether or not a mark gets compared could essentially come down to a luck-of-the-draw of which examiner looks at the case). The second is that many examiners seem to be overly risk-tolerant or overly risk-averse in comparison to their colleagues when making suitability determinations. As can be seen in Fig. 4, a high level of variability was observed in the analysis decision. The NV and VID decisions were not highly reliable, while VEO was never the majority voted decision. There were only 7 marks in the study that were unanimously judged to be NV by all participants who viewed them. For a more in-depth examination of the variability in examiner conclusions, see the section *Consensus between examiners* in Appendix A. Supplementary Material.

3.2. Comparison

Overall, 9,460 comparison decisions were rendered. Each case was viewed by an average of 23 examiners. Inconclusive responses were not considered errors against known ground truth because "Inconclusive" could be the most appropriate response for some comparisons depending on the information that was available, even though ground truth is known to the researchers. Without having an expectation of when "Inconclusive" is the best response that is in some way objective enough to count as "ground truth," there is no fair way to judge the correctness of inconclusive decisions against ground truth. Thus, inconclusive responses were omitted entirely from the false positive and false negative rate calculations in our reported calculations and counted as neither correct nor incorrect responses. A summary of all results of analysis and comparison can be seen in Table 2.

After omitting inconclusive responses,7620 comparison decisions remained, shown in the confusion matrix shown in Fig. 5 (a) and the associated error rates against ground truth (Table 3 and Table 4). Results are given in both tables with and without counting the inconclusive decisions for easy comparison with previous studies. Recall that the "with inconclusive" figures given in Table 3 (b) and Fig. 5 (b) do NOT include any trials in which the analysis decision was VEO. This was to preserve an apples-to-apples comparison with the way the FBI/Noblis study calculated their main false positive rate findings on VID comparisons. The results presented in Table 3 (a) do include trials with VEO determinations and are the data that we will use throughout this article to discuss our results, unless we are making a direct comparison with specific FBI/Noblis results, in which case we will specify what conditions are being compared.

There are two ways to compare our results to those obtained in the FBI/Noblis study. The first is by including the inconclusive responses in our data, as we have done in Table 3 (b). By including the inconclusive responses and comparing our data to the FBI/Noblis data, the resulting figures are a FPR of 0.4% for palms versus 0.1% for fingers and a FNR of 7.7% for palms versus 7.5% for fingers. The other way to compare the results is by removing the inconclusive decisions from the FBI/Noblis data so their results

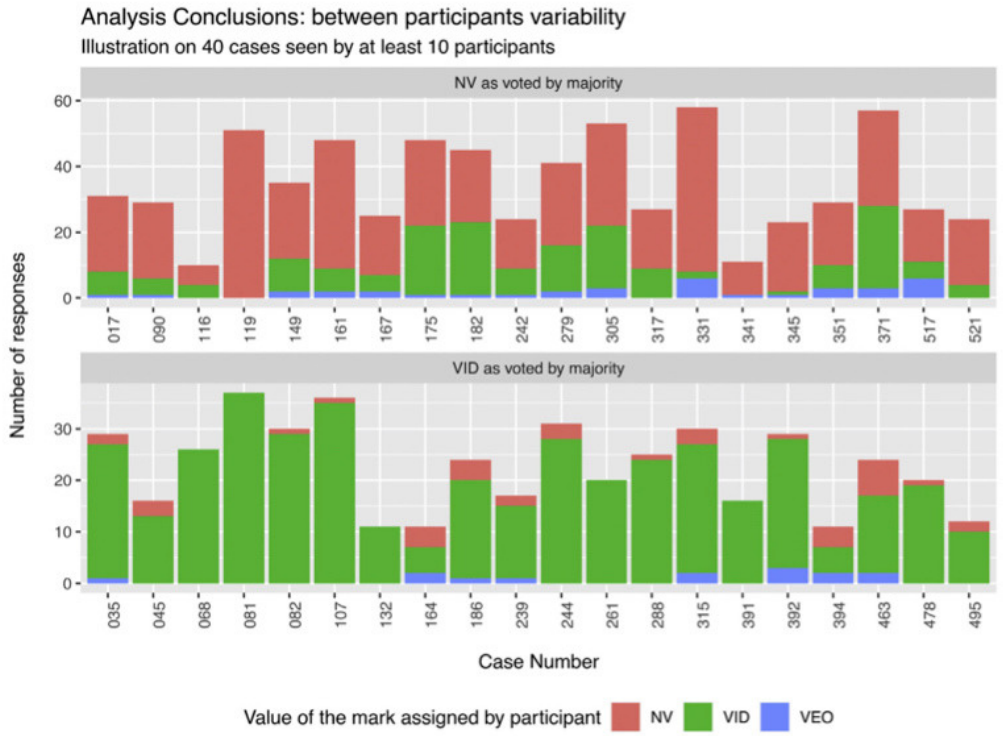


Fig. 4. A randomly selected sample of 40 cases in which at least 10 participants viewed the mark. This selection illustrates the scope of examiner variability observed (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 2
Counts of conclusions reached in comparison as a function of the Analysis conclusion. The VID/VEO cases labelled in comparison as 'not compared' are cases where an analysis conclusion was reached but the comparison was not completed.

Conclusion		Ground Truth		
Comparison	Analysis	Different sources	Same source	**Total**
not compared	NV	789	1 617	**2 406**
	VEO	3	5	**8**
	VID	197	208	**405**
Exclusion	VEO	46	37	**83**
	VID	1 727	515	**2 242**
Identification	VEO	2	39	**41**
	VID	10	5 244	**5 254**
Inconclusive	VEO	42	141	**183**
	VID	733	924	**1 657**
Total	-	**3 549**	**8 730**	**12 279**

are calculated the same way we preferred to report ours. By this calculation, their false positive error rate becomes 0.2% (6/3953) and their false negative error rate becomes 14.2% (611/4314). This is interesting because their false positive error rate for fingers is actually higher than ours (compare to Table 3 (a)) for palms by this reckoning. Although the FPR is not hugely affected by whether inconclusives are included or omitted in both studies, the FNR is. This is likely because between the two studies, there are large differences in the rates of inconclusives for same source and different sources trials. This is just one of many ways in which the two studies should not be directly compared even though they drive at the same basic question. There are just too many unknown variables that differ between finger and palm comparisons and between the two studies.

As would be expected, as the difficulty of the comparison increased, so too did both the false positive and false negative error rates. No erroneous identifications were made when the comparison was judged as easy, although it is worth noting that there were erroneous exclusions made, even on comparisons rated by the examiner as easy. These may well be due to the influence of mindset – where the examiner makes an early decision about the location or orientation of the mark and fails to widen their search parameters after an initial unsuccessful search – which is the topic of a separate publication. It is also worth noting that the positive predictive value remained high, even as the difficulty of the comparisons increased. In other words, in this study, even for comparisons that were judged to be "difficult," when the conclusion was "Identification," 99.5% of the time, the ground truth was that the images originated from the same source.

Note that in casework, the ground truth is not known. Hence as with Analysis, we also measuredparticipant responses as compared to the majority voted comparison conclusion. This leads to the confusion matrix in Fig. 6 and the associated disagreement rates(Tables 5 and 6).

If we focus on the identification conclusions, we obtain a disagreement rate with the majority of 5.9%. Note as seen in Table 6 that the disagreement rate will increase as a function of difficulty, for example, for identifications, it moved from 2.4% when the comparison is qualified as easy to 8.4% when qualified as difficult.

These results can have serious implications for the criminal justice system. Typically, testimony about error rates is given in respect to ground truth – after all, if one is discussing accuracy, the matter of interest is whether the correct conclusion was reached. However, we only have the luxury of knowing ground truth in structured studies like this one. In the real world, we never know ground truth. Thus, constructing an error rate based upon knowledge of ground truth does not give a complete picture about how examiners might perform in the real world, where the

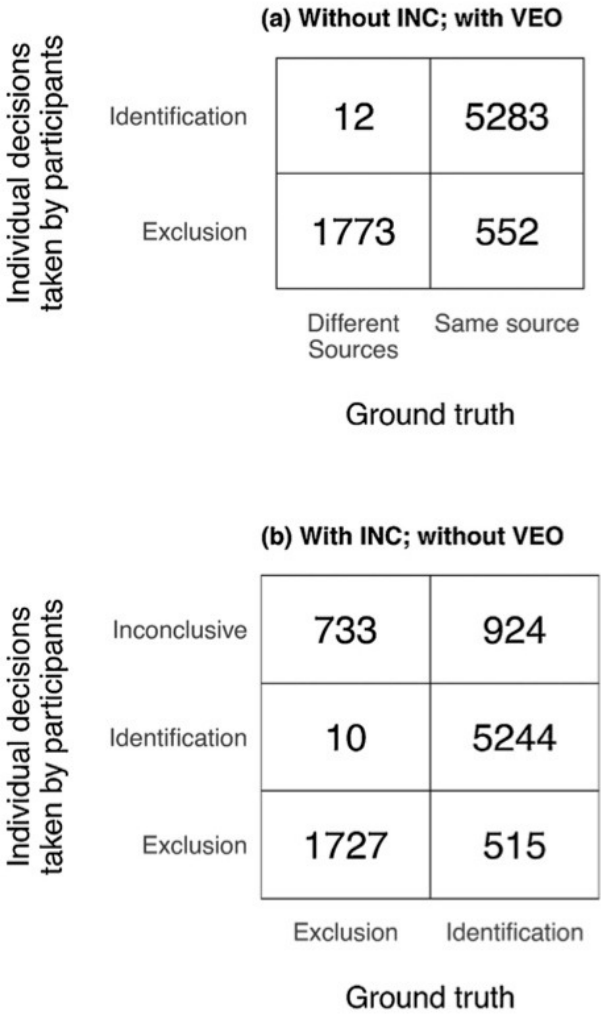


Fig. 5. Confusion matrices obtained in Comparison against the ground truth. (a) without inconclusives; with VEO decisions, (b) with inconclusives; without VEO decisions.

Table 3
Error rates and predictive values against ground truth obtained following Comparison overall as per confusion matrix in Figure 6.

	Conclusion	FPR	FNR	PPV	NPV
(a) Without INC	ID/EXC	0.7%	9.5%	99.8%	76.3%
(b) With INC	ID/EXC	0.4%	7.7%	99.8%	82.7%

truth is uncertain. In the real world, the "rightness" of an answer is generally determined by verification – or whether one or more colleagues agree with the decision.

