The DNA Database Search Controversy

David J. Balding

Department of Applied Statistics, University of Reading,
P.O. Box 240, Earley Gate, Reading RG6 6FN, U.K.
Current address: Department of Epidemiology and Public Health, Imperial College,
St Mary's Campus, Norfolk Place, London W2 1PG, U.K.
e-mail: d.balding@ic.ac.uk

SUMMARY. A recent article in Biometrics (Stockmarr, 1999, 55, 671–677) has generated correspondence (56, 1274–1277; 57, 976–980) reigniting a controversy started by a 1996 report on DNA profile evidence issued by the U.S. National Research Council (NRC). The issue concerns the evidential weight of a DNA profile match when the match results from a search through a profile database. The views of both Stockmarr and the NRC report conflict with those of many statisticians working in the area, and the differing viewpoints lead to dramatically different assessments of evidence. I outline reasons why Stockmarr and the NRC report are wrong. I also briefly discuss possible reasons why forensic applications tend to be problematic for statisticians.

1. Introduction

There are few settings outside the medical arena in which statistical issues have a high public profile. One of these is the census undercount in the United States (cf., Anderson and Fienberg, 2000). Since the introduction of forensic DNA profiling in the late 1980s, statistical issues arising in connection with the presentation of this evidence in court have provided another focus for both public attention and statistical controversy.

In part because of improved technology but also because of a consensus gradually developing following experience and debate, many of the statistical issues are now largely resolved. One of the areas where the consensus remains incomplete concerns the effect of database searches on evidential strength.

Consider two scenarios in which a DNA profile has been recovered from a crime scene and is presumed to have come from the culprit.

Scenario 1. You are told that the defendant's DNA profile matches the crime-scene profile.

Scenario 2. You are told that the defendant's DNA profile matches the crime-scene profile and that the suspect was identified by virtue of his profile providing the only match in a large database of DNA profiles.

What is the effect of the additional information about the database search on the weight of the DNA evidence? Is the DNA evidence stronger in scenario 1 than in scenario 2? Weaker? About the same?

To focus on this issue without undue distraction, I make various simplifying assumptions in the discussion below. These include that the alleged crime actually occurred and had a single perpetrator, whose DNA was recovered from the crime scene, and that neither profiling nor data-handling errors have occurred. I will also ignore here relatedness among the possible culprits. None of these assumptions is strictly valid in general, but deviations from them are, I believe, of little relevance to the present discussion.

2. The Effects of Hypothesis Trawls

Statisticians have well-honed instincts when it comes to hypothesis trawls—we are taught to be cautious of them because of problems of multiple testing. Consequently, many statisticians instinctively feel that the database search weakens the evidence, i.e., the case against the defendant is weaker in scenario 2 than in scenario 1.

Such instincts may serve us well in other settings, but they are misplaced in the present context. A crucial distinction is that we know in advance that exactly one hypothesis of the form “X is the culprit” is true. Our problem is to identify the true one among a (possibly very large) number of such hypotheses, which may include “The defendant's brother Fred is the culprit,” “Joe Bloggs is the culprit,” “The Prime Minister is the culprit,” and so forth.

Our two scenarios can thus be characterized as follows:

Scenario 1. Evidence is presented that a particular one of the hypotheses is true.

Scenario 2. Evidence is presented that a particular one of the hypotheses is true and that many of the other hypotheses are false.

Expressed in these terms, it becomes clear that the additional evidence strengthens the case against the defendant. Informally, if you know that there exists at least one member of a population satisfying your search criterion and you are concerned with whether or not it is unique, then the bigger the
search resulting in a single match, the more you have reason to be convinced that the observed match is unique in the population.

In practice, the magnitude of the effect of the additional evidence is usually modest because relatively few of the alternative possible culprits are included in the database. However, if the database includes most of the possible culprits, then a unique matching profile in the database provides overwhelming evidence against the defendant because (under our simplifying assumptions) the search result eliminates most of the alternative hypotheses. Moreover, intermediate situations imply intermediate conclusions, i.e., every alternative possible culprit eliminated from suspicion because his DNA profile is found not to match implies a (slightly) stronger case against the unique individual found to match. Of course if more than one individual is found to match, we are in a different situation, but one that is not under discussion here.

