Comments on the proposed GAO Forensic Technology algorithms draft
Section 2.3, Subsection PGS

Mark Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA
14 April 2020

2.3 How Forensic Algorithms Work

Probabilistic genotyping software (PGS)

152. PGS is used to assess complex DNA evidence.

Correct.

152-154. PGS provides a likelihood that an individual's DNA is found in the DNA sample, which could include DNA from multiple individuals, collected during an investigation.

Incorrect. A "likelihood" is a probability of evidence conditioned on variable values or a hypothesis. Rather, PGS provides a "likelihood ratio" (LR), which is the statistical support for a person having left their DNA in a sample.

154-155. Traditional DNA analysis cannot do this when the sample is small or includes DNA from more than two people.

Partially correct. Some less sophisticated DNA interpretation methods can give results for an evidence sample containing more than two contributors if there is a clear major component. Also, traditional interpretation methods usually derive little or no identification information from a minor contributor (e.g., under 50% contribution for a two-person mixture).

156-157. The usual first step in PGS analysis is to gather genetic material from both the evidence and the individual in question.

Incorrect. A reference individual is needed in the final step to calculate a match statistic for a suspect. However, in the first step, advanced PGS does not use a genotype from "the individual in question," which ensures greater objectivity. This suspect-independent problem-solving PGS feature enables searching databases for investigative leads.

157-158. Laboratory scientists then separately analyze the samples using a process that cuts DNA into fragments of different lengths.

Incorrect. The "cutting" mentioned here may refer to older Restriction Fragment Length Polymorphism (RFLP) DNA cutting that is no longer used in forensic identification. Rather, for over twenty years, forensic DNA identification has used Short Tandem Repeat (STR) genetic loci that are amplified using Polymerase Chain Reaction (PCR) that can look at very small amounts of DNA. DNA fragments are formed by exponential copying of short STR regions, not by RFLP cutting with restriction enzymes.

158-159. These DNA lengths are then measured and appear as peak heights in a computer graph.

Not exactly; the DNA length does not appear as a peak height. Each STR length variant (or "allele") is a DNA sentence of some length corresponding to the number of DNA nucleotide letters in the sentence. Allele lengths are measured with respect to a within-experiment calibration ladder.

Each allele's fluorescent signal is detected via laser to measure its relative amount, which can be displayed as a data peak. Taller peaks correspond to more amplified DNA, while a shorter peak indicates less allele signal. A set of allele peaks (i.e., their

lengths and heights) can be displayed on a “computer graph” called an electropherogram (EPG).

159-161. Because the resulting mix of fragment lengths differs among individuals, these peak heights represent a profile that can be used to distinguish one person from another (see fig. 1, first and second panels).

Unclear. The word “profile” has many different meanings. Here, it is used to mean the peak heights of the alleles that form an EPG data pattern. That DNA peak pattern (or data “profile,” as it’s called here) can be interpreted for human identification.

However, the pattern by itself does not distinguish people’s genotypes. It must be mathematically compared with many synthetic patterns formed from various proposed genotype combinations in order to assess the relative likelihood of contributing genotypes, or to separate out genotypes of individual contributors.

162. Figure 1. How Probabilistic Genotyping Software Works

The figure is incorrect. For simplicity, one locus should be shown. The individual’s EPG image should have data peaks from one person. The mixture evidence image should show peaks that are weighted combinations of several individuals’ genotypes. The genotype inference step should be added. The match meter should show a likelihood ratio (LR) match strength scale, which is both inclusionary (to the right) and exclusionary (to the left).

Please see a reworked “Figure 1”, provided as a separate file.

164-166. Next, laboratories compare the genetic profile of the evidence gathered with that of the individual in question, or with a database of DNA from known criminal individuals, or with other previously collected DNA samples (shown by the double-headed arrow in fig. 1).

This is incorrect. PGS does not compare an STR data peak pattern (called here a "genetic profile") with other data or a genotype. Genotypes are compared with genotypes. PGS abstracts away from the data level to a statistical genotype level. The inability of pre-PGS methods to compare genotypes contributed to their failure.

