Reversing the legacy of junk science in the courtroom

Matthew Rakola

By Kelly Servick Mar. 7, 2016 , 4:30 PM

On a September afternoon in 2000, a man named Richard Green was shot and wounded in his neighborhood south of Boston. About a year later, police found a loaded pistol in the yard of a nearby house. A detective with the Boston Police Department fired the gun multiple times in a lab and compared the minute grooves and scratches that the firing pin and the interior of the gun left on its cartridge casings with those discovered on casings found at the crime scene. They matched, he would later say at a pretrial hearing, "to the exclusion of every other firearm in the world."

The detective's finding might have bolstered federal racketeering charges for two alleged gang members implicated in various crimes on that street. But the defendants' lawyers challenged its admissibility. The patterns on the cartridges from the lab weren't identical to those from the crime scene, they pointed out. So how could the detective be sure that the shots hadn't been fired from another gun?

Read more of our special package that examines the hurdles and advances in the field of forensics

The short answer, if you ask any statistician, is that he couldn't. There was some unknown chance that a different gun struck a similar pattern. But for decades, forensic examiners have sometimes claimed in court that close but not identical ballistic markings could conclusively link evidence to a suspect—and judges and juries have trusted their expertise. Examiners have made similar statements for other forms of so-called pattern evidence, such as fingerprints, shoeprints, tire tracks, and bite marks.

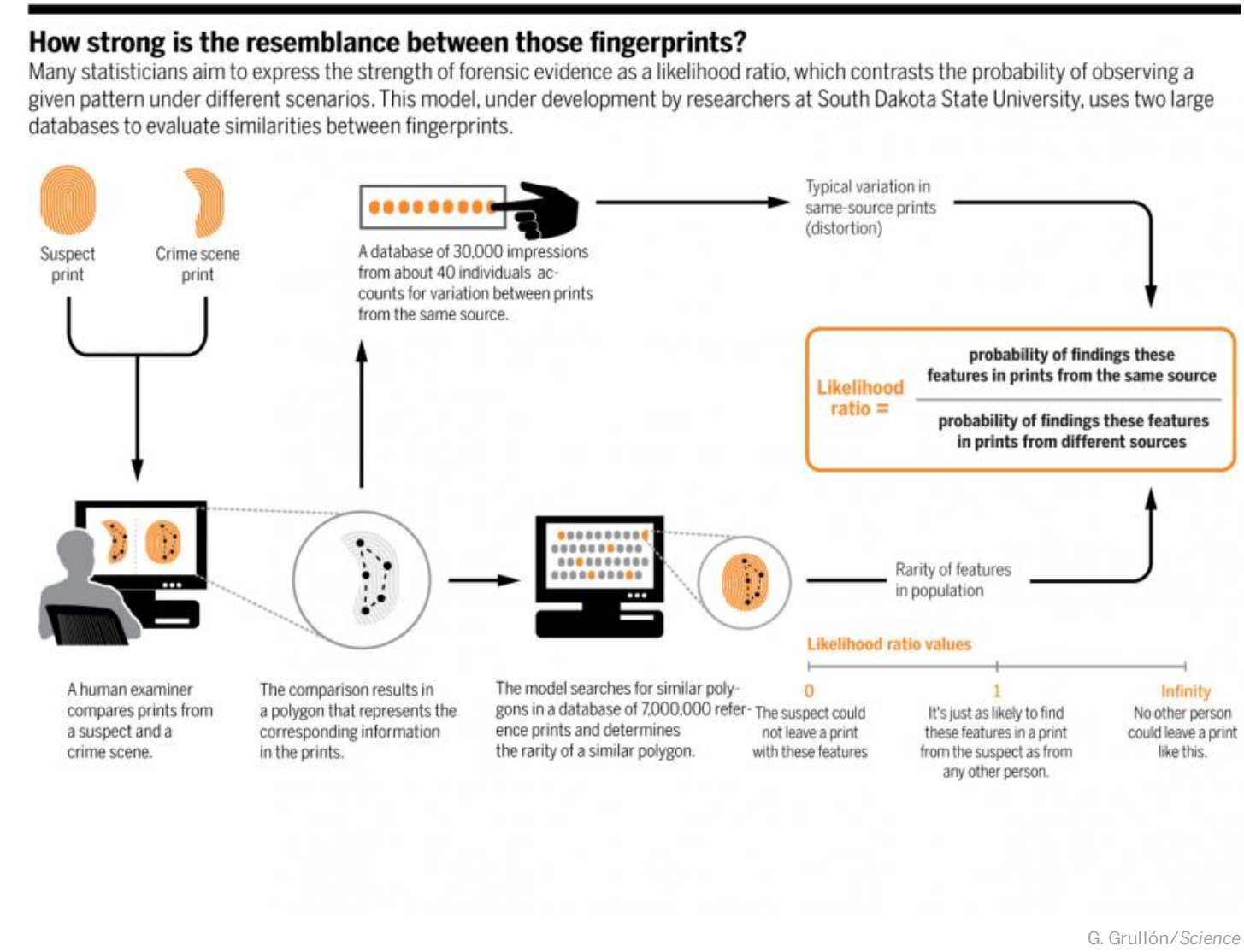
But such claims are ill-founded, a committee at the National Academy of Sciences (NAS) concluded in 2009. "No forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source," the panel wrote. In other words: Judges and juries were sometimes sending people to jail based on bogus science.

