United States Government Accountability Office

Testimony

Before the Subcommittee on Criminal Justice and Counterterrorism, Committee on the Judiciary, U.S. Senate

FORENSIC TECHNOLOGY

Algorithms Offer Benefits for Criminal Investigations, but a Range of Factors Can Affect Outcomes

Statement of Dr. Karen L. Howard,
Director,
Science, Technology Assessment, and Analytics

FORENSIC TECHNOLOGY

Algorithms Offer Benefits for Criminal Investigations, but a Range of 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 partially automate the assessment of evidence collected in a criminal investigation, potentially improving the speed and objectivity of their investigations.

GAO conducted technology assessments on the use of forensic algorithms in law enforcement (GAO-21-435SP and GAO-20-479SP). This statement addresses the benefits and challenges of three algorithm types—probabilistic genotyping, latent print analysis, and facial recognition—along with options policymakers could consider to help address these challenges.

In conducting the prior assessments, GAO interviewed federal officials, select non-federal law enforcement agencies and crime laboratories, algorithm vendors, academic researchers; convened an interdisciplinary meeting of 16 experts; and reviewed relevant agency documentation and literature.

GAO presented policy options in the 2021 report supporting this testimony.

What GAO Found

GAO's technology assessments in 2020 and 2021 found that federal law enforcement agencies primarily used three types of forensic algorithms to help assess whether evidence may have originated from an individual: probabilistic genotyping, latent print analysis, and facial recognition.

Figure. Federal law enforcement primarily uses three types of forensic algorithms.

Source: GAO (illustrations and information). | GAO-24-107206

Probabilistic genotyping algorithms compare collected DNA evidence (e.g., from blood, hair) to DNA samples from persons of interest. Such algorithms can assess a wider range of DNA samples than conventional methods but face some challenges. For example, it can be difficult to interpret or explain the results, and some experts told GAO that insufficient scientific studies have been conducted to fully validate their use for samples containing DNA from multiple people.

Latent print analysis can search larger databases of fingerprints and palm prints faster and more consistently than an analyst working alone. However, human involvement in using the outputs introduces opportunities for error and cognitive biases.

Facial recognition algorithms can also search large databases. They can be more accurate than a human analyst alone, but one study reported that the highest accuracy came from combining algorithms with trained analysts, as is common practice in federal law enforcement. However, human interpretation of algorithm outputs can introduce error or bias, and some law enforcement users may perceive the results as more certain than is warranted.

Furthermore, although some algorithms have high accuracy and very low bias, some law enforcement entities may not have enough information or resources to help them select such algorithms.

GAO's 2021 report described three options that policymakers could consider to help address key challenges to the use of forensic algorithms. Policymakers could support: increased training to improve consistent and objective use of forensic algorithms, standards and policies on appropriate use of forensic algorithms in investigations, and increased transparency related to algorithm testing.

United States Government Accountability Office

Chairman Booker, Ranking Member Cotton, and Members of the Subcommittee:

Thank you for the opportunity to discuss our work on forensic algorithms. My testimony today summarizes our May 2020 and July 2021 technology assessments describing the uses, benefits, and challenges of forensic algorithms for criminal investigations, along with policy options that may help address the challenges.[1]

For more than a century, law enforcement has used fingerprints and other physical evidence to help identify persons of interest, solve cold cases, and find missing or exploited people. Scientific advances are now allowing forensic experts to partially automate the examination of such evidence using algorithms.[2] In our prior reports, we found that federal law enforcement agencies primarily used three types of forensic algorithms to help assess whether evidence collected in a criminal investigation may have originated from an individual: probabilistic genotyping (an automated form of genetic analysis that is helpful for examining small or complex genetic samples), latent print (fingerprint and palm print) analysis, and facial recognition.[3]

My statement today addresses the benefits and challenges of those three algorithm types, along with options policymakers could consider to help address these challenges.

Forensic algorithms can be based on AI (e.g., facial recognition) or on mathematical models (e.g., probabilistic genotyping). With some forensic algorithms, like facial recognition and latent print analysis, numerous vendors offer forensic algorithms; previous testing by the National Institute of Standards and Technology (NIST) has shown that some of these algorithms are very accurate while others are less so. Experts told us that the best algorithms, when combined with human analysts, can offer significant benefits for law enforcement. However, forensic algorithm

---

[1]GAO, *Forensic Technology: Algorithms Used in Federal Law Enforcement,* (Washington, D.C.: May 2020) GAO-21-435SP; and GAO, *Forensic Technology: Algorithms Strengthen Forensic Analysis, but Several Factors Can Affect Outcomes,* (Washington, D.C.: July 2021). GAO-20-479SP

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

[3]Agencies also used algorithms to compare iris images, speech, and handwriting, to a lesser extent.

complexity and the variations between algorithms make it hard to understand and explain the results.

