GAO

United States Government Accountability Office

Report to Congressional Requesters

July 2021

TECHNOLOGY ASSESSMENT

Forensic Technology

Algorithms Strengthen Forensic Analysis, but Several Factors Can Affect Outcomes

GAO-21-435SP

GAO@100
A Century of Non-Partisan Fact-Based Work

The cover image displays a stylized representation of forensic algorithms in use by law enforcement by depicting evidence from DNA, fingerprints, and facial recognition. Algorithmic analysis of each evidence type is assessed by a human reviewer before being used in an investigation.

GAO@100

Highlights of GAO-21-435SP, a report to congressional requesters

July 2021

TECHNOLOGY ASSESSMENT

Forensic Technology

Algorithms Strengthen Forensic Analysis, but Several Factors Can Affect Outcomes

Why GAO did this study

For more than a century, law enforcement agencies have examined physical evidence to help identify persons of interest, solve cold cases, and find missing or exploited people. Forensic experts are now also using algorithms to help assess evidence collected in a criminal investigation, potentially improving the speed and objectivity of their investigations.

GAO was asked to conduct a technology assessment on the use of forensic algorithms in law enforcement. In a prior report (GAO-20-479SP), GAO described algorithms used by federal law enforcement agencies and how they work. This report discusses (1) the key performance metrics for assessing latent print, facial recognition, and probabilistic genotyping algorithms; (2) the strengths of these algorithms compared to related forensic methods; (3) challenges affecting their use; and (4) policy options that may help address these challenges.

In conducting this assessment, GAO interviewed federal officials, select non-federal law enforcement agencies and crime laboratories, algorithm vendors, academic researchers, and nonprofit groups; convened an interdisciplinary meeting of 16 experts with assistance from the National Academies of Sciences, Engineering, and Medicine; and reviewed relevant literature. GAO is identifying policy options in this report.

View GAO-21-435SP. For more information, contact Karen L. Howard at (202) 512-6888 or howardk@gao.gov.

What GAO found

Law enforcement agencies primarily use three kinds of forensic algorithms in criminal investigations: latent print, facial recognition, and probabilistic genotyping. Each offers strengths over related, conventional forensic methods, but analysts and investigators also face challenges when using them to assist in criminal investigations.

Latent print algorithms help analysts compare details in a latent print from a crime scene to prints in a database. These algorithms can search larger databases faster and more consistently than an analyst alone. Accuracy is assessed across a variety of influencing factors, including image quality, number of image features (e.g., ridge patterns) identified, and variations in the feature mark-up completed by analysts. GAO identified several limitations and challenges to the use of these algorithms. For example, poor quality latent or known prints can reduce accuracy.

Facial recognition algorithms help analysts extract digital details from an image and compare them to images in a database. These algorithms can search large databases faster and can be more accurate than analysts. The accuracy of these algorithms is assessed across a variety of influencing factors, including image quality, database size, and demographics. GAO identified several challenges to the use of these algorithms. For example, human involvement can introduce errors, and agencies face challenges in testing and procuring the algorithms that are most accurate and that have minimal differences in performance across demographic groups.

Probabilistic genotyping algorithms help analysts evaluate a wider variety of DNA evidence than conventional analysis—including DNA evidence with multiple contributors or partially degraded DNA—and compare such evidence to DNA samples taken from persons of interest. These algorithms provide a numerical measure of the strength of evidence called the likelihood ratio. To assess these algorithms, law enforcement agencies and others test the influence of several factors on the likelihood ratio, including DNA sample quality, amount of DNA in the sample, number of contributors, and ethnicity or familial relationships. GAO identified two challenges to the use of these algorithms. For example, likelihood ratios are complex and there are no standards for interpreting or communicating the results as they relate to probabilistic genotyping.

Generally, three entities test forensic algorithms to ensure they are reliable for law enforcement use.

Source: GAO. | GAO-21-435SP

GAO developed three policy options that could help address challenges related to law enforcement use of forensic algorithms. The policy options identify possible actions by policymakers, which may include Congress, other elected officials, federal agencies, state and local governments, and industry. See below for details of the policy options and relevant opportunities and considerations.

Policy Options to Help Address Challenges with the Use of Forensic Algorithms

Policy option	Opportunities	Considerations
Increased training (report p. 44) Policymakers could support increased training for analysts and investigators.	• Training on human factors could reduce risks associated with analyst error and decision-making. • Could help users or investigators understand and interpret the results they receive. • For latent print and facial recognition, training on cognitive biases could raise awareness and improve objectivity. • Standards for training or certification of analysts or users could increase consistency and reduce risk of improper use across the various federal and non-federal labs and law enforcement agencies that use algorithms.	• Training materials may need to be developed or made more widely available. • May not be clear what entity should establish standards or certifications of training because multiple groups are involved in developing and disseminating training.
Standards and policies on appropriate use (report p. 45) Policymakers could support development and implementation of standards and policies on appropriate use of algorithms.	• Standards or policies addressing the quality of data inputs could reduce improper use. • Increased consistency across law enforcement agencies could increase public confidence. • Standards for testing and performance of facial recognition algorithms could help to reassure the public and other stakeholders that algorithms are providing reliable results. • Standards or policies for communicating results could help users to better understand the strength of the evidence and come to an informed conclusion.	• May be difficult to implement across different levels of government. • Individual localities or agencies may be reluctant to conform to more universal standards. • May increase the cost of procuring and maintaining forensic algorithms. • Standards creation can be resource-intensive, requiring research and testing as well consensus from public- and private-sector stakeholders.
Increased transparency (report p. 46) Policymakers could support increased transparency related to testing, performance, and use of algorithms.	• The public may be more inclined to trust algorithms if officials provide access to the results of testing, and to information about data sources, how algorithms are used, and for what types of investigations. • Increasing the availability of comparative testing results and presenting them in a way that is easy for non-technical users to understand could make it easier for agencies to select the best-performing algorithms. • For facial recognition algorithms, clearly identifying software versions used in testing could improve public confidence and help agencies choose algorithms. • Making more data sets publicly available for facial recognition algorithm training and testing could improve algorithms and minimize demographic effects.	• Algorithm developers may not want to divulge proprietary information related to training and testing. • Sharing the source of training and testing data may create risks to privacy. • Law enforcement agencies or crime labs may have difficulty finding peer-reviewed journals interested in publishing validation studies from testing.

Source: GAO. | GAO-21-435SP

Table of Contents

Abbreviations

AFIS	Automated Fingerprint Identification System
ANSI	American National Standards Institute
DHS	Department of Homeland Security
DOD	Department of Defense
DOJ	Department of Justice
ELFT-EFS	Evaluation of Latent Fingerprint Technologies: Extended Features Sets
FBI	Federal Bureau of Investigation
FRVT	Face Recognition Vendor Test
NGI	Next Generation Identification
NIST	National Institute of Standards and Technology
OSAC	Organization of Scientific Area Committees
PCAST	President's Council of Advisors on Science and Technology

441 G St. N.W.
Washington, DC 20548

Introduction

July 6, 2021

The Honorable Eddie Bernice Johnson
Chairwoman
The Honorable Frank Lucas
Ranking Member
Committee on Science, Space, and Technology
House of Representatives

The Honorable Mark Takano
House of Representatives

For more than a century, law enforcement agencies have examined certain types of physical evidence—such as fingerprints— to help identify persons of interest, solve cold cases, and find missing or exploited people. Scientific advances are now allowing forensic experts to use algorithms to partially automate the process of assessing such evidence collected in a criminal investigation.[1] Federal law enforcement agencies have adopted or are currently evaluating such algorithms to improve the speed and objectivity of their work.

Based on the emergence of these technologies, you requested that we examine the use of forensic algorithms in federal, state, and local law enforcement. This is the second report in a two-part series of technology assessments responding to this request. The first report, *Forensic Technology: Algorithms Used in Federal Law Enforcement* (GAO-20-479SP),[2] described forensic algorithms that are used by federal law enforcement and how those algorithms work. We found that federal agencies use several different forensic algorithms, but mainly use three: latent print, facial recognition, and probabilistic genotyping. In this second report, we conducted a more in-depth analysis of these three types of algorithms as used by federal law enforcement agencies and selected state and local law enforcement agencies. This report discusses (1) the key performance metrics for assessing latent print, facial recognition, and probabilistic genotyping algorithms; (2) the strengths of these algorithms compared to related forensic methods; (3) the key challenges affecting the use of these algorithms and the associated social and ethical implications; and (4) options policymakers could consider to address these challenges.

To address these objectives, we obtained information from the Department of Commerce's National Institute of Standards and Technology (NIST); federal law enforcement agencies from

[1] An algorithm is a set of rules that a computer follows to compute an outcome.

[2] GAO, *Forensic Technology: Algorithms Used in Federal Law Enforcement*, GAO-20-479SP (Washington, D.C.: May 12, 2020).

the Department of Justice (DOJ), the Department of Homeland Security (DHS), and the Department of Defense (DOD); convened an expert meeting of 16 experts with assistance from the National Academies of Sciences, Engineering, and Medicine; conducted interviews with five non-federal law enforcement agencies or crime laboratories, four forensic algorithm vendors, and additional stakeholders, including industry consultants, nonprofit groups, and academic researchers; conducted literature searches; and reviewed relevant literature, including scientific articles and case law. For more information on our scope and methodology, see appendix I.

We conducted our work from June 2020 to July 2021 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

1 Background

1.1 The use of forensic algorithms

Forensic algorithms are tools used by law enforcement agencies to help determine whether an evidentiary sample (i.e., collected from a crime scene) is or is not associated with a potential source sample (i.e., collected directly from a person of interest), based on the presence of similar patterns, impressions, or other features in the sample and the source.[3] While recent advances in computing have allowed for greater automation, some forensic methods were used in law enforcement over a hundred years ago. For example, in the late 1800s, prisoners in Argentina were identified using prints.[4]

In our previous work, we found that federal law enforcement agencies primarily use three types of forensic algorithms to help determine whether an evidentiary sample is or is not associated with a potential source sample. Each algorithm relies on different features: (1) analysis of latent prints relies on fingerprints or palm prints left behind on various surfaces, (2) facial recognition relies on images (e.g., from photographs) of a face, and (3) probabilistic genotyping analysis relies on profiles extracted from analysis of DNA. In each case, the forensic algorithms compare features from the evidence to features contained within source databases and for probabilistic genotyping the forensic algorithms compare features from evidence to DNA profiles of persons of interest, such as a suspect or victim in a case.[5] Such databases typically contain source features collected under controlled conditions, such as photos (e.g., mugshots), known prints, and DNA previously collected by law enforcement.

1.2 Latent print algorithms

Latent print algorithms can help analysts match details in an evidentiary latent print to known prints contained within source databases to provide a candidate list faster than human analysis.[6] For forensic applications, the source database contains known prints—a set of prints taken under controlled conditions.

To conduct latent print analysis, each latent print is first digitally scanned or photographed to produce a latent print image. An analyst or the algorithm identifies and marks features—called minutiae—on the latent print image.[7] The marked latent print image is uploaded into a software tool, commonly called an automated fingerprint identification system (AFIS). AFIS compares the types and locations of minutiae in a latent print image to those established in a known print database. The algorithm scores the similarities between the

[3]President's Council of Advisors on Science and Technology (PCAST), *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (September 2016).

[4]Non law-enforcement use of similar methods is even older. For example, in 200 B.C., prints were used for general identification in China.

[5]Source databases contain evidentiary records that can allow a link to be made between additional cases with an unknown person of interest. This allows investigators to link multiple cases.

[6] Latent prints are finger-, palm-, or footprints left on items such as those found at crime scenes. A candidate list is a list of possible matches to the latent print image.

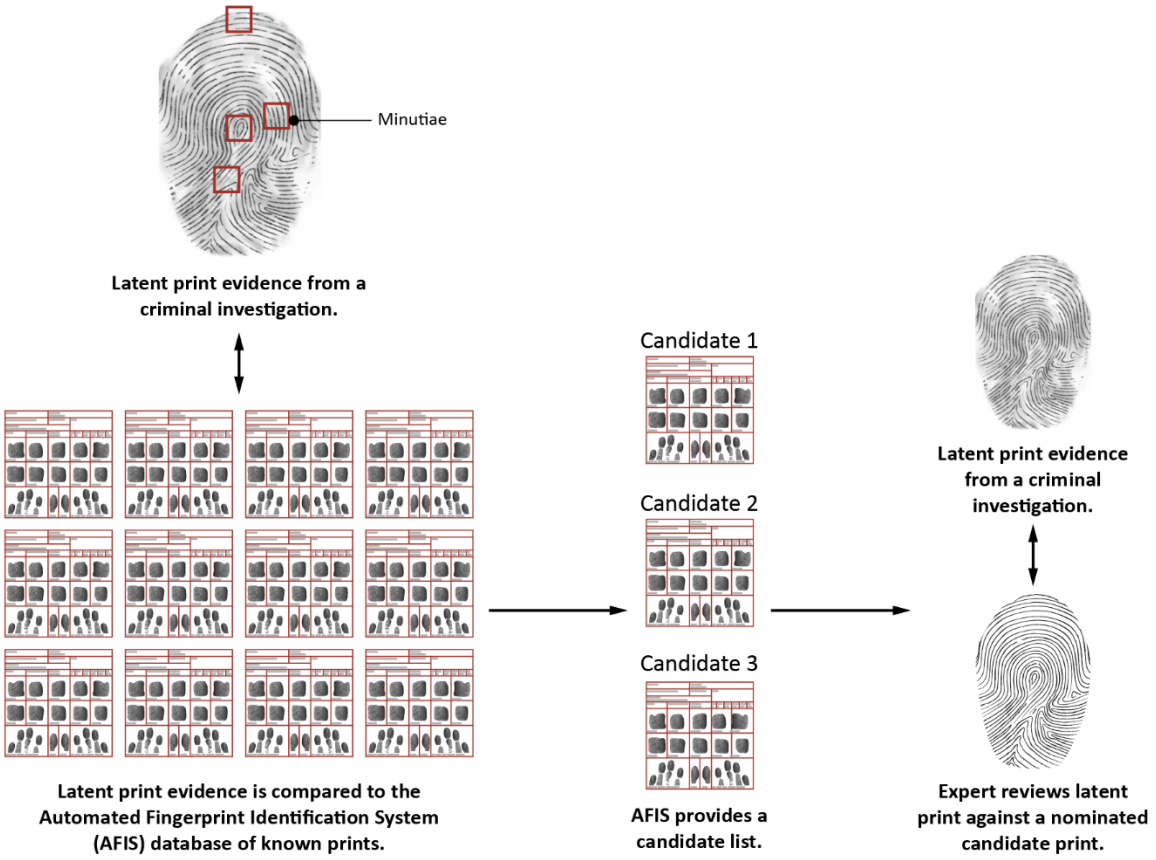
[7]Minutiae refers to specific plot points on a fingerprint. This includes characteristics such as ridge bifurcation or a ridge ending on a fingerprint.

latent print image and known prints in the source database and returns a candidate list of pre-defined length for review. An analyst then compares the relevant known print from each candidate on the list to the latent print to determine whether they are from the same source. In some cases, the system could return no matching candidates if no known prints in the database are found to be sufficiently similar to the latent print image, even if the individual is in the database.

Both federal and non-federal law enforcement agencies use latent print algorithms in criminal investigations. Additionally, identifications made with the assistance of AFIS searches have been used as evidence in criminal cases. The Department of Justice (DOJ) Office of Inspector General reported on one instance in which a subject was misidentified from latent prints; however, this instance of latent print misidentification during an investigation was attributed to human error.[8]

[8]DOJ, Office of the Inspector General, *A Review of the FBI's Handling of the Brandon Mayfield Case* (March 2006).