> **"When somebody tells you, 'I think this is a match or not a match,' they ought to tell you an estimate of the statistical uncertainty about it."**
>
> ***Constantine Gatsonis, Brown University statistician***

The committee's report sent shockwaves through the legal system, and forensic science is now grinding toward reform. A series of expert working groups, assembled by the National Institute of Standards and Technology (NIST) and the Department of Justice, has begun to gather and endorse standards for collecting and evaluating different kinds of evidence. What is needed, says Constantine Gatsonis, a statistician at Brown University, who chaired the NAS committee, is statistical rigor. "When somebody tells you, 'I think this is a match or not a match,' they ought to tell you an estimate of the statistical uncertainty about it," he says.

Last May, NIST awarded $20 million to a team of about 30 statisticians and legal professionals to help develop tools for analyzing the strength of an apparent match. Called the Center for Statistics and Applications in Forensic Evidence (CSAFE), it will collaborate with NIST statisticians to develop statistical methods that describe how strongly a shoeprint in the dirt links the owner of a certain pair of sneakers to a crime scene, for example, or how many fingerprints other than the suspect's might have left a similar pattern on a murder weapon.

The group is staring down a problem of immense complexity. Pattern evidence has historically relied on the trained eyes and subjective judgments of human examiners, not on rigorous statistical analysis. It's not known how much variation exists in the world's population of shoes, guns, or fingerprints, or just how much similarity between two patterns is enough to suggest a common source. "I know some people think we are not going to be able to do this, [that] you cannot put a probability on some types of evidence," says Alicia Carriquiry, a statistician at Iowa State University in Ames who heads CSAFE. "And they may be right, but we need to try."



G. Grullón/*Science*

Many forensic disciplines have been plagued with high-profile errors. An ongoing review of the Federal Bureau of Investigation's (FBI's) microscopic hair comparisons, in which forensic scientists look for distinguishing features such as the thickness, texture, and pigment in a hair strand, has revealed erroneous statements in more than 90% of cases before 2000 in which FBI examiners gave testimony. Often, analysts said that hair could be associated with a specific person—which hair analysis cannot prove. At least five of the cases reviewed so far ended in convictions later reversed with DNA evidence.

The analysis of bite mark patterns has been shown to be so weak scientifically that a state commission in Texas recently recommended banning it from the courtroom. In one high-profile case, a man named Ray Krone was convicted of murder after prosecutors used bite marks on the victim's neck and breast to link Krone to the crime; he served 10 years in prison before DNA evidence showed that he was innocent.

Even more-established methods, such as fingerprint comparison, have faced criticism. Many fingerprint analysts use standard procedures to mark different levels of detail in a suspect's fingerprint and in a "latent print" left at a crime scene. But making a so-called individualization

—a conclusion that the prints are from the same source—is "where it gets a little fuzzy," says Glenn Langenburg, a forensic scientist and fingerprint examiner at the Minnesota Bureau of Criminal Apprehension in St. Paul. After examiners look at enough prints known to be from the same source and from different sources, "their brain gets calibrated" to some internal threshold of similarity, he says.

The fuzziness shows in their findings. One study of 169 fingerprint examiners found 7.5% false negatives—in which examiners concluded that two prints from the same person came from different people—and 0.1% false positives, where two prints were incorrectly said to be from the same source. When some of the examiners were retested on some of the same prints after 7 months, they repeated only about 90% of their exclusions and 89% of their individualizations.

Testing examiner accuracy using known samples can give the judge or jury a sense of general error rates in a field, but it can't describe the level of uncertainty around a specific piece of evidence. Right now, only DNA identification includes that measure of uncertainty. (DNA analyses are based on 13 genetic variants, or alleles, that are statistically independent, and known to vary widely among individuals.) Mixtures of genetic material from multiple people can complicate the analysis, but DNA profiling is "a relatively easy statistical problem to solve," says Nicholas Petraco, an applied mathematician at City University of New York's John Jay College of Criminal Justice in New York City. Pattern evidence doesn't operate under the same rules, he says. "What's an allele on a tool mark?"; "What's an allele on a hair or fiber?"

What not to say in a courtroom

Several phrases often used to describe similarities in evidence are falling out of favor as statisticians begin to influence examiners and judges.

"TO A REASONABLE DEGREE OF SCIENTIFIC CERTAINTY"

A customary way to express the strength of scientific testimony that has "zero meaning," says Carnegie Mellon University in Pittsburgh, Pennsylvania, statistician Stephen Fienberg, a member of the U.S. National Commission on Forensic Science, which is urging the legal community to abandon the phrase.

"IT'S A MATCH."

A misleading fixture of TV crime dramas. "The correct testimony is, 'I've looked at these two things, and they look similar,'" says Brandon Garrett, a law professor at the University of Virginia in Charlottesville.

