2009

Identification, Individualization, Uniqueness

David H. Kaye
Penn State Law

these cases. Consequently, the court rejected Nelson's speedy trial claims under the state and federal constitutions.

The court also considered whether the vanishingly small random-match probabilities should have been admitted. These probabilities, computed with the 'product rule', are routinely admitted when DNA from a suspect, initially identified by non-genetic evidence, is shown to match DNA recovered from a crime scene (Kaye, 2004). The question in *Nelson* was whether this random-match probability no longer applies because the defendant was not selected 'at random' with respect to his DNA type, but precisely because of this type.

1. The putative need to adjust the random-match probability

Two committees of the U.S. National Academy of Sciences (NAS) maintained that a match coming from a trawl is much less impressive than a match in a confirmation case—just as finding a tasty apple on the very first bite is more impressive than pawing through the whole barrel of apples to locate a succulent one. The first committee (National Research Council Committee on DNA Forensic Science, 1992, p. 124) cautioned that:

> The distinction between finding a match between an evidence sample and a suspect sample and finding a match between an evidence sample and one of many entries in a DNA profile databank is important. The chance of finding a match in the second case is considerably higher, because one does not start with a single hypothesis to test (i.e., that the evidence was left by a particular suspect), but instead fishes through the databank, trying out many hypotheses.

This concern about testing multiple hypotheses arises from a frequentist approach to statistical inference. If I flip an unbiased coin 20 times, I am unlikely to observe 20 heads. Such an outcome would happen only about one time in a million trials. Consequently, the observation is very surprising—if the null hypothesis (that the coin is unbiased) is true. This low probability justifies rejecting the null hypothesis. Even if we always encountered unbiased coins, we would falsely reject the null hypothesis only about one time per million for those suspect coins that produced 20 heads in 20 tosses.

But the one-in-a-million probability is a misleading measure of how surprising the result is when we fish through a huge list of the sequences of heads and tails generated by unbiased coins. Suppose we have 300 000 unbiased coins, and we toss each one 20 times to create a database of 300 000 records of sequences of heads and tails. It would not be too surprising if one or more coins in this 'innocent database' had a 'profile' of 20 heads. In fact, if we repeat this experiment of creating and trawling innocent databases, we would expect to find about a quarter of the databases with at least one record of 20 heads.

The NAS committees thought that this coin-tossing analogy applies to DNA databases. To solve the perceived problem of searching for significance, the first committee recommended using only a small set of genetic locations (loci) for the initial testing. Then, '[w]hen a match is obtained between an evidence sample and a databank entry, the match should be confirmed by testing with additional loci. [O]nly the statistical frequency associated with the additional loci should be presented at trial (to prevent the selection bias that is inherent in searching a databank)' (Ibid.). The second committee described another way to inflate the random-match probability, p. It recommended multiplying

p by *n*, the number of profiles searched (National Research Council Committee on DNA Forensic Science: An Update, 1996). In our coin-tossing database, the *np* rule gives an adjusted probability of $300\,000 \times 1/1\,048\,576 = 0.27$. As stated above, about a quarter of all the innocent databases would contain the suspect profile.

These recommendations were not well received in the literature. The adjustments were criticized as needlessly conservative. In a series of papers, David Balding, Peter Donnelly, Phil Dawid and others showed that because the effect of the database search is not merely to identify a single suspect, but to exclude the rest of the database as potential suspects, the match actually is at least slightly more probative than a single-suspect match (for citations and a simple proof, see Walsh & Buckleton, 2005). Adopting a Bayesian framework, all these statisticians concluded that there is no need to adjust the random-match probability.