Figure 1: The workflow for latent print analysis for law enforcement investigations



Source: Adapted from GAO-20-479SP. | GAO-21-435SP

Note: The process is similar for palm prints. There can be more or fewer candidates in the list than the three shown here as an example.

1.3 Facial recognition algorithms

Facial recognition algorithms can help analysts extract digital details in an evidentiary image, also called a probe image, of a person of interest and compare them to images in a source database to provide a candidate list faster than human analysis. Algorithms used by federal agencies analyze the entire set of pixels across the image and generate a mathematical representation of the face in the image (see text box). This mathematical representation of the face is called a "template." Facial recognition algorithms can then compare the template of the probe image to a database of source templates.

The most accurate facial recognition algorithms use artificial intelligence

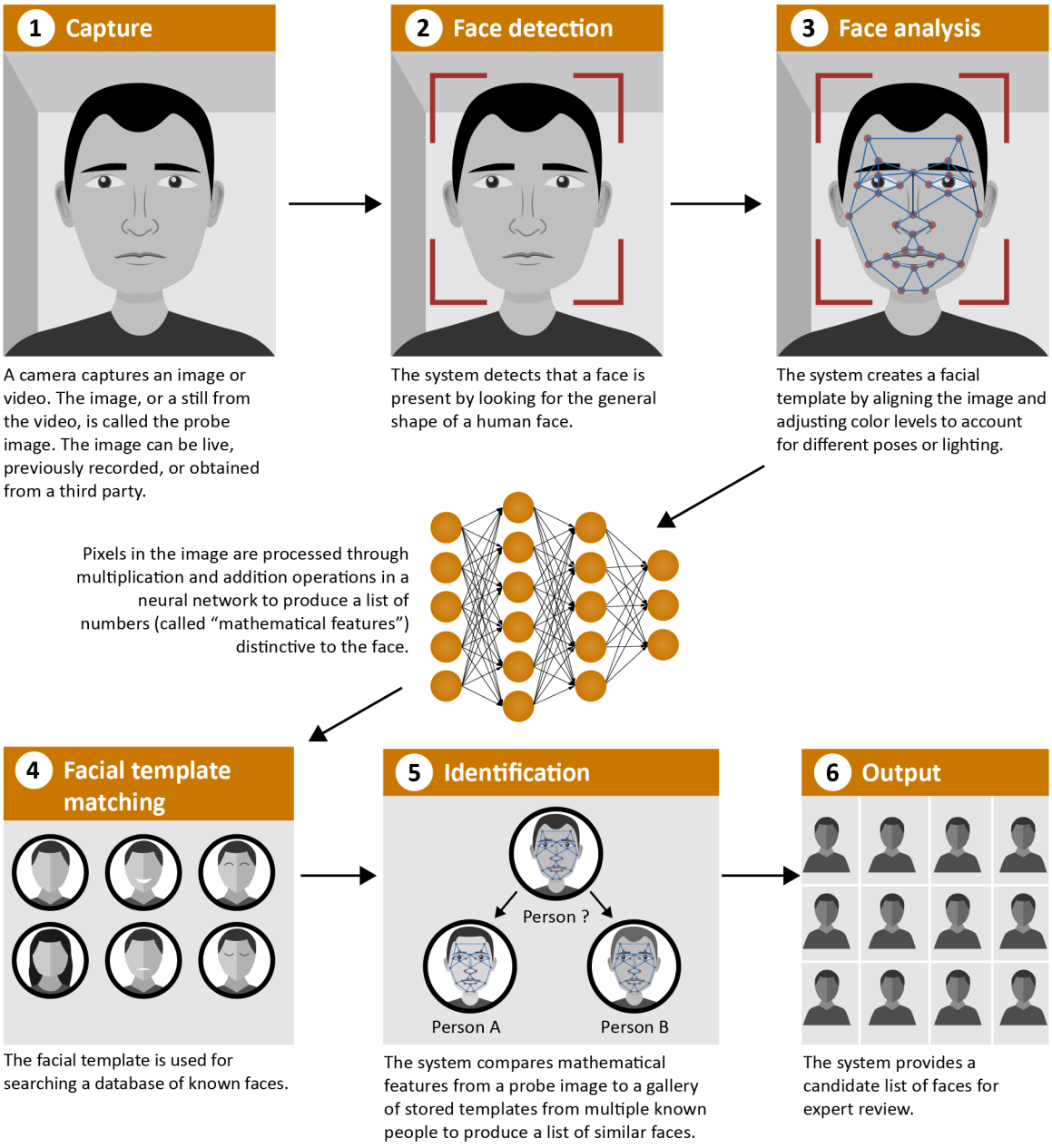
According to the National Institute of Standards and Technology (NIST), the most accurate facial recognition algorithms use an artificial intelligence method called convolutional neural networks. Convolutional neural networks perform multiplication and addition operations on the pixels in a probe image to produce a list of numbers (called vectors) for comparison to similar vectors developed for the source images in the database. They may not explicitly use facial features such as the distance between the eyes or the length of the nose.

Source: GAO analysis of agency documentation and interviews. | GAO-21-435SP

The output of the algorithm is a candidate list of faces from the source database of known persons. If the algorithm is configured to identify multiple candidates, it will generate a list of best-matched photos with a ranking from most to least similar to the probe image. The length of this candidate list is determined by the analyst, agency, or vendor, but typically is between 20 and 100. In some cases, the system could return no candidates if no source database photos are found to be sufficiently similar to the probe image. Similar to latent print analysis, a human analyst makes the final decision as to whether an image from the candidate list is from the same source as the probe image.

Both federal and non-federal law enforcement agencies use facial recognition for criminal investigations. DOJ's Federal Bureau of Investigation (FBI) indicated that they use facial recognition to generate investigative leads. We identified an example of a non-federal law enforcement agency using facial recognition to help identify suspects.

Figure 2: The workflow of facial recognition analysis for law enforcement investigations



Source: Adapted from GAO-20-522 and GAO-20-479SP. | GAO-21-435SP

Note: There may be more or fewer candidates in the list than the 12 shown here as an example.

1.4 Probabilistic genotyping algorithms

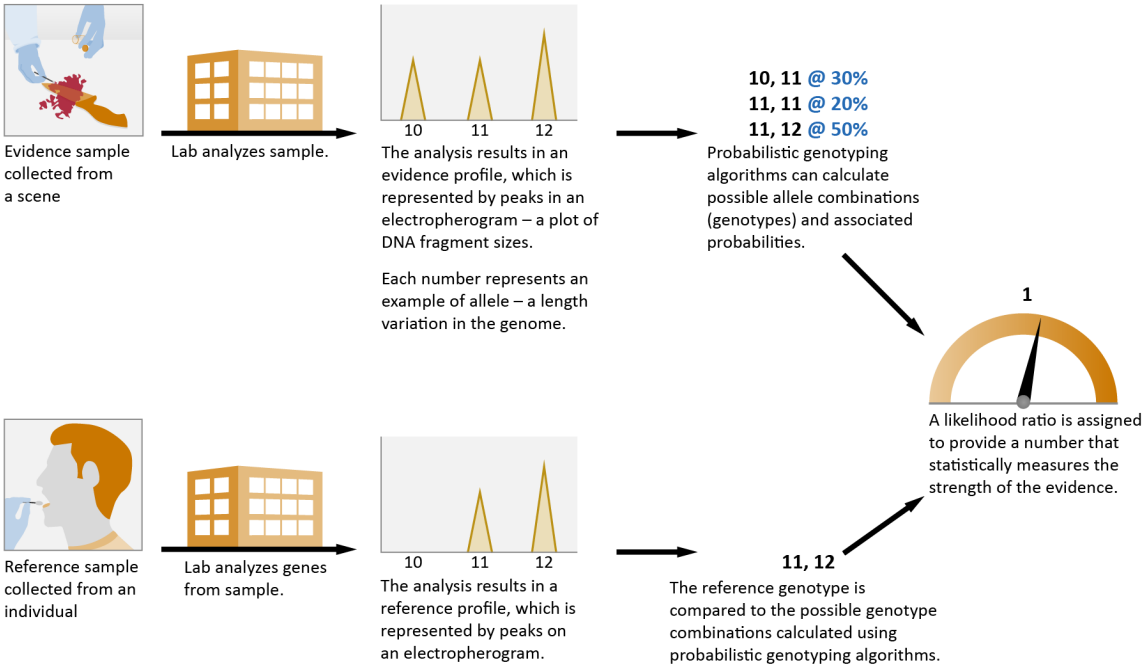
DNA analysis—whether conventional or via probabilistic genotyping algorithms—relies on detecting naturally occurring genetic variations that can be used to help identify individuals. This analysis works by creating profiles from evidentiary samples and comparing them to profiles of samples taken from persons of interest or a source database of known profiles. The profiles, consisting of a set of *genotypes*, are normally represented as a series of *peaks* on a graph known as an *electropherogram*. The peaks represent the quantity of DNA fragments of given lengths. These lengths vary among individuals, which is what makes this method useful for identification.

Probabilistic genotyping algorithms can help analysts evaluate a wider variety of DNA evidence than conventional analysis. This evaluation is dictated by providing a systematically applied interpretation of the genetic profile by both the algorithm and laboratory-developed parameters that are unique to the forensic workflow of the user. The algorithm extracts this information by accounting for the evidentiary profile features such as peak height. Probabilistic genotyping algorithms also simulate many different scenarios with different combinations of people—referred to as possible contributors—to the evidence. This allows the algorithm to account for potential contributors whose profiles are known to investigators, as well as those whose profiles are unknown. According to a federal agency, using these capabilities, probabilistic genotyping algorithms can analyze evidence containing DNA from multiple contributors, or evidence containing small amounts of DNA or partially degraded DNA. In contrast, conventional DNA analysis can produce inconsistent results when the sample has limited DNA, poor quality DNA, or multiple contributors.

The main output of probabilistic genotyping is a number called the likelihood ratio, which is the ratio of two probabilities: the probability that the evidence would appear as it does if the DNA originated from the person of interest, divided by the probability that the evidence would appear as it does if the DNA originated from an unknown (and unrelated) individual. If the likelihood ratio is greater than 1, the results generally support the hypothesis that the person of interest is a contributor. Likelihood ratios of less than 1 generally support the hypothesis that the person of interest is more likely not a contributor. Support for hypotheses can range from low or weak support to high or strong support based on the available information in an evidentiary profile. However, a likelihood ratio is not the same thing as the probability that the individual's DNA is actually contained in a given DNA mixture. For example, a probabilistic genotyping algorithm could return a low likelihood ratio even if a person is a contributor to the sample—if, for example, the sample is degraded or contains very little DNA.

Both federal and non-federal law enforcement use probabilistic genotyping algorithms to determine the strength of evidence that a person of interest has contributed to the DNA evidence. Probabilistic genotyping has been used in criminal investigations and as evidence in criminal court cases.

Figure 3: The workflow of probabilistic genotyping analysis for law enforcement investigations



Source: Adapted from GAO-20-479SP. | GAO-21-435SP

Note: Probabilistic genotyping algorithms would analyze more peaks than the three shown here as an example.

2 Latent Print Algorithms

To ensure latent print algorithms are reliable, law enforcement agencies and others test them for accuracy using two methodologies. They also assess the algorithms across a range of factors that can influence accuracy, such as the quality of the latent print image and the number of minutiae. Such testing suggests, as the experts we interviewed confirmed, that latent print algorithms have two main strengths: speed and consistency. However, algorithm users face limitations and challenges associated with human error and cognitive bias.

2.1 Latent print algorithms are assessed for accuracy across a variety of influencing factors

2.1.1 Latent print algorithms are assessed for accuracy

Accuracy is the key performance metric used to assess latent print algorithms. Accuracy is defined here as the percentage of latent prints that returned the correct print in a specific position or higher on a ranked list (e.g. the accuracy listed for rank-4 refers to the algorithm returning the correct print at rank four or higher).

Three groups perform testing to determine accuracy: algorithms developers and vendors, law enforcement agencies, and NIST, which is a non-regulatory agency in the Department of Commerce and independent of law enforcement and vendors. According to stakeholders, algorithm developers and vendors test the accuracy of their algorithms to confirm they work as expected.[9] Law enforcement agencies told us they conduct their own testing to determine an algorithm's accuracy.[10] NIST stated that they perform algorithm testing to support other agency partners' development of standards and best practices and to inform developers, end users, and policy and decision makers about the capabilities of the technology.

A comparative performance test of latent print algorithms was conducted by NIST in 2012— the Evaluation of Latent Fingerprint Technologies: Extended Feature Sets (ELFT-EFS) study. The FBI provided NIST with a de-identified dataset to aid in this testing, and DHS officials cited NIST testing results as a factor in their decision about which algorithm to use. The ELFT-EFS as well as an internal validation study conducted by the FBI form the primary basis for this discussion.[11] For the algorithms from the five vendors tested in NIST's 2012 ELFT-EFS, the overall average rank-1 search accuracy rates ranged from 0.0 to 49.8 percent across a variety of influencing factors.[12] A more recent 2018 internal

[9]Such testing is considered "developmental validation" as it is tested by the manufacturer.

[10]Such testing is often called "internal validation" testing and is done by the law enforcement associated lab.

[11]The FBI indicated that these results may not be representative of the performance of current generation AFIS because the FBI's Next Generation Identification (NGI) system was not available in 2012. NIST launched a new ELFT evaluation in May 2020.

[12]The NIST ELFT-EFS tested different latent feature sets. The results presented here are from the LG dataset in Table 10B, which is most comparable to AFIS latent features in 2012. One algorithm provided no data, a second algorithm which provided the result of 0.0 percent does not appear to operate properly with this feature set. The remaining three algorithms showed 45.1, 49.3, and 49.8 percent accuracy. As stated in the

validation study conducted by the FBI on their current algorithm showed a 63.3 to 69.6 percent accuracy rate.[13] The accuracy rates from the NIST and FBI studies are not directly comparable because they involved different sets of latent prints. To measure accuracy, a tester selects a latent print image to be compared to a database that contains a known print mate (match), or to a source database in which all the prints are known to be non-mates (no match). Comparing a latent print image to prints in a source database is known as one-to-many testing, or 1:N testing, because the algorithm compares one latent print image against many (N) source prints.[14] For example, NIST conducted 1:N testing in 2012 as part of their ELFT-EFS work, to test five latent print algorithms. Experts and law enforcement organizations we interviewed told us that two of the vendors that submitted algorithms for NIST testing also supply algorithms to law enforcement for use in latent print analysis. However, as vendors supply the algorithms to NIST without software names or versions, NIST cannot report whether the algorithm tested in the ELFT-EFS study is the same algorithm—or the same software version—used by law enforcement.

In NIST's ELFT-EFS test of latent print algorithms, 1:N testing is done using one of the following two methodologies:

- **Rank-based analyses:** This method measures the proportion of latent print images whose mate is reported at a given rank or higher. The results of this type of test show how many latent print images are correctly identified for any rank (e.g., NIST tested to rank 100, the FBI's 2018 internal validation study tested to rank 20). Law enforcement agencies told us they use this method for their own testing.
- **Score-based analyses:** This method uses mated and non-mated searches to report false positive and false negative error rates. Law enforcement do not necessarily use this method to test their algorithms internally, but they may consider NIST evaluation results when selecting an algorithm.

2.1.2 Latent print algorithms are assessed across a variety of influencing factors

Accuracy of latent print algorithms is tested across multiple influencing factors, including image quality, number of minutiae, image orientation, and user input.