At this point, a statistician may well object: If the database size is large, then a match may be reasonably likely even under the hypothesis that everyone in the database is innocent, and hence the DNA evidence cannot reasonably be regarded as strongly incriminating against the matching individual. This objection seems consonant with the standard statistical reasoning that, if the data (a match in the database) is likely under a hypothesis (innocence), then the hypothesis cannot be dismissed. However, the data here is not a match but “the defendant matched and other particular persons did not match.” More generally, the role of the statistician is to assist jurors in assessing the unknown event that one or more potential suspects outside the database also has a matching DNA profile. If the match in the database was likely, then a match among possible culprits outside the database may well also be likely. But the probability of the match in the database per se is of no direct relevance to evidential weight.

To illustrate the above argument, consider a setting in which there are 1 million and 1 possible culprits, of which 1 million are included in the database and one is not. Suppose also that each innocent individual has, independently, probability 1 in 1 million of having a DNA profile matching the crime-scene profile. Then under the assumption that all 1 million individuals in the database are innocent, the expected number of matches among them is one. However, there remains only one unexamined possible culprit who is extremely unlikely (1 in 1 million) to also have a matching profile. Thus, the unique matching profile in the database provides overwhelming evidence against the matching individual, even though a match was expected under the hypothesis of innocence. On the other hand, if the database is not examined but the individual outside the database is profiled and found to match, then the case against this matching individual, absent further evidence, is not overwhelming because we expect that another matching individual is included in the database.

Readers previously unfamiliar with the topic may still be feeling a sense of unease. In scenario 1, we were not told why the authorities decided to profile the individual who subsequently became the defendant. No doubt they had grounds for suspicion that led them to this individual, whereas the authorities in scenario 2 presumably had no such grounds and had to resort to a database trawl. Isn’t the overall case in scenario 1—prior suspicion backed up by a DNA profile match—much stronger than in a case of no a priori reasons for suspicion?

Broadly speaking, such reasoning is valid. However, it does not bear on the weight of the DNA evidence. It is important to keep this distinction clear, among other reasons to avoid falling into the trap of double counting of evidence. If the authorities have valid grounds for suspicion in scenario 1, then this is presumably based on information that may be presented as evidence at trial and assessed in conjunction with the DNA evidence. If the original grounds for suspicion are not based on admissible evidence, then it is not the role of either experts or jurors to speculate on the existence or strength of such information.

Not only do I regard the logical separation of scientific from other evidence as proper, I do not believe that it will somehow inadvertently lead to misinterpretation or error by jurors. Based on my experience of giving expert witness evidence in criminal cases, I do not believe that there is any need for experts to subtly usurp some of the tasks of jurors for fear that otherwise the jurors may become confused in the presence of complex evidence. Two key points that I believe are crucial for courts to convey to jurors are

1. that, to be convinced of a defendant’s guilt, they must as a matter of logic be simultaneously convinced that every other individual on earth—including each of the defendant’s close relatives—is innocent, and
2. that the scientific evidence must be combined with the nonscientific evidence to come to an overall view on guilt or innocence.

Point 1 in effect overcomes the potential for multiple testing problems. The combining of the evidence under point 2 is a matter for jurors, not for any expert.

3. The np Rule

3.1 The NRC Report

The viewpoint that I have sketched above is now in broad terms shared by most statisticians working in the field (Robertson and Vignaux, 1995, Chapter 7; Balding and Donnelly, 1996; Dawid and Mortera, 1996; Weir, 1996; Taroni and Aitken, 1997; Evett and Weir, 1998, Chapter 9; Donnelly and Friedman, 1999; Dawid, 2000). Had history followed a different course, the effect of database searches on identification evidence might now be regarded as merely an interesting statistical phenomenon suitable for textbook examples.

The situation today is otherwise because of the publication in 1996 of the second report into DNA profiling evidence of the U.S. NRC. Its Recommendation 5.1 is that the match probability in the database search setting should be calculated as np, where n is the size of the database and p is the match probability that would apply in a no-search setting. Thus, the NRC recommendation implies that the evidence is n times weaker in the database search setting.

The NRC committee motivated their recommendation by an analogy with repeated coin tossing: 20 heads from a single toss of 20 coins is very surprising, but it is much less surprising to observe this outcome once in many throws of the 20 coins. The inappropriateness of this analogy is manifest. In particular, there is no feature corresponding to the fact that we know in advance that exactly one of the possible culprits is the source of the crime scene DNA. Moreover there is no element of repetition in the database search problem—differ-
ent individuals have their profiles compared with the crime scene profile, which cannot reasonably be compared with repeated tossing of a single set of coins.