Roughly speaking, their argument goes like this. The crime scene DNA originated from someone, but we do not know who this individual is. Before the DNA information is considered, the probability for being the source can be divided up somehow among everyone in the world. Most people will get almost none of the probability mass, but some will get more, depending on what is known about them. For example, people living in the same geographical area might get larger shares of the total probability. The details are not crucial to the argument. The claim is simply that, in principle, a prior probability for everyone could be enumerated. Now, for every individual, there also is a conditional probability that the DNA from the suspect would match assuming that this individual is the actual source. This likelihood will be one (ignoring laboratory or handling error or fraud) when the individual is the suspect. For all individuals who are not very closely related to the suspect, it will be close to zero. So we have a prior probability function and a likelihood function defined over all individuals. These functions can be combined according to Bayes' formula to give the posterior probability that the suspect who matches is the source of the crime scene stain. A single-suspect DNA match shifts a great deal of the probability mass from the world at large to the suspect (for details, see Balding, 2005).

Now, the Bayesian analysis compares this shift in the probabilities to that which results from a database search. The trawl provides more, not less, information. We know not only that the suspect matches but also that everyone else in the database does not match. These individuals can be eliminated entirely, shifting slightly more of the probability mass to the suspect (along with the untested remainder of the population). The frequentist analysis ignores this 'elimination effect' because it defines the null hypothesis to be the claim that the database is innocent—it does not contain the true source of the crime scene DNA. Thus, the frequentist argument treats the name of the matching individual as irrelevant, but the Bayesian analysis focuses on the hypothesis that the named individual is the source. It concludes that as to this hypothesis, the database trawl is more informative than the single-suspect match, and therefore, the random-match probability for the single-suspect match is not unfair to the defendant in a trawl case.

Walsh & Buckleton (2005, pp. 464–465), two statisticians from the Bayesian camp, have observed that '[t]he mathematical arguments given by [each] side appear impressive' and 'there is [so] much confused writing on the subject that it [is] very difficult for a court to make a reasoned decision based on a simple assessment of the literature recommendations'. Although it has been said that '[t]he discussion about these approaches now has somewhat died out (at least in scientific journals), with most scientists in the field preferring the [no-adjustment] approach' (Storvik & Egeland, 2007, p. 922), *Nelson* shows that the debate is very much alive in the courtroom and that an adequately reasoned opinion in that forum remains elusive.

2. The Court of Appeal opinion

On appeal to California's intermediate Court of Appeal, Nelson again contended that 'the DNA evidence should not have been introduced because . . . there is no generally accepted statistical method for explaining the significance of DNA evidence when a suspect is identified through use of a convicted offender databank'. 48 Cal.Rptr.3d 399, 411 (Ct. App. 2006). The Court of Appeal rejected the argument on two grounds. First, it insisted that '[t]he DNA databank search merely identified defendant as a possible candidate as the murderer; it was not the basis for declaring that his DNA matched DNA on the evidentiary samples. The latter determination was made based upon further, complete testing utilizing scientific techniques found to be reliable and admissible under the [general-acceptance] test' (Ibid. at 402). The 'further, complete testing' involved DNA taken from the defendant with a warrant procured as a direct result of the database trawl.

This is a foolish argument. Reliability is important, but repeating the DNA test on a fresh sample of a suspect's DNA is not responsive to the concern about trawling. The argument for adjustment is that it is less surprising to find a match by trawling than by doing a single DNA test at the outset; hence, a cold hit is less of a reason to reject the null hypothesis that the database is innocent. This effect, if it exists, does not disappear just because a second test verifies that the suspect matches. Given the knowledge of Nelson's profile already obtained from the trawl, this result is not a surprise. It is exactly what would be expected when the null hypothesis that the database is composed of all innocent people (and that the match to Nelson was coincidental) is true (or, for that matter, false). Hence, it is not evidence against that hypothesis.

The Court of Appeal's other basis for affirming the conviction was that 'use of the product rule to compute a random-match probability is overwhelmingly accepted by the scientific community as a scientifically reliable means of demonstrating the rarity of a profile in the population. Consequently, the technique satisfies the reliability prong of the [general-acceptance] test and it was for the trial court, not the scientific community, to determine the relevance of the technique to this criminal prosecution' (Ibid. at 418). As we soon shall see, this argument is decidedly better than the court's reliance on replication, but it is incomplete.

