Law, Probability and Risk (2014) **13**, 151–168 doi:10.1093/lpr/mgu002
Advance Access publication on April 2, 2014

The relation between sensitivity, similar non-matches and database size in fingerprint database searches

Thomas Busey†
Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana, USA

Arch Silapiruti
Indiana University, Bloomington, Indiana, USA

and

John Vanderkolk
Indiana State Police Laboratory, Fort Wayne, Indiana, USA

[Received on 12 September 2013; accepted on 11 February 2014]

Searching against larger Automated Fingerprint Identification System (AFIS) databases may increase the likelihood of finding a suspect in the database. However, Dror and Mnookin (2010) have argued that this also leads to an increase in the number of similar non-matching prints, which could lead to an erroneous identification. Using simulations, we explore the relation between database size and two outcome factors: close non-matching prints and overall database sensitivity, which is a measure of discriminability between true matches and close non-matches. We find that larger databases tend to increase both the likelihood of finding the suspect in the database as well as the number of close non-matching prints. However, the former tends to asymptote while the latter increases without bound, and this leads to an initial increase and then a decrease in the sensitivity of the database as more prints are added. This suggests the existence of an optimal database size, and that caution should be observed when interpreting results from larger databases. Quantitative evidentiary techniques such as likelihood ratios have the potential to address some of these concerns, although they too must consider the database size when calculating the likelihood ratio. Implications for practitioners are discussed.

Keywords: AFIS; close non-matching prints; errors; sensitivity; fingerprints.

Latent prints are typically recovered from crime scenes and often submitted to an automated database such as the Automated Fingerprint Information System (AFIS) to discover a set of prints from candidate donors for comparison. The databases vary in size, from tens of thousands in local in-house databases to the IAFIS system of the Federal Bureau of Investigation (FBI), which contains almost a billion fingerprints. In a recent article, Dror and Mnookin (2010) have argued that examiners should pay critical attention to the size of the database from which a candidate print is recovered. They suggest that although a larger database increases the likelihood that the suspect will be in the database, it also increases the likelihood of producing a similar, but non-matching, print.

This situation produces a potential dilemma for an examiner. If a suspect is developed through evidence other than a fingerprint (i.e. through the investigation), the chances of obtaining a close

†Corresponding author. Email: busey@indiana.edu

non-match simply by chance are quite small given the variations in fingerprints. Thus, similarities in detail and appearance between a latent print and the suspect's print in this situation should be considered quite valuable as evidence. However, if the candidate matching print is developed through a query of a database, this may lead to an erroneous identification, as the job of computer database search programmes such as AFIS is to provide similar fingerprints. Dror and Mnookin (2010, p. 58) argued:

> AFIS must, *by design,* increase the chances that the examiner will be presented with quite similar look-alike prints, as compared to those prints presented if suspects were identified through traditional investigative techniques rather than AFIS.

The goal of the present article is to provide a statistical analysis of the argument put forth by Dror and Mnookin, as well as to explore the tradeoffs between finding a suspect and encountering close non-matches in larger databases. In this article, close non-matches are comparisons in which a cursory inspection of the print might lead to a conclusion of similarity, but upon closer inspection it may be noted that it is different. However, this could also lead to an erroneous identification. In particular, we are interested in whether the number of close non-matches increases as the number of prints in the database increases, and if so, how this might affect overall similarity. In this context, sensitivity is the ability of the system (which includes the AFIS database and the examiner) to separate true matches from true non-matches.[1] Because this measure depends on both matching and non-matching prints, we must consider not only the possibility of finding a non-matching print that is sufficiently similar to the latent that it might lead to a false identification, but also the probability that the suspect is in the database. Larger databases might increase the likelihood of a false identification and also increase the likelihood of the suspect being in the database, making these two factors trade off in complex ways. Because both likelihoods grow as the database increase, but at different rates, we may find an optimal database size exists, and larger databases may produce diminishing returns.

One challenge of this analysis is that the AFIS vendors are reluctant to release the details of their matching algorithms. In the absence of details from vendors, we instead work from first principles to develop a simulation of the matching process. The assumptions of the model are spelled out below, and by parametrically varying different model assumptions we can explore the relation between database size, close non-matches, and sensitivity. However, note that without access to the AFIS matching algorithms, we will not have prescriptions at the level of computing the exact probability of making an error, or defining the optimal database size for a given circumstance. Nonetheless, the simulations reveal important relations between the different factors that an examiner may wish to include in their decision making process. We will offer specific recommendations to the practitioner at the end of the manuscript.

Before describing the details of our simulation, we would like to summarize the current state of fingerprint testimony in the USA Typically a latent print examiner would be asked to testify to one of several decisions. These indicate some degree of correspondence, such as 'identified', 'substantial areas of agreement', 'was made by' and similar statements. The examiner could also indicate a lack of correspondence, such as 'excluded', 'few areas of agreement', or 'was not made by'. In some agencies the testimony could include statements about the poor quality of the latent print, such as 'insufficient detail', or 'not of value'. This tradition of offering an option that in many cases constitutes a decision about the nature of the evidence can be contrasted against quantitative statistical techniques that can be

[1] We adopt the term 'sensitivity' from signal detection theory as a means to evaluate the performance of the system. This use of the term should be distinguished from the use in biometrics, where sensitivity refers to the probability that an individual with the trait is correctly classified.

used to measure both the similarity of a latent print to prints in a database, as well as the typicality of the latent print. These procedures have been developed by Neumann, Champod and colleagues (Egli *et al.*, 2007; Neumann *et al.*, 2007, 2006), and model both the within- and between-finger similarity to create a likelihood ratio that expresses the evidentiary value of a latent print given a particular database. Although the inclusion of such testimony is still rare, such techniques show promise to address both problems identified in the current manuscript and more general questions about whether forensic examiners should change how they present testimony in court. However, likelihood ratio approaches are not immune to the challenges presented by large database searches and we discuss how these challenges might be overcome.

1. Details of the simulation

Our simulation relies on a relatively abstract concept of *features*, which we take to be analogous to ending ridge or bifurcation minutiae, or larger features such as general ridge flow. It could even be taken to mean detail such as individual pore elements, ridge textures, creases or scars (Vanderkolk, 2009). In fact, because perceptual information is notoriously difficult to verbalize (Snodgrass *et al.*, 2004; Vanselst and Merikle, 1993), the actual features used by examiners are still unknown to some degree. Regardless, our simulation can still make the reasonable assumption that individual features exist and that examiners use individual features (either separately or in combination) to come to a general conclusion about the degree of similarity or match between a latent print and a candidate matching print.

The precise number of features is relatively unimportant, but to begin we assume that there are 20 features plotted in a fingerprint. The conclusions do not depend greatly on this number, and similar results are found if we assumed 5 or 100 features. This could be considered the typical number of features, and in a later section we discuss the consequences of introducing variability into the number of features.