Thus, it is genuinely concerning that while only 12 false identifications were made against ground truth, there were 45 instances in which someone concluded identification when the majority concluded exclusion. Since in the real world, the majority vote is the only "ground truth" we know, these 45 cases may be assumed in a real laboratory to be true exclusions, in which case the examiner who made the identification would be accused of making a false ID and who is to say they didn't? However, because we know ground truth in this study we can reconstruct how often the identifying examiner was correct and the majority was wrong. It turns out that of the 45 apparent false identifications in Fig. 6, only 9 were actual false identifications from different source trials. (The other three false identifications in the study were in cases where the majority voted Inconclusive, so are not included in the

Table 4
Error rates and predictive values against ground truth obtained following Comparison as a function of the comparison difficulty as reported by the participants. *N* is the number of trials.

	Conclusion	FPR	FNR	PPV	NPV	N
(a) without INC	ID/EXC_Easy	0.0%	3.6%	100.0%	83.5%	2872
	ID/EXC_Moderate	0.8%	12.6%	99.7%	75.1%	3187
	ID/EXC_Difficult	1.1%	16.0%	99.5%	72.4%	1561
(b) with INC	ID/EXC_Easy	0.0%	3.3%	100.0%	84.5%	2933
	ID/EXC_Moderate	0.5%	10.3%	99.8%	80.4%	3724
	ID/EXC_Difficult	0.6%	10.2%	99.7%	84.8%	2496

45). Thus, the other 36 apparent false identifications were actually cases in which the majority vote was wrong. That is, the consensus reached an exclusion decision, but the ground truth was same source. If these were real cases, not only might the examiner who reached an identification conclusion be shamed for an error they didn't make, but the true culprit could potentially go free, able to commit additional crimes.

Despite this potential danger, it is unknown whether a member of the incorrect, exclusion-voting majority would change their mind after being shown the correct ID. In cases where they simply failed to find the correct area or orientation, this is a likely outcome. In cases where there is a real difference of opinion, the outcome is uncertain. Thus, although there were 36 instances in which the majority was incorrect and the person concluding ID was correct, the situation may not be as dire as it initially seems because each case will have different reasons for the initial disagreement and those different reasons will likely lead to different outcomes. Nonetheless, the phenomenon warrants awareness and caution when resolving conflicts of opinion in an operational laboratory. The 6 cases where a ground truth same source pair was voted as an exclusion by the majority are presented in Fig. 7.

Note that the 36 instances in which a same source pair was incorrectly judged to be an exclusion by the majority are distributed among only 4 of the cases presented in this study. The mark in case 368 was judged NV by the majority, but the only vote in comparison went to EXC so that case does not contribute to the misleading false identifications. The same goes for case 100, where there were almost even votes between INC (4) and EXC (6) for a mark judged NV by the majority and nobody identified the mark. The other four cases, however, illustrate well the variations that we may observe between participants with opposing conclusions (ID versus EXC). Case 423 shows an almost even split between ID (28), INC (28) and EXC (29). This case was judged as "difficult" by the majority.

Naturally, there is no easy solution to this problem. Without knowing ground truth, we cannot suggest in casework that an identification should be reported when one examiner makes the identification and multiple other examiners say it was an exclusion. However, it is worth being cautious and acknowledging that the group is not always right.

In Appendix A. Supplementary Material, we present an analysis of the variability in examiner conclusions. As expected, apart from the inconclusive decisions, the level of consensus is much higher in cases qualified as easy compared to cases qualified as more difficult. There is little consensus in the inconclusive decision, regardless of the difficulty of the comparison. Finally we note that unanimity on the identification conclusion is reached by participants in 102 same source trials (25.5%). Conversely unanimous exclusions are obtained in 9 different sources trials (7.1%). This illustrates that it is easier to get consensus on identifications than on exclusions, which could be a result of the relative difficulty of

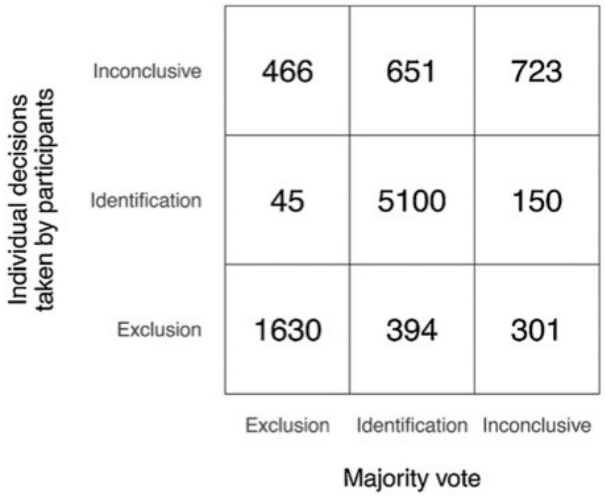


Fig. 6. Confusion matrix obtained in Comparison against the majority voted opinion.

Table 5
Disagreement rates against majority voted opinion obtained following Comparison, overall as per confusion matrix in Figure 7. The Response column indicates the response that is being considered.

Response	ERD_C	MRD_C
Identification	5.9%	17.0%
Exclusion	9.5%	23.9%
Inconclusive	13.5%	38.4%

Table 6
Disagreement rates against majority voted opinion obtained following Comparison, as a function of the comparison difficulty as reported by the participants. *N* is the number of trials for each majority vote conclusion.

Response	Difficulty	ERD_C	MRD_C	N
Identification	Easy	2.4%	4.7%	2449
	Moderate	4.7%	19.3%	2422
	Difficult	8.4%	36.3%	1274
Exclusion	Easy	3.6%	3.8%	453
	Moderate	11.9%	19.2%	1030
	Difficult	13.7%	45.0%	658
Inconclusive	Easy	2.2%	41.2%	80
	Moderate	12.5%	43.5%	386
	Difficult	32.1%	35.3%	708

reaching exclusion decisions, or of differing agency policies that limit for some when an exclusion may be reached.

Among the 102 unanimous identification trials, only 4 involved marks that were declared NV by the majority of examiners. The other 98 marks were voted VID by at least 75% of participants who viewed the trial. The same applies to the unanimous exclusion trials: only 1 involved a mark that was declared NV by the majority of examiners. The other 8 marks were voted VID by at least 84% of participants who viewed the trial. This illustrates that it is easier to reach unanimous comparison conclusions on marks unanimously declared VID.

3.2.1. Identification conclusions

Twelve false positive errors were made out of1785 different source trials where a comparison conclusion was reached (excluding inconclusive decisions), resulting in a false positive error rate of 0.7% (the false positive error rate for the analogous data obtained by FBI/Noblis was 0.2%). The positive predictive value for the study is 99.8%. No two false positive errors were made in the same case (i.e.on the same mark-exemplar pairing), and no false positive errors were made by trainees.

Eight examiners committed false positive errors in this study. Four participants made one false positive error each, while the other four made two false positive errors each. Although the sample size of people committing false positive errors is too small to perform a rigorous statistical analysis, there are some interesting commonalities in the data that are worth observing.

First, and possibly most importantly, although 94.8% of participants reported currently working as latent print examiners, two of the examiners who each made two of the false positive errors answered "No" to the question "Are you currently or have you previously been employed as a latent fingerprint examiner?". Together these two examiners were responsible for 1/3 of the false positive errors made in the study, yet they were from a pool that represented only 2.9% of the study population. Another interesting pattern is that although only 18.1% of the study participants reported working for an agency outside of the U.S., 6 of the 12 false positive errors (50%) were made by participants from non-U.S. agencies, a disproportionate number to their presence in the study population. Altogether, 8 of the 12 false positive errors (66.7%) were made by participants who were either non-active LPEs, non-U.S. examiners, or both. This leaves 4 of the 12 false positive errors that were committed by U.S. examiners who are currently active LPEs. For comparison, the FBI/Noblis study reported 96% participation by current LPEs, but only 1% participation from non-U.S. examiners.

Ideally, it would be helpful to find a pattern to the false identifications that would allow us to understand how they happen or predict when they might happen. Unfortunately, no discernable pattern was evident in the trials themselves for the 12 false identifications in this study. None of the obvious potential factors were able to explain these errors. Of the cases that had false positive errors, 5 came from CNM2 pairings and 7 came from CNM1 pairings. The substrates included paper, plastics, glass, and tape. The development techniques included powder, ninhydrin, CA, and dye stain. Only one of the cases had a partial tonal reversal, and two of them showed light ridges throughout. The cases were nearly evenly distributed between interdigital and thenar, with one that was from the hypothenar. The sizes of the marks were 3 Large, 5 Medium, and 4 Small. None had been rotated prior to being presented in the trial.

Fig. 8presents the conclusions of all the examiners who viewed each case in which a false identification was made. It illustrates that there was no pattern whereby most examiners thought the mark was of value, or most examiners reached an inconclusive decision, or nobody was able to reach the correct exclusion decision or anything that one could point to as the obvious warning sign that this was a mark that was likely to be falsely identified.

3.2.2. Inconclusive conclusions

The inconclusive results in this study exhibit an interesting trend in that nearly double the percentage of inconclusives were reported for different sources trials (30.3%) as for same source trials (15.4%), a trend that is visible in Fig. 2. This is in opposition to previous studies in which inconclusive decisions were more than

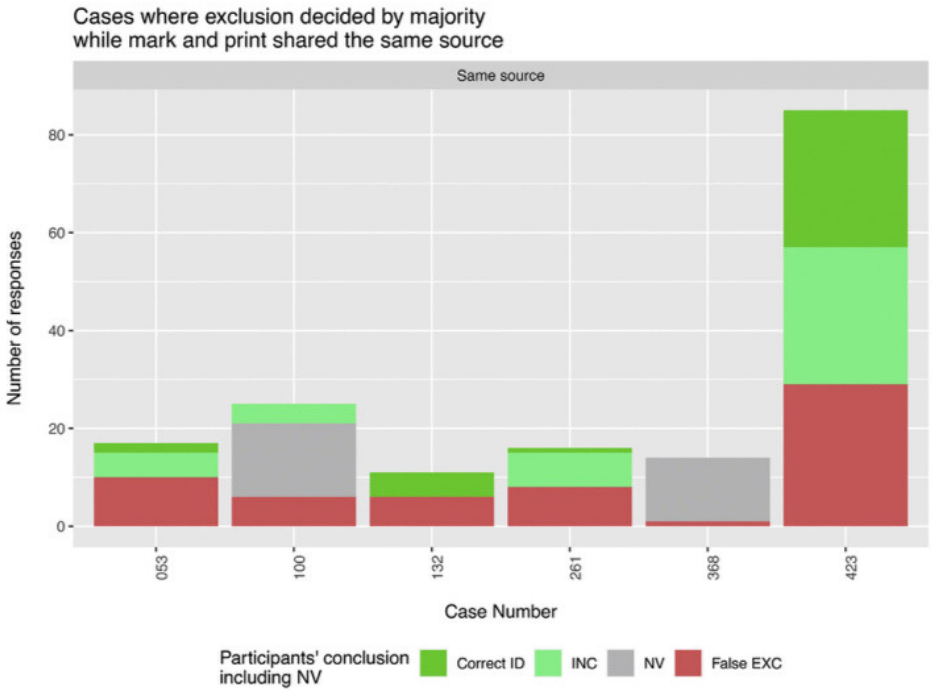


Fig. 7. Conclusions reached by the participants in the six cases where the majority conclusion (EXC) was different from the ground truth (same source) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

twice as likely in same source trials as in different sources trials. For example, in the FBI/Noblis study, the inconclusive rate for same source trials was 47.3% whereas for different sources trials it was only 20.7%. Similarly in Langenburg's informed judgements study [28], the inconclusive rate for same source trials was 26.1%, yet for different sources trials it was only 10.5%. It is uncertain why this reversal of inconclusive rates has occurred in this study; however, it may have something to do with the fact that the two aforementioned studies utilized only finger impressions whereas this study focused exclusively on palmar impressions. The challenges created by frequently missing orientation and location information in palmar impressions may make examiners more hesitant to reach an exclusion decision for these comparison types. Additionally, many agencies now have policies that prohibit making an exclusion decision absent anchor areas and target groups. This may have forced a number of examiners to make more inconclusive decisions on palms, where this information is often absent, than they would on fingers, where it is more often present.

