ERRORS AND MISUNDERSTANDINGS IN THE SECOND NRC REPORT

ERRORS AND MISUNDERSTANDINGS
IN THE SECOND NRC REPORT

David J. Balding*

ABSTRACT: Criticisms are presented of the second NRC report on DNA evidence. The underlying theme is that the report does not focus on evidential weight; consequently much of its discussion is tangential to the issues that matter in court, and in some cases the report is positively misleading. In particular, a recommendation concerning database searches and another concerning small groups or tribes are seriously flawed, erring in the former case in favor of defendants and in the latter case against defendants.

CITATION: David J. Balding, Errors and Misunderstandings in the Second NRC Report, 37 Jurimetrics J. 469-476 (1997).

Given the controversy over DNA evidence, and the hostile reception by many commentators of parts of its predecessor report, it was perhaps unlikely that the second NRC report on DNA evidence[1] ("NRC II") would be universally acclaimed. Aspects of the report have been widely welcomed: the so-called "ceiling principle," proposed by the earlier report, had few friends, and its rejection will provoke little, if any, adverse response. Other welcome features of the new report include the helpful introduction to the underlying technology, the recommendations to improve laboratory performance and the discussion of legal implications.[2] A particularly laudable feature of the report is recom

1. NATIONAL RESEARCH COUNCIL, COMMITTEE ON DNA FORENSIC SCIENCE: AN UPDATE, THE EVALUATION OF FORENSIC DNA EVIDENCE (1996) [hereinafter NRC II]. The predecessor report is NATIONAL RESEARCH COUNCIL, COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE, DNA TECHNOLOGY IN FORENSIC SCIENCE (1992).

2. See, e.g., B.S. Weir, The Second National Research Council Report on Forensic DNA

mendation 6.1, which encourages research on possibilities for misinterpretation of DNA profiling evidence and on methods of presenting expert testimony on DNA.

Nevertheless, many aspects of NRC II are disappointing. The problems stem from a failure to grapple adequately with the fundamental issues of the *interpretation* of DNA evidence: What do reasonable finders of fact need to assess, and how might a forensic scientist or other DNA expert best assist them? With some exceptions, the new report assumes an inappropriate hypothesis testing framework which, although common in scientific discourse, has little validity in legal settings.[3] Consequences of this flawed inferential framework permeate the report, resulting in errors and misleading statements on a number of key issues. In addition, addressing the wrong question leads the report to devote much discussion to relatively unimportant matters while some crucial issues receive scant attention.

I. DATABASE SEARCHES

Perhaps the most surprising feature of NRC II is its handling of the scenario in which the defendant has been identified on the basis of a "trawl" through a database of the DNA profiles of many individuals. The report takes the view that the effect of such a search is dramatically to reduce the evidential strength against an individual found to match. Recommendation 5.1 quantifies this reduction: evidence obtained from searching a database of size n is n times weaker than the same evidence obtained by other means.

Recommendation 5.1 seems to be in tune with a general principle in scientific research that data obtained via an extensive search are, broadly speaking, of little value. Consider the following analogy. Suppose that you present to me statistical evidence that appears to link a particular gene with a certain rare form of cancer. Initially, I am impressed by the evidence, but then I ask how these data came to your attention. You reply that you compared data on a large number of rare cancers with a large number of genes and then selected the

---

Evidence, 59 AM. J. HUM. GENET. 497 (1996). In addition to praising positive features of the report, Weir also alludes to some of the criticisms that I make in more detail below.

3. Alternative interpretations are introduced later, but the discussion of the scientific issues is firmly grounded in an hypothesis testing framework in which strength of evidence is viewed as being equivalent to the probability of the evidence under an hypothesis that the defendant has been chosen randomly in some population. This approach enjoys almost no support in the legal or related literature, and its flaws in this context are widely recognized. *See, e.g.,* I.J. Good, *Weight of Evidence and the Bayesian Likelihood Ratio, in* THE USE OF STATISTICS IN FORENSIC SCIENCE ch. 3 (Colin G.G. Aitken & David A. Stoney eds., 1991); BERNARD ROBERTSON & G.A. VIGNAUX, INTERPRETING EVIDENCE (1995). Among the problems in transplanting the hypothesis testing approach from its scientific environment to the legal setting is that the "randomly chosen" hypothesis is *not* equivalent to "the defendant is innocent." For a further discussion, see D.J. Balding & P. Donnelly, *Inferring Identity from DNA Profile Evidence,* 92 PROC. NAT'L ACAD. SCI. (USA) 11741 (1995).

cancer and gene pair that displayed the strongest association. I am now less impressed. Moreover, the more extensive was your search, the less impressed I am. The reason is that among many comparisons, it is unsurprising that one of them will, "by chance," display results that, on their own, might be regarded as significant. In this setting, the effect of the search is to weaken evidential strength. This phenomenon is widely understood by scientists. Indeed, conveying this idea to the general public is often regarded as a major challenge in advancing the public understanding of science.