1.1 *Modelling individual feature matching*

The next step is to compute a measure of the discrepancy between a feature in the latent print and an analogous feature in the candidate matching print. In AFIS systems, this is typically done by creating feature vectors that describe the positions and orientations of individual features such as minutiae. These in turn can be used to measure the difference, or distance, between any two features (and distance could be both differences in positions between features that appear to correspond, as well as differences in a feature vector space that might code feature appearance). However, AFIS vendors use proprietary and undisclosed feature vectors, and human examiners use an unknown feature set that could include other dimensions such as overall curvature, ridge space, a distortion model inferred from the appearance of ridges and the shapes of pores. Given this, we assume that a feature vector exists that can be used to compute a discrepancy score, but model only the resultant discrepancy score for each feature. The overall discrepancy score between the submitted print and a print in the database is the sum of the individual discrepancy scores calculated independently for each feature. This process can be repeated for each print in the database to develop a rank order of scores, much like the scores returned from an AFIS search. Although we assume that features are independent, breaking this assumption would simply reduce the effective number of features, which would change only the quantitative, not the qualitative nature of our conclusions.

We assume that the discrepancy score for each feature is an exponentially distributed random variable. The assumption of an exponential distribution has the properties that it is bounded at zero and has a long-rightward tail. This is consistent with the idea that similarity values cannot be negative, and there is no upper bound. The exact distributional form of this random variable is not critical; any lower-bounded distribution with a long-upper tail would produce similar qualitative results.

The discrepancy score can be viewed as an overall measure of dis-similarity between the feature and a corresponding feature in a database print. In Fig. 1, the blue (top) graph illustrates the distribution of the individual feature scores. Values close to zero imply that the corresponding features in the two prints will have very similar appearance, while values to the right correspond to dissimilar features. The top graph illustrates how the scores are spread out even for matching prints, which might happen if the latent print is of poor quality. However, if the prints come from the same individual, each feature is likely to produce a discrepancy score close to zero, imply strong correspondence. Thus the distribution of individual features on matching prints has most of its mass below a value of 2.

Prints from different individuals are likely to produce poor matches for each feature, due to variations in appearance across individuals and the fact that many features will not appear in both prints. The lower three curves in Fig. 1 represents the distribution of feature discrepancy scores for prints that come from three different individuals. Although there is some probability of finding similarity between two features (and therefore some discrepancy scores near zero), most of the discrepancy scores are more spread out than those from two impressions of the same finger.

The three green distributions in Fig. 1 allows us to explore different values for prints that come from the same and different individuals. These are analogous to intra- and inter-finger variability. Intra-finger variability comes from the fact that multiple impressions from the same finger will vary in appearance due to distortion or the deposition medium. Inter-finger variability comes from the fact that different fingers will have different features, which tends to increase the distance measure such that it typically will be larger than distance measures computed between impressions from the same finger.

Because we cannot know the exact distribution of discrepancy values of matching and non-matching prints (barring access to the AFIS comparison metric), we chose three different values of the mean of the exponential function that governs the distribution of discrepancy scores for matching and non-matching prints. Critically, the amount of overlap between the matching and non-matching features will determine the quantitative behaviour of the full model, although the qualitative patterns are unaffected by the exact choice of parameters for the curves in Fig. 1. We will return to a discussion of the assumptions underling the discrepancy scores in the Discussion. We will also comment on likelihood ratio metrics that provide estimates of the evidentiary value of a print given a reference database (Egli *et al.*, 2007; Kwan, 1977; Meuwly, 2006; Neumann *et al.*, 2007, 2006).

1.2 *Computing overall fingerprint similarity*

The overall discrepancy score between a latent print and a candidate match simply sums the individual feature scores to create a matching score, by making an assumption of independence between the features. By summing multiple variables (from the individual features), the overall discrepancy value becomes distributed as a gamma function with shape parameter of 20 (the number of features) and some scale parameter determined by the mean of the exponential distributions. The shape of the overall gamma functions is uniquely determined by the mean of the individual feature exponential distributions. For generality, we explore three possible values for the mean of the exponential, corresponding to the three means shown in the lower three panels of Fig. 1.

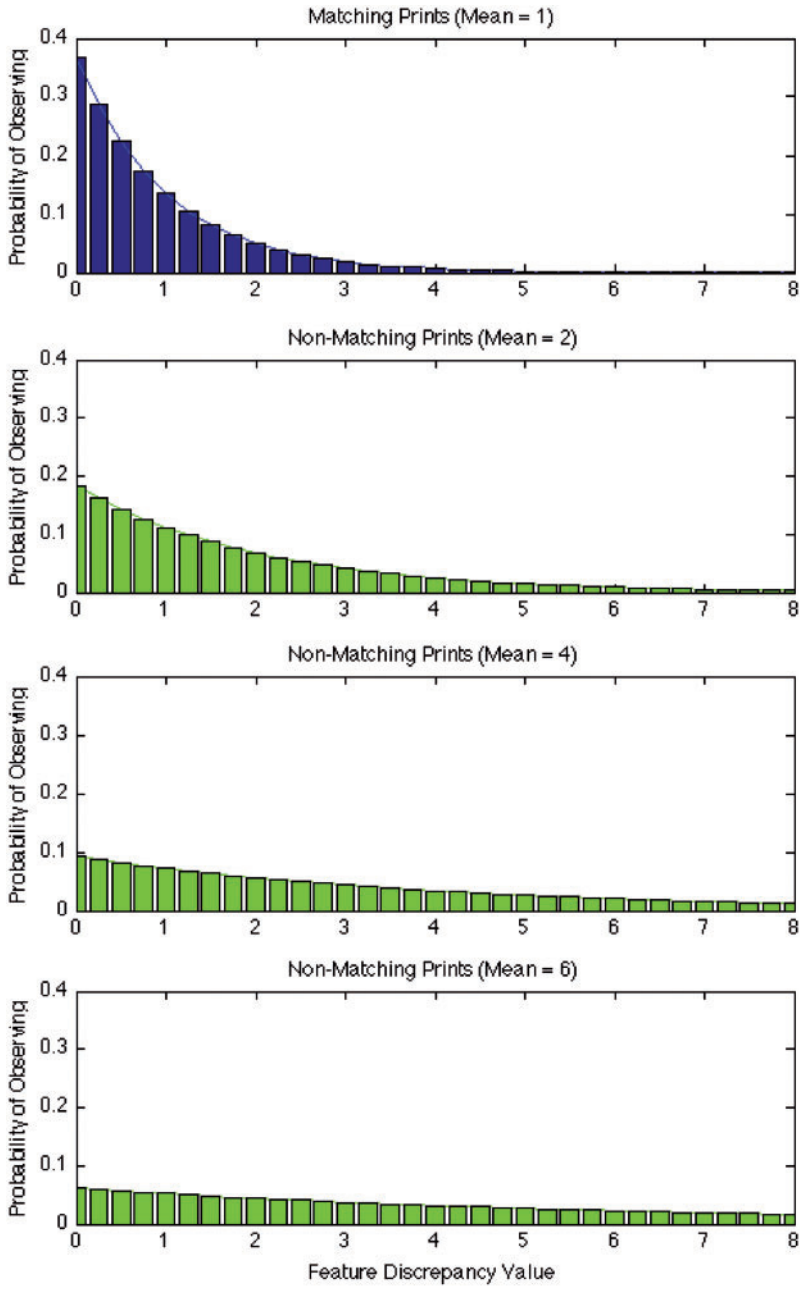

FIG. 1. Top Panel—Distribution of feature discrepancy values for matching prints. Most values cluster to the left, indicating little discrepancy between the features in the latent and ten print images for matching prints. Lower Three Panels—Distributions of feature discrepancy values for non-matching prints. Depending on the assumptions of how much discrepancy a non-matching print will have with a latent print, the discrepancy values take on different distributions. These are determined by the mean of the exponential function, which is listed above each graph.