As an aside, the court also disparaged the Bayesian solution to the trawl problem:

> Another approach that has been suggested is the creation of likelihood ratios through the use of a Bayesian formula. 'Bayesian' refers to the Reverend Thomas Bayes who, in the nineteenth century, created a formula that purports to show the effect of new information on a prior probability. (See McCormick on Evidence (5th ed. 1999) §211, pp. 817–822.) Use of a Bayesian formula requires a quantified prior probability and quantifiable new information. (Ibid.) Bayesian analysis then utilizes a complicated formula to revise the prior probability on the basis of the new information. (Ibid.) Bayesian techniques are inherently confusing and would be difficult, if not impossible, to explain to an average jury. (Ibid.; see also 3 Forensic Sciences (2006 Matthew Bender & Co.) §30.03 et seq.) The end result of a Bayesian analysis is often misleading. (McCormick on Evidence, supra, §211, at p. 819.)

The dictum is peculiar from the statistical standpoint—the likelihood ratios are not derived from Bayes' rule, and neither the ratios nor the formula are especially complicated. It is also a disappointing use of legal authority. The chapter in McCormick (1999, §211) condemns a dubious application of Bayes' theorem in parentage testing cases. With regard to correct applications of the theorem,

I wrote that '[t]here is . . . a strong argument for using a Bayesian approach to help the jury evaluate the evidence' (Ibid.).

3. The Supreme Court opinion

The Supreme Court of California unanimously affirmed. Its opinion, written by Justice Ming Chin, steered clear of the fallacious replication theory or the usefulness of Bayes' rule in explaining DNA matches to juries. Instead, Justice Ming correctly reasoned that the debate in the scientific literature was over a quintessentially legal question—which statistic is relevant to the case—rather than any disputed matter of statistical genetics. Having reached this point, the court then maintained that the random-match probability is relevant as an estimate of 'the rarity of the DNA profile'.

But why is rarity relevant in a trawl case? At this point, the opinion unravels. It contains but a single, short paragraph on this, the core question:

> In a non-cold-hit case, we said that '[i]t is relevant for the jury to know that most persons of at least major portions of the general population could not have left the evidence samples'. (*People* v. *Wilson*, supra, 38 Cal.4th at p. 1245.) We agree with other courts that have considered the question (the Court of Appeal in this case; *People* v. *Johnson*, supra, 139 Cal.App.4th 1135; and *Jenkins*, supra, 887 A.2d 1013) that this remains true even when the suspect is first located through a database search. The database match probability ascertains the probability of a match from a given database. 'But the database is not on trial. Only the defendant is'. ([D.H. Kaye & George Sensabaugh, *DNA Typing*, in 4] Modern Scientific Evidence[: The Law and Science of Expert Testimony] §32:11, pp. 118–119 [David L. Faigman et al. eds. 2006].) Thus, the question of how probable it is that the defendant, not the database, is the source of the crime scene DNA remains relevant. (Id. at p. 119.) The rarity statistic addresses this question.

These remarks on trawling in the *Modern Scientific Evidence* treatise clearly distinguish between the hypothesis that *someone* in the database is the source and the hypothesis that the *named individual* is. The treatise states that the adjusted probability np is relevant when considering the former, but not the latter hypothesis. Since the jury is interested in the hypothesis about the named individual, it concludes that 'asking how probable a search of a particular database is to generate a cold hit on anyone in the database is the wrong question' (Kaye & Sensabaugh, 2006, p. 119).

In agreeing that 'the database is not on trial', the *Nelson* court seemed to be accepting this analysis. Yet, the court turned around and suggested that the database-match probability also is relevant. It wrote that: 'The database match probability statistic might also be admissible. [I]f the database were large enough and the odds shorter than those here, the database match probability statistic might also be probative. Nothing we say prohibits its admission'.