In addition to this motivating analogy, the NRC report gives a number of arguments for its database search recommendation; see Donnelly and Friedman (1999) for a clear exposition of the flaws in each argument. The report cited no support for its recommendation from scientific literature or external expert opinion. The views of statisticians working in the field were not mentioned, although the committee behind the NRC report included two statisticians who were aware of these views.

3.2 Stockmarr (1999)

With the exception of a brief discussion by Morton (1997), the lack of support in the scientific literature for the NRC recommendation on database searches persisted until 1999 when Stockmarr (1999) published a supportive analysis in Biometrics. Briefly, Stockmarr takes the view that courts cannot reasonably address the hypothesis that the defendant is guilty because that is a data-dependent hypothesis that was not explicit when the authorities commenced their search for suspects. He chooses to address instead the hypothesis that the culprit is one of the individuals whose DNA profiles are recorded in the database. Contrasting this with the hypothesis that some other unrelated individual is the culprit, he derives a likelihood ratio (LR) of

\[ LR = \frac{P(1 \text{ match in database} \mid \text{culprit not in database})}{P(1 \text{ match in database} \mid \text{culprit in database})} = np, \]

in accord with the NRC recommendation.

Stockmarr acknowledges that this LR suggests that evidence weakens as the database size increases, which conflicts with the fact that, when the database is large enough to include most or all possible offenders, the evidence must be overwhelmingly strong. Another flaw subsequently pointed out by Donnelly and Friedman (1999) is that \( np \) can exceed unity, implying that the evidence favors the defendant. Thus, if the authorities profile a large number of people and the profile is not extremely rare, then, according to Stockmarr and the NRC report, a single match may well amount to evidence in favor of the innocence of the matching individual!

Stockmarr also offers another analysis appropriate for what he calls the “finite population” case, in which the \( n \) individuals whose profiles are stored in the database are assumed to be drawn uniformly randomly from a population of \( N \) possible culprits. Under these assumptions, he calculates posterior odds on the defendant’s innocence of \((N-n)p\). Stockmarr seems untroubled by the fact that, if the population is modeled as infinite, then, under his chosen measure, evidential strength declines with \( n \), yet if the population is modeled as very large but finite, then evidential strength increases with \( n \).

Further discussion of Stockmarr’s finite population analysis need not detain us here. It can have little practical role in the courtroom since it requires that the expert witness assumes

- knowledge of \( N \), the number of possible culprits,
- that everyone in the database is a possible culprit and each is \textit{a priori} equally likely to be the true culprit, and
- that the database has been drawn uniformly randomly from this population.

These assumptions are both implausible and inappropriate for an expert witness.

Stockmarr’s assumptions and interpretations have been criticized effectively elsewhere (Donnelly and Friedman, 1999; Evett, Foreman, and Weir, 2000a; Dawid, 2001). In particular, the way out of Stockmarr’s dilemma concerning data-dependent hypotheses has been well mapped. When the authorities commence their search for the culprit, they have before them a large number of hypotheses of the form “\( X \) is the culprit.” The authorities gather evidence until it is decided that sufficient evidence has been muster to persuade a jury that a specific one of those hypotheses is true and, equivalently, that each one of the other hypotheses is false. Viewed in this way, the problem of data-dependent hypotheses simply vanishes: All the hypotheses were there all along and all have to be assessed in conjunction to draw a valid conclusion about the guilt of the defendant.

3.3 Devlin (2000)

Devlin (2000) was invited by Biometrics to comment both on the Stockmarr paper and the response to it by Evett, Foreman, and Weir (2000a). He gave some support to both Stockmarr and the NRC recommendation, but his position was seriously undermined by his attribution of the “Prosecutor’s Fallacy” to his opponents (Evett, Foreman, and Weir, 2000b). The Fallacy, which amounts to confusing \( P(A \mid B) \) with \( P(B \mid A) \), can be very misleading in the context of DNA identification evidence and has led to successful appeals in U.K. courts. Even if we set aside this important and elementary misunderstanding, Devlin’s argument is flawed. He is concerned that forensic match probabilities are based on only a handful of the variable loci in the human genome and that there is thus a “(potentially) random draw of loci.” Of course, had additional DNA loci been typed but those favoring the defendant withheld from the court, this would constitute gross dishonesty, but this is distinct from the problem that Devlin raises and has no bearing on the issues under discussion here.