Figure 2 illustrates the gamma functions that derive from matching and mismatching prints, respectively. The blue (left-most) curve comes from two prints from the same individual, while the green (right-most) curves represent prints from different individuals, assuming different ranges of discrepancy scores as shown in Fig. 1.

The values from the blue (left-most) distributions in Fig. 1 are assumed to come from multiple searches where the matching print is recovered and the discrepancy score is noted. The discrepancy score is analogous to the AFIS score, with the exception that most AFIS systems have high values to represent strong matches.

The blue (left-most) distributions in Fig. 2 are spread out because all latent prints have differing qualities and quantities of details, and this introduces variability in the discrepancy score calculation. This has the effect of spreading out the scores. Most of the overall scores these (matching) distribution are clustered around values near 20, which is consistent with the idea that the mean discrepancy value for matching prints was 1.0, and there are 20 features total. Note that even if one feature produces a perfect discrepancy score of zero, it is unlikely that all 20 will produce such extreme values and therefore the distribution of overall discrepancy scores shifts to the right and takes on a bell-shaped appearance. The distribution still has a slight right-ward skew, and theoretically extends to positive infinity. Practically speaking, however, most of the scores are less than 50, implying fairly strong correspondence.

The green (right-most) curves in Fig. 2 illustrate how the discrepancy scores for individual non-matching fingerprints are distributed given different assumptions about how similar individual features appear between two prints that come from different individuals. As with the matching prints, variations in quality and quantity of the latent print produce discrepancy scores that can vary, and in addition there are likely to be regions in the non-matching prints that do not have corresponding minutiae or other features that would produce strong matches with the latent. This further increases the discrepancy scores, pushing the distribution to the right.

The three green (right-most) curves in Fig. 2 are derived from different assumed means for the discrepancy scores of individual features coming from non-matching prints. The top green (right-most) curve shares substantial overlap with the blue (right-most, matching) distribution, suggesting that there may be instances where randomly chosen non-matching prints could produce discrepancy scores that are similar to those produced by matching prints. There appears to be very little overall for the middle and bottom non-matching curves and the matching (blue or left-most) curve. One might be tempted to conclude that there is little chance of confusion if we assume that non-matching prints produce discrepancy distributions with these parameters. However, our modelling will show that even these versions of the model will produce a large number of potential close non-matches when prints are sampled from large databases.

In the next section we explore how these distributions affect the discrepancy scores returned by AFIS as the database increases in size.

1.3 *Modelling errors*

The histograms shown in Fig. 2 represent the distributions of all discrepancy scores for randomly chosen matching and non-matching images. Errors tend to occur when a non-matching print has a very low discrepancy score, similar to those from the matching prints. At first glance, this would seem to be a very rare event for the bottom graph in Fig. 2, and somewhat more likely for the middle and top green graphs but still unlikely. However, prints are not chosen randomly by AFIS, they are selected on the

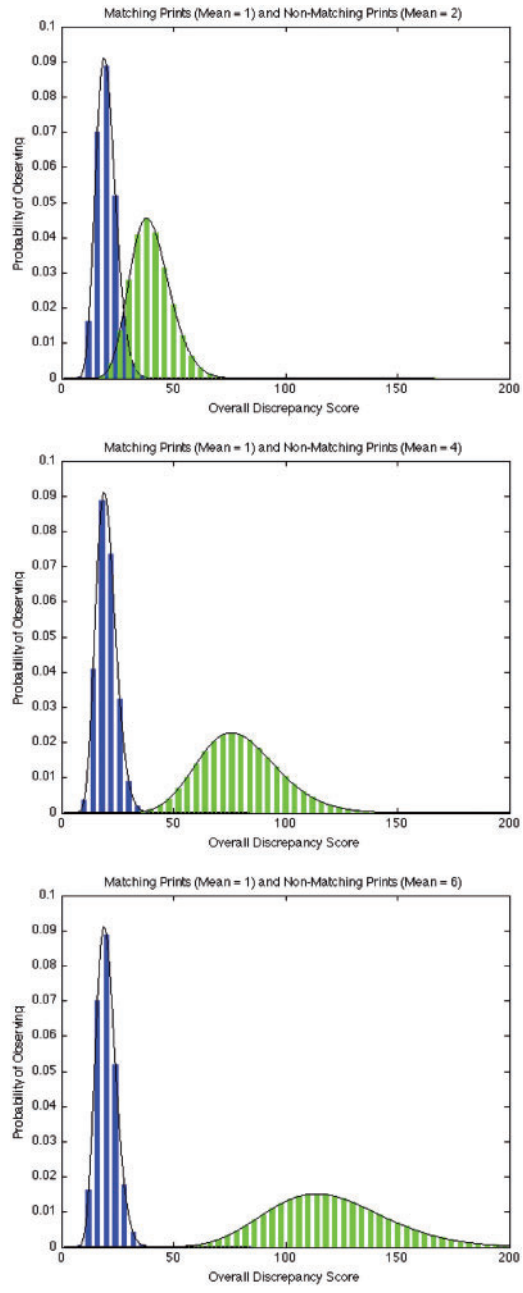

FIG. 2. Distributions of Overall Discrepancy Scores for individual prints. Blue (left-most) scores correspond to scores from matching prints, while green scores come from non-matching prints. The three panels represent different assumed means for the non-matching discrepancy score distribution as in Fig. 1. Note that if a discrepancy score mean of 2 is chosen, there is a great deal of overlap between the two distributions (top panel), implying that non-matching prints are likely to appear somewhat similar to the latent print. However, if a mean of 6 is chosen, there is much less separation between matching and non-matching overall discrepancy scores at least at the level of a randomly chosen non-matching print.

basis of their similarity to the latent print. Thus even though finding a non-matching print with a low discrepancy score (and therefore a high AFIS score) might seem to be a low probability event, there is some small but non-zero chance of it happening for *each* print in the database and thus the likelihood of it happening at least once grows with the number of prints in the database.