Image quality: The image quality of a latent print can be low for several reasons, making algorithm results less accurate. For example, the latent print might come from only a small part of a finger, or be smeared and distorted. In the 2012 ELFT-EFS, to test a range of image quality, NIST acquired latent print images from operational casework and laboratory-collected images. Latent print analysts then

International Organization for Standardization (ISO)/IEC 19795-1: 2006 standard on Biometric Performance Testing and Reporting, accuracy statistics are dependent on the size of the database used to test the algorithm.

[13]The accuracy rates we discuss for both the NIST ELFT-EFS and the FBI's 2018 internal validation study are specific to the rank 1 position.

[14]Another method for testing latent print algorithm accuracy is to compare a latent print image to a single image of another known print. In this case, the algorithm's output is "match" or "no match." This method is called 1:1 testing and is typically used to verify the identity of the print image rather than to investigate possible identities of the probe image.

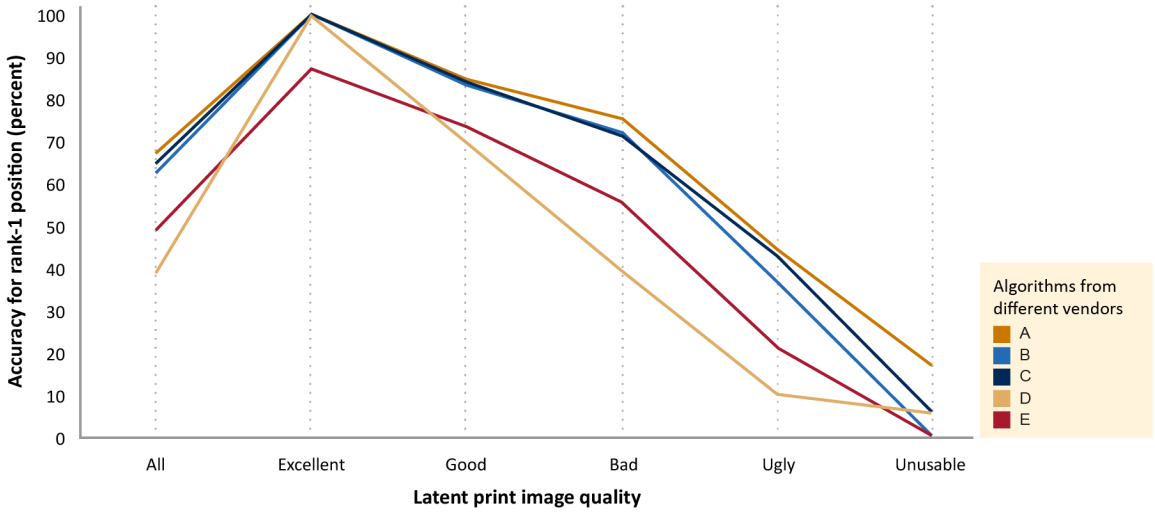
organized the print images into five subjective categories (excellent, good, bad, ugly, and unusable) based on their expert judgement. Across five algorithms tested by NIST, accuracy ranged from approximately 88 to 100 percent with excellent prints, but dropped to approximately 40 to 78 percent with bad print images using the rank-based method, when using image-only data without markup by an analyst.[15]

According to the FBI, the current algorithms used in Next Generation Identification (NGI) are more accurate but were not evaluated in the 2012 ELFT-EFS. The FBI's 2018 internal validation study, which used marked-up latent print images, showed that accuracy within the rank-20 positions varied with print quality between 32.8 percent for latent prints rated "ugly" to 94.4 percent for latent prints rated "good."

[15]These approximate accuracy ranges are drawn from the LA dataset graph in Figure 10 of the 2012 NIST ELFT-EFS. The LA dataset uses the image for analysis without additional markup.

Figure 4: Accuracy of five latent print algorithms based on latent print image quality

The chart shows decreases in algorithm accuracy as latent print image quality worsens.



Source: National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies: Extended Feature Sets 2012. | GAO-21-435SP

Note: Five vendors supplied algorithms (A, B, C, D, and E). Vendors B and C are known to supply latent print algorithms to law enforcement. The X-axis shows the analyst-assessed latent print image quality and the Y-axis shows the accuracy as the percentage of samples for which the algorithm correctly assigned the matching print to the top position in the candidate list. The results shown are from the LA dataset from Figure 10 of the National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies: Extended Feature Sets 2012.

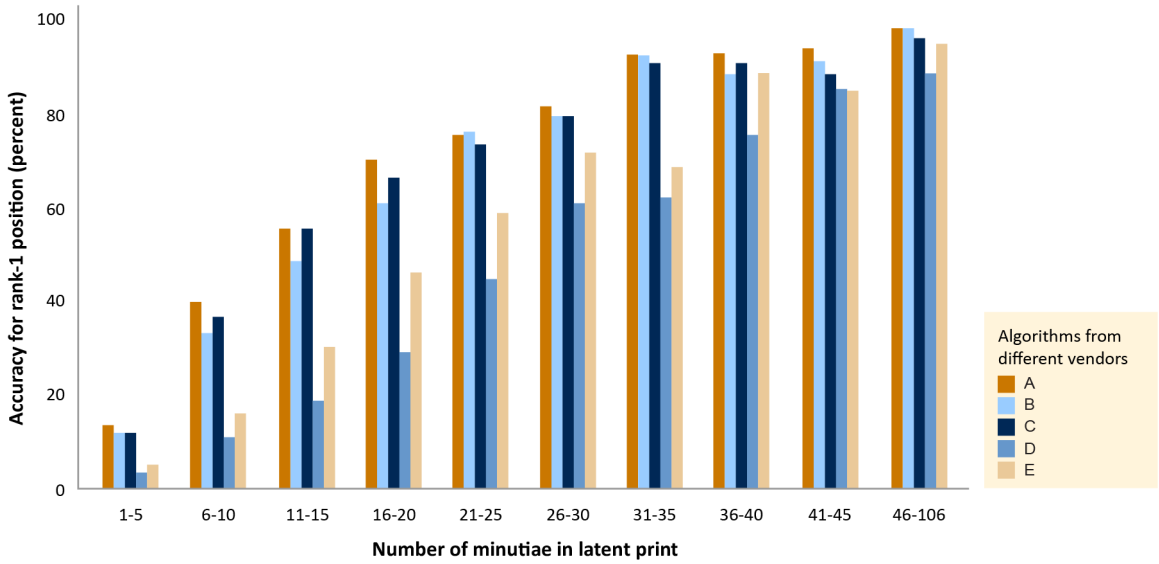
Image Minutiae: Latent print analysis matches *features* of a print image, such as minutiae, to those in a known print. Fewer features generally lead to a decrease in accuracy. NIST's ELFT-EFS study found that the lowest percentage of correctly matched latent print images in rank 1 was due to latent print images with minutiae counts below 10. For example, accuracy ranged from approximately 90 to 100 percent when approximately 46 to 106 minutiae were in the print, but dropped to approximately 40 to 70 percent with 21 to 25 minutiae, and approximately 10 to 40 percent with 6 to 10 minutiae, using the rank-based method.[16] The FBI's 2018 internal validation study showed that the number of minutiae was not as strong an indicator that a correct candidate would be returned at the rank 1 position as score difference.[17] The FBI's 2018 internal validation also showed that score difference had a positive correlation to the quality of the image.

[16]A different study in 2011 reported that 24 countries required a minimum of 12 minutiae for a latent print to be used in analysis. Currently, the U.K. and most U.S. law enforcement agencies use a non-numeric standard that accounts for multiple factors such as the quality of minutiae. A 1995 scientific review of the previous use of a minimum number of minutiae found that there was no scientific basis for selecting a specific number of minutiae. These current standards use a holistic approach that takes into account the quantity of image features as well as the quality of the features as determined by a latent print analyst.

[17]The FBI and NIST studies may not be comparable because they used different datasets. The score is a number generated by the algorithm that orders the candidate list. The higher the score, the more similar the algorithm has determined the latent print and the source print are to each other. Score difference is the difference in score between the top two candidates.

Figure 5: Accuracy of five latent print algorithms based on number of minutiae

The chart generally shows increases in accuracy when more minutiae are present in a print.



Source: National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies: Extended Feature Sets 2012. | GAO-21-435SP

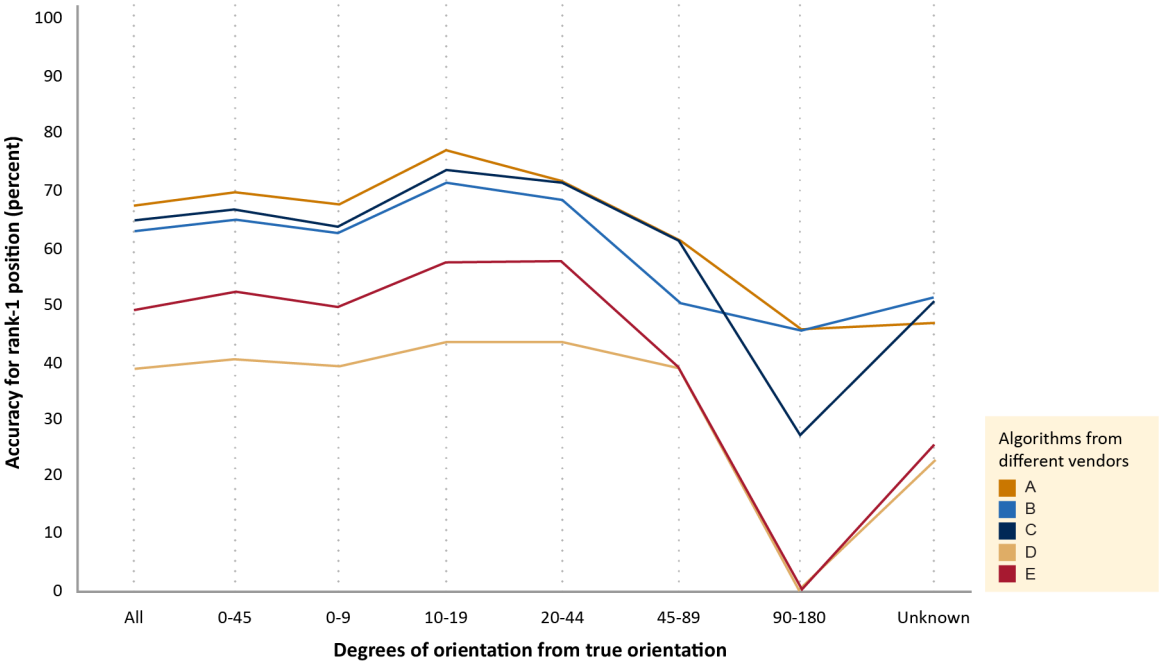
Note: Five vendors supplied algorithms (A, B, C, D, and E). Vendors B and C are known to supply latent print algorithms to law enforcement. The X-axis shows the number of minutiae per latent print image and the Y-axis displays accuracy as the percentage of samples for which the algorithm correctly assigned the matching print to the top position in the candidate list. The results shown are from the LA dataset from Figure 7 of the National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies: Extended Feature Sets 2012.

Image orientation: Prints in a database are typically taken in a controlled manner in which the print is oriented vertically; i.e., the top of the print image corresponds to the tip of the finger. Accuracy is generally higher when the latent print images are more closely oriented in the same way as the known print. NIST reported in its ELFT-EFS study that, in general, the farther the latent print image was from the vertical orientation, the lower the accuracy of the match.

For example, accuracy ranged from approximately 44 to 78 percent when the latent print image was misaligned by 10 to 19 degrees from the source orientation (vertical), but dropped to approximately zero to 48 percent when the latent print image was misaligned by 90 to 180 degrees. The FBI indicated that the current algorithms being used in NGI can accept a print in any orientation.

Figure 6: Accuracy of five latent print algorithms based on image orientation

The chart generally shows higher accuracy when the latent print image is oriented similarly to the known print, as determined by the algorithm for the top position.



Source: National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies: Extended Feature Sets 2012. | GAO-21-435SP

Note: Five vendors supplied algorithms (A, B, C, D, and E). Vendors B and C are known to supply latent print algorithms to law enforcement. The X-axis shows the degree of misalignment between the latent print image and the known print image, as determined by the analyst. The Y-axis displays the accuracy as the percentage of samples for which the algorithm correctly assigned the matching print to the top position in the candidate list. The results shown are from the LA dataset from Figure 6 of the National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies: Extended Feature Sets 2012.

Precision of human input: Accuracy can also be affected by how the latent print image is marked up prior to algorithmic analysis. This markup is sometimes performed by a human and sometimes by an algorithm, or it can be first performed by an algorithm and then checked by a human analyst. The highest accuracy for all algorithms was observed in latent print images that were marked up by human analysts with access to known prints for a given latent print image. For example, accuracy improved by 12 to 15 percent when prints were marked up by latent print analysts with access to the known minutiae from the known print match, compared to those marked up by latent print analysts without access to the known minutiae. NIST states that while an analyst with access to known print mate data represents an operationally impractical scenario, it highlights the importance of the precision of the latent print markup process. In the FBI's 2018 internal validation they assessed the accuracy rate of their algorithm when markup was conducted by the algorithm versus when markup was performed by a latent print analyst. This study showed that when markup was performed by a latent print analyst an identification was made 90.2 percent of the time at rank 1, but was as high as 91.8 percent when encoding was performed by an algorithm. The FBI's 2018 internal validation study therefore suggests that the highest accuracy rate is associated with the encoding method that employed algorithm markup followed by examiner "clean-up."

2.2 Latent print algorithms have two main strengths

Latent print algorithms have the following main strengths, compared to conventional latent print analysis conducted by human analysts performing the search and making conclusions:

- **Faster search and analysis.** An advantage of latent print algorithms is that they are faster than human analysts in searching databases to provide a candidate list.[18] Algorithms essentially do the same thing as an expert analyst, but they can perform those same functions much faster. Algorithms can provide a smaller candidate list from a much larger database, according to a vendor. This enhancement can be critical given the size of some databases. For example, the NGI has over 78 million records from convicted criminals and over 59 million records from non-law-enforcement sources, as of March 2021.[19]
- **Better consistency.** Latent print algorithms also do not suffer fatigue and do the same analysis every time. Law enforcement told us another advantage of latent print algorithms is that they can improve consistency. Human analysts may come to different conclusions when presented with the same latent print images. Algorithms have the potential to increase consistency in latent print analysis.

2.3 Limitations and challenges affecting law enforcement use of latent print algorithms

Latent print algorithms have several technical limitations. One key limitation is that performance is poor when the quality of the evidentiary latent prints or the known prints in the database is poor. Latent prints are collected at crime scenes and are often not of ideal quality—they may contain only a portion of the fingerprint or be distorted. Furthermore, latent prints do not have a standard "capture device"—there are many different techniques for lifting prints, and errors during capture can result in unusable data. The quality of known prints in the database can also adversely affect the accuracy of latent print algorithms. For example, the algorithm cannot make a match between two 10-print cards if the right hand was rolled in the left hand slot when one of the sets of prints was collected, according to FBI officials. However, this limitation would not prevent an algorithmic match being made between a latent print and a known 10-print record of this type. In addition to these technical limitations, law enforcement agencies face several challenges affecting the use of latent print algorithms. Below, we describe three key challenges: human involvement, communicating results, and testing.

2.3.1 Human involvement

Human involvement is necessary for the use of latent print algorithms, which introduces opportunities for human error and cognitive

[18]According to FBI officials, the final identification decision is made by a human analyst.