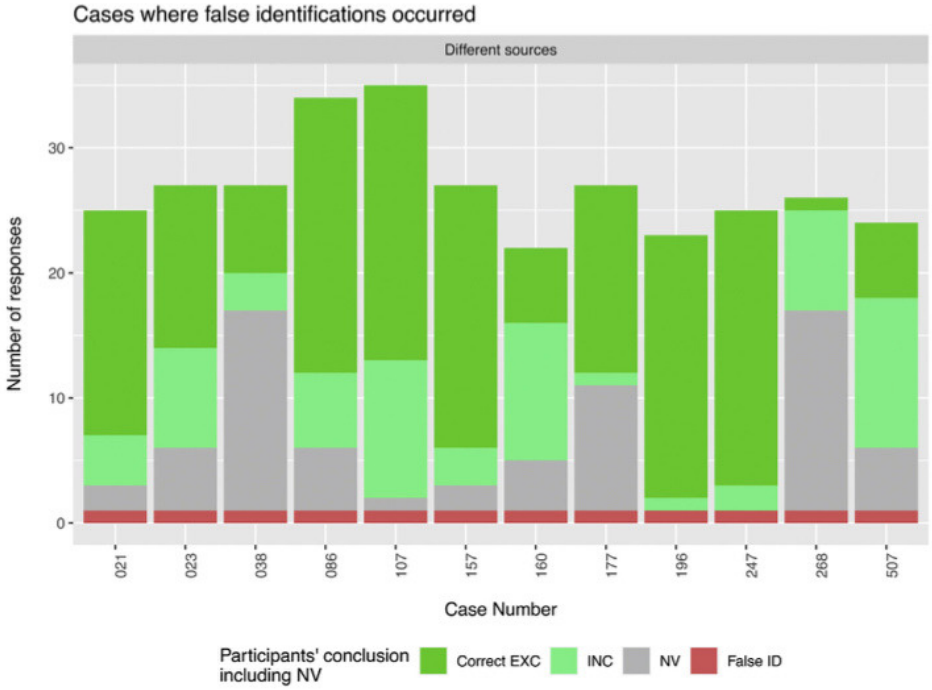


Fig. 8. Conclusions reached by participants in the trials where a false identification was reported (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 7
Conclusions reached respectively in same source and different sources trials including INC.

Trials	Conclusion	N	Rate
Same Source	Correct Identification	5,283	76.6%
	False Exclusion	552	8.0%
	INC_but with corresponding features noted	351	5.1%
	INC_due to insufficient information	591	8.6%
	INC_due to no overlapping area	123	1.8%
Different Sources	Correct Exclusion	1,773	69.3%
	False Identification	12	0.5%
	INC_but with corresponding features noted	56	2.2%
	INC_due to insufficient information	544	21.2%
	INC_due to no overlapping area	175	6.8%

Another possible reason for the shift could be that this study required searching and, often, orienting of the mark, whereas the other studies utilized a 1:1 comparison design where the marks were all correctly oriented. This could make examiners more hesitant to exclude than they might be in studies where there had more confidence that they were looking in the right place. Finally, examiners simply typically receive less training and practice on palmar impressions, which may exacerbate the issue as they may be less likely to pick up on orientation and location cues (even when they are present) than in finger impressions. All these reasons strengthen our argument that a true apples-to-apples comparison cannot be done between the results of this study and previous error rate studies, due to inherent differences between the finger and palm comparison tasks that each tested.

Among the inconclusive decisions, Table 7 shows that on same source trials 351 (5.1%) were reported inconclusive but with corresponding features noted. These represent cases that could have been expressed as providing support for same source that would in most agencies be left out of the criminal justice system.

However, when we look at trials from different sources (Table 7), we have 56 cases (2.2%) when under the same conditions, we would have provided misleading evidence.

It is interesting to observe for the trials where a false identification was reported 6 participants indicated corresponding features as shown in Table 8. These cases, if reported as supporting same source due to their corresponding features, would have been misleading.

3.2.3. Exclusion conclusions

The 552 false negative errors were made out of5835 same source trials where a comparison conclusion was reached (excluding inconclusive decisions), resulting in a false negative error rate of 9.5% (the false negative error rate using the analogous data obtained by FBI/Noblis was 14.2%). The rate of erroneous exclusions made by trainees did not differ from the rate of the general study population. The negative predictive value for the study was 76.3% compared to a negative predictive value of 86.6% using the analogous data from the FBI/Noblis study. It is interesting that our NPV was so much lower than the FBI/Noblis's because our proportion of Inconclusive responses out of all trials inwhich a comparison was completed (19.5%) is also lower than theirs (37.2%). This means that our palm participants were reaching identification or exclusion decisions more often than the FBI/Noblis participants looking at fingers, yet when our participants made an exclusion decision, they were more often incorrect. These data suggest that there are fundamental differences between finger and palm comparisons, both in people's risk tolerance for reaching definitive conclusions and their accuracy in reaching exclusion decisions. However, we must once again caution the reader against putting too much stock in these direct comparisons between the two studies because there may be many unknown variables that differ between the two that impact the results and it is not certain that differences in reported performance measures are due solely to thefact that one study used fingers and the other palms.

Table 8
Reasons selected for Inconclusive decisions (by all participants viewing these trials and reporting Inconclusive) in the trials where a false identification was reported.

Reason given for INC	Total
But with corresponding features noted	6
Due to insufficient information	49
Due to no overlapping area	15

Only 69 of the 204 participants (33.8%) who completed at least one comparison had zero false exclusions, meaning 66.2% of participants who completed at least one comparison made at least one false exclusion error. However, many participants only completed a handful of trials. Taking only participants who completed 10 or more comparisons, the number of participants with zero false exclusions drops to 50 out of 175 (28.6%), meaning that 71.4% of participants who completed 10 or more comparisons committed at least one erroneous exclusion error.

Additionally, of these 50 participants who completed 10 or more comparisons and had zero false exclusions, 11 had 20 or more inconclusive decisions and 5 or fewer true exclusions. This indicates that these participants had a tendency to go inconclusive more readily than they would exclude. Each participant who completed the study received 22 different source trials, thus they had 22 opportunities to make a true exclusion.

Out of 126 different source trials, only 9 received unanimous exclusion decisions from all examiners who viewed them. Out of 400 same source trials, 193 received at least one erroneous exclusion decision.

Fig. 9 shows the distribution of the number of erroneous exclusions made for each same source trial.

Of these cases, 92 had only one erroneous exclusion, indicating that verification of exclusion decisions could drastically reduce the number of erroneous exclusions that are reported in regular casework. On the other hand, 101 of the cases had two or more erroneous exclusions and some had many, many more. These are unlikely to be caught during verification, since multiple people reached the same incorrect conclusion. One way to reduce these erroneous exclusions would be through the addition of policies that regulate when an exclusion decision may be made. Another may be through the use of a case AFIS system as an additional check on exclusion decisions particularly to reduce errors caused by mis-orientation.

In 52 of the cases with erroneous exclusions, a large number of participants (20 or more) reached the correct identification conclusion. We explore this phenomenon of erroneous exclusions when many participants reached the identification conclusion in depth with sample images and commentary in Appendix A. Supplementary Material.

Three hundred fifty-six erroneous exclusions were made when no reliable anchor (defined as a core, delta, primary crease large enough to tell which one it is, thumb bracelet, recurve, or vestige) was present. Some agencies follow a policy where an exclusion may not be made without a reliable anchor – these erroneous exclusions should be significantly reduced by following such a policy. In 108 same source trials, no erroneous exclusions were made, but there were differences of opinion between identification and inconclusive among examiners.

3.2.4. Questionable conclusions

While it is useful to know error rates in relation to ground truth, reporting the majority voted conclusions for each trial also allows

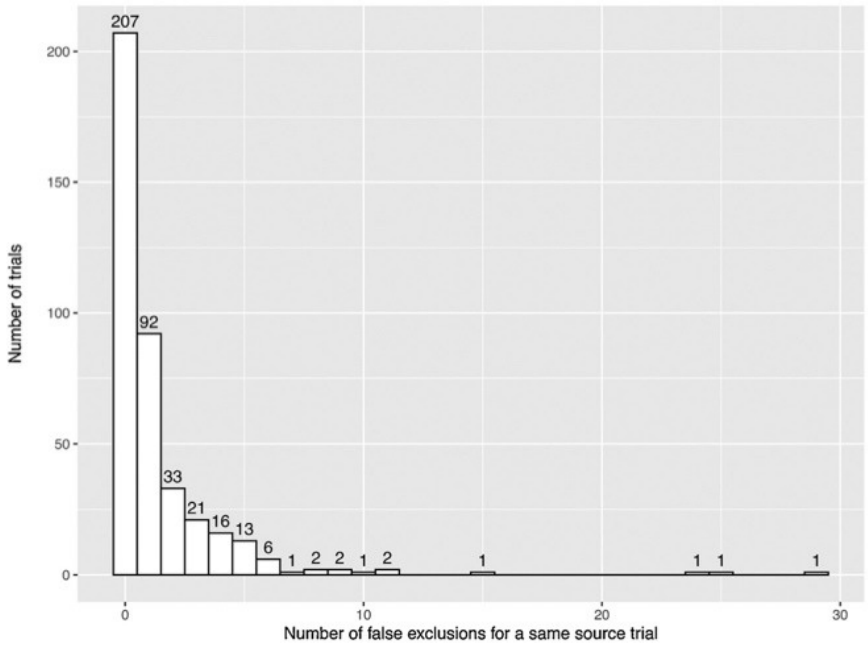


Fig. 9. Distribution of the number of erroneous exclusions made for each same source trial.

us to explore the notion of questionable conclusions. Just because a participant reaches a conclusion that matches ground truth, it does not necessarily follow that the decision was appropriate for the data. If the majority of examiners reach an inconclusive decision, but some examiners reach a definitive decision (identification or exclusion), were those definitive examiners "super-examiners" or were they simply too risk-tolerant and making decisions that were not sufficiently supported by the available data? Conversely, if an examiner reports inconclusive when the majority reached the correct definitive conclusion (i.e.the conclusion that matches ground truth), was the inconclusive examiner the lone voice of reason to exhibit caution, or were they being too risk-averse? Examining these data can help an individual examiner to gauge his or her own sensitivity and see whether they are missing conclusions most of their colleagues would be able to make, or whether they are pushing the envelope, making conclusions that most of their colleagues would not support.