We are testing the proposition that as the size of the database increases, the chances of an extremely small discrepancy score (potentially leading to an erroneous identification) will increase as well. In fact, under some circumstances the discrepancy scores from false matches will be *smaller* than those from true matches. This is particularly problematic when the suspect is not in the database, yet the database produces a candidate print with an extremely small discrepancy score. In this situation the examiner will not have the true matching print for comparison, and may be convinced by the small discrepancy score from the false match. This has been described as the misleading evidence in favour of the prosecution (Champod and Meuwly, 2000). Of course, the examiner does not make a decision based on the discrepancy score (or AFIS score) but instead looks at the print itself. However, assuming a strong correlation between AFIS scores and image similarity, the images with strong AFIS scores are likely to be visually similar and could potentially lead to a false identification.

To characterize the likelihood of this undesirable event, we simulated 10 000 AFIS queries using databases of different sizes (and by database size we mean the number of fingerprints, not the number of individuals). Because we are focusing on erroneous identifications in this first analysis, we assumed that the suspect was always in the database and had only one entry. Later analyses will relax both assumptions when we turn to the relation between database size and overall sensitivity.

To model the influence of these rare close-non-matching prints, on each simulated query we constructed a distribution of discrepancy scores between the latent print and all prints in the database. These discrepancy scores represent the result of computing the distance between the latent print and all prints in the database, which in our model is simulated using the gamma function that represents the distribution of the distance calculations. In an AFIS system these would be computed directly from the feature vectors using whatever distance metric is used (e.g. Euclidian distance). Importantly, we then took all discrepancy scores and found the minimum of the non-matching prints along with the rank position of the matching print. This minimum represents the most similar non-matching print, which is the one most likely to produce an erroneous identification. These values are used in several ways to explore the consequences of database size on the number of close non-matching prints, as described next.

Perhaps the best way to visualize the behaviour of the model given different database sizes is to make an analogy to the ranked list returned by AFIS. In a typical query, the algorithm computes scores for some or all of the prints in the database, returning the most similar prints on the first screen of the programme. Prints that receive poorer scores are returned on subsequent screens, and the examiner can choose how many screens they search through before assuming that the matching print is not in the database. To characterize the behaviour of the model under different assumptions, we assumed that the matching print is in the database and computed the frequency that the matching print is returned in the first position in the list, or if not, how often it fell in one of the other positions near the top.

For example, Fig. 3 illustrates how often the matching print is returned in each of the top 20 positions by the simulation for databases of different sizes. For small databases (top row), most of time searches result in the matching print appearing at or near the top of the list, which is demonstrated by the fact that most of the searches returned scores that rank the matching print in the top 5 candidates. This is especially true for parameters where we assume the non-matching prints tend not to look like matching prints (mean = 6, right column), but less true for the parameters that assume more overlap

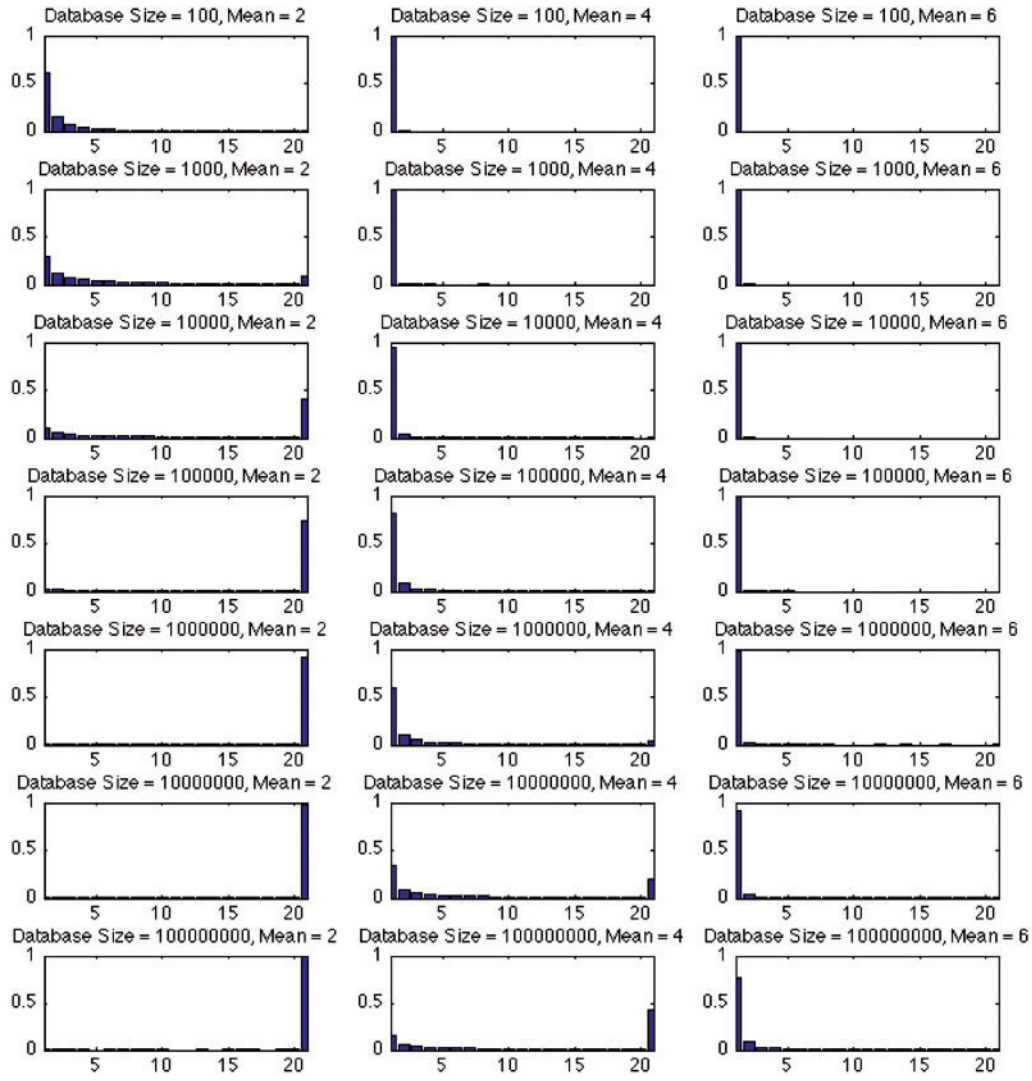

FIG. 3. Simulated AFIS ranks showing the proportion of time that the matching print appears in each of the top 20 slots. Each graph corresponds to one combination of a database size and the mean of the non-match distributions. The abscissa is the rank number, and the ordinate is the frequency of observing the matching print at each position on the AFIS screen. The bar graphs show the frequency of the matching print appearing at each location in the simulated AFIS return screen. The upper-left box shows that many of the simulations result in the matching print appearing on the first slot, often in the top 5, and occasionally lower. However, as the database size increases, the matching print slips off the first 20, as shown as the taller bar at location 21 (which summarizes all of the positions greater than 20).

between matching and non-matching prints (mean = 2, left column) where some of the matches are dropping down to lower positions. However, as the size of the search database increases, the matching print starts dropping to lower and lower positions, finally falling off the top 20 slots (shown in position 21 in the graphs).