To examine these topics, we reviewed documents and interviewed officials from NIST and federal law enforcement agencies including the Department of Justice, the Department of Homeland Security, and the Department of Defense; conducted interviews with five non-federal law enforcement agencies or crime laboratories, four forensic algorithm vendors, and additional stakeholders; and reviewed relevant literature, including scientific articles and case law. We also convened a meeting of 16 experts including federal officials, researchers, legal scholars, defense advocates, and algorithm vendors. Additional information about our scope and methodology can be found in our May 2020 and July 2021 technology assessments. We performed the work on which this testimony is based in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments.

Probabilistic Genotyping

Probabilistic genotyping algorithms compare collected DNA evidence (e.g., from blood, hair) to DNA samples taken from persons of interest. Probabilistic genotyping can assess a wider range of DNA samples than conventional methods, such as when a sample contains DNA from multiple people or when the sample is small or partially degraded. The output is a numerical measure of the probability that the person's DNA is part of the sample—a measure called the likelihood ratio. A high likelihood ratio indicates a higher probability that the person of interest contributed to the DNA sample.

Probabilistic genotyping does have some limitations; notably that the results still depend on the quality and quantity of the DNA. For example, a high-quality DNA sample will generally produce a higher likelihood ratio than a lower-quality sample from the same individuals. Similarly, a larger amount of material will generally yield a higher likelihood ratio than a smaller amount of material from the same source. Furthermore, the likelihood ratio typically decreases for evidence that contains DNA from more individuals.

We also identified two key challenges with the use of probabilistic genotyping algorithms: (1) difficulties with interpreting the meaning of a likelihood ratio and (2) lack of sufficient validation.

Experts told us that interpreting the meaning of a likelihood ratio is challenging for several reasons, including that some law enforcement professionals lack the necessary training about what these likelihood

ratios convey. They may view probabilistic genotyping algorithm results incorrectly as the likelihood that the suspect is guilty or not guilty. However, a probabilistic genotyping algorithm could return a low likelihood ratio if, for example, the sample is degraded or contains very little DNA, or there are multiple contributors. 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 in an investigation. Federal Bureau of Investigation (FBI) officials also explained that sharing technical information with others, including investigators, in an understandable way is the biggest challenge they face in working with probabilistic genotyping algorithms.

Validation, or the process of confirming these algorithms work as intended, is also challenging.

A key report by the 2016 President's Council of Advisors on Science and Technology (PCAST) stated that insufficient scientific studies have been conducted to fully validate the use of these algorithms for complex mixtures. The PCAST report also noted that establishing scientific validity requires independent evaluation, but there have been few such studies. Most of the studies evaluating probabilistic genotyping software have been done by software developers or law enforcement agencies. For example, the FBI has conducted its own studies to address these concerns for the algorithms it uses. Some policymakers have called for NIST to conduct additional studies, testing multiple algorithms across a broader range of variables than has been previously done. NIST is a non-regulatory agency in the Department of Commerce and independent of law enforcement and vendors.

Latent print analysis

Latent print analysis is an automated method to compare fingerprints and palm prints collected during a criminal investigation to a reference database of prints. A latent print can be an incomplete or distorted print left on a surface and then collected during a criminal investigation. The latent print is digitally scanned and its details (or minutiae) are marked by a human examiner. For federal criminal investigations, the marked scan is then uploaded into the Automated Fingerprint Identification System (AFIS), which uses multiple algorithms to analyze the print. These algorithms can improve image quality and read the many minutiae specific to a print, then compare those to a database of palm prints or tenprints—prints from all 10 of an individual's fingers—taken under controlled conditions. The output of this analysis is a candidate list of individuals who may be the source of the latent print found during an investigation, ranked in order by likelihood. An expert independently

compares the recovered print to the prints in this list of candidates and based on their own judgment, reaches an identification, exclusion, or inconclusive decision.

We found in our previous work that latent print algorithms are advantageous because they can search larger databases faster and more consistently than an analyst working alone. Law enforcement officials told us another advantage of these algorithms is that they can improve consistency. Human analysts may come to different conclusions when presented with the same latent print images, and latent print algorithms do not suffer fatigue.