166-167. They often do this through a computer simulation of many different scenarios.

Not exactly. To "compare" genotype with genotype, PGS does not use computer simulation. However, to separate genotypes from mixture data, or assess genotype likelihood relative to evidence data, PGS often does conduct computer simulation.

167-169. The simulation is run by software that is made up of multiple forensic algorithms, which compare the DNA found during a criminal investigation with that of the individual.

Incorrect. It is unclear what is meant here by "multiple forensic" simulation algorithms. This sentence confuses genotype inference (of probability or likelihood) with genotype comparison (for calculating a LR match statistic). Moreover, PGS can compare evidence genotypes from different crime scenes, without any reference individual.

169-171. For example, an algorithm could compare the peak heights of the crime scene evidence to those of the genetic profile from the individual in question.

Incorrect. PGS does not compare evidence peak heights with reference peak heights. Valid statistical methods do not compare data with data; that is done by older failed manual review methods. Valid methods can compare genotype sets with evidence data (likelihood determination), or separated genotypes with one another (likelihood ratio determination).

171-173. The algorithm could then generate a set of hypothetical scenarios in which either the individual is one source of the DNA evidence, or the individual is not a source and the evidence came from other individuals.

Irrelevant. The described approach dates back over two decades. It is commonly used in non-PGS software to calculate LRs, and therefore not a useful distinction here. Moreover, such "hypothetical scenarios" are unnecessary for PGS operation. Typically, PGS is focused on deriving the genotype (probability or likelihood) of an unknown contributor to DNA evidence.

173-174. In these scenarios, the peak heights from other individuals is simulated based on the range of genetic profiles found in the general population.

Incorrect. It is unclear here why PGS would simulate peak heights from other individuals. However, genotypes from "the general population" are used in forensic statistics to represent a random person, but that is true of all DNA match statistics and is not specific to PGS.

174-177. From this comparison, PGS provides a likelihood that the DNA from the individual would have led to the genetic evidence profile that was obtained during an investigation, if the individual in question was a contributor to the evidence (see fig. 1, right panel).

Unclear. A "likelihood" is a probability of evidence conditioned on variable values or a hypothesis. But it is forensically meaningless in isolation. Rather, it is the *ratio* of likelihoods (or of genotype probabilities) that gives the relevant LR strength of match between two genotypes (e.g., evidence and reference), relative to a population.

177-178. This analysis determines the likelihood that the individual's DNA profile contributed to the DNA evidence.

Incorrect or irrelevant. Please see the preceding paragraph.

Conclusions

The proposed GAO section on how PGS works is largely incorrect. The described data and methods are wrong. The "likelihood" concept is used incorrectly, whereas the primary concepts of genotype, likelihood ratio, and probability are not mentioned at all. (This is, after all, a description of *probabilistic genotyping*; the words "probability" and "genotyping" should appear.)

Yes, it is true that older manual methods of DNA mixture interpretation fail. That has been well-known for over a decade. But that conceptual and practical failure are precisely what sophisticated PGS statistical methods rectify. Why dwell on past failures, confusing them with current working solutions?

The word "algorithm" is repeated throughout, but it is not relevant here. PGS is a statistical method, which of course uses data, mathematics, and computers to infer a useful result. But scientists and statisticians generally don't refer to such software systems as an "algorithm." There may be algorithmic components (e.g., Markov chain Monte Carlo statistical search), but the main PGS activity is statistical data analysis.

Key innovative concepts of PGS are missing. For example, manual DNA interpretation methods avoid uncertainty – to work within their cognitive limitations, people discard data and ignore possibilities. But uncertainty is at the heart of the PCR process, having well-understood probabilistic behavior. PGS embraces uncertainty, modeling PCR variation in order to accurately and objectively measure DNA identification information.

A more complete and accurate description of "how PGS works" is provided in a separate "PGS Rewrite" document. The proposed revision is about the same length as the draft section, and is written at a top level. However, unlike the GAO draft, the proposed revision accurately describes PGS and how it works.