There are many situations where the matching print does not always appear in the first position, which implies that there are non-matching prints that the algorithm considers more similar to the latent print than the true match. To illustrate how and why this is a problem, consider the fact that false matches with small discrepancy scores (so-called 'close non-matches') present the biggest problem in the identification process. It matters little whether most of the prints in the database will produce large discrepancy scores and therefore be discarded by the matching algorithm. Instead, the prints with the smallest discrepancy scores are those that are likely to produce false non-matches. How often does occur, and under what circumstances?

Our model provides answers to these questions. In Fig. 4 we plot two distributions. The first comes from true matches, which assumes that the suspect is in the database. The second takes the 'lowest' discrepancy score from the non-matching print comparisons. This is consistent with the idea that consideration of only the most similar-looking non-matching print is relevant for reducing errors, and the most similar-looking non-match will have the lowest discrepancy score. The green curves in Fig. 4 represent the minimum of the non-matching prints, sampled across the 10 000 simulated AFIS queries (equivalent to running 10 000 different latent prints against the database). The blue (darker grey) curves come from the distribution of discrepancy scores from matching prints. For small databases shown in the upper rows of Fig. 4, there is clear separation between the matching prints in blue (darker grey) and the non-matching prints in green. This implies that a close non-match is unlikely to produce a discrepancy value that is in the same range as those produced by matching prints.

However, this situation changes dramatically as the size of the database increases. The lower rows of Fig. 4 illustrate the same two distributions for databases of increasing size, ranging from 10 to 100 million, increasing by factors of 10. As the database size increases, the distribution of matching scores (shown in blue or dark grey) stays the same, while the distribution of the minimum discrepancy score across all non-matching prints moves closer and closer to the matching distribution. For a database of 100 million fingerprints, the two distributions are virtually identical, implying that the best matching and non-matching prints will give discrepancy scores that are approximately equal, at least for a mean of 4 or 6.

The worst-case scenario is in the left-hand column of Fig. 4. For large databases (greater than 10 000), the non-match distribution dominates the match distribution. In this situation, the model predicts that many more non-matching prints will have discrepancy scores that are smaller than true matches (implying larger AFIS scores). This indicates that there are many prints that have the potential to produce false identifications if the examiner does not exercise caution.

The reason that the non-match distribution moves to the left as the size of the database gets bigger results from the fact that with more images to sample from, the chances of finding a very low discrepancy score simply by chance gets larger and larger. The match distribution does not change because there is still only one match in the database to find. This result is entirely consistent with the argument made by Dror and Mnookin (2010), and reinforces their central conclusion: increasing the size of the database search will potentially expose the examiner to more close non-matching prints. There are other benefits to larger databases, such as an increased possibility of finding the suspect in the database, and better estimates of the typicality of a latent print, which are discussed in a later section.

1.4 *Computing sensitivity: receiver operating characteristic curves*

The graphs in Figs 3–4 illustrate that as the size of the database increases, the chance of finding close non-matches also increases. Under some conditions, there could be many close non-matches,

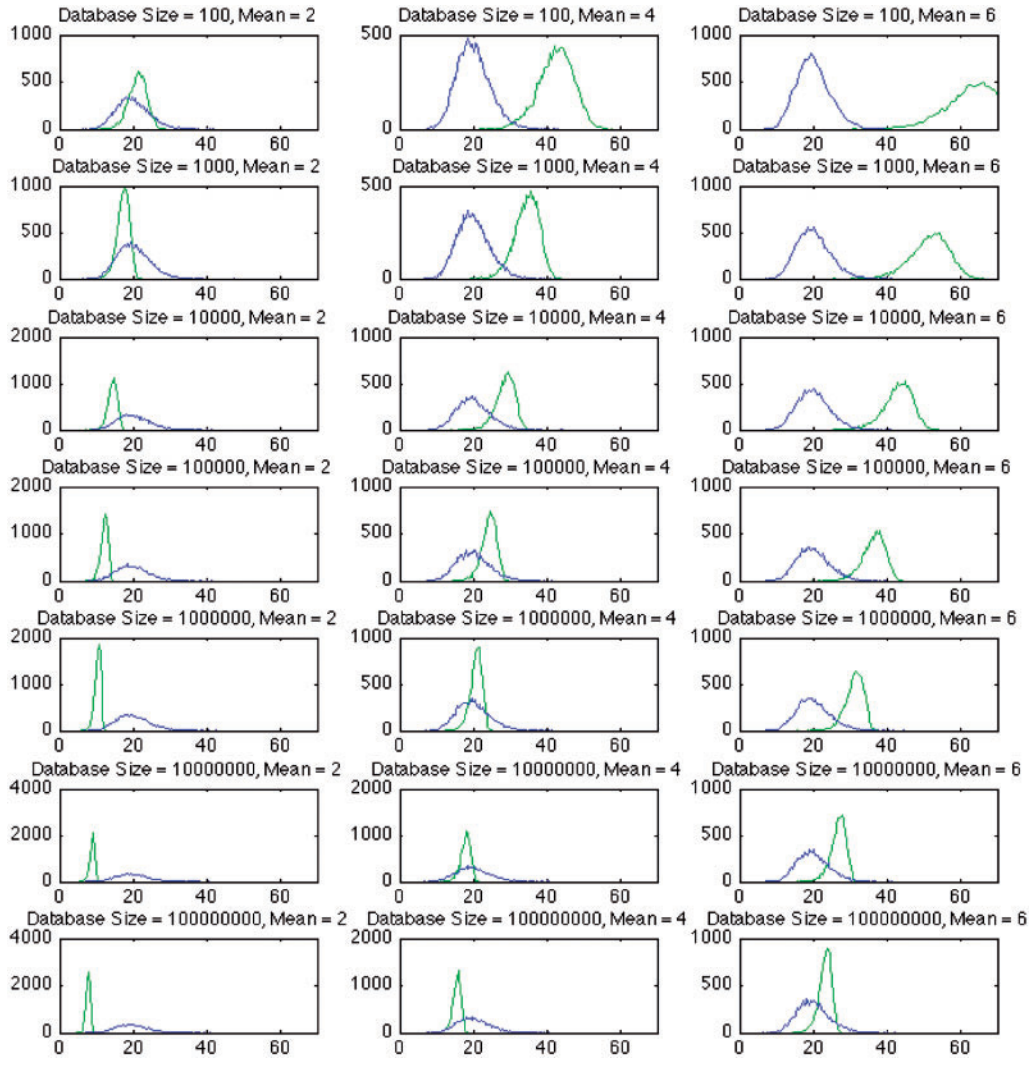

FIG. 4. Distributions of the true matches (blue curves) and the minimum of the non-matching prints (green curves). The abscissa represents the discrepancy score, and the ordinate is the frequency that the minimum value obtains each discrepancy score. Small databases (top rows) show very little overlap between the two different distributions, but as the database size increases (lower rows) the green curve shows more and more overlap, implying that there are increasing numbers of non-matching prints that could be very similar in appearance to the submitted latent print. Although the amount of overlap depends on the assumption of the mean of the distribution of the non-matching prints (different columns), the overall trend is clear: as the database size increases, the number of prints that are as similar or more similar than a true match also increases. This could potentially lead to more erroneous identifications. Please see online version for colour.

effectively pushing the matching print off the top 20 (or even top 50 in other simulations). The degree to which this occurs depends on how much the matching and non-matching prints are assumed to be similar, but, in general, querying large databases will always lead to an increased likelihood of finding close non-matches.