In this section, we will examine three circumstances: cases in which the majority reached an Inconclusive decision while some examiners identified or excluded; cases in which the majority reached the ground truth identification decision whereas a small number reported an inconclusive; and cases in which the majority reached the ground truth exclusion decision whereas a small number reported an inconclusive.

3.2.4.1. Inconclusive majority. First, we will review two cases in which the majority decision was "Inconclusive". Fig. 10 illustrates case_0146, a same source trial. Fig. 11 gives a starting point for the comparison to aid the reader. In this case, the decisions were 32 INC, 11 EXC and 11 ID. It is easy to see how this identification could be missed. There is very little common area actually overlapping between the two impressions; however, there are at least 10 minutiae in common visible once the correct area is located. With a comparison such as this, it is difficult to say what the "correct"

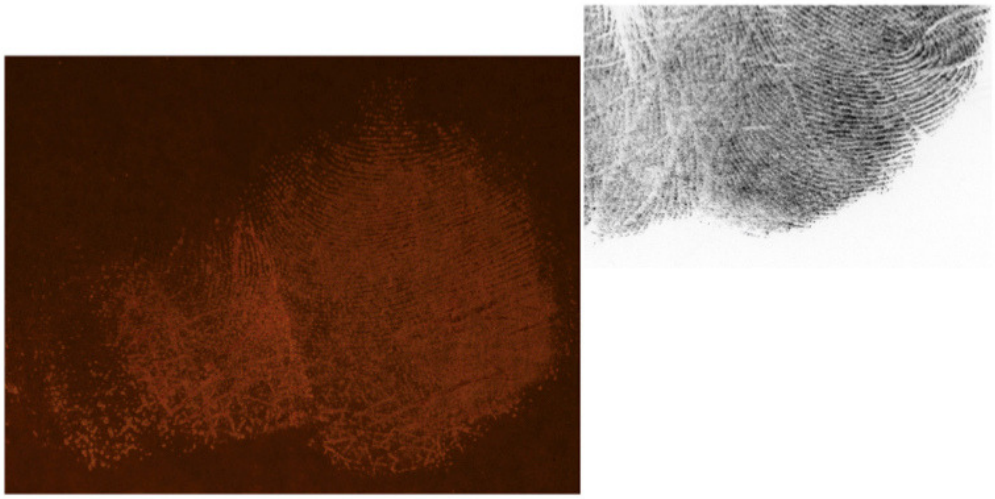
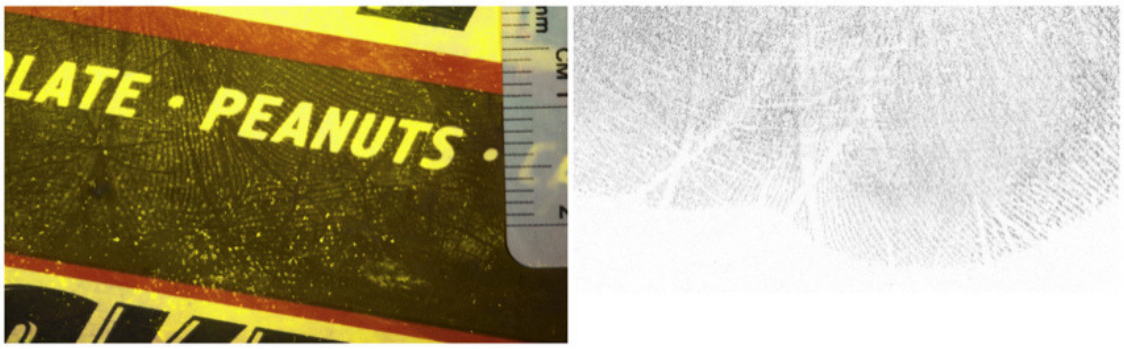


Fig. 10. Case 0146. The mark is presented on the left and the print is on the right. For space and readability, only the portion of the print that was the source of the mark is printed. The mark is presented here in the correct orientation, as it was in the study.



response should have been. Those who are willing to identify on this pair of images may say that the people who failed to make the ID should have looked harder, while those who would report an inconclusive decision may say there was insufficient information to support an identification. These borderline cases quickly devolve into a philosophical debate when there are no clear standards of sufficiency and highlight the difficulties engendered by counting inconclusive responses as either correct responses or errors in error rate studies such as this one. They can, however, be used as an effective barometer of examiners' risk tolerance.

Case_0224, a same source trial, presents a similar challenge and is presented in Figure 12 (clean images) and Fig. 13 (starting point). In this case, the decisions were 33 INC, 2 EXC, and 7 ID. Once again, there is a very narrow band of clear, overlapping area between the two impressions and there are at least 10 minutiae as well as some creases in common visible once the correct area is located.

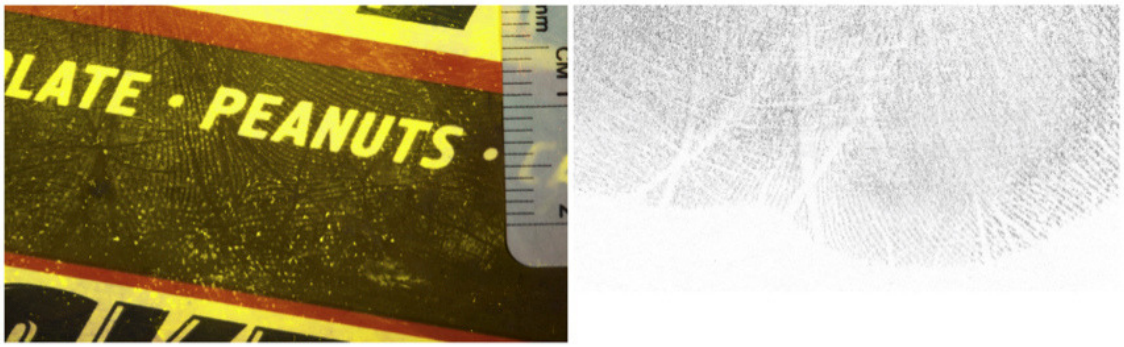


Fig. 12. Case 0224. The mark is presented on the left and the print is on the right. For space and readability, only the portion of the print that was the source of the mark is printed. The mark is presented here in the correct orientation, as it was in the study.

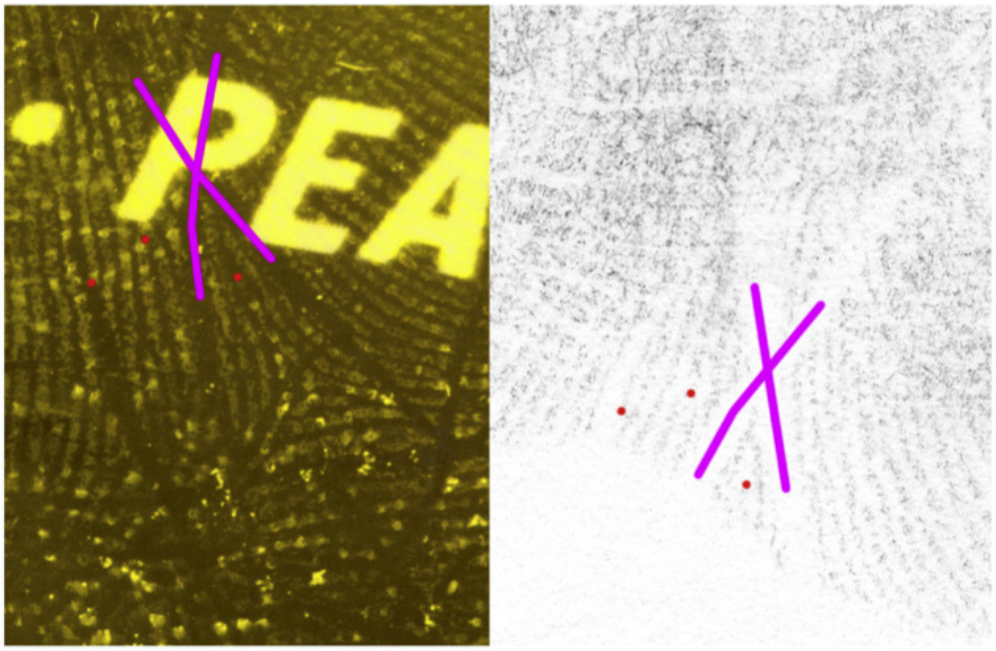


Fig. 13. Comparison starting point for case 0224. Two creases and 3 minutiae in common have been marked out to get the reader started (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

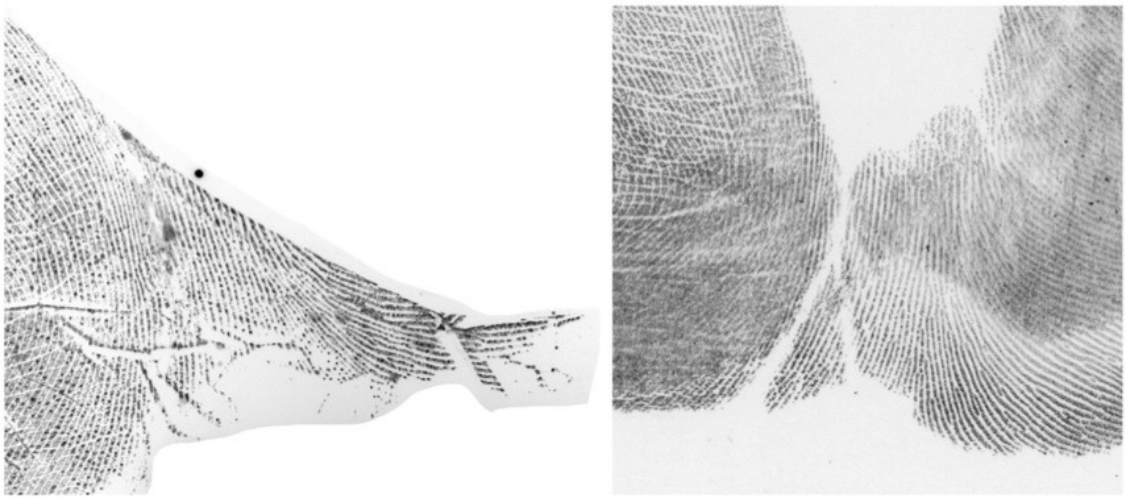


Fig. 14. Case 0458. The mark is presented on the left and the print is on the right. For space and readability, only the portion of the print that was the source of the mark is printed. The mark is in approximately the correct orientation and is printed here as it was presented in the study.

