Law, Probability and Risk (2019) **18**, 1–23 doi:10.1093/lpr/mgz001
Advance Access publication on February 14, 2019

Calculating the posterior odds from a single-match DNA database search

John T. Wixted† and Nicholas J.S. Christenfeld
University of California, San Diego, La Jolla, CA, USA
and
Jeffery N. Rouder
University of California, Irvine

[Received on 19 December 2017; revised on 26 July 2018; accepted on 27 September 2018]

Forensic evidence is often quantified by statisticians in terms of a likelihood ratio. When the evidence consists of a DNA match, the likelihood ratio is equal to the reciprocal of the 'random match probability' (p). When p is small (e.g. 1/10 million), the likelihood ratio is large (e.g. 10 million to 1). However, when a single match is obtained by searching a database, the prior odds that the forensic DNA was deposited by the matcher can be extremely low, in which case the posterior odds can be low as well despite the high likelihood ratio. Unfortunately, prosecutors, judges and jurors are at risk of misinterpreting the likelihood ratio as the posterior odds, a pervasive reasoning error known as base-rate neglect. Here, we propose a solution to that problem, which consists of evaluating the prior odds based on a case-independent estimate of the size of the active criminal population derived from database search statistics. The posterior odds that the forensic DNA belongs to the unidentified single-matcher can then be calculated (avoiding base-rate neglect).

Keywords: DNA database search; likelihood ratio; random match probability; NRC II.

In a criminal investigation, a suspect is sometimes identified when a forensic DNA profile from a single unknown source is compared to known DNA profiles in a local, state or federal database. We call this situation the 'database search' scenario. If only one person's profile in the database is found to match the unknown profile, that person immediately becomes a suspect and risks being convicted based largely on the DNA evidence. Heightened interest in the strength of evidence provided by a match from a database search can be traced to a pair of National Research Council (NRC) reports published in the 1990s (National Research Council, 1992, 1996, henceforth NRC I and NRC II). In both reports, the argument was made that the statistical interpretation of a DNA match obtained by searching a database is complicated by the fact that "A match by chance alone is more likely the larger the number of profiles examined" (NRC II, p. 134; see also NRC I, p. 124). When communicating the implications of a DNA database search to a court of law, it seems important to be clear about the possibility that the single match may have landed on the wrong person. Exactly how to do that remains unresolved and is the focus of this article. We focus on the single-match case because that search outcome excludes everyone else in the database and is therefore especially prejudicial to the matching individual. If the search instead yielded two matches, it would be plainly apparent that the DNA

†Email: jwixted@ucsd.edu

evidence, standing alone, would not strongly implicate either person. By contrast, a single match from a database search can create the impression that the results are definitive even when, in truth, they may not be.

To help prevent misunderstandings that can arise when interpreting the results of a DNA database search, the NRC II report recommended that when a single match occurs, the random match probability (p) should be multiplied by the number of profiles in the database (n). The random match probability is computed from the allele frequencies in the population for each DNA locus in the forensic profile. The value of p represents an estimate of the probability that a person randomly selected from the population would have a DNA profile that matches the DNA profile from the crime scene evidence. A degraded forensic DNA sample with few loci can be associated with a high random match probability (e.g. $p = 1/100$), whereas a fully intact sample with many DNA loci is associated with an infinitesimal random match probability (e.g. $p < 1/100$ trillion). Using the $1 / np$ rule to communicate the 'impact of the DNA evidence' (NRC II, p. 40), the larger n is, the less compelling the single match would be. For example, in a database of 1000 profiles ($n = 1000$), if the random match probability is fairly small (e.g. $p = 1/10$ million), the evidential value of a single match (i.e. the likelihood ratio in Bayesian terminology) would be high ($1 / np = 10$ million$/1000 =$ 10,000 to 1). But if the database were much larger (e.g. $n = 10$ million), the evidential value of a single match would be low ($1/np = 10$ million$/10$ million $= 1$ to 1).

The proposed $1 / np$ rule was controversial from the outset, partly because it was an intuitive solution that was not formally derived. As noted in the NRC II report and in many articles that have addressed the database search scenario since then (e.g. Balding, 2002; Balding and Donnelly, 1996; Berger *et al.*, 2015; Dawid, 2001; Dawid and Morterra, 1996; Donnelly and Friedman, 1999), the implications of a DNA match are provided by Bayes' rule, which holds that:

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

The posterior odds provide the information of interest to the legal system about the two complementary hypotheses under consideration when a single match occurs:

H_1: The forensic DNA belongs to the matcher.
H_2: The forensic DNA belongs to someone else.

Given that a single match has occurred, what are the posterior odds that the forensic DNA belongs to the matcher (H_1) versus someone else (H_2)? That information is obtained by multiplying the prior odds by the likelihood ratio.

The prior odds are the odds that before the DNA search is performed, any one person from the relevant population would have deposited their DNA at the crime scene. 'The relevant population' is a critical consideration because, in the U.S., not just any DNA profile is used to search a database, and not just anyone is a candidate for having their known profile in a DNA database. To be eligible for a database search, a judgement must first be made that the unknown DNA profile recovered from a crime scene likely belongs to the perpetrator (i.e. to an offender). In addition, the known profiles in the database typically belong only to convicted offenders and arrestees, and the forensic profile is compared against all of them. Thus, the relevant population (i.e. the reference class) for computing the prior odds of H_1 versus H_2 does not consist of everyone in the U.S. because not everyone is a plausible candidate for being arrested and having their DNA profile entered into a database. Nor is the relevant

population limited to plausible suspects for having committed the crime in question because all of the profiles in the database are searched, not only the profiles of offenders considered by the prosecution or the defence to be plausible suspects. Instead of having anything to do with plausible suspects for having committed a particular crime, the relevant population for computing the prior odds of H_1 versus H_2 is pre-defined by virtue of how the DNA database search is conducted. Specifically, the relevant population consists of what might be termed the 'active criminal population' (Leary and Pease, 2003; Walsh *et al.*, 2010). For now, we denote the size of this population as N, and we later consider how it can be estimated.

With an estimate of N in hand, the prior odds are easily specified. Let A represent probability that before the DNA search is performed, any one person from the relevant population would have deposited their DNA at the crime scene. If the size of the relevant population is N, then, if each of the N individuals is a priori equally likely to match before any incriminating or exonerating information about any one of them is known, the prior probability of A is given by $P(A) = 1 / N$, from which it follows that the prior odds of A are equal to $P(A) / [1—P(A)] = 1 / (N—1)$. For the competing hypotheses above (H_1 vs. H_2), the prior odds are $P(H_1) / P(H_2) = P(A) / [1 - P(A)] = 1 / (N - 1)$.

According to Bayes' rule, the prior odds are updated by the likelihood ratio to yield the posterior odds. The likelihood ratio captures the impact of the DNA matching evidence (which is how the NRC II report characterized the $1 / np$ rule). Let E_m represent the matching evidence, which consists of a single match in the case of a DNA database search. The likelihood ratio expressed in terms of the competing hypotheses is equal to the probability of E_m under H_1 divided by the probability of E_m under H_2, or $P(E_m|H_1) / P(E_m|H_2)$. In the time since the NRC II report was issued, the mathematics of the single-match DNA database search scenario have been completely worked out, and it is clear that the likelihood ratio for the competing hypotheses specified above does not equal $1 / np$. Instead, when the evidence (E_m) consists of a single match from a DNA database search, the likelihood ratio is now known to be $P(E_m| H_1) / P(E_m| H_2) = (1 / p) \times (N - 1) / (N - n)$ (Balding, 2002; Balding and Donnelly, 1996; Berger *et al.*, 2015; Dawid, 2001; Dawid and Morterra, 1996; Donnelly and Friedman, 1999).

The n profiles in the database are a subset of the larger relevant population N (i.e. not everyone who could be in the database is in the database). Because $N > n$, the multiplying factor $(N - 1) / (N - n)$ in the likelihood ratio must be equal to or greater than 1. Thus, the impact of the matching evidence has a lower bound of $1 / p$ and actually increases from there as n gets larger. As an example, if $n = 0.2N$ (i.e. 20% of the relevant population is in the database), the multiplying factor $(N - 1) / (N - n)$ would be $\sim$1.25. But if n were so large that the database included the DNA profile of everyone in the relevant population ($n = N$), then the forensic DNA profile in the case of a single match would undoubtedly belong to the matching individual (i.e. the multiplying factor, and therefore the likelihood ratio, would explode to infinity). The fact that the likelihood ratio increases with n is exactly the opposite of what the $1 / np$ rule implies. Thus, in the DNA database search scenario, informing a court that the likelihood ratio is $1 / p$ (computed without regard for n) is not misleading and instead provides, if anything, a conservative estimate of the impact of the DNA evidence.

Now that the mathematics of the single-match DNA search scenario have been worked out, it might seem to be a simple matter for the statistician to intelligibly convey the implications of a single match from a database search to a court of law. However, two issues stand in the way of prosecutors, judges and juries understanding the sometimes counterintuitive implications of a single match with respect to H_1 and H_2 above.