[19]According to the FBI, records are known fingerprints and most are associated with ten fingerprints.

biases.[20] Because the algorithms return a candidate list, which is then reviewed by the analyst, human errors and bias can influence the end result, resulting in errors such as a false positives whereby someone is incorrectly identified to be a match. A 2011 study showed that false positives in latent print decisions are rare;[21] however, according to the 2016 President's Council of Advisors on Science and Technology (PCAST) report on Forensic Science in Criminal Courts, false positive results in particular can have negative consequences because they can result in false arrests, investigations, or convictions. A notable example is a 2004 case in which the FBI erroneously arrested and incarcerated Brandon Mayfield for 2 weeks as a result of multiple analysts' errors and cognitive biases.[22] The case illustrates the potential consequences of human error or cognitive bias relating to algorithm use. Law enforcement officials noted that cases such as Brandon Mayfield's misidentification have led to improvements in latent print analysis practices, such as additional education for analysts and "blinded" verifications, in which another analyst with no, or limited contextual information, and no knowledge of the first analyst's conclusion verifies a latent print examination.

Overall usability of latent prints. An analyst determines whether a latent print image is of suitable quality to be run through the algorithm, and this assessment can vary between analysts.[23] This discretion leaves more room for such decisions to be influenced by cognitive biases. Under some circumstances, analysts may run a latent print image with poorer quality than they might have normally accepted, while other analysts may not run certain prints that they determine to be poor quality.[24] Generally, this decision is based on the professional judgement of the analyst.[25]

Quality of analyst-identified features. The quality of the markup by a human analyst prior to algorithm analysis can affect accuracy. Latent print analysis is not a fully autonomous process that can run from beginning to end without human input. As previously described, the precision of human input affects algorithm accuracy. Latent print algorithm vendors typically do not provide training on how to markup features in a print to maximize the capabilities of the system.

[20]PCAST, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (September 2016). According to the PCAST, "cognitive bias refers to ways in which human perceptions and judgments can be shaped by factors other than those relevant to the decision at hand." Examples include contextual bias, where individuals are influenced by irrelevant background information, and confirmation bias, where individuals interpret information, or look for new evidence, in a way that conforms to their pre-existing beliefs or assumptions.

[21]The study found a false positive rate of 0.1 percent among latent print analysts. The study focused on human latent print analysts decisions and did not look at the use of algorithms. B. T. Ulery, R. A. Hicklin, J. Buscaglia, M. A. Roberts, "Accuracy and reliability of forensic latent fingerprint decisions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 19 (2011).

[22]DOJ, Office of the Inspector General, *A Review of the FBI's Handling of the Brandon Mayfield Case* (March 2006). FBI latent print analysts used an AFIS system to conduct the database search during this case.

[23]According to FBI officials, the decision to use an algorithm may also vary by agency because different agencies have different procedures.

[24]Analysts may run latent prints of poorer quality in investigations that have limited leads, are high profile, or urgent, such as terrorism investigations.

[25]Support algorithms have been proposed to assess print quality in a more objective manner, either before the AFIS database search or before human analyst review.

Some latent print algorithms cannot perform certain tasks, such as identifying minutiae and other features in the latent print images. Others, including algorithms used by federal law enforcement, may encode these features automatically, but they may be adjusted by the analyst.

Studies have shown that training improves the performance of latent print analysts; however, stakeholders identified training and certification of analysts as an area for improvement. DOD Defense Forensic Science Center officials stated that the high degree of variability in fully autonomous latent print matching makes it extremely important to include highly trained analysts in the process of performing comparisons. They also stated that analyst training is often the main challenge facing law enforcement agencies because such extensive training is necessary to become proficient. Furthermore, several stakeholders identified room for improvement in analyst training. The Friction Ridge Subcommittee of the Organization of Scientific Area Committees (OSAC) has proposed standards for latent print analyst training; however, these standards have not been universally adopted.[26] There is a certification program for latent print analysts offered by the International Association for Identification, but it is optional. Most crime laboratories require latent print analysts to take periodic proficiency tests, however current tests provide little feedback on analyst skill on more difficult cases or the effectiveness of training.

Quality of final decision-making. The analyst makes the ultimate decision about which print is a match, which risks introducing bias into the process. For example, an exploratory study showed that latent print analysts' match decisions could be influenced by knowledge of another analyst's prior judgement when considering the same prints.[27] In addition, the organization of some criminal justice systems may create a risk of bias toward finding matches. For example, some state and local crime labs are organized within law enforcement agencies, meaning they are not administratively independent. These types of operational conditions create a risk for bias in lab results. Finally, an analyst may become over-reliant on the algorithm and feel compelled to focus on the candidate list provided instead of more objectively assessing the quality of any potential matches, according to stakeholders. Context management procedures that avoid exposing analysts to certain information about a case prematurely or include blind verification by another analyst can mitigate some of these risks.

[26]These proposed standards would require further development through standards developing organizations. Among many requirements, the proposed standards include a general requirement for a latent print analysis trainee to obtain a bachelor's degree with 24 hours in Science, Technology, Engineering, and Math (STEM)-related course work; as well as an understanding of logic, probability, statistics, and human factors affecting the examination process and decision-making, such as cognitive interpretation, and legal issues. See OSAC Friction Ridge Subcommittee, *Standard for Friction Ridge Examination Training Program* V1.0 (December 2017).

[27]I. E. Dror, D. Charlton, A. E. Peron, "Contextual information renders experts vulnerable to making erroneous identifications," *Forensic Science International*, Vol. 156 (2006).

2.3.2 Communicating results

A second key challenge to law enforcement use of latent print algorithms is communicating the analyst's results to investigators and others. The analyst determines the match results from the candidate list provided by the algorithm. The algorithm associates numerical values with each candidate on the list based on similarity; however, these values do not represent an assessment of the strength of evidence of a particular pair of prints being a match. Thus, investigators and others in the criminal justice system do not receive an assessment from the algorithm relaying how confident they can be in the match results. Instead, when analysts communicate the results of their analysis, confidence in the match results is based on factors such as the analyst's experience.[28]

2.3.3 Testing

Finally, existing independent, comparative testing of the performance of these algorithms to help law enforcement agencies understand their capabilities is out of date and does not cover key federal algorithms. As described above, according to NIST's 2012 ELFT-EFS, the overall average rank-1 search accuracy rates for latent print algorithms ranged from 0.0 to 49.8 percent.

According to FBI officials, these accuracy data are out of date. In a 2018 internal validation study, FBI found higher accuracy rates for the algorithm currently used in NGI; however, we were unable to identify reports of comparative testing of latent print algorithms conducted in the intervening years. NIST relaunched its latent print technology research in May 2020.

[28]An exception to this is the DOD Defense Forensic Science Center's FRStat software, which adds a statistical assessment to the strength of the latent print evidence.

3 Facial Recognition Algorithms

To ensure facial recognition algorithms are appropriate for their uses, law enforcement agencies and others test them for accuracy using two methodologies. Accuracy is defined here as the percentage of probe images that returned a candidate source image in a specific position or higher on a ranked list (e.g. rank-4 means any matches returned at ranks 1-4). They also assess the algorithms across a range of factors that can influence accuracy, such as the image quality, database size, and various demographics (such as sex, race, and age). Experts we interviewed, suggest that facial recognition algorithms have two main strengths: speed and accuracy. However, the algorithms face limitations and challenges.

3.1 Facial recognition algorithms are assessed for accuracy across a variety of influencing factors

3.1.1 Facial recognition algorithms are assessed for accuracy

Accuracy is the key performance metric used to assess facial recognition algorithms. To measure accuracy, a tester selects a facial image—called a probe image—of someone whose image is either known to be in a database of source images (a mate) or known to not be in the database (a non-mate). Similar to testing latent print algorithms, this method is called 1:N testing because the algorithm compares the probe image against each image in a source database containing a number (N) of images. Based on our review of agency facial recognition algorithm testing, 1:N testing relies on the following two methodologies for testing accuracy:

- **Identification:** This method quantifies both false positives and false negatives using, respectively, mate and non-mate searches. While not used by law enforcement to test their algorithms internally, law enforcement agencies did say they use the NIST testing to assess the algorithms that they consider for use. Law enforcement agencies are concerned with both false negatives, which could lead to missing a suspect, and false positives, which could lead to a wrongful arrest.
- **Investigation:** Investigation testing is used by law enforcement as well as NIST. This type of testing compares the probe image to a source database and returns a candidate list that contains the source images identified as the closest matches to the probe image. The results of this type of test show how many probe images are correctly identified from rank 1 down to the lowest rank (e.g., NIST tested to rank 50). The size of the candidate list is not uniform across law enforcement. A law enforcement official told us there is no fixed candidate list size, and the number of images returned is specified by the user. For example, FBI officials told us they typically return up to 50 candidates. In contrast, officials from a state law enforcement laboratory told us they may get a candidate list of five images with their facial recognition algorithm. Similar to identification testing, the main metric associated with investigation testing is a false negative

rate. However, according to the NIST Face Recognition Vendor Test (FRVT) report, because this method provides a candidate list with more than one possible candidate, false positives are inherent in the results.[29] Investigation testing quantifies false negative rates for each rank position in the candidate list. For example, if the list has 50 images, and the probe image's known match is not located at the rank 1 position, it would be considered a false negative for that position. This is repeated for each position in the candidate list. The NIST FRVT provides a rank list of the algorithms tested based on false negative rates, which can be found on the agency's website.[30]

- Law enforcement agencies told us they use facial recognition algorithms for investigations, and they use investigative testing methodology when assessing accuracy as this testing is representative of investigative use. Therefore, as false positives are inherently included in candidate lists as described above, the false positive metric is not applicable for investigation testing of algorithms.

NIST FRVT tested over 200 facial recognition algorithms

Since 2018, the National Institute of Standards and Technology (NIST) Facial Recognition Vendor Test (FRVT) has evaluated 286 algorithms from 76 developers, according to NIST officials.. An expert and law enforcement officials told us that two of the vendors tested supply algorithms to law enforcement for use in facial recognition analysis. However, because vendors supply the algorithms to NIST as a "black box", NIST cannot provide details on the exact algorithm, and we cannot confirm whether the algorithm tested in the FRVT is the same algorithm or version used by law enforcement. However, according to the Federal Bureau of Investigation, law enforcement agencies can ask their vendor whether the algorithm provided to them is represented in the NIST testing. Law enforcement officials told us that they collaborate with NIST to assess which facial recognition algorithms to use.

Source: GAO analysis of analysis of agency documentation and stakeholder interviews. | GAO-21-435SP

As with latent print algorithms, three groups perform testing to determine the accuracy of facial recognition algorithms. Algorithm developers and vendors test accuracy. Law enforcement agencies conduct testing to affirm that the algorithm works as intended for their purposes and meets their accuracy demands. NIST, which is independent of law enforcement agencies, , performs algorithm testing to support other agency partners' development of standards and best practices and to inform developers, end users, policymakers, and decision makers about the capabilities of the technology. For example, NIST's FRVT uses investigation testing on multiple algorithms submitted by many developers.

[29]The NIST FRVT states that as a human analyst always reviews the candidate list, what matters is the analyst's decision from the candidate list, which NIST does not test.

[30]The false negative rates for identification and investigation 1:N testing for the facial recognition algorithms submitted to NIST. NIST, *FRVT 1:N Identification,* accessed March 4, 2021, https://pages.nist.gov/frvt/html/frvt1N.html.

3.1.2 Facial recognition algorithms are assessed across a variety of influencing factors

Accuracy of facial recognition algorithms is affected by a range of influencing factors. The following describes how algorithm testing addresses five such factors as identified in the NIST FRVT:

- **Image quality:** Law enforcement officials told us they test the accuracy of their facial recognition algorithms with images exhibiting a range of image quality, from ideal (e.g., mugshots) to lower-quality images such as grainy cell phone images. Similarly, the NIST FRVT used the following image types:[31]
  - **Mugshot:** About 86 percent of the FRVT database consists of frontal mugshot images, which are of relatively high quality based on compliance with established facial recognition standards for quality.[32]
- **Profile view images**: A profile "side" view photo can be searched against a frontal mugshot gallery. Profile view images are common in law enforcement but only a minority of algorithms can successfully perform recognition with them.
- **Webcam images:** Most of the remaining 14 percent of the images were collected using an inexpensive webcam. These images do not meet most of American National Standards Institute (ANSI)/NIST standards for image quality commonly used by law enforcement.[33] As the standard is a frontal facing image the most

Figure 7: Examples of images used in the National Institute of Standards and Technology Facial Recognition Vendor Test

From left to right, the images are a frontal mugshot, profile mugshot, webcam, and wild image.

Source: GAO analysis of National Institute of Standards and Technology (NIST) Facial Recognition Vendor Test. | GAO-21-435SP

[31]The NIST FRVT described mugshot quality as "mostly excellent cooperative live-capture mugshot images collected with an attendant present." Images without a dedicated photographic environment and human or automated quality control checks, may lead to declines in accuracy and are not high quality images.

[32]Per ISO/EC 19794-5:2011, a standard on face image data interchange formats, a mug shot image is defined as a face image type that specifies frontal images with sufficient resolution for human examination as well as reliable computer facial recognition. This face image type includes the full head with all hair in most cases, as well as neck and shoulders. This image type is suitable for permanent storage of the face information, and it is applicable to portraits for passport, driver license, and "mugshot" images.

[33]For more information on data format standard ANSI/NIST-ITL1-2011, see GAO-20-479SP.

common deviations are non-frontal pose, low contrast (e.g., due to varying and intense background lights), and poor resolution (e.g., due to inexpensive camera optics). The images are sometimes also overly compressed.

- **Kiosk-style:** The FRVT also uses lower quality images captured with wide field-of-view cameras mounted in immigration kiosks. The face of the subject is often cropped and not oriented toward the camera.
- **Wild Images:** This additional group makes up less than 1 percent of the images used in the FRVT and includes many photojournalism-style photos. Images are provided to the algorithm using a variable tight crop of the head. Resolution varies widely. Facial poses and expressions also vary widely, and faces can be partially blocked, for example, by hair or hands.

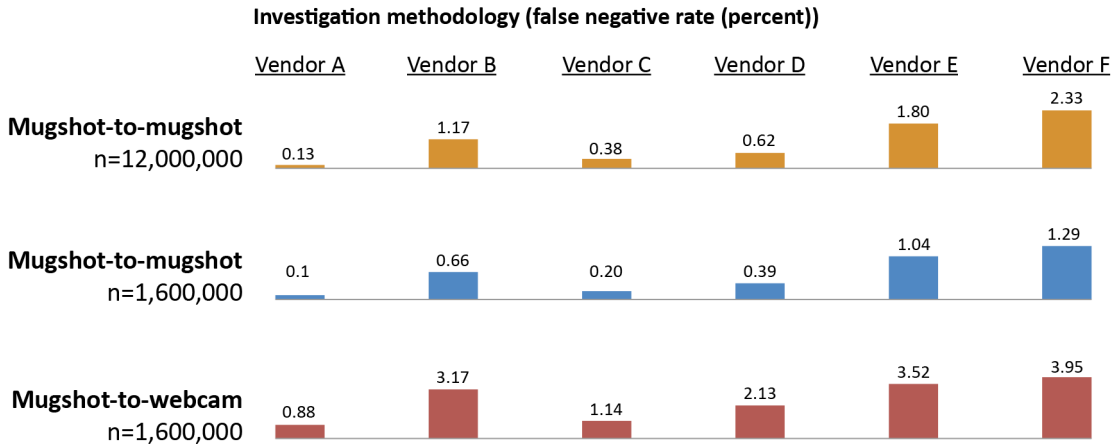
The NIST FRVT report concluded that regardless of how well an algorithm performs with ideal image quality, there is a decrease in accuracy when the image quality decreases. For example, using investigation testing, NIST assessed algorithms from two vendors who supply facial recognition to federal law enforcement and reported false negative rates of 0.1-0.66 percent when comparing a mugshot probe image to a 1.6-million-person mugshot database. These algorithms were less accurate with a webcam image—resulting in false negative rates of 0.88-3.17 percent.