There is presumably almost always a random element in the lines of enquiry that forensic authorities choose to pursue. Assuming that no evidence has been suppressed, this raises no problem of logic: Potential sources of evidence that might have been investigated but were not are irrelevant to the weight of the evidence actually observed. Moreover, forensic authorities invest substantially in testing the reliability of their typing systems and hence in practice alter the loci investigated only very rarely. Even if there were a practical or logical problem to address, Devlin’s solution to it makes little sense, based as it is on the number of individuals in the database and not on the number of loci.

4. Discussion: Statisticians in the Courtroom

It is regrettable that, in the few areas in which statistical issues are prominent in the public eye, we seem to be noticed more for our disagreements than for our positive contributions to society. Perhaps this is inevitable since the need for statistical expertise is only likely to be recognized when complexities and subtleties are involved.

One noticeable feature of the controversy under discussion
is that Stockmarr and the statisticians behind the NRC report had little practical experience of DNA profile evidence in the courtroom prior to making their contributions to the debate. The criminal legal setting has some customs and requirements that appear to conflict with established statistical conventions. For example, the notion that evidence that has led to the identification of the suspect should not subsequently be used as evidence in court is analogous with some modes of statistical reasoning. But it is inconsistent with legal practice and would, I believe, be regarded as absurd by legal commentators.

A statistician may in some instances reasonably take the view that it is the legal practice that is flawed, not the statistical conventions. Stockmarr (1999) seems to take this view when he says “The decision problem of the court should take the implications of statistical hypotheses for data description into account, and not the other way around.” We statisticians would be wise to proceed more cautiously. Many legal systems, although not without fault, have evolved over centuries to fit their purpose. A wise statistician would also consider seriously the possibility that legal conventions may contain useful lessons for us. Donnelly and Friedman (1999) argue that legal systems should be critical consumers of statistical and scientific advice rather than accepting scientific conventions that may not be appropriate for the courtroom.

One useful message from the forensic database search controversy is that experience of the field of application and respect for its established practices are likely to be helpful. Another conclusion is that, although likelihood ratios are useful in assessing evidence, it is important to remember that an LR compares only two hypotheses, whereas in forensic identification settings, there are typically many competing hypotheses. In general, an LR comparing “the defendant is guilty” with “X is guilty” should be considered for every possible culprit X. Taking all the competing hypotheses into account overcomes Stockmarr’s difficulty about data-dependent hypothesis, Devlin’s worry about other evidence that might have been collected but was not, and the NRC’s concern related to multiple tosses of a set of coins.

The key issue for jurors confronted with DNA profile evidence concerns the possibility that one or more potential suspects whose DNA profiles are not available (such as those outside the database) also have a matching DNA profile. For a fixed number of possible culprits outside the database, the size of the database is essentially irrelevant to this question, as is the probability of a match in the database. To assess the probability of a match outside the database, jurors will typically need expert assistance concerning the match probability for the crime-scene profile. But they need no such assistance in their task of assessing the number of possible culprits outside the database, nor do they need an expert to confuse matters by multiplying the match probability by an irrelevant quantity, the database size.

RéSUMÉ
Un article récent dans Biometrics (Stockmarr, A., 55, 671–677) a entraîné une correspondance (56, 1274–1277; 57, 976–980) réactivant une controverse ayant débuté par un rapport de 1996 du National Research Council (NRC) américain sur l’évidence de preuve apporté par l’analyse de profils d’ADN. Le problème concerne le poids de la preuve apporté par un appariement d’un profil d’ADN, lorsque cet appariement résulte d’une recherche sur une base de données de profils d’ADN. Les avis d’à la fois Stockmarr et du rapport du NRC sont en contradiction avec ceux de nombreux statisticiens travaillant dans le domaine, et ces différents points de vue conduisent à des évaluations extrêmement différentes de l’évidence de preuve. Je souligne les raisons pour lesquelles Stockmarr et le rapport du NRC sont dans l’erreur. Je discute aussi brièvement les possibles raisons pour lesquelles les applications médico-légales tendent à être problématique pour les statisticiens.

REFERENCES
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, 976–980.

Accepted August 2001.