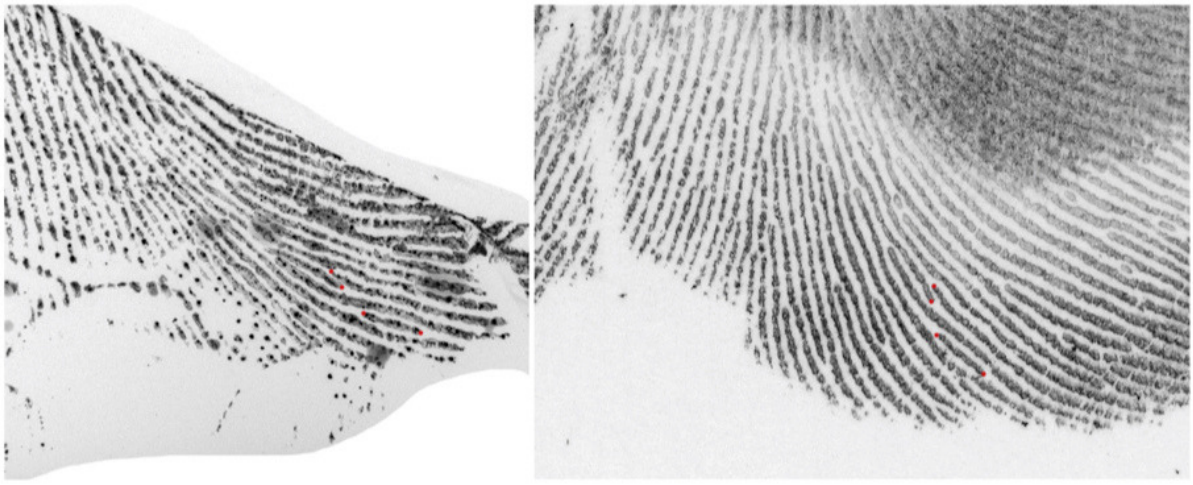


Fig. 15. Comparison starting point for case 0458. Four minutiae in common have been marked in red to get the reader started (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

However, in this case, the notes provided by some participants help to illustrate the range of examiners' comfort with making an identification decision. While not everyone left written notes on this comparison, 15 inconclusive participants wrote comments that indicated that the comparison was incomplete due to the quality of the print or that they would request better standards in casework. Three inconclusive participants' comments made it clear that they simply missed the overlapping area (for example, one participant commented, "knownare not legible and does not show the area needed for the comparison"). Another 6 inconclusive participants' comments made it clear that they found some features in common, but did not consider them sufficient to support an identification decision. Among the participants who concluded Identification, 2 expressed that they would have preferred better known(e.g., "ID can be made with this known, but I still would have asked for a better one.") while 1 claimed to have made the identification using only crease information.

Interestingly, 3 participants made almost identical annotations; all 3 chose the same 10 minutiae (although one also annotated 2 additional minutiae) and all 3 annotated the same creases. Yet, two of them identified while the third reported an inconclusive. The participant who reported an inconclusive decision remarked, "Additional knowns required (poor quality); 3rd level in agreement; insufficient 2nd level detail" while the 2 who identified commented, "Creases helped for comparing 2nd level detail. Exemplar is terrible" and "The exemplars were not that good but fortunately 3rd level detail was present" respectively. All 3 participants essentially agreed that the exemplars were poor, that the third level detail was in agreement, and that there wasn't a whole lot of second level detail in agreement, but for 2 of them, this was sufficient, and for the other, it was not.

3.2.4.2. Identification majority. The cases in which the majority reached the ground truth identification decision are interesting because there is a wide range of difficulty represented in these cases, which could affect the appropriateness of the majority decision, particularly when examined in conjunction with the participants' work habits. The first case we will review is Case_0458, a same source trial in which the decisions were 10 ID and 1 INC. Of the participants viewing this case, 7 judged it as easy or very easy. Fig. 14 presents Case_0458. While it may be difficult for some to locate the area in common, once a target group is located (Fig. 15), there is ample clear information in common between the two impressions.

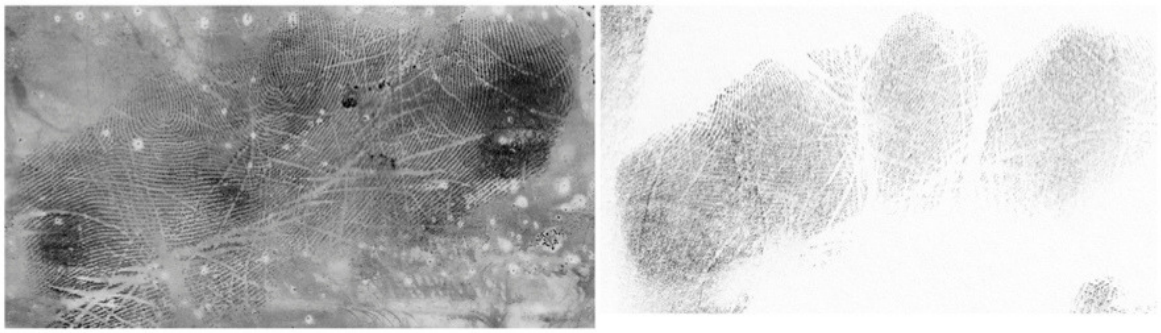


Fig. 16. Case 0488. The mark is presented on the left and the print is on the right. For space and readability, only the portion of the print that was the source of the mark is printed. The mark is presented here in the correct orientation, as it was in the study.

In the second case we will review, it is less clear which are the questionable conclusions. Fig. 16 presents Case_0488, a same source trial in which the decisions were 15 ID and 4 INC. Although in this case, there was still a clear majority of participants who reached an identification conclusion, the comparison is more ambiguous, as are the work habits of the participants who declared an ID. Seven of the participants who declared an ID annotated all of their minutiae in the mark after having begun the comparison and seen the print, a practice that is generally considered to be risky and not recommended.

An additional 6 identifying participants annotated between half and all but 1 of their minutiae on the mark after beginning the comparison. Two identifying participants did not annotate any minutiae at all in support of their decision. Although annotating minutiae was not required in this study, this was a challenging comparison (declared as moderate or difficult by respectively 8 and 9 participants) and annotating minutiae during comparison would assist in reaching a reliable conclusion.

In this case is it less clear whether the people who reached an Inconclusive decision were too risk-averse to make what was a reliable ID, or whether the people who reached an ID decision relied too much on circular reasoning to support their conclusion and should have gone with the more defensible inconclusive decision. In this case, of course, the print was the more challenging and degraded of the two images, which makes it easy to understand why participants did not feel it necessary to begin annotating minutiae until after they had seen the print. If one gives the benefit of the doubt that they properly worked from the lower quality image (the print) to the higher quality image (the mark), then the identifications are supportable and justified, and it is the inconclusive participants who were too risk-averse in comparison to the majority of their peers.

3.2.4.3. Exclusion majority. There were also cases in which the majority reached the ground truth exclusion decision, but 1 or 2 participants reported an inconclusive. In most of these cases, there were clear minutiae present in the mark and the corresponding areas of the print were available and sufficiently clear. We recognize that in some cases, agency policy may have prohibited an exclusion decision if no anchor was present (as we ourselves have advocated elsewhere in this article). However, in many cases, there were examiners who did not reach the exclusion decision even despite the presence of clear anchors. We present here one such example. Case_0500 (Fig. 17) received 25 EXC decisions and 1 INC decision. Due to privacy requirements, we are only able to reproduce the mark here, but the reader will note that there are multiple clear anchors and clear minutiae present, allowing this mark to be easily oriented and located as the interdigital area. The corresponding area of the print (not pictured) was very high quality, both clear and complete. In these types of cases, the exclusion decision should be easy to reach and an inconclusive decision is questionable.

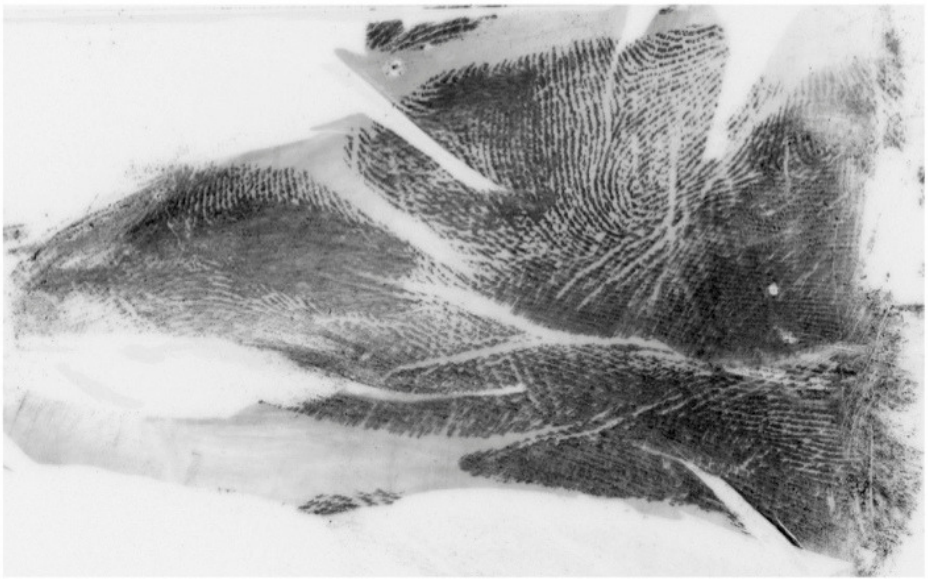


Fig. 17. Case 0500. This mark was presented in the correct orientation. Due to privacy concerns, the print is not shown.