Although accepted in other settings, this principle is misleading in the database search context. Imagine now that, in response to my question about how the data were obtained, you told me that previous studies had established that this cancer was controlled by a single gene and so you decided to search the 23 human chromosomes in increasing numerical order, looking for candidate locations of the gene.[4] After finding nothing significant on chromosomes 1 through $i-1$, you encountered the strong evidence of association with a locus on chromosome i. Since your research is expensive and you felt it highly likely that chromosome i was the correct location of the (unique) gene, you decided to terminate the search at that point.

This search scenario is different from the one previously outlined, primarily because it is known in advance that the target of the search exists. Consequently, the more you search, the fewer the possibilities for the correct location. If the candidate location is discovered on the first chromosome investigated ($i = 1$), your case for having identified the correct location is weaker than if, say, $i = 17$, in which case chromosomes 1 through 16 have been searched and excluded as possible locations. The effect of the search on evidential strength is thus reversed compared with the previous scenario: the longer the search the *stronger is* the evidence for the cancer-gene link.

Database searches for DNA profile matches are more like the second of these gene-hunting scenarios than the first, because it is (usually) known that there is a unique perpetrator of the offence. If individuals are successively searched and excluded from consideration as possible suspects, the effect of these exclusions is to increase the probability that an individual subsequently found to match is the culprit,[5] just as searching and excluding chromosomes increases the likelihood that a candidate link uncovered on a subsequent chromosome reflects the actual location of the gene.

The correct intuition becomes clear if one considers the extreme situation in which the database includes the DNA profile of everyone who might have committed the crime. If only one match is observed, and laboratory or other

---

4. This story is intended to illustrate a statistical point and does not accurately reflect how scientists actually search for disease genes.

5. In most cases, only a small proportion of the population of possible culprits is included in the database, so the increase in evidential strength is negligible. The crucial point is that evidence is not dramatically weakened, as recommendation 5.1 implies.

error is exceedingly unlikely, then the case is overwhelming. The logic of recommendation 5.1 is that the evidence is very weak because the number of individuals searched is so large. The same intuition applies, with less dramatic effect, in the case of databases that contain the DNA profiles of only a small proportion of the population of possible culprits.

NRC II acknowledges this counter-example but does not concede its effect of undermining recommendation 5.1. Instead, it suggests that, although initially evidence is weakened as the length of the search increases, at some point this effect is reversed and very extensive searches result in extremely strong evidence.[6] Readers are given no advice on how to determine the point at which the switch from ever weaker evidence to extremely strong evidence occurs. The absurdity of the report's position here clearly weakens its credibility on other issues.[7]

How does NRC II make such a dramatic error, in which not merely the magnitude of the effect, but its direction (i.e., weaker rather than stronger) is wrong? The reason seems to be a false analogy with tossing 20 coins and interpreting the outcome "all heads" as evidence that the coins are biased. The report asserts that "[t]he initial identification of a suspect through a search of a DNA database is analogous to performing the coin toss experiment many times,"[8] but it makes no attempt to explain this analogy. In database searches, each possible suspect is searched just once and not, like the coins, many times: there is no repetition involved. Moreover, we know in forensic settings that there exists a culprit, which is not the case for the coin tossing model. A valid analogy would be with many sets of 20 coins, among which one set is known to be biased. If some of the sets are tossed and precisely one of them produces "all heads," then the evidence that the biased set has been found increases with the number of sets tossed.

If, as I have suggested, the database search error was inspired by a false analogy, why was the error not corrected in the report's analyses (such as in Appendix 5B)? The reason is that, rather than basing analyses on the directly relevant question

> Given that a database search has been conducted, how strong is the evidence
> against the unique individual found to match?

the report addresses instead the question

---

6. NRC II, *supra* note 1, at 40 ("If the database searched includes a large proportion of the population, the analysis must take that into account. In the extreme case, a search of the whole population should, of course, provide a definitive answer.").

7. It is disappointing that the report chooses to ignore a body of literature discussing the correct analysis of the database search scenario. *See, e.g.,* A.P. Dawid & J. Mortera, *Coherent Analysis of Forensic Identification Evidence,* 58 J. ROYAL STAT. SOC'Y (Series B) 425 (1996), and references therein.