The first issue is the consensus view that providing an estimate of the likelihood ratio is the best way for a statistician to assist a judge or jury in assessing the implications of forensic evidence (e.g. Aitken

et al., 2011; Lindley, 1977; Neumann *et al.*, 2016). The computation of the likelihood ratio involves measurement-based, 'case-independent' calculations that unambiguously fall within the expertise of the forensic statistician. Case-independent calculations are based on information that is unrelated to the identity of the accused and the nature of the crime (e.g. information about the measured population frequency of different alleles). By contrast, estimating the size of the relevant population of plausible suspects for having committed the crime in question (from which the prior odds are computed) typically involves measurement-free, 'case-dependent' calculations about which the prosecution and defence will likely disagree. Case-dependent calculations are based on information that is directly related to the identity of the accused and the nature of the crime (e.g. how likely a putative motive is to spur the criminal act). As a general rule, given the absence of any measurement-based, case-independent information that could be used to estimate the size of the relevant population, estimating the prior odds falls to the judges and jury, not the statistician.

The second issue that complicates the interpretation of a single match from a database search is known as 'base-rate neglect', which is the pervasive human tendency to simply disregard the prior probability (and, therefore, the prior odds) of an event (e.g. Gaissmaier and Gigerenzer, 2011; Gigerenzer and Edwards, 2003; Gigerenzer *et al.*, 2008; Kahneman and Tversky, 1973; Lyon and Slovic, 1976; Meehl and Rosen, 1955). Disregarding the prior odds is problematic because the likelihood ratio—the probability of matching under different hypotheses—is not what is important for a just adjudication of the case. Instead, it is the posterior odds that provide the information that judges and jurors need to know. The posterior odds are equal to the probability that the DNA belongs to the matcher versus someone else in light of the match, or $P(H_1|E_m) / P(H_2|E_m)$.

If the prior odds are ignored, the posterior odds of a single match from a database search will seem vastly larger than they actually are. For example, if $p = 1 / 10$ million (the random match probability) and $N - n = 10$ million (the number of active criminals not in the database), then the posterior odds would be 1. However, if the statistician reports that the likelihood ratio = 10 million to 1, and if the prior odds are ignored, the posterior odds that the forensic DNA belongs to the matcher would be mistakenly evaluated by the jury to be 10 million to 1. In other words, ignoring the prior odds yields a result seven orders of magnitude away from the truth.

Conceivably, despite being characterized as a way to quantify the 'impact of the DNA evidence' (words that ordinarily correspond to the likelihood ratio), the $1 / np$ rule recommended by the NRC II may have been an attempt to specify the posterior odds of a single match database search. However, even if the $1 / np$ rule was intended to capture the posterior odds, the equation would still not be right. If the prior odds are $1 / (N - 1)$ and the likelihood ratio is $(1 / p) \times (N - 1) / (N - n)$, then the posterior odds come to $(1 / p) / (N - n)$. Thus, it is not the large size of n that can yield inconclusive results (as implied by the $1 / np$ rule); instead, the results can be inconclusive when the *difference* between N and n is large (i.e. when much of the relevant population is not in the database). Continuing with the aforementioned example, if $p = 1 / 10$ million, and if the difference between N and n happens to be about 10 million, then the true posterior odds would be close to 1, in which case the match would be inconclusive despite the low value of p.

Aside from the ill-fated $1 / np$ rule, the only other suggested solution to the problem of base-rate neglect in the DNA database search scenario consists of providing a range of hypothetical priors in conjunction with the likelihood ratio to illustrate how different prior odds affect the measure of interest, namely, the posterior odds (Meester and Sjerps, 2004; NRC II, 1996). Indeed, Meester and Sjerps (2004) recommend this educational strategy precisely because jurors are prone to neglect base rates when presented with the likelihood ratio alone. However, in the DNA database search scenario,

the posterior odds that the DNA belongs to the matcher can differ so dramatically from what the likelihood ratio implies that simply providing a table illustrating how the posterior odds are affected by a range of hypothetical priors seems likely be of little assistance. For example, imagine that $p = 1 / 10$ million, and the concern is that jurors might misinterpret the likelihood ratio provided by the statistician to mean that the odds that the DNA belongs to the matcher are 10 million to 1. To address that problem, a table could be provided showing if the prior odds are 1-to-10 million, then the posterior odds that the forensic DNA belongs to the matcher would be 1-to-1, but if the prior odds are 1-to-1000, the posterior odds would instead be 10 000 to 1. It is hard for us to imagine that such a table would help given that jurors are in no position to have any sense at all about which hypothetical prior would be the best one to use. Recall that the prior odds for the competing hypotheses presented earlier are $1 / (N - 1)$. Even a forensic scientist serving as a juror might fail to appreciate that, in the DNA database search scenario, N in this equation has nothing at all to do with the number of plausible suspects for having committed the crime in question and instead represents the size of the active criminal population.

Here, we propose a new solution to this problem. Our solution consists of showing how the implications of a single match from a DNA database search can be quantified in terms of a measurement-based, case-independent statistical evaluation of the prior odds, which, when multiplied by the likelihood ratio, yields a case-independent evaluation of the posterior odds. The novel feature of our proposed solution consists of estimating of the hard-to-count population of N individuals who are plausible candidates for having their DNA profiles in the database (*not* the population of plausible suspects for having committed the crime in question). With that estimate in hand, and together with n and p, Bayes' rule can be used to calculate the case-independent posterior odds that the forensic DNA belongs to the as-yet identified matcher. By providing the posterior odds (rather than providing just the likelihood ratio), the statistician eliminates the possibility that prosecutors, judges and jurors will catastrophically misinterpret the implications of a single match due to base-rate neglect.[1]

Later, the matcher's identity will be revealed, and triers of fact will update those case-independent odds using case-dependent information based on subjective information that will be contested by prosecutors and defence attorneys. Thus, adjudicating the guilt or innocence of a particular identified matcher would remain the responsibility of judges and juries. To illustrate how these issues play out in the DNA database search scenario, we begin by first considering the implications of a DNA match in the simpler 'probable cause' scenario. We consider the probable cause scenario, where base rate neglect is a relatively minor problem, only to contrast it with our main focus, the DNA database search scenario, where base rate neglect can lead to a catastrophic miscalculation of the posterior odds.

1. The probable cause scenario

In the probable cause scenario, the police have already pinpointed a suspect based on non-DNA evidence. To determine if the forensic DNA belongs to the suspect, investigators will compare the suspect's DNA profile (e.g. from a saliva sample) to the profile of the forensic DNA sample. Before the DNA matching evidence is taken into consideration, the following two complementary hypotheses are under consideration by the police:

H_1: The suspect is guilty.
H_2: The suspect is innocent.

[1] As we explain in a subsequent section, this approach also addresses the cognitive error known as the 'prosecutor's fallacy' (Thompson and Schumann, 1987), which is related to, but not identical to, base rate neglect.

Note that these hypotheses differ from the competing hypotheses considered earlier because the known data (evidence) becomes available sequentially. In the probable cause scenario involving DNA evidence, the initial evidence consists of the 'other' (non-DNA) evidence (E_o) that leads the police to suspect a particular individual of having committed the crime. For example, the police might detain a male suspect who was found walking along a street a few blocks from where a robbery occurred and who matches a vague description provided by an eyewitness. This evidence can be used to compute the posterior odds that the suspect is the one who committed the crime, before taking into account the DNA evidence.

The posterior odds of H_1 versus H_2 conditional on E_o are represented as $P(H_1|E_o) / P(H_2|E_o)$ and are given by Bayes rule:

$$\frac{P(H_1|E_o)}{P(H_2|E_o)} = \frac{P(E_o|H_1)}{P(E_o|H_2)} \mathrm{x} \frac{P(H_1)}{P(H_2)} \tag{1}$$

As noted earlier, the prior odds, $P(H_1) / P(H_2)$, are the odds that, before any evidence is known, a person randomly selected from some relevant population of size N would turn out to be the one who is guilty of having committed the crime in question. Assuming that only one person in the relevant population committed the crime in question, the prior probability of guilt (also known as the base rate of guilt), $P(H_1)$, is equal to $1 / N$. Because the two competing hypotheses are complementary, $P(H_2) = 1–P(H_1) = 1–1 / N$. Thus, prior *odds* of guilt, $P(H_1) / P(H_2)$, come to $1 / (N – 1)$.

Critically, in the probable cause scenario, N is not statistically estimated based on measurements but is instead subjectively inferred based on a consideration of the crime that was committed. The prosecution and the defence would be motivated to assume that N is small or large, respectively. For example, consider a robbery that was committed in downtown St. Louis. The prosecution might be inclined to argue the crime was likely committed by someone from the small population of people who live in the neighbourhood where the crime occurred and where the suspect also lives (high prior odds). If the prior odds are high, then the updating factor would not have to be very large (i.e. the other evidence would not have to be extremely compelling) to yield posterior odds consistent with a verdict of guilt. By contrast, the defence might be inclined to argue that the crime could have been committed by anyone from the large population of adults who live in the city (low prior odds). If the prior odds are low, then the updating factor would have to be quite large (i.e. the other evidence would not have to be quite compelling) to yield posterior odds consistent with a verdict of guilt. Thus, the evaluation of the relevant population (N)—and, therefore, the evaluation of the prior odds—is case-dependent. Because it involves subjective considerations about which the prosecution and defence might disagree, the evaluation of the prior odds falls to the judge and jury and not to a forensic statistician.