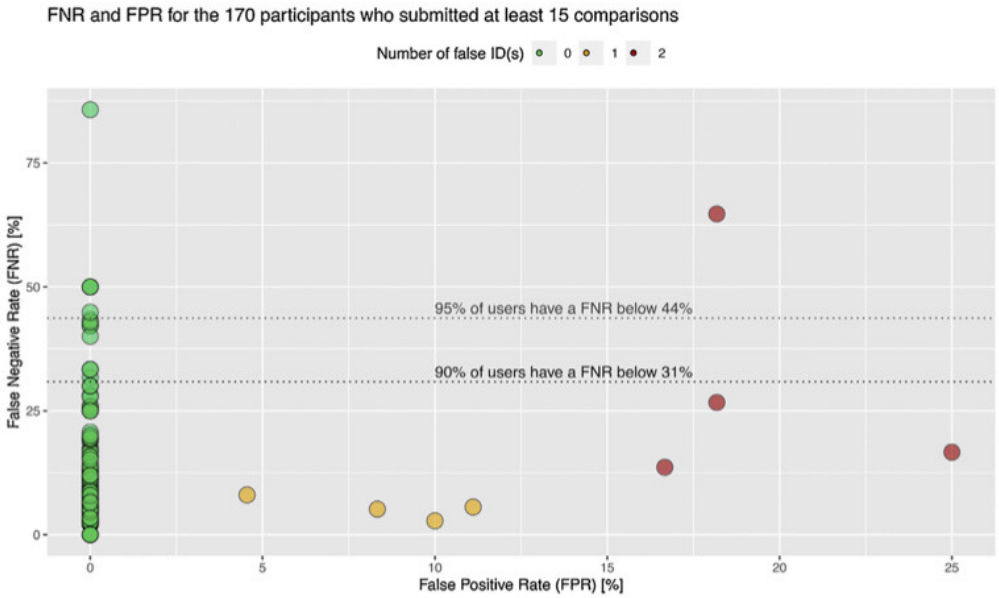


Fig. 18. False positive rate (FPR) against false negative rate (FNR) for participants who submitted at least 15 comparisons (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 9
False Negative errors and error rates as a function of the size of the palm mark.

Size	False negative errors	Same source trials	False negative error rate
L	177	1884	9.4%
M	272	3624	7.5%
S	103	1392	7.4%

Table 10
False Negative errors and error rates as a function of the difficulty of the palm mark as set by one of the authors (HE).

Difficulty	False negative errors	Same source trials	False negative error rate
NV	56	435	12.9%
Inc	17	228	7.5%
Easy	51	1604	3.2%
Medium	123	1851	6.6%
Hard	268	2549	10.5%
Very hard	37	233	15.9%

3.2.5. False positive and false negative error rates for individual participants

The overall error rates in the previous sections consider all the participants' performance jointly. However, all participants are not equal when it comes to their error rates. Fig. 18 presents the false negative rate (FNR) and false positive rate (FPR) associated with each participant. Here only the 170 participants who submitted at least 15 comparisons are plotted. We can see that one participant had a false positive error rate of 25% with 2 false identifications. This is an unusual case as is the participant with a false negative rate above 75%. We note that 95% of the participants have a false negative rate below 44% and 90% of them are below 31%. It is also important to note that of these 170 participants, 46 (27%) made no errors, hence they had perfect accuracy in these trials.

3.2.6. False negative error rates stratified by size, difficulty, and palm area

While historically only a single false positive and false negative error rate have been reported for discipline error rate studies, an aim of this research was to examine whether different error rates were observed depending on the size of the mark, the difficulty of the comparison, or the area of the palm from which the mark originated. This information could be of use to examiners testifying in court, who would be able to cite the error rates that most closely resembled the conditions in the case at hand. Unfortunately, as only 12 false positive errors were made in this study, it would not be feasible or responsible to try to calculate error rates using such small numerators, once those 12 errors had been parsed into sub-categories. Thus, we only address the stratified error rates of false negative errors within this paper.

In the following tables, the "Same Source Trials" columns include only cases in which a comparison conclusion was rendered and exclude inconclusive decisions. The summary of false negative error rates stratified by the size of the mark is shown in Table 9. It is apparent from these data that the size of the mark is not a determining factor in whether or not a false negative error will be made.

The summary of false negative error rates stratified by the difficulty of the comparison (as rated by one of the authors) is given in Table 10. Here more of an effect is observed. It is clear that as the difficulty of the comparison increases, so does the false negative error rate. These data support the idea of defining thresholds for comparison difficulty and documenting these levels in case notes.

The summary of false negative error rates stratified by the area of the palm from which the mark originated is given in Table 11. Again, a distinct effect of palm area is noted. There are clearly areas of the palm that pose a greater challenge to examiners in locating marks and those that pose less challenge.

3.3. Relationship between FNR and participants' information

In order to explore the relationship between information associated with the participant and their potential false negative rate (FNR), we decided to test whether we could reasonably predict a range of FNR given the participant's information used as predictors. Our objective is not to predict the FNR in a regression mode, between 0 and 100%, but to use participant information as a potential detector for examiners who may perform at a high FNR. Indeed, if based on examiner's information, we can reasonably predict if examiners may operate at a high FNR, we could suggest conditions that could favour a

Table 11
False Negative errors and error rates as a function of the palm area of the mark submitted.

Palm area	False negative errors	Same source trials	False negative error rate
Bottom half	6	232	2.6%
Carpal delta	10	181	5.5%
Center	11	78	14.1%
Full palm	1	82	1.2%
Hypothenar	95	1201	7.9%
Int	161	3008	5.4%
Int/Center	1	20	5.0%
Int/Hypo	10	195	5.1%
Int/Thenar	5	53	9.4%
Thenar	220	1522	14.5%
Thenar web	4	38	10.5%
Writer	28	290	9.7%

reduced FNR. FNR has been divided in two classes based on a threshold set at 5%. Below that threshold (BelowT), the examiner is performing with a FNR between 0 and 5%, above it (AboveT), the examiner is operating with a FNR above 5%.

Taking advantage of the `caret` package, we trained a series of classifiers based on the 154 examiners who submitted more than 15 comparisons and for whom we had associated personal information. We used a 10-time repeated 10-fold cross-validation method. The information associated with the participants constitutes 46 predictors. The retained classifiers ranged from simple tree model (CART), to K-Nearest Neighbors, Partial Least Squares, Penalized Multinomial Regression, Generalized Linear Model, Random Forest, 3 types of Neural Networks, boosted trees (GBM, XGBoost, C5.0) and Support Vector Machines.

All classifiers performed with an accuracy between 52% (AvNN) and 59% (CART). Conscious that we were operating with a large number of predictors, we have applied a Recursive Feature Elimination (RFE) procedure[29] using the Random Forest model. It raised the Random Forest model accuracy from 54% to 65%. Note that if we had classified the examiners' FNR (BelowT verus AboveT) by tossing a fair coin, we would have obtained an accuracy of 50%. The fact that the highest accuracy we achieved was 65% shows that we were unable to derive a robust set of personal information that could act as good predictors. In other words, we cannot predict whether or not an examiner will perform with a FNR rate higher or lower than 5% based on their recorded personal information.

For the best performing model (Random Forest with an accuracy of 65% based on an FRE-reduced set of four predictors), an analysis of variable importance also puts accreditation first, followed by whether or not the examiner followed a formal training program, agency policy towards exclusions, and proficiency testing. Besides the reassuring observation that accredition, formal training, and proficiency testing are not harming the performance of examiners, the limited predictive accuracy of the model does not allow us to derive any strong operational recommendations.

3.4. Relationship between image quality measures and accuracy

All submitted marks have been measured by quality measure algorithms. It is of interest to test if, based on these quality measures used as predictors, we can predict the majority voted analysis decisions.

We used a 50-time repeated 10-fold cross-validation method to optimize and select the best model among the ML models tested previously. Model training was done on a training set made of 50% of the data points (randomly drawn). An initial set of predictors was chosen by removing predictors that were fully correlated to another or showing pairwise correlations above 0.8. LFIQ2 predictors (lfiq2_1 and lfiq2_2) were retained alongside with a set of predictors from LQMetrics: AreaOfImpression, AreaOfGoodLevel3, LargestContiguousAreaOfGoodRidgeFlow, AutomatedMinutiaeGreenOrBetter, AutomatedMinutiaeYellowOrBetter, OverallQuality, OverallClarity.

Obtained accuracies ranged from 85% (Neural Networks) to 87% (Gradient Boosting Machine – GBM). We applied an RFE procedure to the GBM model to reduce the number of predictors while maintaining or increasing accuracy. An accuracy of 88% was reached using only 6 predictors (dropping LargestContiguousAreaOfGoodRidgeFlow, AutomatedMinutiaeYellowOrBetter and OverallQuality). The most important variable is lfiq2_1 followed by AreaOfGoodLevel3, as shown in Fig. 19.

We have applied the optimized GBM model to the remaining 50% of the datapoints to test the model. This led to an accuracy of 86%. Fig. 20 illustrates the predictions obtained between two panels for each of the analysis conclusions as voted by majority. The cases are distributed using on the*x*-axis the level of consensus reached by the participants and on the y-axis the log10 of the LFIQ2 quality metric (lfiq2_1). We can see that the GBM model is very efficient when the consensus between participants is above 0.75 and the image quality is high quality (lfiq2_1 above 1 in log10). Below these levels, the modelstruggles to assign the appropriate analysis conclusion, as would the participants. Recall that all cases below a consensus of 1 would have some level of disagreement

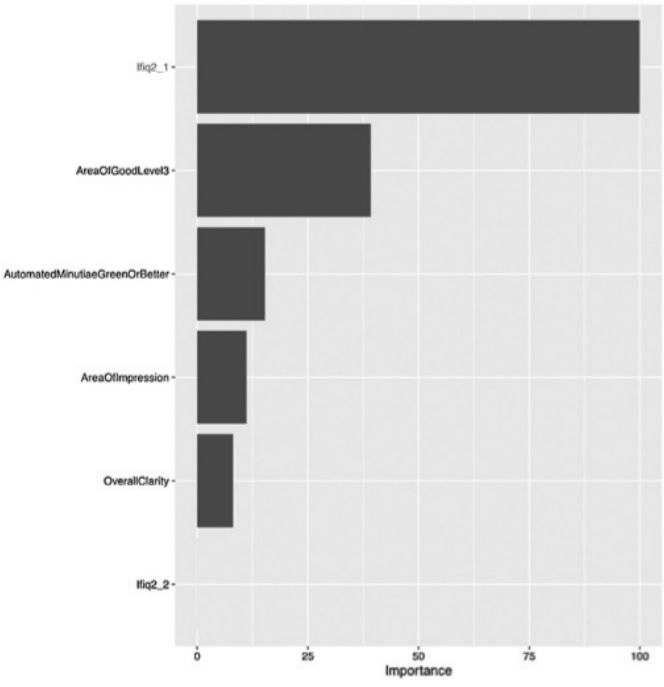


Fig. 19. Variable importance for the best performing GBM model trained with 6 predictors (accuracy = 88%).