8. NRC II, *supra* note 1, at 134.

How likely is it that someone in the database would match if all were innocent?[9]

Not only is recommendation 5.1 based on flawed intuition and misconceived analyses that address the wrong question, but its implications are problematic for the effective use of DNA evidence. In a particular case, it will be difficult to establish the details of any, possibly informal, search that may have occurred as part of the investigation leading to the arrest of the defendant. The supposed evidence-weakening effect of database searches presumably also applies to other searches, so that fair prosecutions based on DNA evidence would become almost impossible if the logic of recommendation 5.1 were to be applied rigorously.[10]

II. POPULATION GENETICS

NRC II was written by a committee that included some of the most prominent and internationally-renowned population geneticists. Unfortunately, its impressive population genetics expertise was not used to best effect because of a failure, as in the database search setting discussed above, to address the directly relevant questions.

A court told that a defendant has the culprit's DNA type is naturally concerned with how many other individuals among the population of possible culprits might be expected also to share this DNA type. Answering this question is complicated by the phenomenon of genetic correlations due to shared ancestry. Although a particular DNA profile might be unlikely to occur, the observation of one copy of the profile makes it more likely that other copies exist. At the level of individual genes, this effect has traditionally been described by population geneticists in terms of a quantity often known as F_{ST}. For entire profiles, consisting of two genes at each of several loci, approximate match probabilities can be derived in terms of F_{ST}.[11]

Note that F_{ST} measures correlations *between* distinct individuals. Another genetic coefficient, known as F_{IT}, measures the correlation of the two genes *within* an individual at a locus. In some situations,[12] F_{IT} is likely to be similar in value to F_{ST}. NRC II devotes many pages to discussing within-person correlations, and in particular their effect on the profile frequency,[13] while

---

9. The report considers the probability of the event M, that at least one of the database profiles matches the evidence sample. *Id.* at 134. Appendix 5B does address strength of evidence rather than merely the probability of matching. However, it focuses on the event that someone in the database is guilty, rather than the event that the defendant is guilty.

10. For further discussion, see D.J. Balding & P. Donnelly, *Evaluating DNA Profile Evidence when the Suspect is Identified Through a Database Search*, 41 J. FORENSIC SCI. 603 (1996).

11. Single-locus match probabilities are given in NRC II as equations (4.10a) and (4.10b), but the recommendations concerning these probabilities are misleading, as discussed below.

12. NRC II, *supra* note 1, at 102-03.

13. *Id.* at 98-112.

almost ignoring the between-person correlations and their effect on match probabilities. Confusingly, the report inappropriately identifies profile frequencies with match probabilities: the two are the same only if between-person correlations are neglected.[14] In devoting so much attention to arriving at the conclusion that within-person correlations typically have little effect on profile frequencies, the report painstakingly builds up and knocks down a straw man, while leaving the central population genetics issues unaddressed.

In practical cases, the pool of possible culprits usually contains individuals with differing levels of ancestry shared with the defendant and therefore differing between-person correlations. The task is then to present to the court the plausible range of match probabilities in a helpful and fair way. Instead, the report recommends allowing for between-person correlations only in the extreme case in which every possible culprit is in the same subpopulation as the defendant.[15] In the general run of cases in which some, but not all, possible culprits have a relatively high level of ancestry shared with the defendant, the report recommends ignoring the between-person correlations due to shared ancestry.[16]

That between-person correlations matter when every possible culprit is in the same subpopulation as the defendant, but can be ignored when this applies to most, but not all, possible culprits makes little sense. In fact, simple numerical illustrations show that same-subpopulation possible culprits can dominate evidential weight even when they are only a small minority of all possible culprits,[17] so that the report's recommendations routinely will be very unfair to defendants.

Similarly, just one brother among the possible culprits can outweigh the effect of very many unrelated men for realistic values of the match probabilities. Consider a simple numerical illustration: the possible culprits are the defendant, one brother of the defendant and one thousand unrelated men. Only the DNA profile of the defendant is available. The match probabilities are 1/100 for the brother and 1/1 million for the other men. If a juror took the view that the weight of the non-DNA evidence was the same for all the alternative possible culprits, then in assessing the probability of the defendant's guilt the weight of the one brother (1/100) is much larger than the combined weight the 1,000 unrelated men (1,000/1 million = 1/1,000). The effect of relatives on the profile frequency may well be small, it is their contribution to evidential weight that matters, and this may be relatively large.