The likelihood ratio in Equation 1, $P(E_o|H_1) / P(E_o|H_2)$, which is the factor by which the prior odds are updated based on the non-DNA evidence (henceforth, the 'updating factor'), reflects the odds of observing the non-DNA evidence if the suspect is guilty. In that case, the updating factor would be the probability of observing the data (i.e. a person who matches a vague description provided by an eyewitness and who was found in the vicinity of the crime) if the suspect is guilty, $P(E_o|H_1)$, divided by probability of observing such data if he is innocent, $P(E_o|H_2)$. Evaluating the updating factor does not involve calculations based on any empirical measurements but instead involves subjective

calculations based on case-dependent information that will likely be contested by the prosecution and defence. Thus, like the evaluation of the prior odds, in the probable cause scenario, its evaluation falls to the judge and jury, not the statistician.

After taking into consideration the prior odds and the updating factor based on the non-DNA evidence, the posterior odds from Equation 1 would be updated using the DNA matching evidence (E_m). Here, we consider only the simplest case where the assumption is made that if the suspect is guilty (H_1), the forensic DNA belongs to the suspect, and if the suspect is innocent (H_2), the forensic DNA belongs to someone else. This is not necessarily true, of course, because the forensic DNA might belong to the suspect even if he is innocent. For example, the innocent suspect might have deposited his DNA at the crime scene the day before the crime, or the innocent suspect's DNA might have been deposited on the forensic evidence due to a contamination even in the lab. Allowing for such possibilities complicates the Bayesian analysis of the probable cause scenario without changing the essence of our main argument concerning the statistical estimation of the posterior odds that the forensic DNA belongs to the matcher in the single-match DNA database search scenario. Evett *et al.* (2002) and, more recently, Gittelson *et al.* (2016) can be consulted for examples of how to use various reasoning aids (such as 'the hierarchy of propositions' and Bayesian networks) to draw valid inferences under the more complex scenarios that often arise in real-world cases. Here, we only consider the simplest scenario for illustrative purposes.

With our simplifying assumption in place, the competing hypotheses we presented earlier concerning whether or not the forensic DNA belongs to the matcher become equivalent to the competing hypotheses under consideration here concerning whether or not the suspect is guilty. Thus:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(E_m|H_1)}{P(E_m|H_2)} \text{x} \left[\frac{P(E_o|H_1)}{P(E_o|H_2)} \text{x} \frac{P(H_1)}{P(H_2)} \right], \tag{2}$$

where D represents all of the evidence (E_o and E_m). The bracketed term on the right represents the evaluation of the posterior odds using the evidence that initially led the police to the suspect (Equation 1). Those posterior odds now effectively serve as the prior odds to be updated by the DNA evidence. The updating factor from the DNA evidence in Equation 2, $P(E_m|H_1) / P(E_m|H_2)$, represents the odds of a match if the forensic DNA belongs to the suspect, in which case he is guilty (in our simplified example). If the suspect is guilty (H_1), a match is virtually certain (assuming no lab errors). Thus, $P(E_m|H_1) = 1.0$. If the suspect is innocent (H_2), the forensic DNA belongs to someone else, in which case the probability of a false match is equal to the random match probability. That is, $P(E_m|H_2) = p$. Thus, in the probable cause scenario, $P(E_m|H_1) / P(E_m|H_2) = 1 / p$.

The determination of p relies on assessing the frequency of certain genetic markers in the suspect's broad racial group, which means that it does not depend on any information about the suspect's possible involvement in the crime in question. In that sense, the estimate of p is based on case-independent measurements, and its evaluation clearly falls within the expertise of the statistician. As noted earlier, there is considerable agreement that a statistician should evaluate the updating factor (i.e., the likelihood ratio) to advise a court of law about the strength of forensic evidence (e.g. Aitken *et al.*, 2011). According to that school of thought, the statistician would present for the court's consideration the random match probability in the form of a case-independent likelihood ratio, $1 / p$.

In this example, we made the assumption that, in the first step (Equation 1), jurors would ideally update their prior beliefs by multiplying the prior odds by the likelihood ratio. In reality, jurors would be more likely to underweight the prior odds if not disregard them altogether. In other words, at least

some of the jurors (if not all of them) would likely fall prey to base-rate neglect. As noted earlier, base-rate neglect is a general failing of human statistical reasoning (e.g. Gaissmaier and Gigerenzer, 2011; Kahneman and Tversky, 1973; Lyon and Slovic, 1976; Meehl and Rosen, 1955). Because the base rate of guilt for the crime in question, $1 / N$, is directly related to the prior odds of guilt, $1 / (N - 1)$, base-rate neglect is tantamount to ignoring the prior odds that a person randomly selected from the relevant population would be guilty of having committed the crime in question. Given its pervasive nature, it seems unlikely that jurors would be immune from making this reasoning error.

Despite the high risk of jurors falling prey to base rate neglect, statisticians are not in a position to evaluate the prior odds and to then use that information to calculate the posterior odds themselves because doing so would involve taking into account case-dependent information that is likely to be contested by the prosecution and defence. Nevertheless, for the probable cause scenario, whether or not the prior odds are taken into consideration is unlikely to matter very much. Suppose, for the sake of argument, that the relevant population is inferred to consist of $N = 201$ individuals such that $P(H_1) / P(H_2) = 1 / (N - 1) = 1 / 200$. And suppose that the updating factor based on non-DNA evidence is evaluated by the jury to be $P(E_o|H_1) / P(E_o|H_2) = 2$. Thus, after taking into account the non-DNA evidence but before taking into account the DNA evidence, the posterior odds of guilt according to Equation 1 would be only 1-to-100 that the suspect is guilty, not 2-to-1, as the jury would assume by ignoring the prior odds. An odds of guilt equal to 1 / 100 before taking into account E_m seems like a reasonable lower bound on what its value might be when the police have probable cause to suspect someone of having committed the crime. Recall that the DNA-match updating factor for incorporating the DNA match evidence, $P(E_m|H_1) / P(E_m|H_2)$, is equal to the reciprocal of the random match probability ($1 / p$). Suppose the random match probability for a forensic sample is $p = 1 / 10$ million such that the updating factor (provided by a statistician because it is computed from case-independent information) is 10 million to 1. In that case, the posterior odds after update are equal to 10 million $\times$ $(1 / 100) = 100\,000$-in-1 in favour of guilt. As this example illustrates, so long as p is reasonably small (e.g. no larger than 1 / 20 000), falling prey to base-rate neglect in the probable cause scenario will usually result in judges and jurors mistakenly believing that the odds of guilt are extraordinarily high (e.g. 20 million to 1) when in fact they are only extremely high (e.g. 100 thousand to 1). Under such conditions, the error would be largely inconsequential (see NRC II, p. 132, for a similar example). By contrast, in the DNA database search scenario, base-rate neglect can easily result in a catastrophic and consequential computational error even when p is as low as 1 / 100 million.

2. The database search scenario

For some crimes, the police do not have a suspect, but they do have an unknown DNA profile recovered from a crime scene. To find out who committed the crime, the unknown profile can be compared to known profiles in a DNA database. Critically, when no suspect has been established, the first piece of evidence against anyone is the DNA match from the database search. Thus, in the database search scenario, the order in which the evidence becomes known (DNA evidence first, other evidence second) is the reverse of the probable cause scenario.

In the U.S., the main DNA database is known as the Combined DNA Index System (CODIS), and it consists of a network of local, state and national DNA databases. Critically, 'No names or other personal identifiers of the offenders, arrestees, or detainees are stored using the CODIS software'. Only later, after the matcher is identified through a separate process (see section 6, National DNA Index System Operational Procedures Manual, 2018) does non-DNA evidence become known.

The fact that the database search is not limited to plausible suspects for having committed the crime in question (indeed, given the current structure of CODIS, it could not be limited in that way), is important to keep in mind when an attempt is made to estimate the prior odds of a match.

After the single database match, but before the matcher is identified, the hypotheses of interest are:

H_1: The forensic DNA belongs to the matcher.
H_2: The forensic DNA belongs to someone else.

The equations relating the elements of Bayes' rule to the random match probability (p), the number of profiles in the database (n) and the relevant population (N) have been previously specified (Balding, 2002; Balding and Donnelly, 1996; Berger *et al.*, 2015; Dawid, 2001; Dawid and Morterra, 1996; Donnelly and Friedman, 1999). We presented those equations earlier and do so again using more formal expressions. In light of the matching evidence (E_m), the posterior odds that the search has landed on the person who deposited the DNA at the crime scene, $P(H_1|E_m) / P(H_2|E_m)$, are given by:

$$\frac{P(H_1|E_m)}{P(H_2|E_m)} = \frac{P(E_m|H_1)}{P(E_m|H_2)} \mathrm{x} \frac{P(H_1)}{P(H_2)}. \quad (3)$$

Once again, the prior odds, $P(H_1) / P(H_2)$, are the odds that, before the DNA search is performed and the match becomes known, any one person from the relevant population would have deposited their DNA at the crime scene. As in the probable cause scenario, $P(H_1) / P(H_2) = 1 / (N - 1)$. The relevant population in the database search scenario is different from the relevant population in the probable cause scenario. In the database search scenario, N represents the number of plausible candidates for having their profiles in a DNA database, not the number of plausible suspects for having committed the crime in question. As a general rule, DNA profiles in a database, all of which are searched, come from convicted offenders and individuals arrested for a crime (usually a felony). Thus, N represents the active criminal population. Because the size of that population has nothing to do with the crime that yielded the forensic DNA profile, an estimate of N would remain unchanged whether or not the crime in question occurred. It therefore follows that N would be based on case-independent information.