We also identified several limitations and challenges to the use of these algorithms. For example, performance is poor when the quality of the evidentiary latent prints is poor. Furthermore, errors during collection can result in unusable data. Importantly, the candidate list of results will only include individuals whose prints are in the comparison database. This means that the actual person of interest may not appear in the candidate list produced by the algorithm.

In addition to these limitations, law enforcement agencies face three key challenges in their use of these algorithms: human involvement, communicating results, and testing. Human involvement is necessary to assess the results of latent print algorithms, but this also introduces opportunities for human error and cognitive biases. Because the algorithms return a candidate list, which is then reviewed by the analyst, human errors and bias can influence the end result. A 2011 study showed that false positives in latent print decisions are rare. But, as the PCAST report noted, false positive results 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 a suspect for 2 weeks as a result of multiple analysts' errors and cognitive biases. One exploratory study showed that analysts' decisions when reviewing results could be influenced by knowledge of another analyst's prior judgement when considering the same prints.[4] Law enforcement officials noted that other cases have led to improvements in latent print analysis practices, such as additional education for analysts to limit the challenge of human error and bias.

---

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

A second key challenge is communicating the results to investigators and others. The results from a latent print algorithm do not include an assessment of the strength of evidence for or against a particular pair of prints being a match. Thus, when analysts communicate the results of their analysis, confidence in the results is generally based on factors such as the analyst's experience.

A third challenge is that the existing independent, comparative testing of the accuracy of these algorithms is out of date and does not always include key algorithms used by federal agencies. NIST last conducted comparative performance testing of latent print algorithms in 2012.[5] NIST currently conducts a latent print algorithm test relaunched in 2020 and last updated their test plan on March 9th, 2022. In addition, vendors supply their algorithms to NIST without software names or versions, so law enforcement agencies cannot use the NIST results to gauge the accuracy of a specific algorithm. A 2018 internal validation study conducted by the FBI on its current algorithm showed a 63.3 to 69.6 percent accuracy rate.[6] The accuracy rates from the NIST and FBI studies are not directly comparable because they involved different sets of latent prints.

Facial recognition

Facial recognition algorithms use AI to help analysts extract digital details from an image of a person's face collected as evidence (i.e., a 'probe photograph') and compare those details to images in a database. During federal criminal investigations that use facial recognition, the probe photograph is compared against the photos of known individuals in the FBI's Next Generation Identification System. Federal law enforcement may also use commercial facial recognition services with their own databases.[7] The algorithm generates a candidate list of individuals from the database, with a ranking from most to least similar to the probe photograph.

[5]According to FBI officials, these 2012 accuracy data are out of date. In a 2018 internal validation study, FBI found higher accuracy rates for the algorithm currently used in the Next Generation Identification (NGI) System; 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.

[6] Accuracy for latent print algorithms is measured as the percentage of attempts that returned the correct individual in the candidate list identified a possible source of the print.

[7] GAO, Facial Recognition Services: Federal Law Enforcement Agencies Should Take Actions to Implement Training, and Policies for Civil Liberties (Washington, D.C.: September 5, 2023), GAO-23-105607.

These algorithms are advantageous because they can search large databases faster and can be more accurate than analysts. For example, one vendor told us for our July, 2021 report that one of their algorithms returned a candidate list in 5 seconds from a test database of 363 people.

One study reported that an algorithm was more accurate than 73 percent of trained human analysts, and also reported that the highest accuracy came from combining the most accurate algorithm with a trained human analyst.[8]

One key limitation of facial recognition algorithms, as with latent print algorithms, is that the candidate list of results will only include individuals whose images are in the comparison database. This means that the actual perpetrator may not appear in the candidate list produced by the algorithm. As with other algorithms, the quality of the evidence (i.e., the probe image) will also affect the accuracy of the results. Our July, 2021 report also identified four challenges to law enforcement use of facial recognition algorithms: human involvement, testing and procurement, demographic effects, and public confidence.

Human involvement is an important aspect in the process of using these algorithms, according to stakeholders representing both law enforcement users and defense advocates. However, even with a highly accurate algorithm, human involvement can introduce errors. For example, a legal expert told us that analysts sometimes alter a low-quality probe image to increase the chances of getting a result, which can affect the candidate list produced by the algorithm. Such alterations can consist of adjustments of color contrast, rotating the face to the front, or more drastic edits such as adding open eyes over closed eyes.[9] Humans can also introduce bias or errors when interpreting the list. For example, a 2020 study showed that seeing the algorithm outputs can influence how analysts interpret other evidence.[10] The study found that prior identity

---

[8]See 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).