- **Database size:** Accuracy can also change with the size of the database being searched. For example, when algorithms from two vendors used by a federal agency were used to compare a mugshot probe image to a mugshot database with 1.6 million images, the rank 1 false negative rate was 0.1-0.66 percent. But the rank-1 false negative rate was 0.13-1.17 percent when the database contained 12 million images. (See fig. 8 for these results (Vendors A and B) plus results for four other vendors.) NIST states that "as the number of enrolled subjects grows, some mates [i.e., correct matches] are displaced from rank-1, decreasing accuracy. For N up to 12 million, false negative rates generally rise slowly with population size.[34] Given that the database sizes are continually expanding, if all else is held constant, we would expect accuracy to slowly decrease over time. However, in reality, other factors cannot be held constant since algorithms are always changing. (The real-world benefit of a larger database would be a higher likelihood that the person of interest is in the database, a benefit not captured by the accuracy metric.)

[34]The NIST FRVT data presented here provide a limited but informative assessment of the effect of database size on accuracy. The change in accuracy at the rank-1 position indicates that the match was displaced from the rank-1 position, not that it is no longer in the candidate list. The FBI states that it worked with NIST to decide which algorithm to use, and the one they utilize had a false negative rate of 0.88 percent at rank-1 and 0.28 percent at rank-50, given that the FBI database contains over 45 million images.

Figure 8: False Negative Rates of Six Algorithms, Measured by Investigation Testing

This graph shows the effect of database size on accuracy by comparing graphs of mugshots at N is 12,000,000 (top) and 1,600,000 (middle), and the effect of image quality on accuracy comparing graphs of mugshots (middle) and webcam (bottom).



Source: GAO analysis of National Institute of Standards and Technology (NIST) Facial Recognition Vendor Test. | GAO-21-435SP

Note: Vendors A and B provide facial recognition algorithms to federal law enforcement. Vendors C, D, E, and F do not provide facial recognition algorithms to federal law enforcement.

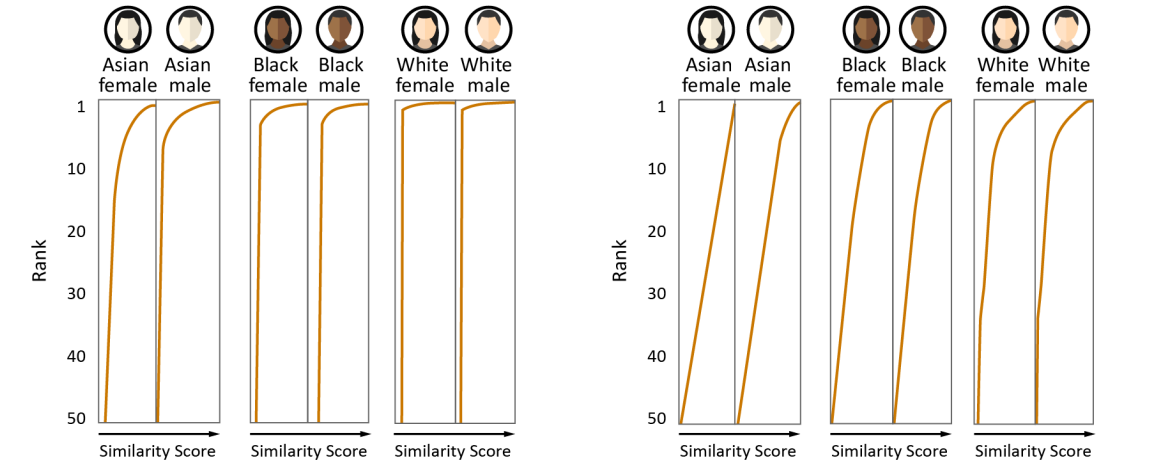
- **Image content (morphed images, glasses, unconstrained images, masks, etc.):** NIST's FRVT used wild and webcam images to test the effects of images taken under non-controlled conditions on facial recognition algorithm accuracy. Such images may show the individual wearing glasses or makeup, or the image may be taken from an extreme angle; these variations can affect algorithm accuracy. For example, an algorithm from a vendor used by law enforcement had a false negative rate of 5.1 percent when using wild probe images, 3.2 percent when using webcam probe images, and 0.7 percent when using mugshot probe images.[35] In a separate 2020 test, according to DHS's Science and Technology Directorate officials, they tested the accuracy of six commercial facial recognition camera systems with 10 commercial face recognition algorithms in a configuration similar to a security checkpoint where ID verification is required. The test evaluated the performance of the technologies on people who were and were not wearing face masks meant to protect from COVID-19. In total, DHS tested 60 face recognition system combinations. DHS officials also stated that they found that without masks, the median facial recognition system being tested (a combination of a face recognition camera system and a face recognition algorithm) had a 93 percent true identification rate, and the best system correctly identified individuals close to 100 percent of the time. With masks, performance declined, the median system had 77 percent true identification rate, and the best-performing system approximately 96

[35]All three tests used investigation methodology and indicated a false negative rate of the rank 1 position.

percent. Identifications can fail due to errors in different biometric system components. Identification errors were largely due to failures of facial recognition camera systems, not algorithms' ability to match acquired photos.

- **Demographics and ethnicity:** NIST published the first study of demographic bias, that is, differences in performance for different demographic groups, for facial recognition algorithms in 2019, finding that algorithms had different levels of accuracy for different demographic groups. NIST states that false positive and false negative rates are of importance to different communities. For example, in a one-to-many deportee detection algorithm, a false negative would present a security problem due to its failure to identify a formerly deported individual, and a false positive would flag legitimate visitors. NIST found that false positive rates varied by as much as 100-fold across demographics. For example, false positives were between two and five times higher for women than men. One algorithm—from a vendor that supplies algorithms to a federal agency—had a false positive rate three times higher in white women (0.00851 percent) than in white men (0.00275 percent), although both rates are still quite low. False negative rates—where an algorithm incorrectly fails to match two images when they are from the same person—did not vary as widely as false positives. However, NIST found that a few algorithms had small differences in accuracy across race, including algorithms from a vendor who supplies to a federal agency. Figure 9 compares the accuracy of two different algorithms based on demographic factors.

Figure 9: Demographic bias varies between algorithms

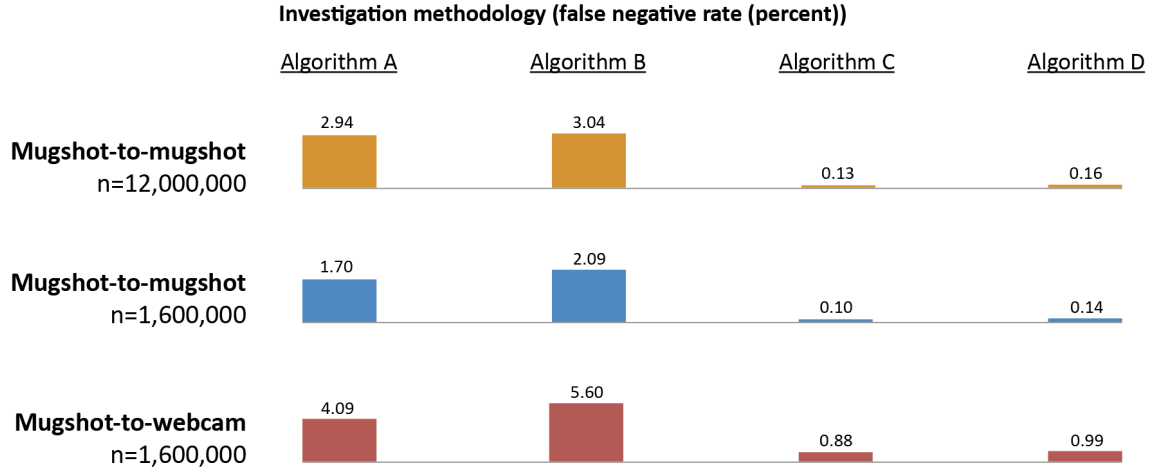


Source: GAO analysis of Figures 26 and 27 of the National Institute of Standards and Technology (NIST) Facial Recognition Vendor Test. | GAO-21-435SP

Note: The orange line represents similarity between the probe image and the matching image in the database for each rank (1-50) in the returned candidate list. The left-most algorithm is from a vendor who supplies facial recognition to federal law enforcement, while the right-most algorithm vendor does not provide facial recognition algorithms to federal law enforcement. The more vertical the orange line the higher the accuracy at higher ranks. The left most algorithm shows higher similarity scores at ranks approaching 1 indicating the probe image match is more often correctly identified in the top positions than in the right algorithm.

- **Versions and vendors:** Vendors provide algorithms for NIST testing as "black boxes"—without specific algorithm information. Thus, we cannot confirm whether they are the same versions of algorithms being used by law enforcement, only that federal law enforcement is using algorithms developed by those vendors. And because accuracy is algorithm-specific, we cannot confirm the actual accuracy of the algorithms being used. For example, analysis of the NIST FRVT results show four algorithms from the same vendor with false negative rates ranging from 0.10 to 2.09 when using a mugshot probe image (see fig. 12). According to the FBI, law enforcement agencies can ask their vendor whether the specific algorithm provided to them is represented in the NIST testing to determine the NIST-tested accuracy of the algorithm used by that law enforcement agency.

Figure 10: Variability in accuracy among four different algorithms developed by a single vendor



Source: GAO analysis of National Institute of Standards and Technology (NIST) Facial Recognition Vendor Test. | GAO-21-435SP

Note: Comparing the top and middle graphs also shows the effect of database size on accuracy; comparing the middle and bottom graphs shows the effect of image quality on accuracy. According to NIST officials, date of algorithm submission can also affect accuracy.

3.2 Facial recognition algorithms have two main strengths

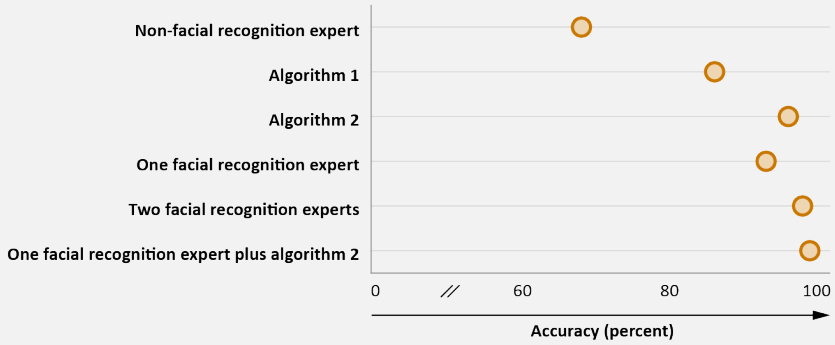
Independent experts and officials from law enforcement agencies identified two main strengths of facial recognition algorithms. First, they can search large databases much faster than human analysts. For example, according to one facial recognition vendor one of their algorithms returns a candidate list in 5 seconds from a test database of 363 people.

Second, facial recognition algorithms can be more accurate than human analysts. For example, one study reported that one facial recognition algorithm was more accurate than 73 percent of trained human analysts. This study also reported that the highest accuracy occurred when the most accurate algorithm was combined with a trained human analyst. This combination is standard practice for forensic investigations, according to law enforcement officials. Thus, the combination of algorithm and an expert human analyst could lead to identifying persons of interest more quickly and accurately when used appropriately.

Algorithm versus human-only performance

Facial recognition algorithm results can be improved when combined with input from a facial recognition expert. Researchers at the National Institute of Standards and Technology (NIST) and the University of Maryland compared the accuracy of different human analysts, algorithms, and an algorithm plus human analysts. Using 1:1 methodology, they found that the best performing algorithm was more accurate than facial recognition experts alone. The best algorithm had above a 95 percent true positive rate, while facial recognition experts were closer to 93 percent true positive rate, and an untrained analyst showed closer to 68 percent true positive rate. The researchers also found that the most accurate results are obtained when a facial recognition expert is combined with a top performing algorithm. The combination of a human expert analyst with the best algorithm had near 100 percent true positive rate, compared to two human expert analysts, combined with about 98 percent true positive rate.

Figure 11: A combination of algorithm and human expert analysis provides for the highest level of facial recognition accuracy



Source: GAO analysis of P. J. Phillips, et. al, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms." Proceedings of the National Academy of Sciences vol. 115, no. 24 (2018). | GAO-21-435SP

3.3 Challenges affecting law enforcement use of facial recognition algorithms

Law enforcement users face challenges affecting the use of facial recognition algorithms. While there have been improvements in the technology as a whole, there is wide variation in accuracy across different facial recognition algorithms.[36] We review four types of challenges: human involvement, testing and procurement, demographic effects, and public confidence.

3.3.1 Human involvement

One challenge to law enforcement use of facial recognition algorithms is the need for human involvement, which can introduce risks for errors and misuse at several points in a criminal investigation. For example, a human analyst selects a probe image to use with an algorithm and reviews the resulting candidate list.

Stakeholders representing both law enforcement users and defense advocates told us that human involvement is an important aspect in the process of using these algorithms. However, even when using a highly accurate algorithm, human involvement can also introduce errors at several points in the process. For example, an analyst sometimes alters a probe image to increase the chances of getting a candidate image from the database, according to a legal expert. Such alterations can consist of adjustments of color contrast, rotating the face so that it faces the front, or more drastic edits such as adding features, like open eyes over closed eyes.[37] This can introduce noise into the original probe image and may affect the resulting candidate list, according to the legal expert. Humans can also introduce bias or errors when interpreting the candidate list. For example, a 2020 study showed that algorithm outputs can cognitively bias analysts.[38] The study found that prior identity decisions, by either a computer or another human, influenced human decisions on whether a face pair was matching or non-matching.

A related challenge is that, despite such pitfalls, some law enforcement users may perceive facial recognition algorithms to provide more certainty in their results than is warranted. For example, some users may not understand the dependency of accuracy on high-quality probe images. Additionally, they may not understand how enhancements or modifications to the probe image might affect results. These algorithms can return a candidate list regardless of image quality or other factors that may affect accuracy. For example, if a user assumes the candidate list is automatically reliable, the user risks

[36] This statement refers to the hundreds of facial recognition algorithms reviewed by NIST, but the NIST review did not include all such algorithms, and not all the reviewed algorithms are being used by law enforcement.

[37] Rotating the face so that it aligns to the front, known as frontalization, uses facial landmarks to model a frontal image of the face. Frontal-facing images may improve accuracy of facial recognition algorithms.

[38] J. J. Howard, L. R. Rabbitt, and Y. B. Sirotin, "Human-algorithm Teaming in Face Recognition: How Algorithm Outcomes Cognitively Bias Human Decision-making," *PLOS ONE*, vol. 15, no. 8, (2020).

identifying the wrong individual as a person of interest.

Despite the importance of human involvement in the use of these algorithms, there is currently no standardized training or certification program for the individual who reviews facial recognition candidates against the probe image, according to stakeholders.[39] To fill this gap, the Facial Identification Scientific Working Group (FISWG) has developed recommendations for training of facial recognition algorithm users. Additionally, according to the FBI, the Facial Identification Subcommittee of OSAC is currently developing standards and guidelines for training of facial examiners, reviewers, and assessors. Likewise, the International Association of Identification is currently developing a certification program for facial examiners. Some agencies have adopted these recommendations. For example, the FBI requires that law enforcement users complete training prior to conducting facial recognition searches of NGI. This training must be consistent with guidelines outlined by the working group. However, law enforcement agencies using non-FBI databases for facial recognition algorithms do not always have this requirement. Some—but not all—state and local law enforcement agencies have specialized facial recognition analysts, but even then, law enforcement may not have standardized training or certification for appropriate use of the algorithms.