The idea that N – and therefore the prior odds – can be appropriately estimated by the statistician based on case-independent measurements is a novel departure from previous analyses of the database search scenario. It is also the lynchpin of our proposed solution to base-rate neglect when a database search yields a single match. Instead of attempting to estimate N, statisticians have assumed that it is a quantity that cannot be reasonably estimated and that, even if it could be estimated, using it to calculate the prior odds would entail implausible simplifying assumptions (e.g. a uniform prior) that would introduce unwarranted statistical error (e.g. Balding, 2002). Here, we argue that N can be reasonably estimated from case-independent measurements (namely, database search statistics) and that the statistical error associated with using it to calculate the prior odds (and ultimately the posterior odds) is of minor consequence compared to the catastrophic computational error that can result from base rate neglect. Before considering how a case-independent estimate of N can be obtained, we first complete the formal presentation of the equations relating p, n and N to the Bayesian analysis of the single-match search result.

As specified in prior work (Balding, 2002; Balding and Donnelly, 1996; Berger *et al.*, 2015; Dawid, 2001; Dawid and Morterra, 1996; Donnelly and Friedman, 1999), the updating factor (likelihood ratio) for the single-match DNA database search scenario is equal to:

$$P(E_m|H_1)/P(E_m|H_2) = (1/p) \times (N-1)/(N-n). \quad (4)$$

Again, the impact of the matching evidence is always equal to or greater than $1 / p$ given that N is always greater than n. If N is much larger than n, as would be true when a database is just beginning to add profiles, then the quantity $(N - 1) / (N - n)$ is close to 1.0 and drops out such that $P(E_m|H_1) / P(E_m|H_2) \approx 1 / p$. Over time, the size of a database grows as more and more profiles are added to it. As n approaches N, the updating factor tends toward infinity. In the limit, when $n = N$ (i.e. when everyone in the relevant population is in the database, including the perpetrator), a single match would be conclusive because everyone but the perpetrator would be ruled out. Hence, the single-match DNA search increases the evidential value of a match relative to the probable cause scenario, more so as the size of the database grows because of the increasing number of potential suspects who are ruled out (Balding, 2002).

Multiplying the impact factor, $P(E_m|H_1) / P(E_m|H_2) = (1 / p) \times (N - 1) / (N - n)$, by the prior odds, $P(H_1) / P(H_2) = 1 / (N - 1)$, yields the expression for the posterior odds for the single-match database search scenario:

$$P(H_1|E_m)/P(H_2|E_m) = (1/p)/(N-n) \quad (5)$$

Equation 5 specifies the posterior odds that the forensic DNA belongs to the as-yet unidentified matcher. Recall that n is a known quantity and p is also a known quantity, so the only additional quantity that is needed to compute the posterior odds is N. Because all n of the profiles in the database are included in the search without regard for case-dependent information, even if the number of plausible suspects for having committed the crime in question could somehow be objectively estimated (call it $N_{plausible}$), it would be of no help. It would be on no help because the database search was not limited to plausible suspects and thus $N_{plausible}$ would not be equal to N in Equation 5.

Although estimating the number of plausible suspects for having committed the crime in question is of no use, database search statistics can be used to estimate the size of N, the number of plausible candidates for having their DNA profiles entered into the DNA database (Leary and Pease, 2003; Walsh *et al.*, 2010). Under the simplest assumption, n would be conceptualized as a random subset of N. We consider that simple scenario below, using n in combination with database search statistics to estimate the uncounted population of interest (namely, N). A more accurate estimate of N would have to take into account a variety of complexities (e.g. n may be a subset of, but not necessarily a random subset of, N, the size of N may be in constant flux, etc.). The key point, however, is that the estimate of N, like the estimate of p, is a case-independent statistical estimate, not an estimate that is in any way tied to subjective defense-oriented or prosecution-oriented theories about the number of plausible suspects for having committed the crime that yielded the forensic DNA profile. N is the size of the (partially hidden and therefore hard to count) active criminal population, and that population does not change depending on the specifics of any particular crime. Moreover, the statistical methodology for estimating the size of hidden populations is well developed.

3. Estimating the size of hidden populations

3.1 *Estimating wildlife populations*

Estimating hard-to-count populations has been a major focus of wildlife population research in the statistical literature for decades (Seber, 1982). Imagine, for example, trying to estimate the number of grizzly bears living in a particular geographical region (Boulanger *et al.*, 2004). There is no practical way to find and count them all (which is also true of the active criminal population), but some bears can be found, marked and then released back into the wild. At some later point, another set of bears can be found, and the proportion of those bears who are already marked provides an estimate of the total population of bears. This method of estimating the size of a hidden population is known as the 'mark and recapture' method, and its origins date back to the late 1800s and early 1900s (Lincoln, 1930; Peterson, 1895). Current methods are far more sophisticated than the simple approach used by the originators of the mark-and-recapture method, but we use the simplest approach to illustrate the basic idea.

Consider an unknown population of grizzly bears of size N in a specified geographical region. At first, 10 bears might be 'captured' and marked (note that non-invasive methods, such as photographing a bear or analysing DNA from a bear's feces can be used in place of actually capturing them). Once 10 bears have been captured and marked, researchers would have a database of size $n = 10$. After releasing the bears and allowing them sufficient time to randomly reintegrate into the wild population, another 10 bears are captured, and the number of those bears who are already marked (i.e. already in the database) is counted. In the simplest case, the population of interest is fixed (e.g. no births, deaths, immigration into the region or emigration out of the region), each bear is equally likely to be captured, and having been captured once does not change the probability of being captured again. Under those conditions, the expected proportion of bears that are marked upon recapture is n / N. Thus, if 10 bears were initially marked ($n = 10$) and if half the bears in the recaptured set are found to have been already marked, then $10 / N = 0.5$, which is to say that the estimated size of the population is equal to $N = 20$.

Obviously, this example involves highly simplified assumptions. In reality, all of the simplifying assumptions would likely be violated, and the implications of those violations would need to be taken into consideration. Indeed, much of the mark-and-recapture statistical literature has focused precisely on the more complicated task of estimating the size of hidden populations when the simplifying assumptions are relaxed (e.g. Chao *et al.*, 1992; McCrea and Morgan, 2015; Pollock, 1981, 1982; Pollock *et al.*, 1990). However, the fact that an accurate estimate is a complex problem does not change the fact that the estimate is purely statistical and measurement-based in nature, one that is unrelated to partisan theories about the number of plausible suspects for the crime that yielded the unknown DNA profile. Moreover, any error in its estimation likely pales in comparison to the error introduced by base rate neglect in the DNA database search scenario. We next consider how the simplest approach can be applied to the DNA database search scenario to estimate the size of the active criminal population.

3.2 *Estimating the active criminal population*

DNA database search statistics can be used in a similar way to estimate size of the active criminal population (Leary and Pease, 2003). The profiles in the DNA database correspond to the captured bears in the wildlife example above. That is, the individuals in the database are 'marked', and the number of such individuals is n (the size of the database). In the recapture phase, DNA profiles recovered from crime scene evidence are entered into the database. In the simplest case (including

the assumption that no false matches occur), a true match would be expected to occur in n / N cases, where N now represents the size of the active criminal population. For example, if the perpetrator is always in the database, a match would always occur, and we would conclude that $n / N = 1.0$, which would mean that $N = n$. However, if a match occurs with probability 0.50, then $n / N = 0.50$, and it would mean that N is about twice the size of the database, n (i.e. about half the people currently prone to committing serious crimes are already in the database).

Empirically, how often does a match occur when a DNA database is searched? In the U.S., the CODIS is a network of local, state and national DNA databases. If a crime scene DNA profile meets certain minimum requirements (e.g. a random match probability of 1 / 10 million or less), a search that fails to yield a hit at one level can proceed to the next level until the entire national network of profiles is searched, if need be. The National DNA Index System (NDIS) is the collective database of DNA profiles in the CODIS system that satisfy the minimum requirements. Some prior research has investigated how often a search of the NDIS database matches an existing profile.

In a prospective study of DNA collected from burglary crime scenes (Roman *et al.*, 2008), 42.5% of unknown profiles entered into the NDIS database yielded a hit. As noted earlier, in the simplest case, the population of active criminals is assumed to be fixed—that is, no births, deaths, immigration into the state of criminality or emigration from the state of criminality (e.g. due to incarceration). In addition, criminals are equally likely to be captured, regardless of whether they have been previously captured. Given those highly simplified assumptions, and ignoring the possibility of false matches, the estimate of N would be $n / 0.425 = 2.35n$.