[9]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.

[10]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).

decisions, by either a computer or another human, influenced human decisions on whether a face pair was matching or non-matching.

Related to this, some law enforcement users may perceive the results of facial recognition algorithms as more certain than is warranted. For example, some users may not understand the extent to which the accuracy of the results depend 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 identifying the wrong individual as a person of interest.

A second challenge is that law enforcement agencies may face difficulty procuring the most accurate, least biased algorithms. NIST currently has an ongoing facial recognition test known as the Face Recognition Vendor Test (FRVT). Federal law enforcement agencies have generally procured algorithms found by NIST to have the highest accuracy and limited or undetectable demographic bias. However, some agencies, including those at the state and local level, may not have the information or budget needed to procure the most accurate algorithms. One local law enforcement agency told us they selected an algorithm because the cost was relatively low.

A third challenge is that some facial recognition algorithms perform less accurately on certain demographic groups. According to the NIST testing, these demographic effects are small in the highest-performing algorithms. As we reported on July, 2021, there is no consensus on the exact cause or interaction of multiple causes of performance differences between demographic groups; 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. Second, algorithm developers and vendors do not have access to representative databases to train facial recognition algorithms to accurately identify faces. Third, image quality can exacerbate these demographic effects. A 2019 study demonstrated that the magnitude of

demographic effects can depend on the system used for image capture, which can affect image quality.[11]

The fourth challenge is lower public confidence in facial recognition algorithms. 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. 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 tool for generating leads in criminal investigations.

According to federal law enforcement officials, 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. 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. A law enforcement official 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. 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.[12] Further, some are concerned that facial recognition use could lead to general law enforcement surveillance of the public. Another potential cause for lower public confidence is the wide variation in standards and policies related to law enforcement use of facial recognition algorithms. An FBI official told us that non-federal law

---

[11]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).

[12]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); GAO-22-106100 Facial Recognition Technology: Federal Agencies' Use and Related Privacy Protections (Washington, D.C.: June 29, 2022); and GAO-23-105607 Facial Recognition Services: Federal Law Enforcement Agencies Should Take Actions to Implement Training, and Policies for Civil Liberties (Washington, D.C.: Sep. 5, 2023)

enforcement agencies may use probe images that the FBI would reject as not meeting its higher image quality standards.

Policy options

Our July, 2021 report described three policy options that may help address key challenges to the use of forensic algorithms.[13] The relevant policymakers could include Congress, other elected officials, federal agencies, state and local governments, academic research institutions, and industry. Specifically, policymakers could:

• Support increased training for law enforcement analysts and investigators to improve consistent and objective use of forensic algorithms and understanding of results. This could help reduce errors and bias and address challenges with interpretation of results.

• Support the development and implementation of standards and policies related to law enforcement testing, procurement, and use of these algorithms. This could help reduce errors, bias, and public mistrust, and help address challenges with interpretation of results. Some standards related to forensic algorithms already exist and others are under development. One step that may facilitate the development of new standards and policies may be to create a new forensic oversight body at the federal level, as recommended by a 2009 National Research Council report. Another option could be to assign a greater role to NIST and other federal agencies, as recommended by the 2016 PCAST report.

• Support increased transparency related to testing, performance, and use of forensic algorithms by law enforcement agencies. Such transparency could improve stakeholder and public knowledge and provide more useful information to law enforcement agencies.

In conclusion, forensic algorithms have expanded the capabilities of law enforcement and can improve the speed and objectivity of evidence analysis in investigations. However, use of these algorithms also poses challenges, including difficulties with understanding and interpreting the results and shortcomings in testing and validation. Fortunately, policymakers do have options for addressing these challenges.

---

[13]GAO-21-435SP

Chairman Booker, Ranking Member Cotton, and Members of the Subcommittee, this concludes my statement. I would be pleased to respond to any questions you or other Members may have.

If you or your staff have any questions about this testimony, please contact Dr. Karen L. Howard 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 statement. Key contributors to this testimony include Hayden Huang (Assistant Director), Eliot Fletcher, and Rebecca Parkhurst. Additional contributors to the prior work on which this testimony is based are listed in our May 2020 and July 2021 reports.

Appendix I: Expert Meeting Participation

With the assistance of the National Academies of Sciences, Engineering, and Medicine, we convened 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