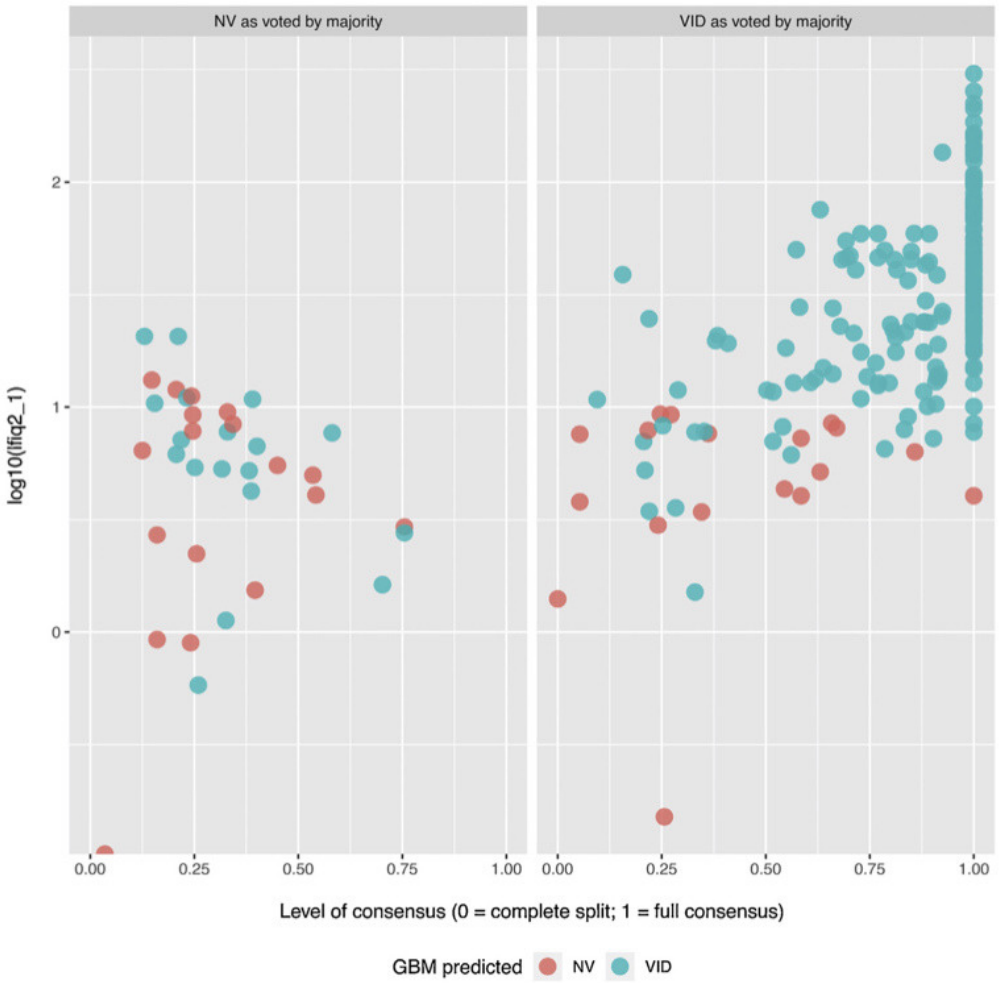


Fig. 20. Variable importance for the best performing GBM model trained with 6 predictors (accuracy = 88%) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

between participants, thus some of them are not reaching the correct conclusion. The accuracy of participants when we compare their own individual conclusions with the voted consensus conclusion reached 87%. This means that the GBM model is as accurate as the participants in assigning the analysis conclusion of a mark. These results illustrate how machine learning models may assist fingerprint examiners in the analysis phase.

3.5. Confidence intervals

PCAST [8] stressed the need to report error rates with associated confidence bounds. Indeed, observing no errors over a limited number of trials (say 100) does not mean that the error rate is zero, even though the observed proportion is actually zero (0/100). This is because we may have obtained this value simply because of the limited number of trials carried out. Observing zero errors may be due to the fact that we experimentally obtained a limited sample out of a larger population where the number of errors is unknown but could be different from zero. The more trials we conduct without errors, the more confident we would be in claiming that the error rate, estimated by the proportion of the number of errors over the number of trials, tends towards 0. But with a limited number of trials such as 100, we have evidence (no errors observed) suggesting a small error rate, but the observed proportion (0/100) is not telling the full story. Thus, it is appropriate to qualify these error rates (or any proportion derived from the data obtained in this study) along with these bounds.

However PCAST indicated that (pp.152–153): "currently, for technical reasons, there is no single, universally agreed method for calculating these confidence intervals." It is true that the statistical debate is complicated by a philosophical but fundamental difference between the frequentist approach and theBayesian approach that can be used to compute these bounds. From a frequenstist perspective we talk about *confidence intervals* that are computed directly from the observed data counts that led to the proportion of interest. These intervals will cover the true proportion in the long run. Hence, if we could repeat the experiment, a 99% confidence interval will cover the true value 99% of the time. Statisticians espousing the Bayesian perspective will compute *credible intervals* that are based on the observed empirical data but also consider some prior belief regarding the true value. These credible intervals are easier to interpret as they will more completely reflect on the probability that the true value (in our case a proportion) will fall within the computed interval. A 99% credible interval means indeed that the true but unknown value has a 99% probability of lying between the lower and the upper limits defined by the interval.

There is abundant literature contrasting these two schools of thought, refer to[30] for a review, but for this project we have taken the position to compare both approaches. We aim at showing that given reasonable conditions and enough data, they tend to converge. The difference of interpretation of their meanings is also well explained in[30], but the risk of misunderstanding posed by the usual confusion between confidence and credible intervals is reduced. In other words, we argue that the debate is

philosophically important but practically not decisive when some conditions are met as we will show.

The package *proportion*[18,19] was useful as it allowed us to compute the confidence and credible intervals using a range of statistical techniques. An in-depth discussion of the statistical techniques used and the data generated from this analysis are provided in Appendix A. Supplementary Material; however, for these data, a reported upper bound of 1% would be appropriate by either frequentist or Bayesian methods. The dedicated shiny application `https://cchampod.shinyapps.io/app_CI/` allows the reader to compute all intervals for any proportion obtained in the study, either on the results of all participants or on the individual results.

Although the large number of trials completed in this study allows for us to claim the upper bound of 1% for the study population taken in aggregate, if we want to focus on the error rates of an individual who took part in the study, we need to look at the matter in more detail. We will provide an example to illustrate this.

If we take the participant (User-0101) who concluded all comparisons in line with the ground truth and delivered the largest number of exclusions (19), the proportion of false identifications is zero (0/19). Given that we have only 19 trials where a possibility for a false identification existed, the observation of no errors speaks in favor of a low error rate but not very strongly. In fact, the upper limits of the credible or confidence intervals are quite high, as shown in Table S4 in Appendix A. Supplementary Material. The minimum on the computed upper limits is 0.0%, whereas their maximum is 28.9%.

If we take the methods allowing the smallest Root Mean Square Error (RMSE), we have somewhat of a convergence with respectively an upper limit of 17.6% for the frequentist method (Exact_1) and 13.9% for the Bayesian method (HPD). Given that credible intervals speak directly to the true value of interest, we suggest adopting the Bayesian method.

We would like to go further than just reporting these numbers, especially when dealing with individual participants. We don't dispute that data gathered from the conducted experiments lead to an upper bound for the proportion of false positives of about 14%. However, this participant excelled in the study: besides the 19 correct exclusions, 37 correct identifications were also reported.

We could legitimately wonder if the above false positive rate is a fair representation of the risk posed by this participant. Does it give credit to the participant's proficiency to quote a false positive rate of about 14%? We believe that such a value is not a fair account of the professional competency of that individual. We mathematically end up in this situation because the credible interval has been computed only on the data from the study without making any prior assessment of the performance of the participant before entering the study. In other words, we stated (technically by the adoption of an uninformative prior) that the error rate may be anywhere between 0% and 100%. Only the acquired data are used as evidence towards the estimate of the error rate. Because we have a limited number of trials for that individual, the boundaries are very large. Fortunately, the Bayesian approach allows us to overcome this limit by introducing consideration of past performance in the computation of the credible interval. For example, if the examiner has a track record of proficiency tests where he performed as well as in the study (say no errors recorded over 100 trials), we can use that knowledge to inform a priori the estimate of the error rate and recompute the credible interval informed by the study data. In this case, we obtained the credible interval shown in Table 12.

Table 12
User-0101 – Confidence and credible intervals associated with the false positive rate for a coverage of 0.95 but allowing the consideration of past proficiency with 0 error over 100 trials.

Method	LowerLimit	UpperLimit	RMSE	Minimum	Stat
Quantile	0	0.030	0.303		
HPD	0	0.025	0.181	***	Bayesian

The upper limit for the method with the lowest RMSE of 2.5% is now a fairer representation of the upper limit that should be considered. This value takes into account three elements: the participant's past performance, their performance in this study and the fact that we have only a sample of their performance. Given the above, the estimate of the 95%upper bound for the probability of a false positive error from this individual is 2.5%. Advocates of the frequentist approach would object to accounting for previous information in the treatment of the study data. We disagree with that view. However, that said, prior experiences cannot be set so extremely that the acquired data have no way to change the prior view. For example, it would be odd to consider as prior knowledge 0 errors over 10,000 trials. In such a case, the 19 different sources trials do not have the potential to sway this prior belief (even if the examiner made plenty of mistakes), hence our illustrative choice of 100 past trials. The shiny app allows the exploration of any chosen numbers of prior counts.

4. Conclusion

A black box study was undertaken to establish a discipline-wide error rate estimate for the comparison of palmar friction ridge impressions. This conclusion provides a brief summary of the main results, followed by recommendations for the improvement of the field.

4.1. Summary of findings

A total of 226 latent print examiners completed a combined 12,279 examinations of palmar friction ridge impressions using a dedicated web-based interface (PiAnoS). Of these,2406 no value determinations were made and these marks were not compared. An additional 413 marks were found to be of value, but the comparisons were never completed, leaving9460 trials in which a comparison conclusion was reached.

Analysis decisions were found to be highly variable between examiners, but this variability was dependent upon the quality of the image. Quality metrics were applied to the marks and consensus between examiners on suitability decisions was lower on marks of low quality than marks of higher quality. Consensus was also much lower on no value decisions than on VID decisions. A GBM model was able to predict the majority voted suitability determination with an accuracy of 86%, based on quality metric scores.

Inconclusive decisions were removed from the dataset prior to calculation of false positive and false negative rates against ground truth, leaving7620 comparison decisions. 12 false identifications were reported, yielding a false positive rate (FPR) of 0.7%.

A total of 552 false exclusions were reported, yielding a false negative rate (FNR) of 9.5%. In both cases, the error rate depended on the difficulty of the comparison, as rated by the participants, with higher difficulty comparisons resulting in higher error rates. Consensus between participants in comparison decisions was also dependent upon the difficulty of the comparison, as judged by the participants.

For false exclusions, the area of the palm from which the mark originated also had an effect on the error rate with some areas of the palm presenting a greater challenge than others.

False exclusions were prevalent in this study and 66.2% of participants who completed at least one comparison made at least one false exclusion, whereas 71.4% of participants who completed 10 or more comparisons made at least one false exclusion. Furthermore, out of 400 same source trials, 193 received at least 1 erroneous exclusion decision and the number of false exclusions reported per case ranged widely from 0 to 29. In cases where at least 20

participants reached the ground truth identification decision, the number of false exclusions reported per case still ranged from 0 to 25.