Of course, the whole reason for building larger databases is to increase the likelihood that the suspect will be in the database. We know that as the database gets larger, the likelihood of finding a close non-match will increase. However, the likelihood of finding the suspect in the database will likely increase with larger databases. How do these two factors trade off, and how does the overall sensitivity change as a function of database size?

To answer these questions, we must define sensitivity, and in this case we refer to the ability of the system to separate the distribution of intra-finger distance values from the smallest inter-finger distance values (the blue and green curves in Fig. 4). There are other techniques that may be used to characterize the performance of a database search, such as the specificity of the test, which relates to the probability of correctly classifying a print as having come from the same source as the latent print. Similar approaches have been documented in drug testing, where the base rate of drug uses plays a role in the overall accuracy of a screening test (Gastwirth, 1987). The methods below allow a straightforward way to generalize from the curves in Fig. 4 to an overall measure of performance that is dependent upon both the database size as well as the likelihood of the suspect having been enrolled in the database, while simultaneously removing the influence of any one decision criterion.

The challenge of relying solely on one outcome (say, avoiding erroneous identifications) is that this depends critically on the choice of the decision criterion. For example, an examiner could go through their entire career and never make an erroneous identification by simply having an unreasonably high standard for what constitutes evidence for a match. As a result, they would miss many potential correct identifications because they would pronounce them as insufficient or inconclusive. Alternatively, an examiner could have an unreasonably low standard for a match, which would lead to many more correct identifications but also potentially more erroneous identifications.

The choice of the decision criterion depends on a number of factors, including the individual examiner's abilities, experience and training, as well as the value that society places on the different outcomes. However, for the purposes of analysing the relation between sensitivity and database size, we would like to remove the influence of the choice of decision criterion using Receiver Operating Characteristic (ROC) curves (Macmillan and Creelman, 2005; Zou, 2012). Assume for the moment that a discrepancy score of less than a value of 10 would be judged as a match or identification. Discrepancy scores between two impressions from the same finger of less than 10 would constitute a correct identification, and discrepancy scores between two impressions from different fingers would constitute an erroneous identification. Of course examiners would not simply take the AFIS score at face value when making a decision, but for the purposes of exploring the relation between database size and sensitivity it is helpful to start with this assumption. Given this decision criterion, we can compute the number of true identifications and false identifications that would be made. This is simply the number of simulations that returned a discrepancy value from the matching print that is less than 10, as well as the number of simulations that returned the maximum of the non-matching distributions as less than 10. If we divide both numbers by the number of simulations, we produce the correct and erroneous identification rates, respectively.

If we repeat this process for different decision criteria (say 5, or 15), we can produce pairs of points that represent the theoretical correct and erroneous identification rates for all possible decision criteria. These are more formally known as hit and false alarm rates, and for each pair we can plot a point as a scatterplot. These pairs allow us to map out the entire space to produce what is known as the ROC (Bamber, 1975; Egan, 1975). We construct an ROC curve from the curves in Fig. 4 by sweeping the decision criteria along the abscissa to produce a wide set of pairs of points. The functions in Fig. 5 are a result of sweeping across the entire range of discrepancy thresholds to produce pairs of points. This

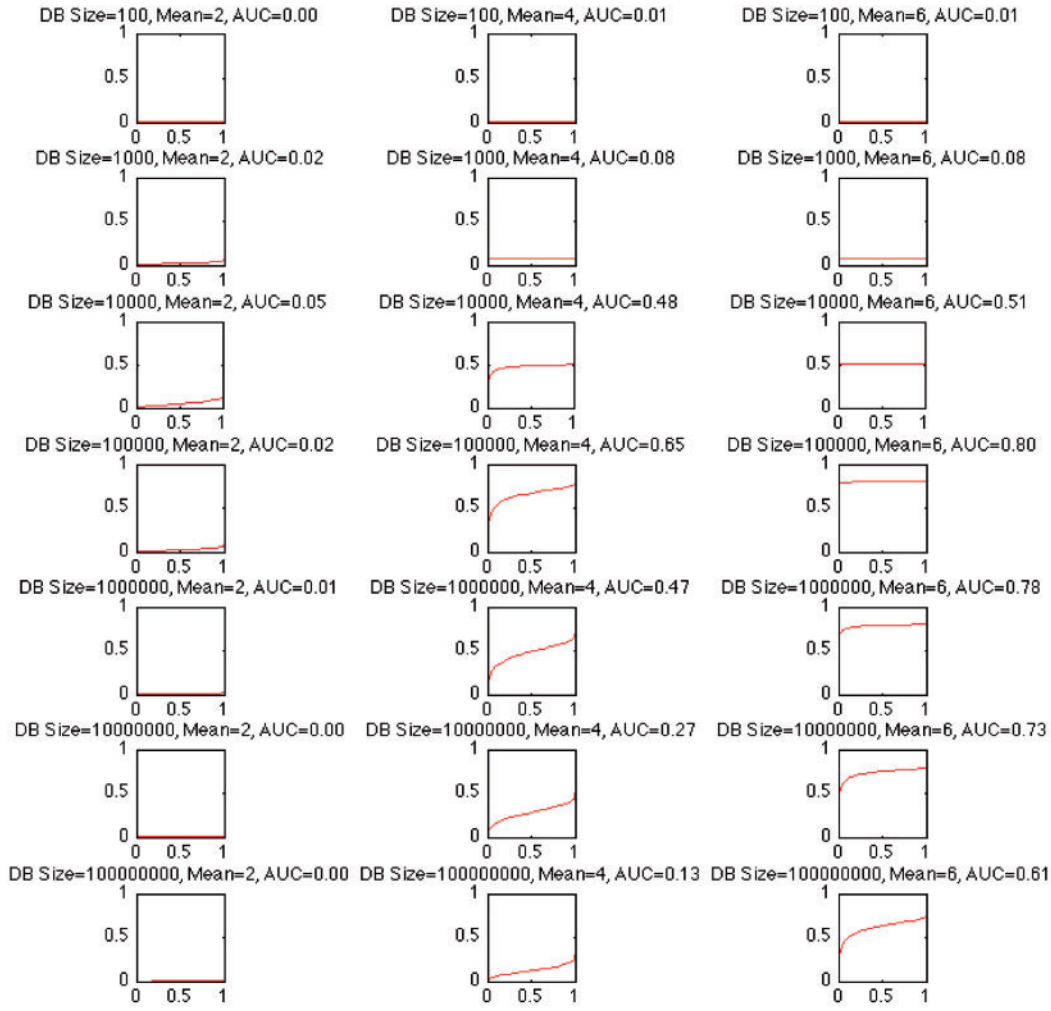

FIG. 5. ROC curves for databases of different sizes. The ordinate of each graph is the correct identification rate, and the abscissa is the erroneous identification rate. As in Fig. 3, we plot the ROC curve for each database size for different assumptions of the overall between the matching and non-matching prints (see Fig. 1). AUC is Area Under the Curve. See text for details.