This question of the admissibility of np at the behest of the defense is likely to arise in future cases. As soon as the opinion was issued, 'Deputy Los Angeles Public Defender Jennifer Friedman . . . said the ruling was better than the case law that preceded it because now the prosecution "will have a difficult time preventing the defense from using" the database probability' (Dolan & Felch, 2008).

4. A narrower opinion?

The opinion in *Nelson* is broader than it needed to be. The case could have been decided easily without ruling that *p* is generally admissible. The defendant did not challenge the many zeroes in the random-match probability. Having rejected the defense argument about general acceptance, the court could simply have observed that the choice between *np* and *p* could not have affected the outcome of the case. So what if the adjusted probability is $np \approx 10^{-19}$ rather than the prosecution's figure of $p \approx 10^{-24}$? The court could have disposed of the case without reaching the question of whether the random-match probability is admissible in a trawl case, but it endorsed the 10^{-24} figure anyway.[1]

Perhaps, the court chose to reach out in this way because it wanted to avoid future litigation over the statistic with 'odds shorter than those here'. A cold-hit case of this sort had just received notoriety in the state from enterprising reporters for the *Los Angeles Times*. Not long before the oral argument in *Nelson*, Felch & Dolan (2008) had proclaimed that:

> Prosecutors and crime labs across the country routinely use numbers that exaggerate the significance of DNA matches in 'cold hit' cases, in which a suspect is identified through a database search.

> Jurors are often told that the odds of a coincidental match are hundreds of thousands of times more remote than they actually are, according to a review of scientific literature and interviews with leading authorities in the field.

The case that was said to epitomize this 'national problem' involved an adjusted match probability of about 1/3, which the defendant was not allowed to present. Diana Sylvester, a '22-year-old San Francisco nurse had been sexually assaulted and stabbed in the heart' in her San Francisco apartment over 30 years ago. A DNA database match from a highly degraded semen sample led investigators to 'John Puckett, an obese, wheelchair-bound 70-year-old with a history of rape'. The jury heard that the random-match probability for the match at five or so loci was about one in 1.1 million. It did not learn that the California database had 338 000 profiles in it, making *np* almost one in three— a number that would render the match almost worthless to the prosecution (and that the reporters mischaracterized as 'the probability that the database search had hit upon an innocent person').

5. Other *np*-like rules

If logic were the life of the law, the *np* statistic would not be permitted. The figure of $np = 1/3$ in *People* v. *Puckett*, for instance, is an estimate of the probability that a database of profiles of $n = 338\,000$ individuals would yield a hit to someone (not necessarily Puckett) if it were composed exclusively of individuals who are not the source of the crime scene DNA (and who are not identical twins of the true source). Unlike the random-match probability of $p = 1/1\,100\,000$, this number is not part of a likelihood ratio that is of interest to the jury. The legal issue, as the Supreme Court stated in *Nelson*, is not whether the database is innocent, but only whether the one defendant named Puckett is guilty or innocent. The likelihood ratio for the match with respect to Puckett as compared

---

[1] Another way to handle cases like *Nelson* would be to have the expert testify that the 15-locus DNA profile is very likely unique to the defendant (and any identical twin he might have). If the jury agrees, then why worry about how the defendant came to be identified?

to a randomly selected individual is closer to 1 100 000 than to three (Kaye, 2008). Thus, it is hard to see how the 1/3 figure is of much benefit to a juror seeking a reasonable explanation of the probative force of the evidence.

Nevertheless, the innocent database-match probability is not completely irrelevant. After all, if the database is innocent, then so is the individual who became a suspect because his profile was in it. Although each side in the statistical debate over adjustment claims that only its favourite probability should be admissible, perhaps the introduction of the innocent database probability on top of the random-match probability can be rationalized as a Solomonic sort of compromise, giving each side half of what it wants and trusting that their further statements will clarify the matter for the jury. This let-it-all-in ruling would allow individuals who were charged as a result of a database trawl to introduce testimony about the chance that an innocent database would have within it a profile that matches the crime scene DNA. Although introducing both the adjusted and the unadjusted number seems much more confusing than edifying, it might well appeal to judges who, understandably, find it hard to decide whether the frequentist perspective is appropriate here.