In contrast, NRC II says: "Because one or a few relatives in a large population will have only a very slight effect on [profile frequency], we believe

---

14. Effectively the same point is made by Weir, who uses the term "conditional frequencies" to refer to match probabilities. Weir, *supra* note 2, at 498.
15. NCR II, *supra* note 1, at 122 (recommendation 4.2).
16. *Id.* at 114.
17. *See* Balding & Donnelly, *supra* note 3.

that the importance of unknown relatives has been exaggerated."[18] Once again, the report misunderstands the issues and focuses on the wrong question—the profile frequency rather than evidential weight.

The most troubling consequence of the report's failure to address adequately population genetics issues is recommendation 4.3, which concerns the situation in which the defendant is from a group or tribe for which no adequate database exists. In such a scenario, between-person correlations are most likely to be important, yet the recommendation implies that they should be ignored. This recommendation will result in gross unfairness to defendants when individuals from the same group or tribe as the defendant are among the alternative possible culprits.

Between-person correlations do arise implicitly in the report at equation 5.1c.[19] However, instead of discussing appropriate models and plausible values for these correlations, the report assumes without discussion that they do not exist. No explanation is given for devoting so much attention to the relatively unimportant within-person correlations and their effect on profile frequencies, while the correlations that are crucial to match probabilities and hence the fair interpretation of DNA evidence are assumed away without discussion.

III. POSSIBLE LABORATORY OR HANDLING ERROR

The effect of possible laboratory or handling error on evidential weight is well understood in principle. The observed DNA profile match could have arisen in at least two ways consistent with innocence: (a) suspect and culprit happen to have matching DNA profiles and no error occurred; and (b) suspect and culprit have distinct DNA profiles, and the observation of matching profiles is due to an error in one or both recorded profiles. A reasonable juror must assess both of these possibilities on the basis of the information presented to them in court. Some immediate consequences of this observation are the following: (1) to achieve a satisfactory conviction based primarily on DNA evidence, the prosecution needs to persuade the jury that the relevant error probabilities are small; (2) if the probability of error (b) is much greater than the probability of matching profiles (a), then the latter probability is effectively irrelevant to evidential weight; and (3) what matters are not the probabilities of *any* profiling or handling errors, but only the probabilities of errors which could have led to the observed DNA profile match.

The report does not give guidelines on how these important points should be conveyed to juries. Instead, it undertakes another exercise of building up and demolishing straw men by discussing the question of whether or not experts should report a match probability which adds error rates to profile

---

18. NRC II, *supra* note 1, at 113.
19. This equation occurs on page 128, after the population genetics chapter.

frequencies.[20] Such a practice would clearly be unacceptable since overall error rates are not directly relevant: jurors must assess on the basis of the evidence presented to them the chance that an error has occurred in the particular case at hand.[21] Having rightly dismissed this practice, however, the discussion is terminated without guidance on how experts *should* convey to jurors a fair assessment of the evidence in view of the possibilities of both "chance match" and handling or laboratory error. Allowing defendants the possibility of retesting samples does, of course, provide some guarantee against possible laboratory error, but it cannot completely eliminate the need to assess possible handling errors before the samples are divided.

———————— ·••◉●••· ————————

Most of the flaws in NRC II stem from an inappropriate inferential frame-work. In the most widely-accepted theory of inference for forensic evidence,[22] strength of evidence is measured by its effect on the probability of guilt.[23] The role of database searches, population genetics, possible laboratory or handling error and other factors are then all assessed in terms of their effect on the issue that matters in court—whether or not the defendant is guilty. The discussion of these issues in the report, however, lacks this logical focus with the result that much of it is only tangential to evidential weight, and some of it is positively misguided.[24]

It is perhaps fortunate that the report's errors are not all in the same direction: the error involving database searches provides a substantial boon to some defendants, whereas the other errors and misleading statements tend to be unfair to defendants, sometimes substantially so. While these errors may in some sense "average out" over many cases, in particular cases substantial injustices will arise in one direction or the other. On balance, the report probably errs in favor of prosecutions more than defenses, which may mean that it attracts less adverse comment than its predecessor. Nevertheless, a thorough and fair study of the interpretation of DNA evidence is still awaited.

---

20. NRC II, *supra* note 1, at 85-87.

21. Error rates observed in blind trials may well be helpful to jurors.

22. See, for example, the three texts reviewed in Richard D. Friedman, *Assessing Evidence*, 94 MICH. L. REV. 1810 (1996).

23. In many cases, but not always, "is the source of the sample" is effectively equivalent to "is guilty."

24. The evaluation of probability of guilt is discussed in the report, for example, at pages 132-33 and 199-202. In fact, the report concedes that "what we would like to know and most easily interpret is the probability that the suspect contributed the DNA in the evidence sample." *Id.* at 132. However, this viewpoint does not underpin the scientific discussion in the preceding pages.