3.3.2 Testing and procurement

As described previously, the highest-performing algorithms can be more accurate than human analysts, and exhibit limited demographic bias. However, law enforcement agencies face challenges in testing and procuring the most accurate, least biased algorithms.

One such challenge is obtaining sufficient resources to procure and implement algorithms. According to one facial recognition algorithm vendor, resource-constrained law enforcement agencies may face barriers to procuring the highest-performing algorithms. These agencies could benefit from the use of these algorithms, but may not have the time or budget needed to procure them, according to the vendor. For example, one local law enforcement agency told us they selected an algorithm because the cost was relatively low and therefore within the budget of a small, non-federal agency.[40] Another local law enforcement agency estimated that replacing its current system would cost $1.5 million.

Another challenge is that agencies may not have enough information—such as comparative information on available algorithms and operational testing information—to help them select an algorithm. While NIST reviews over 200 voluntarily-submitted algorithms, not all algorithms used by law enforcement are submitted to NIST, and there are no requirements for vendors to do so. Federal

[39]Latent print and probabilistic genotyping algorithms have certifications for analysts who review the results or candidate list, but they are not required.

[40]Other factors in this local law enforcement agency's choice of algorithm included the ability to manage and track their own data, conduct audits of algorithm use on their own, and create their own user interface.

law enforcement agencies have generally procured algorithms from vendors that have submitted to NIST for testing, are found to have the highest accuracy, and have limited or undetectable demographic bias.[41] Some state and local law enforcement agencies do the same, but not all. Of the two local law enforcement agencies whose officials we interviewed who use facial recognition algorithms, one currently uses an algorithm from a vendor that submits to NIST for testing and the other previously used an algorithm from a vendor that does not submit to NIST. Furthermore, according to DHS Science and Technology Directorate officials, if a law enforcement agency identifies a preferred algorithm through testing, it may not be the same as the versions available for procurement. In this situation, the agency may not have detailed information about available algorithms to assure agency officials the algorithm will perform as desired. Additionally, the FRVT presents its testing information in a large and complex document that, according to stakeholders, may be difficult for law enforcement agencies to understand.[42]

Furthermore, NIST testing does not specifically address whether algorithms are valid for law enforcement use. FBI and DHS provide anonymized operational data for NIST testing; however testing does not necessarily reflect operational conditions, according to experts. The FRVT is performed in a controlled environment, so while the probe and database images are real-world data, the controlled algorithm test environment does not fully represent the real-world use case that includes human in the loop.. Law enforcement officials at some agencies told us that they conduct operational testing, which would provide a better indication of real-world algorithm performance. However, the results from operational testing are not easily accessible outside of the agencies, so the public and other law enforcement agencies may not be aware of the accuracy of algorithms that law enforcement agencies use.[43]

3.3.3 Demographic effects

Some recent reports have noted that some facial recognition algorithms perform less accurately on different demographic groups;[44] however, testing by NIST shows that the magnitude of these effects varies across algorithms. According to the NIST FRVT report, these demographic effects are small in the highest-performing algorithms. Although federal agencies are primarily using the highest-performing algorithms, state and local law enforcement may not. Since vendors can choose not to submit their algorithms to studies such as FRVT, there is no information

[41]Some federal law enforcement agencies also report submitting photos to an algorithm from a vendor that does not submit to NIST for testing.

[42]NIST officials told us they offer to help law enforcement users understand the results.

[43]We interviewed non-federal officials for this report but did not conduct a generalizable survey, and thus cannot determine the prevalence of operational testing. Non-federal law enforcement we interviewed conducted operational testing, but this cannot be generalized to all law enforcement using facial recognition algorithms.

[44]For an example of such reports, see C. Garvie, A. M. Bedoya, J. Frankle, *The Perpetual Line-up: Unregulated Police Face Recognition in America* (Washington, D.C.: Georgetown University Law School, 2018).

on demographic effects across all algorithms available to law enforcement agencies.

As we recently reported, there is no consensus on the exact cause or interaction of multiple causes of performance differences between demographic groups;[45] however, we identified three possible factors specific to law enforcement use. First, people of color are disproportionately enrolled in the source mugshot databases searched by these algorithms. Incorrect matches will therefore tend to fall disproportionately on people of color, potentially resulting in a higher rate of false arrests or other negative interactions with law enforcement.

Second, algorithm developers and vendors do not have access to representative databases to "train" facial recognition algorithms to accurately identify faces. Training data has a large effect on the accuracy of facial recognition algorithms and larger, more representative data sets are crucial to addressing performance differences. One algorithm vendor said that for unbiased results, there needs to be balance between races, genders, and ethnicities in these data. However, cost and limited access to a large body of usable photographs can reduce vendors' ability to build representative data sets for training, according to experts. For example, one industry expert said the widely available data sets are not broadly representative of the U.S. population and that such data are costly to collect. Another expert noted that developers do not have access to the government data for training algorithms.

Third, image quality can exacerbate these demographic effects. As previously discussed, algorithm accuracy generally decreases with decreasing image quality. A 2019 study demonstrated that the magnitude of demographic effects can depend on the system used for image capture, which can affect image quality.[46] Similar to the FRVT, the study found these effects varied between the tested algorithms, and that performance differences between demographic groups were lowest with the highest performing algorithms.

3.3.4 Public confidence in facial recognition algorithms

Lower public confidence in facial recognition algorithms presents a challenge to both law enforcement users who want access to a powerful tool to aid investigations and the public who seek accountability and transparency from these agencies. Public mistrust of facial recognition algorithms can pose a challenge to law enforcement users if it leads to policies that restrict the use of the technology. For example, several localities have passed laws limiting or banning the use of facial recognition technology, due to concerns with privacy and misuse. As explained previously, the combination of a human expert analyst and top-performing algorithm can be more accurate than humans alone, and thus algorithms can be a powerful

[45]GAO, *Facial Recognition Technology: Privacy and Accuracy Issues Related to Commercial Uses,* GAO-20-522 (Washington, D.C.: July 13, 2020).

[46]C. M. Cook; J. J. Howard, Y. B. Sirotin, J. L. Tipton, A. R. Vemury, "Demographic Effects in Facial Recognition and their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science,* vol. 1 no. 1 (2019).

tool for generating leads in criminal investigations.

In some cases, the general public may not fully understand the types of controls that certain agencies have in place to govern use of the technology or its capabilities. According to federal officials, for example, a key misperception among the public is that the algorithms operate with little to no human oversight. In fact, as described above, current algorithms require human involvement. The algorithms provide a candidate list to human analysts and investigators, who make the decision about which suspects to investigate. An official from one law enforcement agency said that their algorithms simply replaced the act of searching through a series of mugshot books and selecting mugshots that looked similar to the suspect. Additionally, both law enforcement users and the public may assume that all facial recognition algorithms have the same capabilities and performance. However, an important conclusion from the NIST FRVT is that performance can vary significantly between algorithms.

Factors related to privacy and the images used for running searches may also reduce public confidence. Stakeholders we spoke with and literature we reviewed identified several sources of privacy concerns.[47] One concern is the use of photo sources such as driver's license databases and social media. While such databases may be more representative of the population and thus less likely to be a source of demographic bias compared to a mugshot database, some believe the use of these photos in criminal investigations represents an expansion of law enforcement access to personal data, to which the public did not consent. Both mugshots and driver's license photos come from the government and therefore are an authoritative source of identity, according to stakeholders. However, social media photos, which some facial recognition algorithms use for their database, are not authoritative and present another privacy concern. In addition, some express concern that it is difficult or impossible for people to unenroll or opt-out once they are enrolled in a database used for facial recognition. Further, some are concerned that facial recognition use could lead to general law enforcement surveillance of the public.[48]

Another potential cause for lower public confidence is the wide variation in standards and policies related to law enforcement use of facial recognition algorithms. For example, standards for the quality of probe images vary. An FBI official told us that non-federal law enforcement agencies may use probe images that the FBI would reject due to its higher image quality standards. As discussed previously, some law enforcement agencies may allow editing of probe images (see sec. 3.3.1). Some researchers have expressed concern that the use of low-quality images or

[47]We have also previously discussed facial recognition privacy concerns in several reports: GAO-20-522; *Facial Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains*, GAO-19-579T (Washington, D.C.: June 4, 2019); and *FACE Recognition Technology: FBI Should Better Ensure Privacy and Accuracy*, GAO-16-267 (Washington, D.C.: May 16, 2016).

[48]Surveillance recognition conducts continuous live capture of all individuals in a camera's view and continuously compares them in real time to a list of individuals created by law enforcement. If the program finds a match, it alerts law enforcement. According to the FBI, surveillance technology of this sort is not known to be in use in the U.S. by law enforcement.

edited photos could increase the risk of identifying the wrong individual as a person of interest. Agencies may also have different standards for what sources are permissible for database images. For example, some states allow law enforcement agencies to search DMV photos, while others do not. Additionally, some agencies may limit searches to official government photos, while other agencies may use non-governmental source images, such as social media photos. In addition, policies surrounding the use of facial recognition in investigations can vary.

For example, FBI policy for NGI users prohibits the use of photos as positive identification and photos cannot serve as the sole basis for law enforcement action. Many law enforcement agencies have a similar policy; however, these agencies may not provide clear guidance to investigators on what additional evidence is sufficient to corroborate a potential facial recognition match, according to a 2019 report.[49]

[49]C. Garvie, *Garbage in, Garbage Out: Face Recognition on Flawed Data* (Georgetown University Law School, 2019).

4 Probabilistic Genotyping Algorithms

To assess probabilistic genotyping algorithms, law enforcement agencies and others test the influence of several factors on the likelihood ratio, such as the quality of the DNA sample and the number of contributors. Such testing, along with interviews with agencies and experts, suggest that probabilistic genotyping algorithms' three main strengths lie in their ability to analyze samples with multiple contributors, analyze samples with partially degraded DNA, and provide a numerical measure of the strength of the evidence. However, the algorithms face limitations and challenges.

4.1 Probabilistic genotyping algorithms are assessed using likelihood ratios that account for influencing factors

4.1.1 The key metric for assessing probabilistic genotyping algorithms is the likelihood ratio

FBI and other law enforcement agencies assess the reliability of a probabilistic genotyping algorithm by reviewing the likelihood ratios it generates from known DNA samples. The likelihood ratio is a statistical measure that quantifies the strength of the genetic evidence. Probabilistic genotyping algorithms are assessed using known profiles to account for limitations of the algorithms and the uncertainty of the likelihood ratio calculations. In one such assessment, a peer-reviewed internal validation study conducted in part by the FBI showed that when a single-source sample (i.e., with only one DNA contributor) was compared to a known profile from that individual, the manually-calculated likelihood ratios were identical to likelihood ratios produced by the algorithm that the FBI uses, indicating that the algorithm correctly detected and quantified the presence of DNA from the known contributor.[50] Testing samples against the profile of a known non-contributor provided likelihood ratios approaching 0, indicating the algorithm correctly supported ruling out the presence of DNA from the non-contributor (see textbox). The study showed that the probabilistic genotyping algorithm was able to discriminate between contributors and non-contributors under those test conditions.

[50] T. R. Moretti, R. S. Just, S. C. Kehl, L. E. Willis, J. S. Buckleton, J.-A. Bright, D. A. Taylor, A. J. Onorato, "Internal validation of STRmix for the interpretation of single source and mixed DNA profiles." *Forensic Science International: Genetics,* vol. 29 (2017).

Similarly, another internal validation study—conducted by the Northeast Regional Forensic Institute, the New York State Police, and the algorithm vendor on a different probabilistic genotyping algorithm—found very high likelihood ratios for tests run with known contributors in the sample and very low likelihood ratios for tests with known non-contributors in the sample. Therefore, the study states, this is generally consistent with what one would expect if this specific algorithm is reliable.[51]

[51]The study conducted in part by the FBI and Northeast Regional Forensic Institute/New York State Police tests were conducted under different conditions at different times and are thus not comparable to each other.

Likelihood ratio ranges

Typical likelihood ratios of algorithms we examined that are used by law enforcement ranged from close to 0 to more than two quintillion. For example, an expert told us that some labs have decided to use an "inconclusive" range of 0.001 to 1,000. However the Scientific Working Group on DNA Analysis Methods (SWGDAM) has recommended against using this range as it may make some likelihood ratios sound ambiguous when the results actually indicate that the profile is not likely to contain DNA from the person of interest.[a] Comparatively, the DOJ stated that analysts shall not report a likelihood ratio as inconclusive. Among its duties, the working group recommends and conducts research to develop and validate forensic biology methods, and has promulgated a verbal equivalent for describing likelihood ratios, which can be used to convey the relative strength associated with a given likelihood ratio. While the working group recommends standards, it has not created specific numerical standards for probabilistic genotyping algorithms. Another source of standards for validating probabilistic genotyping in crime labs is the July 2020 ANSI/ASB Standard 018—Standard for Validation of Probabilistic Genotyping Systems.

One study conducted in part by the FBI on the algorithm it uses showed likelihood ratios greater than 45,000 for samples with a known contributor. In a separate study involving data from 31 labs, the FBI found that of the 28,250,000 known non-contributor samples tested, 20 of them had likelihood ratios above 10,000, ranging from 10,298 to 505,924. The remaining samples gave likelihood ratios below 10,000 for a known non-contributor, and according to the FBI the vast majority of these non-contributor tests yielded likelihood ratios of less than 1. These two studies indicate a general range of likelihood ratios that the FBI may encounter with the algorithm it uses and are consistent with what is predicted based upon mathematical theory. That is, the evidence suggests the study algorithm performs as it is expected. Resulting likelihood ratios will be dependent on the specific probabilistic genotyping algorithm being used, the parameters input by the user, and the specific forensic workflow used to generate the genetic profile.

[a]The Scientific Working Group on DNA Analysis Methods (SWGDAM) was created in 1988 by forensic scientists to engage scientists involved in validating new DNA technology. The DNA Identification Act of 1994, authorized the FBI to issue Quality Assurance Standards governing forensic DNA testing laboratories. See Pub. L. No. 103–322, tit. XXI, § 210303, 108 Stat. 2065, 2068 (codified as amended at 34 U.S.C. § 12591 (2021)). The working group recommends revisions to the FBI's Quality Assurance Standards for DNA analysis. Adherence to these Quality Assurance Standards is required by Federal law as a condition of a laboratory's participation in the National DNA Index System, which is the FBI's DNA source database. See 34 U.S.C. § 12592 (2021).

Source: GAO analysis of expert discussions; T. R. Moretti, R. S. Just, S. C. Kehl, L. E. Willis, J. S. Buckleton, J.-A. Bright, D. A. Taylor, A. J. Onorato, "Internal validation of STRmix for the interpretation of single source and mixed DNA profiles." Forensic Science International: Genetics, vol. 29 (2017); and J.-A. Bright, et al. "Internal validation of STRmix™ – A multi laboratory response to PCAST." Forensic Science International: Genetics, vol. 34 (2018). | GAO-21-435SP

Likelihood ratio is a measure of the strength of DNA evidence

Probabilistic genotyping algorithms account for many influencing factors—such as peak height, number of contributors, and total DNA amount—in the likelihood ratio calculation. Therefore, a failure to detect a contributor or rule out a non-contributor may not indicate an error or failure in the algorithm. For example, in a test sample in which it is known that someone is a contributor, a high-quality sample generally results in a high likelihood ratio from probabilistic genotyping analysis. Conversely, if the same sample is degraded, the likelihood ratio is likely to be low, despite the fact that the person was a contributor, because less evidence was available in the sample.

Conventional methods of DNA analysis provide a specific probability that the genotype of a person of interest would appear in a population with a given frequency. In conventional DNA analysis methods, if even one genetic marker is different or missing between the evidence profile and the source sample, this requires an exclusion—that the person of interest is not matched to the DNA in the evidentiary sample. This can occur, even if the person of interest is in fact a contributor but the sample is degraded and results in partial profiles. Such a sample is more likely to return a likelihood ratio of close to 0 in a probabilistic genotyping algorithm.