technique allows us to specify the overall sensitivity of the system regardless of the choice of decision criteria. A final summary score can be computed by computing the area under the ROC curve, which is the overall sensitivity of the database search (Birdsall and Roberts, 1965; Egan, 1975). Note that this sensitivity measure is different than the sensitivity value referred to in biometric applications, where sensitivity refers to the probability than an individual with the trait is correctly classified. Instead we use sensitivity in the signal detection sense, which relates to the distance between the intra-finger discrepancy scores (blue curves in Fig. 4) and inter-finger discrepancy scores (green curves in Fig. 4), both of which factor into the area under the ROC curve along with their respective variances.

Note that this procedure for measuring the sensitivity of a database does not actually depend on any one decision criterion, or even on the concept of a decision criteria. It simply measures the statistical value of a database search and its ability to separate intra-finger discrepancy scores from the smallest inter-finger discrepancy scores. This represents the evidential strength extracted from the discrepancy scores computed between the latent print and all prints in the reference database.

There is one final correction that must be applied to the hit rate values to compute the probability of finding the suspect in the database. The issue revolves around the fact that not all suspects are going to be in the database. Thus even if your town has 100 000 residents and you are query a national database with 10 million, there is no guarantee that the suspect will be in the database. All criminals have a first crime, and not all crime is local. Thus we need to estimate the probability that the suspect is present in the database, which will affect the correct identification rate. We computed this probability using this equation:

$$P(SuspectInDatabase) = \gamma(1 - e^{-n/\varsigma}) \quad (1)$$

where γ is an upper asymptote to account for the fact that all criminals have a first crime (set to 0.8 in our simulation), and ς is a scaling factor that is set to 10 000. The scale factor essentially describes the size of the database that is required to achieve approximately a 50% likelihood of the suspect being in the database. This probability is computed for each database and then used to scale the y-axis values in Fig. 5.

The overall sensitivity of the system is computed by taking the area under the ROC curve. Typically it is greater than 0.5, although in the present case this is not true as given by the area under the curve (AUC) values in Fig. 5. The very low values result from the fact that the non-match distributions tend to dominate the match distributions for large databases and for situations with a great deal of overlap between matching and non-matching distributions (i.e. mean = 2). However, the general interpretation of the area still holds: larger areas under the curve correspond to better sensitivity, and below we explore the relation between this and database size.

1.5 *The relation between database size and sensitivity*

A central question revolves around how sensitivity changes as the database increases. We have seen how the number of close non-matches will increase as the database size increases, but also how the larger databases are more likely to contain the suspect (see equation 1). How do these two factors trade off to determine sensitivity? Are bigger databases always better, or is there an optimal database size given the tradeoffs between the likelihood of the suspect being in the database and the number of close non-matches?

Figure 6 plots the sensitivity values (AUC) for different database sizes, with the three amounts of overlap between the matching and non-matching distributions as curve parameters. Interestingly, sensitivity initially rises as the database size grows, but ultimately plateaus and then drops for very large databases. This implies that there is an optimal point at which a database has the highest sensitivity, and going to a larger database will overwhelm the user with close non-matches while doing little to increase the likelihood of the suspect being in the database.

A very similar result is found if equation 1 is replaced by a simple function that computes the probability of the suspect being in the database as simply the size of the database divided by the population size, capped at 0.8 if the database exceeds the population. This function has similar properties as equation 1, in that it grows initially and then asymptotes at some value, because even

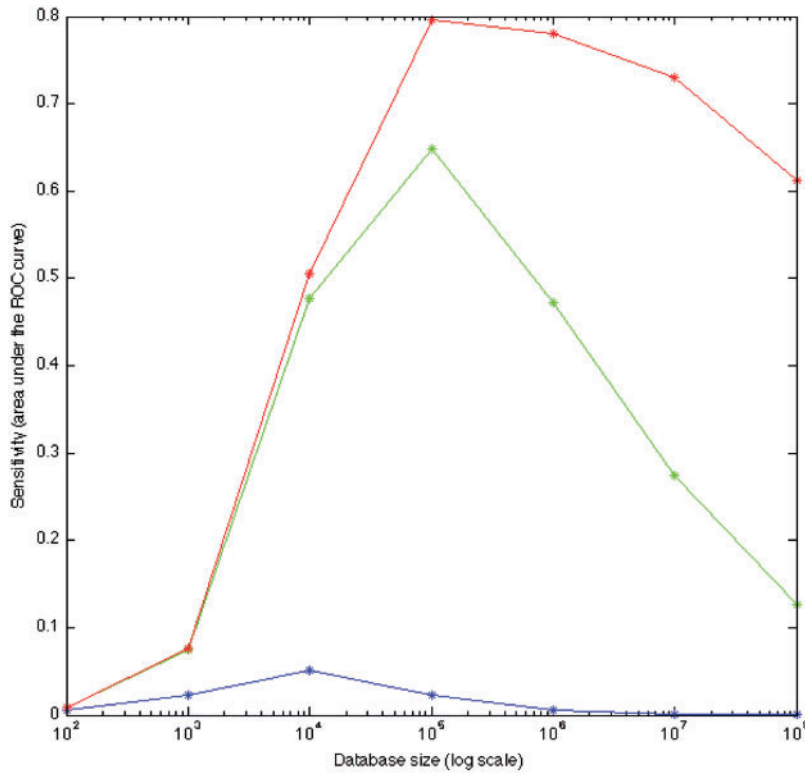

FIG. 6. Sensitivity (which intuitively combines both the likelihood of finding the suspect in the database along with the number of nuisance close non-matches) as a function of the database size for the three levels of non-match means. Although sensitivity is higher overall for the condition in which there is less overlap between the matching and non-matching prints (red or top curve, mean = 6), all three curves illustrate the effect of diminishing returns: sensitivity grows as the size of the database increases, until the likelihood of having close non-matches that are more similar than the target begins to overwhelm the benefit of larger databases. Then sensitivity declines for all three conditions.

having a database that is larger than the population does not guarantee that the suspect is in the database. However, it grows linearly with the size of the database before abruptly coming to an asymptote at 0.8.

The conclusion from the sensitivity analysis is that sensitivity has a non-monotonic relation with database size. Sensitivity initially increases, but once the database reaches a critical size there is a 'decrease' in overall sensitivity. This results from the fact that the number of close non-matches begins to overwhelm the likelihood of finding the suspect in the database. The implications of this finding are discussed below.