"THERE IS A 0% ERROR RATE."

Never true, says Alicia Carriquiry, a statistician at Iowa State University in Ames: "There's always a chance, even if it's infinitesimal, that there's a different explanation for your evidence."

"IDENTIFICATION"/"INDIVIDUALIZATION"

Often used in fingerprint analyses, these terms, too, "imply absolute certainty of the conclusion," the U.S. Army's Defense Forensic Science Center explained last year when it announced that it would no longer use them.

G. Grullón/*Science*

To estimate how frequently a given feature occurs in pattern evidence, researchers will need large databases. Carriquiry and her CSAFE colleagues will begin by exploring digital collections, such as images of bullet and casing marks assembled by NIST researchers, and one of the world's largest collections of crime scene shoeprints, kept by the Israeli police force. The team must also decide what aspects of an image are relevant for comparison. For example, sole patterns indicating the brand and model of a shoe may not be as informative for a comparison as acquired characteristics such as damage or wear patterns.

A large database and a set of rules for feature selection could then feed a statistical model that describes how unusual the set of similarities between two samples really is, relative to similarities between two randomly selected samples from the population. Ideally, says Carriquiry, the model would produce a "likelihood ratio." That would allow an examiner to say, for example, that the similarities between two fingerprints are 10,000 times more likely to occur if they came from the same finger than if they came from different ones.

For fingerprints, that kind of assessment seems within reach. A model under development by forensic scientist Cedric Neumann and statistician Christopher Saunders at South Dakota State University in Brookings can estimate a likelihood ratio for prints once a trained examiner marks their similarities (see diagram, above). The approach still isn't quite ready for use in court, says Neumann, in part because its results vary too widely depending on which features an examiner selects as relevant. Tighter standards for examiners could resolve the problem, he says.

For other types of evidence, the approach may never work, some scientists say. For instance, a relevant database of shoeprints might not be practical, says Lesley Hammer, a forensic scientist in Anchorage, Alaska, who specializes in footwear and tire track analysis. The database would have to keep up with an ever-changing market of brand-name and counterfeit products, document distinctive features like wear or damage patterns, and possibly even account for regional variations in shoe frequency—the likelihood of a snow boot turning up in Hawaii versus North Dakota, for example.

What statisticians manage to compute with their new models will have little value if forensic examiners, jurors, judges, and lawyers don't know how to interpret statistical claims. That's why CSAFE collaborator Brandon Garrett, a law professor at the University of Virginia in Charlottesville, has begun to study how jurors perceive a forensic examiner's testimony.

In a 2013 study, for instance, online participants had to rate the likelihood of a defendant's guilt in a hypothetical robbery based on different kinds of testimony from a fingerprint examiner. It didn't seem to matter whether they were simply told that a print at the scene "matched" or was "individualized" to the defendant, or whether the examiner offered further justification—the chance of an error is "so remote that it is considered to be a practical impossibility," for example. In all those cases, jurors rated the likelihood of guilt at about 4.5

on a 7-point scale. “As a lawyer, I would have thought the specific wording would have mattered more than it did,” Garrett says. But if subjects were told that the print could have come from someone else, they seemed to discount the fingerprint evidence altogether.

“When Neumann and his colleagues tested out their fingerprint likelihood ratios on mock jurors, participants recognized that making an “identification” was fundamentally different from providing a probability statement. But they didn’t seem to distinguish between a strong likelihood ratio (one in 100,000) and a weaker one (one in 1000) when estimating the probability that a suspect was the source of a print. Neumann suspects that numbers can still be useful for describing testimony, but that lawyers and cognitive psychologists will have to team up to figure out the best presentation.

The final decision about what kinds of statements jurors can ponder, though, is up to judges, who often confer with lawyers and forensic examiners to decide what evidence is admissible. CSAFE aims to reach all these players through a campaign to boost statistical literacy. Last week, the statisticians conducted training for practitioners across Florida crime labs at the Palm Beach County Sheriff’s Office, and they intend to launch similar courses around the United States.

Some judges are already pretty savvy about statistics. In the Boston racketeering case, federal district court judge Nancy Gertner found the detective’s conclusion that only one gun on the entire planet could have produced the imprints on the bullet cartridges “preposterous.” She believed the evidence should have been excluded completely. But Gertner—now a professor at Harvard University—feared that an appeals court would reverse that move, so she “reluctantly” ruled that the detective could describe ways in which the bullet casings looked similar, but not conclude that they came from the same pistol. Ultimately, a jury said there was no evidence of a racketeering operation; Gertner cleared the defendants of the more serious federal charges and their cases were moved to state court.

What’s troubling, Gertner says, is that when judges accept junk science, an appeals court rarely overrules them. Attaching a numerical probability to evidence, as CSAFE hopes to do, “would certainly be interesting,” she says. But even a standard practice of critically evaluating evidence would be a step forward. “The pattern now is that the judges who care about these issues are enforcing them, and the judges who don’t care about these issues are not.”

Posted in: **Forensics**

DOI: 10.1126/science.aaf4158

Kelly Servick

Kelly is a staff writer at *Science*.

Email Kelly | Twitter