Source: GAO analysis of T. R. Moretti, R. S. Just, S. C. Kehl, L. E. Willis, J. S. Buckleton, J.-A. Bright, D. A. Taylor, A. J. Onorato, "Internal validation of STRmix for the interpretation of source and mixed DNA profiles." Forensic Science International: Genetics, vol. 29 (2017); and B. Stiffelman, "No Longer the Gold Standard: Probabilistic Genotyping is Changing the Nature of DNA Evidence in Criminal Trials," Berkeley Journal of Criminal Law, vol. 24, no. 1 (2019). | GAO-21-435SP

4.1.2 Probabilistic genotyping algorithms are assessed across a variety of influencing factors

Law enforcement agencies test for the influence of several factors on the likelihood ratios that probabilistic genotyping algorithms produce. The purpose of this testing is to confirm that the algorithms are functioning as expected when integrated into the agency's processes. For example, studies showing that likelihood ratios for samples with a known contributor increase along with the quality and quantity of DNA are consistent with the expectation that better evidence leads to higher likelihood ratios. For example, a study conducted by the FBI showed that likelihood ratios for samples with a known contributor trend downward when the sample is repeatedly degraded.[52] Such results give the agency greater confidence in the systems they have acquired. Algorithm developers also test these influencing factors, as a similar check on their systems and to help them make improvements if needed.

Several factors influence the strength of genetic evidence that comes from probabilistic genotyping algorithms (as measured by the likelihood ratio). These factors, described below, can act together to influence the likelihood ratio and can also affect the likelihood ratio independent of each other. For example, a large amount of DNA could still result in a low likelihood ratio if the DNA quality is low.

Six factors influence the likelihood ratio:

- **Quality of DNA sample:** For a sample with a known contributor, a low-quality DNA sample will generally have a lower likelihood ratio compared to a high-quality DNA sample because there will be less information available for the probabilistic genotyping algorithm to analyze. To estimate the magnitude of this reduction, laboratories can prepare degraded DNA test samples by exposing DNA samples to ultraviolet light. One study published in part by the FBI demonstrated that, when comparing a non-degraded sample with a known contributor to a degraded sample, the likelihood ratio decreased by a factor of approximately 10,000,000 and the decreases in the likelihood ratio correlated with the level of sample degradation.[53]
- **Amount of DNA in the sample:** For an evidentiary sample with a known contributor, larger amounts of DNA generally yield higher likelihood ratios. For example, a peer-reviewed study conducted in part by the FBI showed that reducing the amount of DNA in a sample from 1 nanogram (ng) to about 0.03 ng led to a decrease of the

[52] T. R. Moretti, R. S. Just, S. C. Kehl, L. E. Willis, J. S. Buckleton, J.-A. Bright, D. A. Taylor, A. J. Onorato, "Internal validation of STRmix for the interpretation of single source and mixed DNA profiles." *Forensic Science International: Genetics,* vol. 29 (2017).

[53] Moretti, et al., "Internal validation of STRmix for the interpretation of single source and mixed DNA profiles."

likelihood ratio by a factor of over 50 trillion.[54]

- **Number of contributors to the sample:** As the number of contributors to an evidentiary DNA sample increases, the likelihood ratio returned by the probabilistic genotyping algorithm for a given contributor typically decreases. To measure this effect, labs use samples that simulate evidence with mixtures of DNA from two, three, four, or five people. In a published internal validation study conducted in part by the FBI, the likelihood ratio for one contributor to a two-person sample decreased by a factor of 1,000 when a third person's DNA was added to the sample. Officials at law enforcement agencies we interviewed told us that they do not use probabilistic genotyping algorithms for mixtures of more than four or five people because of the low likelihood ratios that result. According to the FBI, not all laboratories base policies on avoiding low likelihood ratios and the extreme computer memory and processor capabilities needed to deconvolute the complex four to five person mixtures may be what steer laboratories away from using algorithms on such mixtures.

Human analysts can decide the number of contributors for probabilistic genotyping analysis.

Analysts can determine the number of contributors on a case-by-case basis prior to running the probabilistic genotyping analysis. The higher the number of contributors the analyst selects the more this decreases the likelihood ratio. According to the FBI, depending on the case scenario and the comparison performed (e.g., defendant is the person of interest), decreases in the likelihood ratio are typically favorable to defendants, as the lower the likelihood ratio, the more support for the hypothesis that the person of interest is not a contributor to the evidentiary sample.

Source: GAO analysis of agency interviews and documentation. | GAO-21-435SP

- **Ratio of DNA per contributor:** When a contributor's DNA makes up a lower ratio of the total evidentiary DNA in a sample, the likelihood ratio returned by the probabilistic genotyping algorithm may be lower for the contributor with a lower share of DNA in the sample. As with the other factors, this effect can be tested with prepared samples, in which the share of DNA from each contributor is known. When DNA evidence is used to create an electropherogram, each contributor's profile is determined to be a certain ratio of the total DNA present in the evidence. An internal validation study conducted in part by the FBI on the probabilistic genotyping algorithm it uses confirmed that, when the ratio of a given contributor decreases, the likelihood ratio also decreases. For example, in the 64 three-person mixtures analyzed for that study, a four-fold increase in the ratio of a contributor to the sample resulted

[54]Moretti, et al., "Internal validation of STRmix for the interpretation of single source and mixed DNA profiles." A nanogram is one billionth of a gram. It is a typical unit of measure for working with DNA in a laboratory.

in a roughly 1,000,000,000-fold increase in the likelihood ratio.

- **Artifacts:** Artifacts, which are data in an electropherogram that cannot be associated with a genotype, can also reduce the likelihood ratio. Artifacts can be introduced as a by-product of the process used to amplify or detect DNA, among other things. Prior to the electropherogram evidence being analyzed by probabilistic genotyping algorithms, a human analyst reviews it to identify artifact peaks and can remove them from the analysis. Officials from the FBI told us that if the analyst is unable to remove all artifacts, likelihood ratios may decrease because the algorithm may start treating the artifacts as peaks from a contributor.
- **Genetic relationships:** When contributors to DNA evidence have close genetic relationships, known as *allele sharing*, likelihood ratios are reduced.[55] Allele sharing can make it harder to discern one contributor's profile from another because the profiles tend to produce some of the same peaks. According to the FBI, when two or more contributors to a DNA mixture share the same allele, it appears on the electropherogram as a single peak, making it more difficult to distinguish one contributor's profile from another. To test the effects of allele sharing on the ability to discern mixture contributors, DNA from different individuals that share alleles is mixed in various amounts and relative proportions. Testing the effects of allele sharing is challenging because the results are confounded by other factors, such as the amount of DNA a contributor has donated, mixture proportions, and artifacts. An internal validation test in which the FBI participated of the probabilistic genotyping algorithm that the agency uses showed that, in the 2,825 samples analyzed with more allele sharing, the likelihood ratio averaged about 1 million, whereas a sample with less allele sharing produced average likelihood ratios of about 100 billion.

4.2 Probabilistic genotyping algorithms have three main strengths

Probabilistic genotyping algorithms have three main strengths compared to conventional DNA forensic methods, including the following:

- **Analyze samples with multiple contributors.** The primary advantage to probabilistic genotyping algorithms is their ability to analyze complex evidentiary samples that an expert may not be able to otherwise analyze. Probabilistic genotyping algorithms can simplify, or *deconvolute*, electropherograms into all possible combinations of donors. A state police official told us that deconvolution of complex DNA is generally too computationally intensive for a human expert.

[55] An allele is a variant of a gene found in DNA.

- **Analyze samples with partially degraded DNA.** Probabilistic genotyping algorithms can analyze evidentiary samples with partially degraded DNA by using more information in an electropherogram than conventional methods of DNA analysis. By using more information inherent in the sample, probabilistic genotyping algorithms can account for missing information mathematically and provide a quantitative interpretation of the strength of the evidence. Likelihood ratios, even for degraded samples, may also be more useful than those obtained by conventional manual methods.
- **Provide a numerical measure of strength of evidence.** Probabilistic genotyping algorithms can provide a numerical strength of evidence by using more information from a profile, which investigators and others can use to determine how that evidence should be factored into their analysis of a case. Because the algorithms can use more of the information inherent in a profile, the likelihood ratio is likely to be more accurate than with conventional manual methods of analysis.

4.3 Challenges affecting law enforcement use of probabilistic genotyping algorithms

Law enforcement faces two main challenges affecting its use of probabilistic genotyping algorithms. First, analysts and investigators have a difficult task of interpreting and communicating the technically complex results of probabilistic genotyping algorithms. Second, there are challenges around validation, such as the lack of independent evaluation of validation studies.

4.3.1 Interpreting and communicating results

According to stakeholders and the literature, interpreting and communicating the results from probabilistic genotyping algorithms presents a challenging task for analysts and investigators. First, the analyst interprets the meaning of the likelihood ratio output by the probabilistic genotyping algorithm. Additionally, in some cases the analyst makes subjective decisions when conducting the analysis, such as assigning the number of contributors or identifying artifacts, according to stakeholders.

Another aspect of this challenge is communicating the meaning of a likelihood ratio, which can be prone to misinterpretation. According to experts, some consumers of DNA reports, including law enforcement professionals, lack the statistical education about what these likelihood ratios are conveying, viewing probabilistic genotyping algorithm results incorrectly as the chance that the suspect is guilty or not guilty. One agency official said that investigators who receive reports from probabilistic genotyping algorithms generally just look at the bottom line—whether an individual can be excluded or included. FBI officials explained that sharing technical information with others, including investigators, is the biggest challenge they face in working with probabilistic genotyping algorithms. One way they mitigate this challenge is through a mentoring process to help trainees learn

from more experienced analysts how to explain technical information.[56]

4.3.2 Validation

Law enforcement users face several challenges related to validation, which is the process used to improve algorithms and affirm that they work as intended:

Validation for complex mixtures. The authors of the 2016 PCAST report expressed concern that too few scientific studies have been conducted on the validity of probabilistic genotyping algorithms for complex mixtures.[57] The report noted that scientists consider objective methods of analysis to have been established only under very limited circumstances—namely, "a three-person mixture in which the person of interest constitutes at least 20 percent of the intact DNA in the mixture." PCAST only considered analyses under such limited circumstances to be reliable because those circumstances had sufficient numbers of published studies. As a result, the PCAST report urged forensic scientists to submit additional high-quality studies to leading scientific journals. The report noted that it is likely possible to extend the range over which scientific validity has been established to include more challenging samples. Prior to and continuing since the publication of the PCAST report, the FBI and others had begun to address these concerns by publishing validation papers on the probabilistic genotyping algorithms they use, aiming to increase confidence in the use of this technology. These include a 31-lab validation study conducted by FBI and other law enforcement agencies which found that the algorithms performed as expected, as described in section 4.1.1.[58] But some policymakers have called for NIST to conduct additional studies, testing multiple algorithms across a broader range of variables than has been previously done.[59] As described above, the Scientific Working Group on DNA Analysis Methods (SWGDAM) provides guidance on probabilistic genotyping validation. Another source of standards for validating probabilistic genotyping in crime labs is the July 2020 ANSI/ASB Standard 018—Standard for Validation of Probabilistic Genotyping Systems.

Lack of independent review. Most of the studies evaluating probabilistic genotyping software have been undertaken by software developers themselves or by law enforcement agencies. According to the PCAST report, establishing scientific validity also requires independent evaluation, but there have been few such studies. For example, NIST has not done a comparative study on probabilistic genotyping algorithms, as it has for facial and latent

[56]Communicating results is not just an issue between analysts and investigators; several stakeholders stated that the likelihood ratio is a difficult concept to explain to juries, judges, attorneys, and advocates.

[57]PCAST, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (September 2016).

[58]J.-A. Bright, et al. "Internal validation of STRmix™ – A multi laboratory response to PCAST," *Forensic Science International: Genetics,* vol. 34 (2018).

[59]See Justice in Forensic Algorithms Act of 2021, H.R. 2438, 117th Cong. (2021). This bill was introduced in the U.S. House of Representatives in April 2021, but has not been enacted into law.

print algorithms.[60] In the absence of such studies, these algorithms could lack sufficient independent verification for general acceptance.[61] Additionally, an expert told us some academics may face challenges obtaining research licenses to conduct independent validation studies. This expert said that one probabilistic genotyping algorithm vendor told a colleague that it does not provide research licenses.

Publishing validation studies. Both the authors of the 2016 PCAST report and the Scientific Working Group on DNA Analysis Methods (SWGDAM) have stated that validation studies should be published in a peer-reviewed journal. One agency official agreed that publishing peer-reviewed validation studies is a good step towards transparency; however, the official said many labs find they are not able to publish these studies in a timely fashion. According to an expert and the 31-lab validation study, scientific journal editors said they do not want to publish internal validation studies because they are not novel.[62]

According to NIST, empirical data that could be used to assess performance could also be made publicly available outside of journal articles. However, one expert noted that genetic privacy laws may prohibit making DNA profiles available, according to an expert. Software upgrades may also require new validation studies, which could add to an agency's burden (see text box).

Upgrades to software

A potential difficulty noted by one expert is the possibility that the output of forensic algorithms might change after algorithm upgrades. The Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines for validation require agencies to conduct another internal validation if there have been significant upgrades to a probabilistic genotyping algorithm. For probabilistic genotyping algorithms, this expert suggested part of the internal validation could include running past cases to see if the same results are reached to help ensure consistency. Moreover, SWGDAM recommends that data used during the initial validation may be re-evaluated as a performance check or for subsequent validation assessment. Given that probabilistic genotyping algorithms are probabilistic, the resulting likelihood ratios will vary from run to run. However, the variation in the resulting likelihood ratios from different runs is generally limited.

Source: GAO analysis of expert discussions. | GAO-21-435SP

[60]NIST sponsored interlaboratory studies where they presented different labs with the same set of data from DNA mixtures to interpret. But this is not the same as testing different algorithms under a range of influencing factors. J. M. Butler, M. C. Kline, M. D. Coble, "NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): Variation observed and lessons learned," *Forensic Science International: Genetics,* vol. 37 (2018). In June 2021, NIST published *DNA Mixture Interpretation: A NIST Scientific Foundation Review* as a draft report to receive public comment. Among other findings, the draft report states there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices, including the use of probabilistic genotyping algorithms. The NIST news release is available at: https://www.nist.gov/news-events/news/2021/06/nist-publishes-review-dna-mixture-interpretation-methods. The 250-page draft report can be accessed at https://doi.org/10.6028/NIST.IR.8351-draft and will be finalized after considering public comments received.

[61]Some groups stated that the lack of available source code for probabilistic genotyping algorithms is another challenge related to their use by law enforcement. (Some probabilistic genotyping algorithms are open-source.) Specifically in the context of court cases, they argued that source code should be available for defendants to determine whether the algorithm performs accurately. Some probabilistic genotyping algorithm vendors allow access to their software's source code, while others do not, claiming the information is proprietary. However, according to NIST officials, both developmental and internal validation have not historically used access to the source code Several stakeholders told us they believe validation testing is sufficient to identify coding errors.

[62]J.-A. Bright, et al. "Internal validation of STRmix™" p23.

5 Policy Options to Help Address Challenges with the Use of Forensic Algorithms

Forensic algorithms continue to advance. By automating assessment of evidence collected in criminal investigations, forensic algorithms can expand the capabilities of law enforcement and improve objectivity in investigations. For example, an algorithm's ability to analyze complex samples that an expert might not be able to feasibly analyze can help law enforcement identify individuals who may have been involved in a crime and exclude individuals who likely were not involved. However, use of these algorithms also poses challenges if the status quo continues, as described in this report.