Even if defendants are permitted to introduce an adjusted probability, however, this adjustment should not be made with the nominal size n of the database. Instead, a smaller, 'effective database size', n', should be used.[2] The frequentist perspective suggests that the jury needs to know how often innocent defendants would be charged if the null hypothesis were always true. If individuals who clearly could not have committed the crime will not be charged, then there is no risk that they will be falsely convicted. Therefore, to the extent that non-viable suspects are included in the database, np does not provide the risk of a false decision even when the null hypothesis is true. Instead of multiplying by n, we should be multiplying by the number n' of potentially realistic suspects within the database.

With convicted offender databases, n' can be substantially smaller than n because these databases include samples from offenders who are behind bars and could not realistically be the perpetrators of new crimes in the outside world. They may even include individuals who are dead but whose records have not been purged. When evidence from a decades-old crime is tested, young offenders who could not possibly have committed the crime will be in the database.

In *People* v. *Puckett*, for example, many of the 338 000 individuals in the 2004 database were not even alive when Diana Sylvester was killed over 30 years ago. Had the matching software been designed to skip over their profiles, the outcome would have been no more (or less) probative, but the adjusted np probability would have been smaller, indicating a more probative match. Nothing useful is accomplished by complicating the database searches so that they exclude impossible suspects before the trawl or look first at the most plausible suspects.[3] If the defendant is allowed to introduce an

---

[2] I am grateful to John Hartmann for raising this point.

[3] Anticipating that some defendants will introduce the np figure to moderate the prosecution's p, the state might be tempted to adopt a more complicated search strategy. It could arrange the database in order of decreasing suspicion. The least suspicious people in the database would be those who were incarcerated, not yet born or very young at the time that the crime was committed. They would be at the bottom of the list. Those living in the locale of the crime and being of the same sex, age, and race of the perpetrator (when such information is available) would be at the top of the list. Instead of searching the entire list, the state could confine (at least initially) the trawl to the most likely candidates. Rather than trawl all 338 000 profiles in *Puckett*, for instance, California might have trawled only (let us say) 11 000 profiles and stopped if it obtained a match in that group of prime suspects. In this situation, np would have been $11/1100 = 1/100$. Or, it could have gone down the whole, ordered list until it obtained a cold hit, then stopped. If the hit occurred at profile number 110, for instance, it might seem that the adjusted p would be $110/1\,100\,000 = 1/1000$. However, this method of adjusting p would not be correct if the

adjusted random-match probability, an $n'p$ statistic is more appropriate than the np figure discussed in *Nelson*.[4]

Although the relevance of even the $n'p$ figure is doubtful, it is loosely related to another approach that places the random-match probability p in perspective. A defendant might try to argue that even though the DNA profile in question is not likely to be shared by many other unrelated people, it could be present in at least a few other individuals in a large population of potential suspects. Instead of asking how many other people in the database could be considered viable suspects, we could ask how many people in world fall into this category. Perhaps, the entire male population of the San Francisco Bay area between certain ages approximates the size of the initial population of potential suspects. If we call this number n^*, then n^*p is an estimate of the number of unrelated individuals within this population who share the DNA profile taken from the victim's body. Census data suggest that there were over 2 million men between the ages of 18 and 64 years in the Bay area in 2000 (Metropolitan Transportation Commission and Association of Bay Area Governments, 2003). Since p is about $1/(1.1$ million$)$, n^*p is approximately two, indicating that Puckett is not only man in the region who would have the requisite DNA profile. Refinements on this logic are possible, including an explicitly Bayesian presentation (Kaye *et al.*, 2004; Kaye, 2008). Whatever reasonable variant the defense might use, it remains open to the prosecution to argue that the defendant is much more likely than the other possible members of the suspect population—as whittled down by the DNA evidence—to be the source. But if the prosecution has no evidence or reasonable arguments that can distinguish the defendant from the genetic cohorts, then the jury lacks a posterior probability large enough to warrant a conviction.