A somewhat larger estimate of N results from statistics associated with forensic profiles submitted to the NDIS database by police departments across the nation. As of March 2018, NDIS contained 13 247 189 offender profiles, 3 020 223 arrestee profiles and 837 348 forensic profiles (a forensic profile is one that was recovered from a crime scene and entered into the database but did not match a known profile). That is, $n = 13\,247\,189 + 3\,020\,223 + 837\,348 = 17\,104\,760$. The probability of a match to a known profile when NDIS is searched using an unknown profile recovered from a crime scene is provided by a statistic called 'investigations aided'. Based on prior searches of the NDIS database, a match to a profile has occurred 410 968 times. When the database is searched and a match to a known profile does not occur, the profile remains unknown and is added to the database as a 'forensic profile', which means that 837 348 are now in the database as unknown forensic profiles because that number of searches was performed without yielding a match to an existing profile. Thus, out of 410 968 successful searches plus 837 348 unsuccessful searches = 1 248 316 total searches, a match to an existing profile occurred with probability of approximately $410\,968 / 1\,248\,316 = 0.329$. Using the same highly simplified assumptions as before, the estimate N would be $n / 0.329 = 3.13n$. It seems reasonable to assume that this is an overestimate because the size of the database has grown enormously since its inception in 1994 (Walsh *et al.*, 2010). Thus, many of the failures to match presumably occurred when the size of the database was much smaller than it is today (when a higher proportion of the database population was in prison compared to today).

If we conservatively assume $N = 2n$, it means that our initial estimate of $N = 2 \times 17\,104\,760 = 34\,209\,520$, or about 10% of the current U.S. population. This estimate is not wildly different from other seemingly reasonable but even more conservative approaches to estimating the active criminal population. For example, Shannon *et al.* (2017) estimated the size of the U.S. criminal population at 19.8 million (6% of the current U.S. population) as of 2010 based on the number of people who have ever been convicted of a felony, whether or not they served time in prison. Using their number as an estimate of the active criminal population seems conservative in that it does not include many potential

criminals who were arrested for but not ultimately convicted of a felony and many others who have committed felonies but were never arrested. Moreover, given that the data exhibit a linear rise in the size of the felony population between 2000 and 2010, extrapolating, one would estimate that as of 2018, this figure would be approximately 25 million. Using this number as an estimate of the active criminal population, a very conservative estimate of N would be approximately $1.5n$.

4. The case-independent posterior odds in the DNA database search scenario

If these estimates of the active criminal population are anywhere near accurate, the implications would be profound. With an estimate of N in hand, Equation 5 could be used to estimate the posterior odds independent of case-specific information, thereby avoiding base-rate neglect. Consider, for example, what the posterior odds would be for an unknown forensic profile with a random match probability of 1 / 10 million in the probable cause scenario versus the database search scenario. For the probable cause scenario in which (for example) the prior odds might be ≈ 1 (or close enough to this value so as not matter much), the posterior odds would be $1 \times 1/p = 1$ / (1 / 10 million) = 10 million. That is, the posterior odds would be 10 million-to-1 in favour of H_1 (the suspect is guilty). The same estimate of the posterior odds would be made following a search of a national DNA database if decision-makers fell prey to base-rate neglect (ignoring the prior odds). In reality, the posterior odds would be radically different when the same match is obtained from a national database search. For example, assuming $N = 2n$, the posterior odds for the database search scenario are equal to $1 / [p(N - n)] = 1 / \{(1 / 10$ million) [2(17 104 760) – 17 104 760]} = 0.58 / 1.[2] In other words, instead of the DNA match indicating 10 million-to-1 odds in favour of guilt, the match from a database search indicates slightly 'less than even odds' that the matcher is guilty of having deposited his DNA at the crime scene. This is true even though the probability of a false match, p, is extremely low and is the same in both scenarios, which is the key point of this article.

A concern might be that any statistical estimate of N based on case-independent information would not be sufficiently precise because of the simplifying assumptions involved (e.g. Balding, 2002). However, given the catastrophic misinterpretation of a single match that arises from neglecting base rate information, any error in the posterior odds arising from error in the estimate of N is trivial by comparison. Figure 1 illustrates this point. Using a log scale, the x-axis represents the random match probability (p) ranging from 10^{-3} (1 / 1 thousand) to 10^{-12} (1 / 1 trillion). The y-axis represents odds, also using a log scale. The upper shaded region depicts the likelihood ratio as a function of p computed from Equation 4, with $n = 17$ million, and bounded by $N = 1.1n$ to $N = 10n$. This range of N is much larger than any reasonable estimate of N (exaggerating its imprecision) to illustrate why having an exact estimate is not essential. The vertical dashed line is placed at $p = 1 / 10$ million. At that point, the likelihood ratio (upper shaded region) exceeds 10 million to 1 for $N = 10n$ (the solid lower boundary of the shaded region) and is even higher than that as n approaches N, that is, for $N = 1.1n$ (the dashed upper boundary).

For $p = 1 / 10$ million, the posterior odds (lower shaded region) are less than even for $N = 10n$ (the solid lower boundary of the shaded region) and only reaches about 6 to 1 as N approaches n, that is, for $N = 1.1n$ (the dashed upper boundary). Clearly, the error in the posterior odds is miniscule relative to

[2] As an interesting aside, if we assume that $N = 2n$, then, according to Equation 5, the posterior odds = $(1/p) / (N - n) = (1/p) / (2n - n) = 1 / np$, which corresponds exactly to the equation recommended by NRC II for interpreting the outcome of a single-match DNA database search. Thus, if the NRC II intended that equation to apply to the posterior odds despite how they verbally characterized it, then it was a reasonable approximation.

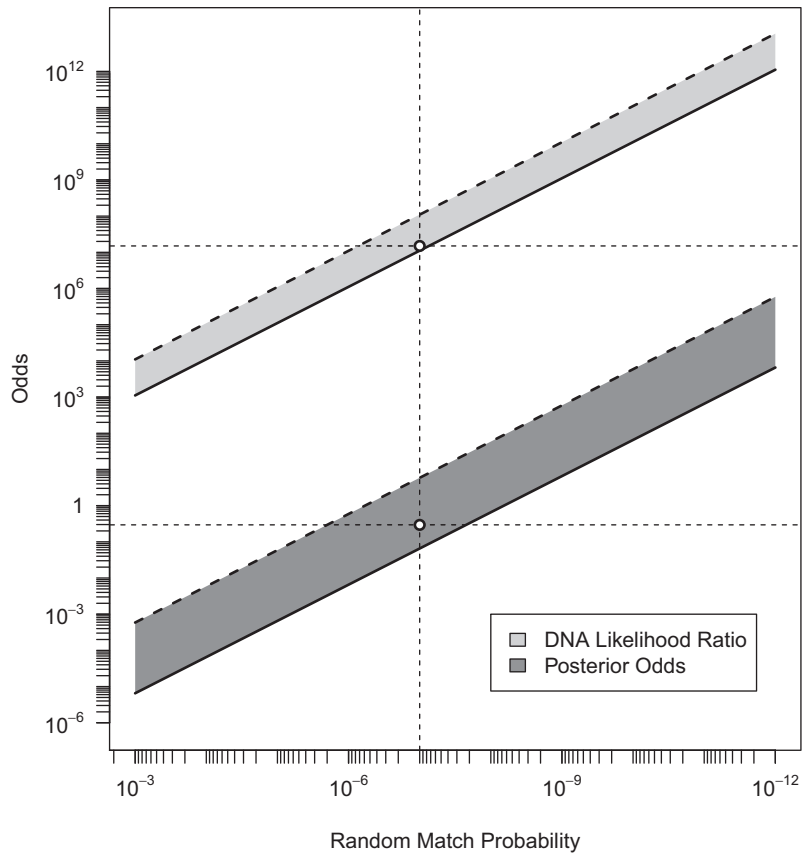

FIG. 1. Relationship between the random match probability and two values: (1) the likelihood ratio computed from Equation 4 (upper region shaded in light grey) and (2) the posterior odds computed from Equation 5 (lower region shaded in dark grey). For both, $n = 17$ million and the shaded regions span a range encompassed by estimates of N ranging from a low of $N = 1.1n$ (upper dashed boundary) to a high of $N = 10n$ (lower solid boundary). The vertical dashed line corresponds to a random match probability of 1 / 10 million, and the two white points corresponds to $N = 3n$.

the error that would occur if the prior odds were disregarded altogether. Disregarding the prior odds would result in an estimate of the posterior odds equalling the likelihood ratio (upper shaded region), which is many orders of magnitude larger. The magnitude of the cognitive error is even larger than it appears to be in Fig. 1 because log scales are used. The point is that even assuming implausibly large error in the statistical estimate of N based on case-independent information, the resulting posterior odds based on an estimate of N provide a vastly closer approximation to the truth than what would be assumed by someone who ignores the prior odds. Figure 1 also illustrates that if p is low enough (e.g. 1 / 10^{-12}), the posterior odds are very high ($\sim 10^{5}$) despite being many orders of magnitude less than the likelihood ratio ($\sim 10^{12}$). But for values of p no smaller than $\sim 10^{-9}$ (1 / 1 billion), the posterior odds are not necessarily conclusive. For example, if $p = 10^{-9}$, the posterior odds range from about 6 to 600 (and are approximately 60 to 1 for $N = 2n$).