We describe three policy options that policymakers could consider to address these challenges. The relevant policymakers could include Congress, other elected officials, federal agencies, state and local governments, academic research institutions, and industry. These policy options are:

- Support increased training to improve the use and understanding of forensic algorithms
- Support standards and policies on appropriate use of forensic algorithms in investigations
- Support increased transparency related to testing, performance, and use of forensic algorithms

For each option, we describe potential opportunities and considerations. The options address the major challenges we identified, but they are not intended to be exhaustive. Rather, we intend to provide a policy-focused base of information to aid in decision-making.

The options are neither recommendations to federal agencies nor matters for congressional consideration. We did not rank the options in any way. We are not suggesting that they be done individually or combined in any particular manner. We did not conduct work to assess how effective the options may be, and express no view regarding whether legal changes would be needed to implement them.

Policymakers could support increased training

Policymakers could support increased training for law enforcement analysts and investigators to improve their use of forensic algorithms and their understanding of results. Increased training could lead to more consistent and objective use of those algorithms. This could help address the challenges we identified related to human involvement, interpretation and communication of results, and training needs.

Opportunities	• Training on human factors could reduce risks associated with analyst error and decision-making. For all algorithms, training around system capabilities and limitations could reduce the potential for use of these algorithms on inappropriate evidentiary samples (e.g., low-quality latent print images or probe images). For example, increased training could help users better understand what evidentiary samples are of sufficient quality to be analyzed by forensic algorithms. • Additionally, increased training could help users or investigators understand and interpret the strength of the results they receive. For probabilistic genotyping, training around interpreting and communicating results could ensure that investigators better communicate the meaning of likelihood ratios. • For both latent print and facial recognition algorithms, training on cognitive biases could raise awareness of such biases and improve the objectivity of algorithm use in investigations. • Standards for training or certification of analysts or users could increase consistency and reduce risk of improper use across the various federal and non-federal labs and law enforcement agencies that use these algorithms. For example, in latent print analysis, where extensive training is required to achieve proficiency, standards such as those proposed by the forensic science standards group OSAC could ensure that latent print analysts across state and local labs and law enforcement agencies receive sufficient training. Well-enforced training standards may also improve use of facial recognition, given that not all law enforcement users follow existing training recommendations.
Considerations	• Certain training materials may need to be developed or made more widely available. For example, some non-federal law enforcement agencies may provide users of their algorithms with training, but smaller agencies may not have the resources (i.e., funding or personnel) to develop these materials on their own. • It may not be clear what entity should decide standards or certifications of training because multiple groups are involved in developing and disseminating training.

Source: GAO. | GAO-21-435SP

Policymakers could support standards and policies on appropriate use

Policymakers could support the development and implementation of standards and policies related to law enforcement testing, procurement, and use to improve consistency and reduce the risk of misuse. This could help address the challenges we identified related to human involvement, public confidence, and interpreting and communicating results. Some standards related to forensic algorithms already exist, and others are under development by standards groups, including NIST and OSAC. For example, agencies use standards to facilitate transmission of data for analysis by forensic algorithms. One step that may improve the development of new standards and policies may be to create a new forensic oversight body at the federal level, as recommended by the 2009 National Research Council report, or to assign a greater role to NIST and other federal agencies, as recommended by the 2016 PCAST report.

Opportunities

- Standards or policies that address the quality of data inputs—such as guidelines on what alterations of the probe image are acceptable, if any—could reduce improper use. Reducing improper use could in turn improve public confidence in forensic algorithms.
- Increasing the consistency of the standards and policies used by law enforcement agencies could also increase public confidence. For facial recognition algorithms, there is currently a wide variation in standards and policies on factors such as probe image quality and source.
- Setting further standards for testing facial recognition algorithms and performance thresholds for algorithms used by law enforcement could also help to reassure the public and other stakeholders that algorithms are providing reliable leads for further human investigation. Testing standards might require, for example, that algorithms show a minimum level of accuracy under law-enforcement-relevant conditions before an agency can procure them.
- Standards or policies for communicating results from forensic algorithms could help analysts, investigators, and other users better understand the strength of the evidence and come to an informed conclusion. For probabilistic genotyping algorithms, this could include best practices for communicating the meaning of likelihood ratios to investigators and other stakeholders.

Considerations

- Standards and policies may be difficult to implement across different levels of government, since federal and non-federal law enforcement agencies and crime labs answer to a variety of stakeholders. For example, they may be required to conform to different federal, state, and local laws and regulations.
- A patchwork of standards and policies already exists, and individual localities or agencies may be reluctant to conform to new standards. For example, some localities have already banned the use of facial recognition by law enforcement and it may be difficult to convince policymakers or the public to reinstate its use, even under new standards or policies.
- Implementing standards and policies may increase the cost of procuring and maintaining forensic algorithms to the point that law enforcement agencies with fewer resources can no longer afford them.
- The creation of standards can be resource-intensive, requiring research and testing as well as consensus from many public- and private-sector stakeholders. Standards development can require multiple iterations and take anywhere from 18 months to a decade to complete. Standards development organizations generally use a consensus-based process that requires careful coordination and collaboration across a myriad of stakeholders.

Source: GAO. | GAO-21-435SP

Policymakers could support increased transparency

Policymakers could support increased transparency related to testing, performance, and use of forensic algorithms by law enforcement agencies. This could improve stakeholder and public knowledge and help address the challenges we identified related to public confidence, testing and procurement, and demographic effects.

Opportunities	• The public may be more inclined to trust algorithms if officials provide free and easy access to results of operational testing, and to information about data sources, how algorithms are used, and for what types of investigations. For example, making operational testing results public could increase confidence in the accuracy and fairness of the algorithms. • Increasing the availability of comparative testing results and presenting them in a way that is easy for law enforcement and other non-technical users to understand could make it easier for agencies to select the best performing algorithms. NIST has conducted such comparative testing for latent print and facial recognition algorithms, but not for probabilistic genotyping. • For facial recognition algorithms, clearly identifying the software versions used in testing could also improve public confidence and help agencies choose algorithms. NIST provides the only publicly available test data, but the data do not include easily identifiable software versions, and some of the algorithms tested may not be commercially available. • Making more data sets publicly available could help developers improve algorithms and minimize undesirable demographic effects. In particular, for the training of facial recognition algorithms, there is a limited supply of usable photographs that are both representative of law enforcement conditions and sufficiently representative of all demographic groups.
Considerations	• Algorithm developers may not want to divulge proprietary information about databases they use for training and testing. • There may be privacy issues with sharing information about training and testing data. • Law enforcement agencies or crime labs may have difficulty finding peer-reviewed journals interested in publishing validation studies from testing.

Source: GAO. | GAO-21-435SP

6 Agency and Expert Comments

We provided a draft of this report to the Departments of Justice, Homeland Security, Defense, and Commerce with a request for technical comments. We incorporated agency comments into this report as appropriate.

We also provided a draft of this report to participants from our expert meeting and one additional industry stakeholder for review, and incorporated comments received as appropriate.

We are sending copies of the report to the appropriate congressional committees, relevant federal agencies, and other interested parties. In addition, the report will be available at no charge on the GAO website at https://www.gao.gov.

If you or your staff have any questions concerning this report, please contact me at (202) 512-6888 or howardk@gao.gov. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. Key contributors to this report are listed in appendix III.

Karen L. Howard, PhD
Director
Science, Technology Assessment, and Analytics

Appendix I: Objectives, Scope, and Methodology

We describe our scope and methodology for addressing the four objectives outlined below:

1. What are the key performance metrics for assessing latent print, face recognition, and probabilistic genotyping algorithms?
2. What are the strengths of these algorithms compared to related forensic methods?
3. What are the key challenges affecting the use of these algorithms and the associated legal, social, and ethical implications?
4. What options could policymakers consider to address these challenges?

To address all research objectives, we reviewed data and information about selected latent print, and facial recognition, and probabilistic genotyping algorithms used by federal and selected non-federal (state or local) law enforcement agencies for civilian criminal investigations as well as their strengths and limitations and challenges associated with their use. To do so, we conducted interviews with relevant federal agencies listed below; convened an expert meeting with assistance from the National Academies of Sciences, Engineering, and Medicine; conducted interviews with additional stakeholders, including selected non-federal agencies, nonprofit groups, and academic researchers; conducted a literature search; and reviewed relevant literature and case law.

Interviews

We interviewed key stakeholders in the field of forensic algorithms, including representatives or officials from:

- relevant federal agencies including the Department of Justice's (DOJ) Federal Bureau of Investigation (FBI), Drug Enforcement Agency (DEA), and National Institute of Justice (NIJ); the Department of Homeland Security's (DHS) Office of Biometric Identity Management (OBIM) and Science and Technology Directorate; the Department of Defense's (DOD) Defense Forensic Science Center; and the Department of Commerce's National Institute of Standards and Technology (NIST).
- five non-federal law enforcement agencies or crime laboratories;
- four vendors that develop forensic algorithms, spanning the three types of algorithms included within our scope;
- three academic researchers, including a legal scholar and two scientists;
- two industry consultants; and
- three nonprofit groups.

To select non-federal agencies to interview, we reviewed scientific articles, media articles, and case law to identify non-federal agencies that are using these algorithms and which algorithms they are using. We also reviewed information gathered from federal agencies. We created a list of non-federal agencies using algorithms from the same vendors as federal agencies and those using algorithms

from different vendors.[63] From this list, we interviewed representatives from five non-federal agencies that use one or more of the three types of algorithms and were geographically diverse, including agencies from the Northeast, Mid-Atlantic, Southeast, Midwest, and Northwest regions of the U.S. Selected non-federal agencies do not constitute a generalizable sample of non-federal law enforcement use of these algorithms.

Expert Meeting

We convened an expert meeting in collaboration with the National Academies to support this and our prior work. This 1½-day meeting included 16 experts on forensic algorithms used by law enforcement. We worked with the National Academies' staff to identify experts from a range of stakeholder groups, including federal agencies, academia, and industry. We evaluated the experts for any conflicts of interest. A conflict of interest was considered to be any current financial or other interest (such as an organizational position) that could (1) impair objectivity or (2) create an unfair competitive advantage for any person or organization. The 16 experts were determined to be free of reported conflicts of interest, except those that were outside the scope of the forum or where the overall design of our panel and methodology was sufficient to address them, and the group as a whole was determined to not have any inappropriate biases. (See app. II for a list of these experts and their affiliations.) The comments of these experts generally represented the views of the experts themselves and not the agency, university, or company with which they were affiliated, and are not generalizable to the views of others in the field.

We divided the meeting into five moderated discussion sessions based on key questions we provided on the following topics: (1) overview of forensic algorithms and their operational use; (2) characterizing the accuracy of forensic algorithms; (3) strengths and limitations of forensic algorithms; (4) key issues affecting usage of forensic algorithms; and (5) policy options relevant to the use of forensic algorithms. For sessions two through five, the discussion focused on latent print, probabilistic genotyping, and facial recognition algorithms. The meeting was transcribed to ensure that we accurately captured the experts' statements. After the meeting, we reviewed the transcripts to characterize their responses and to inform our understanding of forensic algorithms. Following the meeting, we continued to seek the experts' advice to clarify and expand on what we had heard. Consistent with GAO's Quality Assurance Framework, we provided the experts with a draft of our report and solicited their feedback, which we incorporated as appropriate.

Literature search and selected review

We conducted a literature search in support of all objectives. We conducted the search using a variety of databases, including ProQuest, EBSCO, Scopus, and NCJRS. In addition to the names of the types of algorithms in our scope we used search terms

[63] Conducting a full 50-state survey was outside the scope of this review.

such as “accuracy,” “validity,” “strength”, “limit” in support of objectives one through three. In support of objective four, we also used search terms such as “policy” and “rulemaking”. We narrowed our search to articles published within the last five years. For these searches, results could originate from scholarly or peer reviewed material, government reports, dissertations, working papers, books, and legislative materials. We selected the most relevant articles for further review based on our objectives. We also reviewed additional articles suggested to us through agency and stakeholder interviews.

Policy options

We formulated policy options to address challenges affecting the use of forensic algorithms and analyzed each policy option by identifying and discussing opportunities and considerations of their implementation. The policy options and analyses were supported by documentary and testimonial evidence from sources including interviews, the expert meeting, and the literature search.

We conducted our work from June 2020 through July 2021 in accordance with all sections of GAO’s Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

Appendix II: Expert Meeting Participation

We collaborated with the National Academies of Sciences, Engineering, and Medicine to convene a 1½-day meeting of 16 experts on forensic algorithms used in federal law enforcement. The meeting was held on January 15-16, 2020 in Washington, D.C. Many of these experts provided us with additional assistance throughout our work, including sending additional information for our review or reviewing our draft report for technical accuracy. The experts who participated in this meeting are listed below.

Sarah Chu
Senior Advisor on Forensic Science Policy
Innocence Project

Michael Coble
Associate Director of the Center for Human Identification
University of North Texas Health Science Center

Robert English
Special Counsel, Science and Technology Branch
Federal Bureau of Investigation

Tamara Giwa
Attorney, Assistant Federal Defender
Federal Defenders of New York

Patrick Grother
Scientist, Information Technology Laboratory, Information Access Division, Image Group
National Institute of Standards and Technology

William Guthrie
Division Chief, Statistical Engineering Division
National Institute of Standards and Technology

Karen Kafadar
Commonwealth Professor and Chair of Statistics
University of Virginia

Dan E. Krane
Professor and Interim Dean
Wright State University

James Loudermilk
Senior Director, Innovation and Customer Solutions
IDEMIA National Security Solutions

Anne May

Biometric Support Center Program Manager, Office of Biometric Identity Management

Department of Homeland Security

Mark Perlin

Chief Scientific and Executive Officer

Cybergenetics

Peter M. Vallone

Scientist, Biomolecular Measurement Division

National Institute of Standards and Technology

Kit Walsh

Senior Staff Attorney

Electronic Frontier Foundation

James L. Wayman

Editor-in-Chief

IET Biometrics Journal

Rebecca Wexler

Assistant Professor University of California, Berkeley School of Law

Michael Yates

Senior Technical Advisor on Biometrics, Science and Technology Branch

Federal Bureau of Investigation

Appendix III: GAO Contacts and Staff Acknowledgments

GAO contacts

Karen L. Howard, PhD, (202) 512-6888 or howardk@gao.gov

Staff acknowledgments

In addition to the contact named above, Hayden Huang (Assistant Director), Eleni Orphanides (Analyst-in-Charge), Rebecca Parkhurst (Analyst-in-Charge), Nora Adkins, Mariel Alper, Virginia Chanley, Eliot Fletcher, Darren Grant, Anika McMillon, Nikasha Patel, and Ben Shouse made key contributions to this report. Frederick K. Childers, Paul Kazemersky, Eric Larson, and Sean Manzano also contributed to this report.

(104389)

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website (https://www.gao.gov). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to https://www.gao.gov and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, https://www.gao.gov/ordering.htm.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or
TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on Facebook, Flickr, Twitter, and YouTube.
Subscribe to our RSS Feeds or E-mail Updates.
Listen to our Podcasts and read The Watchblog.
Visit GAO on the web at https://www.gao.gov.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact: Website: https://www.gao.gov/fraudnet/fraudnet.htm
Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Orice Williams Brown, Managing Director, WilliamsO@gao.gov, (202) 512-4400,
U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, YoungC1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149, Washington, DC 20548

Strategic Planning and External Liaison

Stephen Sanford, Acting Managing Director, spel@gao.gov, (202) 512-9715
U.S. Government Accountability Office, 441 G Street NW, Room 7B37N, Washington, DC 20548