Acknowledgements

REFERENCES

BALDING, D. J. (2005). *Weight-of-Evidence for Forensic DNA Profiles*. Chichester, UK: John Wiley and Sons, Ltd.

COOK, T. D. (2002). "P-Value Adjustment in Sequential Clinical Trials." *Biometrics* **58**, 1005–1111.

DOLAN, M. & FELCH, J. (2008). "California Supreme Court Ruling Allows 'Rarity' Statistic in DNA Cases." *Los Angeles Times*: June 17, available at http://www.latimes.com/news/science/la-me-dna17-2008jun17,0,3313471.story

FELCH, J. & DOLAN, M. (2008). "DNA Matches Aren't Always a Lock." *Los Angeles Times*: May 3, available at http://www.latimes.com/news/local/la-me-dna4-2008may04,0,6156934,full.story

stopping rule would have allowed more profiles to be compared had the cold hit not occurred at that point. See Siegmund, 1985; Cook, 2002 (discussing the effect of stopping rules in clinical trials in medicine).

[4] Cf. Storvik & Egeland (2007) (describing but not endorsing a substitute for the np rule that incorporates the size of a realistic population of potential suspects).

KAYE, D. H. (2004). "Logical Relevance: Problems with the Reference Population and DNA Mixtures in *People* v. *Pizarro.*" *Law, Probability and Risk* 3, 211–220.

KAYE, D. H. (2008). "Rounding Up the Usual Suspects: A Logical and Legal Analysis of DNA Trawling Cases." *North Carolina Law Review*: in press. Prepublication draft available at SSRN: http://ssrn.com/abstract=1134205

KAYE, D. H., BERNSTEIN, D. E. & MNOOKIN, J. L. (2004). *The New Wigmore, A Treatise on Evidence: Expert Evidence.* New York, NY: Aspen.

KAYE, D. H. & SENSABAUGH, G. (2006). "DNA Typing." In David L. Faigman *et al.*, eds., *Modern Scientific Evidence: The Law and Science of Expert Testimony.* Eagan, MN: Thompson-West. vol. 4 §32:11, 110–119.

MCCORMICK, C. T. (1999). *Handbook of the Law of Evidence.* John Strong, ed. St Paul, MN: West Publishing, 5th edition.

METROPOLITAN TRANSPORTATION COMMISSION AND ASSOCIATION OF BAY AREA GOVERNMENTS. (2003). *Census Data,* available at http://www.bayareacensus.ca.gov/bayarea.htm.

NATIONAL RESEARCH COUNCIL COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE. (1992). *DNA Technology in Forensic Science.* Washington, DC: National Academies Press.

NATIONAL RESEARCH COUNCIL COMMITTEE ON DNA FORENSIC SCIENCE: AN UPDATE. (1996). *The Evaluation of Forensic DNA Evidence.* Washington, DC: National Academies Press.

People v. *Nelson* (2006) 78 Cal.Rptr.3d 69 (Cal. 2008), *aff'g*, 48 Cal.Rptr.3d 399 (Ct. App. 2006).

People v. *Jenkins* (2005) 887 A.2d 1013 (D.C. 2005).

SIEGMUND, D. O. (1985). *Sequential Analysis: Tests and Confidence Intervals.* New York, NY: Springer-Verlag.

STORVIK, G. & EGELAND, T. (2007). "The DNA Database Search Controversy Revisited: Bridging the Bayesian-Frequentist Gap." *Biometrics* 63, 922–925.

WALSH, S. J. & BUCKLETON, J. S. (2005). "DNA Intelligence Databases." In John S. Buckleton, Chris M. Triggs and Simon J. Walsh, eds. *Forensic DNA Evidence Interpretation.* Boca Raton, FL: CRC Press, 439–469.