Of the inconclusive decisions rendered on same source trials, 5.1% were reported as being made "with corresponding features." These cases represent instances in which a "support for same source" conclusion could provide some information to the criminal justice system. However, 2.2% of inconclusive decisions on different sources trials were also reported as being made "with corresponding features." These cases, if reported in the real world as "support for same source" would provide misleading information.

Participant comparison conclusions were also compared to decisions according to the majority vote because in casework, ground truth is not known, so the majority vote may be taken as the "correct" response. In 45 instances, an individual examiner concluded ID while the majority concluded exclusion. Of these, 9 were actual false IDs against ground truth while the other 36 represented instances in which the ground truth was same source and the majority vote was incorrect. Although this happened 36 times, those cases were distributed amongst only 4 cases. Nonetheless, this observation can have implications for verification in casework because if someone made these ground truth identifications in the real world, they may have been incorrectly judged to have made an erroneous ID depending on the reason for the incorrect exclusion by another examiner.

This leads to the concept of "questionable conclusions" in which a conclusion that matched ground truth (e.g., an ID on a same source trial) may not have been the most appropriate response, given that the majority voted differently and there may not have been sufficient data available to support the ground truth conclusion. These questionable conclusions were observed when the majority vote was identification, inconclusive, and exclusion.

FPR and FNR both varied greatly per examiner and thus the aggregate results of this study cannot be viewed as indicative of the performance of any particular examiner. Likewise, the demographic information associated with individual examiner performance was only very weakly predictive. Working in an accredited laboratory and completing a formal training program had some effect on increasing examiner accuracy, but the accuracy of the model was not high enough to support making operational recommendations.

Confidence and credible intervals were computed for all error rate proportions in the study and for individual performance. These intervals are informative when the whole study results are considered. However, when computed for each participant individually, we have shown that these intervals have a limited capacity to inform on the performance of a particular examiner as they are based on very small sample sizes at the individual level and do not tell the full story.

4.2. Recommendations

Overall, the results of this study have been encouraging. Although they cannot be directly compared, the FPR and FNR reported here are not vastly different from those reported in the FBI/Noblis black box study[6] and support that the friction ridge comparison task is fairly robust when it comes to identification decisions. Examiner accuracy on exclusion decisions is less robust and improvements can and should be made to reduce this number.

Measuring errors provides information that is needed to work toward the improvement of the field and documentation of the comparison process further assists in the improvement of the field by allowing errors to be reconstructed and understood. Although this was a black box study and documentation was not required, the documentation that was provided in some cases allowed us the kind of insight into errors that would allow for corrective action plans to be developed for the future reduction of errors.

Taking into consideration the totality of the results reported in this article, the authors would like to offer the following recommendations for the improvement of the friction ridge comparison discipline and reduction of errors and variability:

- Introduce the use of automatic quality metrics to assist in suitability determinations.
- Develop suitability criteria for analysis conclusions that are declared and part of the standard operating procedures.
- Utilize a feature consensus panel on low quality impressions.
- Document the information used to support decisions to allow meaningful review of errors
- Verify exclusion decisions. Full verification may be unachievable given limited resources, but at least a sampling process should be in place.
- Develop criteria that must be met (and documented) to reach an exclusion decision.
- Laboratories that do not already do so should define "inconclusive" to encompass the situation where an examiner can't find the mark within the print but does not have clear differences between the two impressions at the correct anatomical source and orientation or the correct anatomical source and orientation are unknown. This situation does not justify an exclusion decision.
- Use difficulty of the comparison as assessed by the examiner to drive blind verification schemes.
- Develop and implement training, competency testing, and proficiency testing specifically geared toward palmar impression comparisons. It is clear from our data that expertise in fingerprint and palm comparisons is not entirely equal and that palmar comparisons deserve dedicated attention and resources.

Funding

This project was supported by Award No. 2017-DN-BX-0170, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

The authors would like to thank all the participants of this study who took the time to complete the trials. We would also like to thank the Arizona Department of Public Safety Forensic Laboratory; Columbus, MS Forensic Laboratory; Douglas County, Nebraska Sheriff's Office Forensic Laboratory; City of Durham, NC Police Department Forensic Laboratory; Illinois State Police Forensic Laboratory; and University of Lausanne School of Criminal Justice for providing the impressions used in this study.

Appendix A. Supplementary Material

Supplementary materials associated with this article can be found, in the online version, at https://doi.org/10.1016/j.forsciint.2020.110457. These supplementary materials include additional data about the participants and impressions in the study; and additional discussion about consensus between examiners, erroneous exclusions, confidence intervals, and comparison decisions made based on creases alone.

References

[1] Jennifer L. Mnookin, Fingerprint evidence in an age of DNA profiling, Brooklyn Law Rev. 67 (1) (2001) 14–71.

[2] Jennifer L. Mnookin, The validity of latent fingerprint identification: confessions of a fingerprinting moderate, Law Probabil. Risk 7 (2) (2008) 127–141.

[3] Simon A. Cole, More than zero: accounting for error in latent fingerprint identification, J. Crim. Law Criminol. 95 (3) (2005) 985–1078.

[4] Michael J. Saks, Jonathan J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (5 August 2005) (2005) 892–895.

[5] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C, 2009.

[6] Bradford T. Ulery, R. Austin Hicklin, JoAnn Buscaglia, Maria Antonia Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. USA 108 (19) (2011) 7733–7738.

[7] Igor Pacheco, Brian Cerchiai, Stephanie Stoiloff, Miami-Dade Research Study for the Reliability of the Ace-V Process: Accuracy & Precision in Latent Fingerprint Examinations, Report, National Institute of Justice, (2014) . https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf.

[8] Advisors on Science, President's Council of, and Technology (PCAST). 2016. Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. Report. Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

[9] H. Eldridge, M. De Donno, C. Champod, Mind-Set - how bias leads to errors in friction ridge comparisons, Forensic Sci. Int. (2020), doi:http://dx.doi.org/10.1016/j.forsciint.2020.110545.

[10] Soweon Yoon, Eryun Liu, Anil K. Jain, On Latent Fingerprint Image Quality, Conference Proceedings., In Proceedings of the 5th International Workshop on Computational Forensics (2012).

[11] Soweon Yoon, Kai Cao, Eryun Liu, Anil K. Jain, LFIQ: Latent Fingerprint Image Quality, Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on (2013) 1–8.

[12] R. Austin Hicklin, JoAnn Buscaglia, Maria Antonia Roberts, Assessing the clarity of friction ridge impressions, Forensic Sci. Int. 226 (2013) 106–117.

[13] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018. https://www.R-project.org/.

[14] RStudio Team, RStudio: Integrated Development Environment for R, RStudio, Inc, Boston, MA, 2015. http://www.rstudio.com/.

[15] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al., Welcome to the tidyverse, J. Open Source Softw. 4 (43) (2019) 1686, doi:http://dx.doi.org/10.21105/joss.01686.

[16] Max Kuhn, Caret: Classification and Regression Training, (2020) . https://CRAN.R-project.org/package=caret.

[17] Brandon Greenwell, Brad Boehmke, Bernie Gray, Vip: Variable Importance Plots, (2019) . https://CRAN.R-project.org/package=vip.

[18] M. Subbiah, V. Rajeswaran, Proportion: Inference on Single Binomial Proportion and Bayesian Computations, (2017) . https://CRAN.R-project.org/package=proportion.

[19] M. Subbiah, V. Rajeswaran, Proportion: a comprehensive r package for inference on single binomial proportion and bayesian computations, SoftwareX 6 (2017) 36–41, doi:http://dx.doi.org/10.1016/j.softx.2017.01.001.

[20] Yihui Xie, Knitr: A General-Purpose Package for Dynamic Report Generation in R, (2020) https://yihui.org/knitr/.

[21] Hao Zhu, KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax, (2019) https://CRAN.R-project.org/package=kableExtra.

[22] Winston Chang, Joe Cheng, J.J. Allaire, Yihui Xie, Jonathan McPherson, Shiny: Web Application Framework for R, (2019) . https://CRAN.R-project.org/package=shiny.

[23] Winston Chang, Barbara Borges Ribeiro, Shinydashboard: Create Dashboards with 'Shiny', (2018) . https://CRAN.R-project.org/package=shinydashboard.

[24] Dean Attali, Shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds, (2020) https://CRAN.R-project.org/package=shinyjs.

[25] Eric Bailey, ShinyBS: Twitter Bootstrap Components for Shiny, (2015) . https://CRAN.R-project.org/package=shinyBS.

[26] Carl Ganz, rintrojs: A Wrapper for the Intro.js Library, J. Open Source Softw. 1 (6) (2016), doi:http://dx.doi.org/10.21105/joss.00063.

[27] Jonathan Owen, Rhandsontable: Interface to the 'Handsontable. js' Library, (2018) https://CRAN.R-project.org/package=rhandsontable.

[28] Glenn Langenburg, Christophe Champod, Thibault Genessay, Informing the Judgments of fingerprint analysts using quality metric and statistical assessment tools, Forensic Sci. Int. 219 (1-3) (2012) 183–198.

[29] Max Kuhn, Kjell Johnson, Applied Predictive Modeling. Book, Springer-Verlag, New York, 2013.

[30] Richard D. Morey, Rink Hoekstra, Jeffrey N. Rouder, Michael D. Lee, Eric-Jan Wagenmakers, The fallacy of placing confidence in confidence intervals, Psychon. Bull. Rev. 23 (1) (2016) 103–123, doi:http://dx.doi.org/10.3758/s13423-015-0947-8.

TENNESSEE DIVISION

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION

ADDITIONAL INFORMATION

Research and Innovation:

If anyone has a research project that they have worked on, an article they have written, or anything else that other members of the TNIAI might find interesting feel free to send it to tennesseeiai@gmail.com and we will add them to the newsletters.

About Us:

The Tennessee Division for the International Association for Identification (TNIAI) is comprised of forensic scientists, tenprint examiners,detectives, crime scene technicians, police officers and others dedicated to the advancing of education and training for crime scene processing and forensic identification disciplines. The TNIAI hosts an educational conference featuring current topics and distinguished speakers.

For additional training opportunities, visit our website at https://www.tniai.org/training.

Connect with us online:

Facebook: https://www.facebook.com/TennesseeIAI

Website: https://www.tniai.org

Instagram: @tennesseedivisioniai

Missing something?

If you or any other TNIAI member in your agency has received any sort of promotion, award, or achievement, please submit the story to us for inclusion in the next newsletter!

Got any ideas or suggestions for the newsletter? Send them our way! We would love to hear from you.

Website: https://www.tniai.org/forum

Email: tennesseeiai@gmail.com

EVIDENCE TECHNOLOGY MAGAZINE