5. The case-dependent posterior odds of guilt for the identified matcher

Another concern might be that by following our recommended approach, we have taken the adjudication of guilt or innocence out of the hands of the judge and jury and put it into the hands of the statistician. However, this is not the case because the estimated posterior odds pertain to the odds that the DNA belongs to the matcher (not that the matcher is guilty). The judge and jury will use the posterior odds provided by the statistician only as a starting point for their case-dependent calculations based on information that becomes available after the matcher is identified. Two simplified examples will illustrate one way they might go about doing that. For both examples, we assume that a single match was obtained by searching the NDIS database and that, for both, the random match probability was $p = 1 / 10$ million (the maximum value for searching the NDIS database). Assume further that the statistician's best estimate of N is $2n$. As noted earlier, for a search of the NDIS database ($n = 17\ 104\ 760$), this would come to $N = 34\ 209\ 520$. Before taking into consideration case-dependent information, the hypotheses are:

H_1: The forensic DNA belongs to the matcher.
H_2: The forensic DNA belongs to someone else.

The posterior odds based on the DNA matching evidence (E_m) are given by:

$$\frac{P(H_1|E_m)}{P(H_2|E_m)} = \frac{P(E_m|H_1)}{P(E_m|H_2)} \mathrm{x} \frac{P(H_1)}{P(H_2)} \tag{6}$$

As noted earlier, for the single-match database search scenario, $P(H_1) / P(H_2) = 1 / (N - 1)$, and $P(E_m|H_1) / P(E_m|H_2) = (1 / p) \times (N - 1) / (N - n)$.

Plugging in the values listed above, the posterior odds based on case-independent information comes to 0.58. In other words, the odds are less than even that the forensic DNA belongs to the as-yet unidentified matcher. This value is much larger than the prior odds of approximately 1 / 34 million (i.e. the forensic DNA is now much more likely to belong to the matcher than to a person randomly selected from the relevant population), but the evidence remains inconclusive nonetheless. Providing this estimate of the posterior odds as a starting point for their deliberations would protect prosecutors and jurors from engaging in base rate neglect. Their subsequent case-dependent calculations designed to evaluate guilt or innocence would be based on information that becomes available once the matcher is identified. How that information is integrated with the DNA matching evidence would be up to them.

We assume that jurors would consider the evidence in chronological order, first considering how the suspect was identified—namely, the forensic DNA profile matched the defendant's known profile (and no other profile) in a database search—and then taking into consideration all evidence that has since come to light. As we did for the probable cause scenario, we again consider the simplest possible case where the jury might assume that if the forensic DNA belongs to the suspect, the suspect is guilty (H_1), but if the forensic DNA belongs to someone else, the suspect is innocent (H_2). Under those conditions, H_1 and H_2 specified above also correspond to guilt versus innocence, respectively. Note that we consider only this simple scenario because our purpose is not to explain how jurors should use the DNA matching evidence to determine whether or not the suspect is guilty. That is a complex problem that has been the focus of a considerable body of recent work (e.g. Gittelson *et al.*, 2016). Instead, our purpose is to argue that judges and jurors should be provided with a statistical estimate of the posterior

odds that the forensic DNA belongs to the matcher before any other information pertaining to the (now identified) suspect is taken into consideration. Next, we consider how that information would be updated given the simplifying assumption that guilt or innocence map onto whether the forensic DNA belongs to the suspect or not.

Example 1: Imagine that the matcher turns out to be Mr. Jones, a felon who, as it turns out, was in prison when the crime occurred. With no further help from the statistician, the jury would use the new information about Mr. Jones to update the prior odds and compute the posterior odds of guilt. The competing hypotheses are now personalized:

H_1: Mr. Jones is guilty.
H_2: Mr. Jones is innocent.

To compute the posterior odds that Mr. Jones is guilty, the jury would use the new case-specific evidence (E_o)—the observation that the matcher, Mr. Jones, was in prison when the crime was committed—to formulate a subjective updating factor consisting of the true positive rate, $P(E_o|H_1)$, divided by the false positive rate, $P(E_o|H_2)$, and then multiply that ratio by the case-independent posterior odds provided by the statistician (the term in brackets below):

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(E_o|H_1)}{P(E_o|H_2)} \text{x} \left[\frac{P(E_m|H_1)}{P(E_m|H_2)} \text{x} \frac{P(H_1)}{P(H_2)} \right], \tag{7}$$

D represents all of the evidence, that is, the other evidence plus the DNA matching evidence (i.e. $D = E_o + E_m$). Conceptually, $P(E_o|H_1)$ corresponds to the proportion of DNA database searches that, in the past, yielded a single match to a guilty person whose DNA was in fact deposited on the crime scene evidence and who was in prison at the time of the crime. Although such information would not actually be available, imagine that of the M_1 single-match true positives in past searches, m_1 of the guilty matchers were in prison when the crime was committed. In that case, $P(E_o|H_1)$, would equal m_1 / M_1. This value would obviously be very close to 0 because it is extremely unlikely that someone who is in prison would have deposited their DNA on the crime scene evidence. $P(E_o|H_2)$ corresponds to the proportion of DNA database searches that, in the past, yielded a single match to an innocent person whose DNA was *not* deposited on the crime scene evidence (i.e. the crime scene DNA belongs to someone else despite the match) and who was in prison at the time of the crime. Imagine that of those M_2 single-match false positives, m_2 of them were in prison when the crime was committed. In that case, $P(E_o|H_2)$, would equal m_2 / M_2. Because the people with profiles in the database come mainly from the active criminal population, it would not be very surprising to discover that some of these false matchers were in prison when the crime was committed. Thus, although $P(E_o|H_2)$ might be small, it would presumably be much larger than $P(E_o|H_1)$.

Imagine that the jury subjectively estimates that $P(E_o|H_1) = 0.00001$ and $P(E_o|H_2) = 0.1$, in which case the updating factor would be $P(E_o|H_1) / P(E_o|H_2) = 0.00001 / 0.1 = 0.0001$. Multiplying the updating factor by the prior odds of 0.58 provided by the statistician yields a posterior odds that Mr. Jones is guilty of $P(H_1|D) / P(H_2|D) = 0.000058$. The posterior odds weigh heavily against guilt, so (absent other incriminating evidence) the jury would presumably find Mr. Jones to be not guilty. Note that the jury was never given the opportunity to fall prey to base rate neglect.

Instead of following the steps outlined above, imagine instead that current practice were followed such that the statistician only advised the jury about the updating factor associated with the DNA evidence, $P(E_m|H_1) / P(E_m|H_2)$. The jury would be told by the statistician that $P(E_m|H_1) / P(E_m|H_2)$ is at least $1 / p = 10$ million. Further assume that the jury evaluated $P(E_o|H_1) / P(E_o|H_2)$ to be 0.0001 and then ignored the prior odds, using the following equation (which omits the prior odds) to compute the wrong (W) posterior odds:

$$W = \frac{P(E_o|H_1)}{P(E_o|H_2)} \text{x} \frac{P(E_m|H_1)}{P(E_m|H_2)}. \quad (8)$$

The wrong posterior odds of guilt would come to 0.0001×10 million $= 1000$. The posterior odds of 1000-to-1 weigh heavily, but inappropriately, in favour of the hypothesis that Mr. Jones is guilty.

Example 2: Consider next a different hypothetical example where the case-dependent evidence, when properly evaluated, would have the opposite effect compared to the previous example. For this example, assume that the case-independent posterior odds from the database search is the same as before (0.58). Further assume that the crime in question is a sexual assault and that the matcher is revealed to be Mr. Smith. The competing hypotheses are:

H_1: Mr. Smith is guilty.
H_2: Mr. Smith is innocent.

Now that he has been identified, case-dependent calculations can be performed based on the new evidence that becomes available (E_o). Imagine that of all the places in the country where he might live, it turns out that Mr. Smith lives one block away from the victim. Suppose it also turns out that he has a history of sexually assaulting women. Based on this new information, the jury would then update the prior odds of guilt to determine the posterior odds of guilt.

As before, $P(E_o|H_1)$ corresponds to the proportion of DNA database searches that, in the past, yielded a single match to a guilty person whose DNA was in fact, deposited on the crime scene evidence and who lived near their victims and had a history of committing similar crimes. Imagine that of those M_1 true positives, m_1 of them fell into that category. In that case, $P(E_o |H_1)$, would equal m_1 / M_1. This value would probably be fairly high.

$P(E_o|H_2)$ corresponds to the proportion of DNA database searches that, in the past, yielded a single match to an innocent person whose DNA was not deposited on the crime scene evidence and who lived near their victims and had a history committing similar crimes. Imagine that of those M_2 false positives, m_2 of them fell into that category. In that case, $P(E_o |H_2)$, would equal m_2 / M_2. The probability that a false matcher, who could live anywhere in the country, would coincidentally live close to the victim of a crime committed by someone else is extremely low.

Imagine that the jury subjectively estimates that $P(E_o|H_1) = 0.2$ and that $P(E_o|H_2) = 0.0001$, in which case the likelihood ratio for updating the prior odds would be $P(E_o|H_1) / P(E_o|H_2) = 0.2 / 0.0001 = 2000$. Multiplying this updating factor by the prior odds of 0.58 yields a posterior odds that Mr. Smith is guilty of $P(H_1|D) / P(H_2|D) = 1160$. The posterior odds weigh heavily in favour of the hypothesis that the crime scene DNA belongs to Mr. Smith, so the jury would therefore likely conclude that it belongs to him and (absent other exonerating evidence) presumably find Mr. Smith guilty.