1.6 *Recommendations for practitioners*

The implications of these simulations are clear: as the database size grows, the likelihood of the suspect having been enrolled in the database goes up, but so too does the likelihood of finding close non-matches. In fact, the number of potential false non-matches tends to overwhelm the likelihood of the suspect being in the database for large database sizes. This results in a sensitivity curve that initially

grows for small databases, but then plateaus and then ultimately decreases for very large databases. Although the interpretation of these simulations depends in part on the assumptions underlying the model, there are a wide range of models that will produce similar outcomes.

The first result confirms a proposal previously argued by Dror and Mnookin (2010), and quantitatively demonstrates that as the size of the database increases the chance of finding close non-matches will also increase. Because these are the prints that lead to potential false identifications, examiners should be cautious when going to national databases. We recommend that practitioners view the discrepancy scores from automated matching systems such as AFIS with a critical eye, and consider adjusting their criterion for what constitutes an impressive discrepancy score as the size of the database increases. For example, if a suspect is developed based on non-fingerprint evidence, a close match between the latent and candidate print is very unlikely by chance, and should be viewed as strong evidence of an identification. However, if the candidate print is developed after a search against a large database, even a very small discrepancy score should be viewed with suspicion because the likelihood of this occurring simply due to chance go up with the size of the database. For example, labs might adopt policies that require more matching features than usual for prints recovered from a large database search, or an examiner may adjust their own internal criterion for the amount of evidence that they consider to be sufficient for an identification.

The second result (that sensitivity will drop for very large databases) is even more important, because it demonstrates that in addition to dealing with more close non-matching prints with larger databases, there is a diminishing returns element that contributes to an eventual drop in the sensitivity of the database search. That is, the gains from larger databases do not offset the problem of an increased number of close non-matches. A complete treatment of this relation would probably require a geographic model of crime or a consideration of crime type, but given that most crime is local, the recommendation is for an investigator to query the largest database that they think the criminal would reasonably be enrolled in. Should the examiner go to a larger database such as a national database, extreme care should be exercised to avoid erroneous identifications from close non-matches. For example, if a print is run against local jurisdiction with no obvious match, going to national database has two problems. First, it mainly adds people outside your area, which makes them less likely to be the criminal, and second it exposes the examiner to many more potential close non-matches.

One solution to these problems might be to adopt a likelihood model approach to quantifying the evidentiary value of a latent print (Egli *et al.*, 2007; Neumann *et al.*, 2007, 2006). With this approach, the goal of the forensic identification is to quantify the strength of the evidence rather than render a decision about the nature of that evidence. In practice, the likelihood ratio could be translated into a rough linguistic equivalent (e.g. ‘strong evidence in favor of the same source’ or ‘weak evidence in favor of the same source’).

A complete critique of the methods and applications of likelihood ratio approaches is beyond the scope of this article, but there is one important conclusion that applies to the likelihood ratio approach as well as the traditional decision-based methodology. The advantage of larger databases from the perspective of likelihood ratio models is that they provide better estimates of the typicality of the latent print. This value (expressed by the denominator of the likelihood ratio) comes from the discrepancy values between the latent print and all prints in the database. However, as the database grows, the typicality will decrease for all prints. This results from the fact that larger databases tend to discover prints that are more similar to the latent print than previously found. This will tend to increase the denominator of the likelihood ratio (because the probability that the latent print comes from the non-match distribution increases), which decreases the overall likelihood ratio. This is not a fatal flaw for

likelihood ratio models, but it does suggest that the likelihood ratio must be a function of the size of the reference database, because the likelihood value will drop as the database increases.

Alternatives to the likelihood approach exist and could be considered by the community. For example, an agency could adopt a policy in which multiple impressions or a greater number of minutia should be required before the results of an extremely large database search are accepted as compelling evidence of an identification.

We hope that these conclusions elicit caution amongst examiners when querying large databases, and also provokes discussion within the biometric and latent print examiner communities about the need to explore these issues within the ecosystem of an actual AFIS environment. This would allow an assessment of how database size affects one particular distance function and feature set, as well as allow the exploration of how likelihood ratios are affected by database size. The answer to which set of assumptions from Fig. 2 best reflects reality can easily be answered by those with access to the underlying AFIS algorithms from different vendors. This may require pressure from outside groups to encourage the vendors to open up the algorithms for inspection, which can still be protected by patents and copyrights.

Funding

This work was supported by National Institute of Justice grants #2005-MU-BX-K076 and #2009-DN-BX-K226.

References

Bamber, D. (1975). Area above ordinal dominance graph and area below receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12** (4), 387–415. Doi 10.1016/0022-2496(75)90001-2.

Birdsall, T. G. & Roberts, R. A. (1965). On the theory of signal detectability - an optimum nonsequential observation-decision procedure. *Ieee Transactions on Information Theory*, **11** (2), 195–204.

Champod, C. & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, **31** (2–3), 193–203. Doi 10.1016/S0167-6393(99)00078-3.

Dror, I. E. & Mnookin, J. L. (2010). The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law, Probability and Risk*, **9** (1), 47.

Egan, J. (1975). *Signal Detection Theory and Roc Analysis*. Academic Press, New York.

Egli, N. M., Champod, C. & Margot, P. (2007). Evidence evaluation in fingerprint comparison and automated fingerprint identification systems–modelling within finger variability. *Forensic Science International*, **167** (2–3), 189–195. Doi 10.1016/j.forsciint.2006.06.054.

Gastwirth, J. L. (1987). The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. *Statistical Science*, 213–222.

Kwan, Q. Y. (1977). *Inference of Identity of Source*. (Ph.D. thesis of Criminology), University of California, Berkeley, CA.

Macmillan, N. A. & Creelman, C. D. (2005). *Detection Theory : A User's Guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Meuwly, D. (2006). Forensic individualisation from biometric data. *Science Justice*, **46** (4), 205–213. Doi 10.1016/S1355-0306(06)71600-8.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A. & Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences*, **52** (1), 54–64. Doi 10.1111/J.1556-4029.2006.00327.X.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Meuwly, D. & Bromage-Griffiths, A. (2006). Computation of likelihood ratios in fingerprint identification for configurations of three minutiae. *Journal of Forensic Sciences*, **51** (6), 1255–1266. Doi 10.1111/j.1556-4029.2006.00266.x.

Snodgrass, M., Bernat, E. & Shevrin, H. (2004). Unconscious perception at the objective detection threshold exists. *Perception & Psychophysics*, **66** (5), 888–895.

Vanderkolk, J. R. (2009). *Forensic Comparative Science: Qualitative Quantitative Source Determination of Unique Impressions, Images, and Objects*. Elsevier, Burlington, MA.

Vanselst, M. & Merikle, P. M. (1993). Perception below the objective threshold. *Consciousness and Cognition*, **2** (3), 194–203.

Zou, K. H. (2012). *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. CRC Press, Boca Raton.