6. General discussion

Although intuition may suggest that a reported single match between a DNA profile of a suspect and a forensic DNA profile is compelling, especially when a single match is obtained with a low random match probability, the implications can be far more equivocal. To assist the legal system in the interpretation of such a match, a statistician may be called upon to offer expert advice, and the nature of that advice is what concerns us here. According to a common perspective, the goal of the statistician is to provide an estimate of the strength of the DNA evidence 'standing alone'. Standing alone, the strength of evidence is provided by an estimate of the likelihood ratio, which quantifies how much a reported DNA match should amend the prior odds, whether the prior odds are low (as in the database search scenario) or high (as in the probable cause scenario). However, judges and jurors are ultimately interested in the posterior odds: given the reported match, what are the odds that the forensic DNA belongs to the suspect?

Any estimate of the prior odds is generally thought to be based on an estimate of the number of plausible suspects for having committed a particular crime and is therefore an inherently subjective judgment about which the prosecution and defence may disagree. Because a statistician's subjective judgment about those matters is no more authoritative than that of a juror's, there is no reason for the statistician to weigh in on that issue. In light of these considerations, a consensus has emerged that the statistician has no business proving an estimate of the prior odds (e.g. Dawid, 2001; Donnelly and Friedman, 1999; Neumann *et al.*, 2016). As we illustrated earlier, in the probable cause scenario, such an estimate is usually not needed because a DNA match typically provides compelling evidence whether or not the prior odds are taken into consideration. In the database search scenario, by contrast, ignoring the prior odds (base rate neglect) can result in a catastrophic misinterpretation of the implications of a single match.

6.1 *Base-rate neglect is a pervasive problem*

Unfortunately, without statistical guidance, the jury is likely to engage in base rate neglect, ignoring the prior odds altogether. Given the large body of evidence showing that people are prone to base-rate neglect across a wide variety of circumstances (Gaissmaier and Gigerenzer, 2011; Gigerenzer and Edwards, 2003; Gigerenzer *et al.*, 2008; Kahneman and Tversky, 1973; Koehler, 1996; Lyon and Slovic, 1976; Meehl and Rosen, 1955), there is no reason to believe that judges, jurors and prosecutors are immune to making that error. Indeed, Koehler (1993) documented the occurrence of base-rate neglect by judges, expert witnesses, and (as a possible proxy for jurors) reporters in the popular press. Research on how people interpret a DNA match in the probable cause scenario versus the database search scenario also suggests that they are prone to base-rate neglect as well. For example, Lindsey *et al.* (2003) tested 127 advanced law students and 27 professional jurists in Germany to evaluate two realistic rape cases. For a given random match probability of 1 in 100 000, the participants were almost equally likely to vote to convict based on a DNA match from a probable cause scenario involving eyewitness identification evidence (53%) as they were based solely on a match from a database search (48%). Similarly, Scurich and John (2011) found that, given equal random match probabilities of 1 in 200 million, participants were about as likely to believe that the suspect was guilty based solely on a DNA match from a database search (∼75%) as they were based on a match from a probable cause scenario (61–81%, depending on whether the strength of the non-DNA evidence against the suspect was weak or strong). Given how the DNA matching evidence was interpreted in the probable cause

scenario, the DNA matching evidence in the database search scenario (with no other evidence involved) should have been discounted to a very considerable degree. Yet the participants in both studies exhibited base-rate neglect by not discounting the posterior odds of guilt much at all.

6.2 Previously proposed solutions

In addition to recommending the 1 / *np* rule, the NRC II report briefly considered three ways of presenting the prior odds (pp. 201–202): (1) relying on an expert statistician to estimate the prior odds, (2) allowing the jury to work out the prior odds for themselves, and (3) presenting the jury with a range of possible prior odds in a table, showing how each prior odds value, when multiplied by the likelihood ratio, yields a different estimate of the posterior odds. As already noted, the problem with the first solution is that an estimate of the prior odds that the forensic DNA belongs to an accused person is thought to involve case-dependent calculations that fall outside the purview of the statistician. The problem with the second solution is that jurors are at risk of falling prey to base rate neglect, as almost everyone does. Thus, the NRC II endorsed the third solution, as have others (e.g. Kaye, 1993; Meester and Sjerps, 2003).

Although providing an educational table designed to teach a jury about the role of the prior odds seems far superior to simply hoping that jurors will not engage in base rate neglect, it is not clear that it would actually facilitate the average juror's understanding of Bayes' theorem. Moreover, the prior odds truly depend on *N*. How should that fact be communicated to the jury? Meester and Sjerps (2003) offered the following suggestion: 'For this, the size *N* of the relevant population is needed, and this value may be discussed with the court or, alternatively, multiple tables for various values of *N* can be reported' (p. 731). This approach at least attempts to grapple with the known importance of *N*, but it is not clear that it would effectively address the problem of base rate neglect.

6.3 Our solution differs from previously proposed solutions

Here, we have proposed a different solution. Because prosecutors, judges and jurors are at risk of misinterpreting the likelihood ratio (i.e. 1 / *p*) as the posterior odds, it makes sense for statisticians to estimate *N* (the active criminal population) using methods that have been developed to estimate hard-to-count wildlife populations. Having an estimate of *N* permits the statistician to calculate the case-independent prior odds which, in turn, allows the calculation of the posterior odds—independent of case-specific information—using Equation 5. Critically, this evaluation of the posterior odds would be many orders of magnitude closer to what jurors left to their own devices would conclude if they engaged in base rate neglect. These case-independent posterior odds would then serve as the prior odds to be updated by the jury's subjective updating factor based on case-dependent information pertaining to the named defendant and the crime in question. Thus, it would be the judge and jury, not the statistician, who would decide the guilt or innocence of the named defendant.

6.4 The prosecutor's fallacy

Although we have focused here on base-rate neglect (ignoring the prior probability of guilt) and its close relative (ignoring the prior odds), our recommended approach bears on another well-known cognitive error known as the 'prosecutor's fallacy' (Thompson and Schumann, 1987). Applied to the database search scenario, the prosecutor's fallacy consists of misinterpreting the false positive rate, $P(H_2|E_m) = p$, as the probability that the DNA does not belong to the matching individual despite the

match. In other words, the fallacy consists of transposing the conditional by assuming that $P(E_m|H_2) = P(H_2|E_m)$.

The cognitive error in the prosecutor's fallacy (transposing the conditional) is not identical to the cognitive error known as base rate neglect. However, mathematically, the difference between these two errors is negligible when p is small (as it almost always is for DNA evidence) and the statistician estimates the false positive rate to be at least as small as p. As noted earlier, if the statistician provides the relevant information in the form of a likelihood ratio ($1 / p$), and if the prior odds are ignored, the posterior *odds* of guilt in the single-match database search scenario are mistakenly assumed to equal $1 / p$. If, instead, the prosecutor's fallacy occurs in response to the statistician's estimate of p, the posterior *probability* of guilt is mistakenly assumed to equal to $1 - p$. In odds form, the prosecutor's fallacy is given by $(1 - p) / [1 - (1 - p)] = (1 - p) / p$. Note that as p becomes small, its contribution to the numerator of this equation becomes increasingly negligible, such that $(1 - p) / p \approx 1 / p$. Thus, in a purely mathematical sense, the prosecutor's fallacy is tantamount to base rate neglect. In both cases, in the single-match DNA database search scenario, the posterior odds of guilt are mistakenly judged to be approximately $1 / p$. However, if the statistician evaluates the case-independent posterior odds and communicates that result to the prosecutor (instead of communicating only p or $1 / p$, in accordance with current practice), as we have recommended here, the prosecutor would never have an opportunity to catastrophically misinterpret the probability of guilt as being $1 - p$. Thus, our proposed solution addresses the prosecutor's fallacy as well. In the absence of other compelling evidence, the prosecutor might decide not to bring charges against the matcher.

6.5 State and local DNA databases

Although our main focus has been on the implications of a single match obtained from a search of the NDIS database (which limits searches to forensic DNA samples with a random match probability of 1 / 10 million or less), the problem we seek to address is even more concerning for state and local DNA database searches. As a general rule, the search criteria ' . . . are stricter when moving from the local (i.e., LDIS) to the national (i.e., NDIS) DNA index systems' (Roman *et al.*, 2009, p. 347). In some states (e.g. California, Colorado and Arizona), a single match from a search of a state database is sufficient grounds for an arrest warrant to be issued (Roman *et al.*, 2009). California currently has 2 799 262 profiles in its state database. If we assume that $N_{California} = 2n_{California}$, then $N_{California} = 5\,598\,524$. If the random match probability is $p = 1 / 1$ million, then, according to Equation 5, the posterior odds of a single match would come to 0.357. In other words, 3 out of 4 people arrested on that basis would be innocent.

6.6 Case-independent posterior odds in a court of law

Presenting an evaluation of the case-independent posterior odds to a jury would involve additional considerations because the admissibility of a statistical argument by an expert is a complex legal issue. For example, Federal Rule of Evidence 702 states that if specialized knowledge will help judges and jurors understand the evidence, then an expert witness may testify so long as: (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case (Bernstein and Jackson, 2004). Equation 5, which quantifies the relevant posterior odds, directly applies to the facts of

any case in which a suspect was identified through a single-match DNA database search, and few would dispute that it was derived using sound principles and methods. Moreover, although the Bayesian statistical analysis of the DNA database search scenario was once a controversial issue, the proper equation for computing the posterior odds (Equation 5) now enjoys a wide consensus, thereby satisfying the *Frye* test of general acceptance in the scientific community. Another potential concern is that a presentation of the case-independent posterior odds might be confusing or misleading to jurors, thereby running afoul of Federal Rule of Evidence Rule 403. However, as noted by Koehler (1996), 'Although they have scientific merit, likelihood ratios – which are the ratios of conditional probabilities – are not easy to understand' (p. 877). In addition, given how prone judges and jurors already are to seriously misinterpret the likelihood ratio or p presented alone (due to the general human proclivity to engage in base rate neglect), presenting the posterior odds using Equation 5 would seem to represent a potentially dramatic improvement over the current state of affairs.

Although Equation 5 is generally accepted in the scientific community and yields an understandable result, its use in court would depend on a statistical estimate of a partially hidden and therefore uncounted population, namely, the active criminal population. Would such testimony be based on 'sufficient facts or data', as required by Federal Rule of Evidence 702? We assume so given that the estimated posterior odds are relatively insensitive to the size of N over a very wide range (Fig. 1). Indeed, any reasonable estimate of N yields an estimate of the posterior odds that is 'many orders of magnitude closer to the truth' compared to an estimate based on ignoring the prior odds (or falling prey to the prosecutor's fallacy).

As noted earlier, Balding (2002) briefly considered estimating N (although without taking into account the mark-and-recapture approach) and asserted that the simplifying assumptions that would be needed are both 'implausible and inappropriate for an expert witness' (p. 243). However, in our view, the expert statistician is the only one who is in a position to provide a reasonable estimate of N. On their own, judges and jurors are unlikely to appreciate either the size or the importance of N. A statistician, by contrast, has useful information to provide about both issues. Moreover, the estimates would be purely statistical and would have nothing at all to do with the details of any particular case. Finally, as noted above, the evaluation of the case-independent posterior odds based on a statistical estimate of N would, despite the imprecision of that estimate, result in a much closer approximation to the truth than would otherwise be the case.

Although only a ballpark estimate of N is needed to dramatically improve a jury's understanding of the posterior odds associated with a single-match DNA database search (Fig. 1), there is no reason why statisticians should not seek ever more accurate estimates. More accurate estimates of N would take into account immigration into the state of criminality (i.e. criminally inclined individuals coming of age), emigration from the state of criminality (i.e. criminals dying, or being incarcerated), re-immigration into the state of criminality (upon release from prison), criminals having different probabilities of being captured, the changing size of the database over the years, and the fact that not all forensic profiles entered in the database are from criminals and some may be duplicates (Walsh *et al.*, 2010). These are the same kinds of complexities that are involved in attempts to estimate hidden wildlife populations. However, these complexities do not argue in favour of statisticians leaving the work of statistically estimating N to prosecutors, judges and juries. Quite the opposite, in our view. Who better than a statistician to estimate N, and the posterior odds, independent of subjective considerations pertaining to a specific case involving named defendant?

References

Aitken, C, Berger, C. E. H., Buckleton, J. S., Champod, C., Curran, J., Dawid, A. and Jackson G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice*, **51**, 1–2.

Balding, D. J. (2002). The DNA database search controversy. *Biometrics*, **58**, 241–244.

Balding, D. J. and Donnelly, P. (1996). Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Science*, **41**, 603–607.

Berger, C. E. H., Vergeer, P. and Buckleton, J. S. (2015). A more straightforward derivation of the LR for a database search. *Forensic Science International: Genetics*, **14**, 156–160.

Bernstein, D. E. and Jackson, J. D. (2004). The Daubert trilogy in the States. *Jurimetrics*, **44**, 351–366.

Boulanger, J., Himmer, S. and Swan, C. (2004). Monitoring of grizzly bear population trends and demography using DNA mark–recapture methods in the Owikeno Lake area of British Columbia. *Canadian Journal of Zoology*, **82**, 267–1277.

Chao, A., Lee, S. M. and Jeng, S. L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, **48**, 201–216.

Dawid, A. P. (2001). Comment on Stockmarr's 'Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search'. *Biometrics*, **57**, 976–978.

Dawid, A. P. and Morterra, J. (1996). Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society, Series B*, **58**, 425–443.

Donnelly, P. and Friedman, R. D. (1999). DNA database searches and the legal consumption of scientific evidence. *Michigan Law Review*, **97**, 931–984.

Evett, I. W., Gill, P. D., Jackson, G., Whitaker, J. and Champod, C. (2002). Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks. *Journal of Forensic Science*, **47**, 520–530.

Gaissmaier, W. and Gigerenzer, G. (2011). When misinformed patients try to make informed health decisions. In Gigerenzer and Muir Gray (Eds.), *Better Doctors, Better Patients, Better Decisions: Envisioning Health Care 2020*, MIT Press, Cambridge MA.

Gigerenzer, G. and Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ : British Medical Journal*, **327**, 741–744.

Gigerenzer, G., Gaissmaier, W., Kurze-Milcke, E., Schwartz, L. M. and Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, **8**, 53–96.

Gittelson, S., Kalafut, T., Myers, S., Taylor, D., Hicks, T., Taroni, F., Evett, I. W., Bright, J. A. and Buckleton, J. (2016). A practical guide for the formulation of propositions in the Bayesian approach to DNA evidence interpretation in an adversarial environment. *Journal of Forensic Sciences*, **61**, 186–195.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237–251.

Kaye, D. H. (1993). DNA evidence: Probability, population genetics, and the courts. *Harvard Journal of Law and Technology*, **7**, 101–172.

Koehler, J. J. (1993). Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics Journal*, **34**, 21–39.

Koehler, J. J. (1996). On conveying the probative value of DNA evidence: Frequencies, likelihood ratios, and error rates. *University of Colorado Law Review*, **67**, 859–886.

Leary, D. and Pease, K. (2003). DNA and the active criminal population. *Crime Prevention and Community Safety: An International Journal*, **5**, 7–12.

Lincoln, F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. *United States Department of Agriculture Circular*, **118**, 1–4.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika*, **64**, 207–213.

Lindsey, S., Hertwig, R. and Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics Journal*, **43**, 147–163.

Lyon, D. and Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, **40**, 287–298.

McCrea, R. S. and Morgan, B. J. T. (2015). *Analysis of Capture-recapture Data*. CRC Press, Boca Raton, FL.
Meehl, P. E. and Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, **52**, 194–216.
Meester, R. and Sjerps, M. (2003). The evidential value in the DNA database search controversy and the two-stain problem. *Biometrics*, **59**, 727–732.
Meester, R. and Sjerps, M. (2004). Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence. *Law, Probability and Risk*, **3**, 51–62.
National DNA Index System (NDIS) Operational Procedures Manual (2018). FBI Laboratory Version 7: Effective 06/01/2018. https://www.fbi.gov/file-repository/ndis-operational-procedures-manual.pdf
National Research Council Committee on DNA Technology in Forensic Science, DNA technology in forensic science, National Academy Press, Washington DC, 1992.
National Research Council Committee on DNA Forensic Science: An Update, The evaluation of forensic DNA evidence, National Academy Press, Washington DC, 1996.
Neumann, C., Kaye, D., Jackson, G., Reyna, V. F. and Ranadive, A. (2016). Presenting qualitative information on forensic science evidence in the court room. *CHANCE*, **29**, 37–43.
Petersen, C. G. J. (1895). The yearly immigration of young plaice into the limfjord from the German Sea. *Report of the Danish Biological Station*, **6**, 5–84.
Pollock, K. H. (1981). Capture recapture models allowing for age dependent survival and capture rates. *Biometrics*, **37**, 521–529.
Pollock, K. H. (1982). A capture recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, **46**, 752–757.
Pollock, K. H., Nichols, J. D., Brownie, C. and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Society Monographs* #**107**.
Roman, J. K., Reid, S., Reid, J., Chalfin, A., Adams, W. and Knight, C. (2008). The DNA field experiment: A randomised experiment of the cost-effectiveness of using DNA to solve property crimes. *Journal of Experimental Criminology*, **5**, 345–369.
Scurich, N. and John, R. S. (2011). Trawling genetic databases: When a DNA match is just a naked statistic. *Journal of Empirical Legal Studies*, **8**, 49–71.
Seber, G. A. F. (1982). *The Estimation of Animal Abundance*, 2nd edition, Griffin, London.
Shannon, S., Uggen, C., Thompson, M., Schnittker, J. and Massoglia, M. (2017). The growth, scope, and spatial distribution of people with felony records in the United States, 1948 to 2010. *Demography*, **54**, 1795–1818.
Thompson, W. C. and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, **11**, 167–187.
Walsh, S. J., Curran, J. and Buckleton, J. S. (2010). Modeling forensic DNA database performance. *Journal of Forensic Science*, **55**, 1174–1183.