DNA database matches: A p versus np problem

Meester, R. W.J.; Slooten, K.

published in
Forensic Science International: Genetics
2020

DOI (link to publisher)
10.1016/j.fsigen.2019.102229

document version
Publisher's PDF, also known as Version of record

Forensic Science International: Genetics 46 (2020) 102229

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

DNA database matches: A *p* versus *np* problem

R.W.J. Meester[a], K. Slooten[a,b,*]

[a] VU University Amsterdam, The Netherlands
[b] Netherlands Forensic Institute, The Netherlands

ARTICLE INFO

Keywords:
DNA database
Likelihood ratio
Evidential value
Criminal casework
Database controversy

ABSTRACT

The evidential value of a unique DNA database match has been extensively discussed. In principle the matter has been mathematically resolved, since the posterior odds on the match being with the trace donor are unambiguously defined. There are multiple ways to express these odds as a product of likelihood ratio and prior odds, and so the mathematics do not immediately tell us what to do in concrete cases, in particular which likelihood ratio to choose for reporting. With p the random match probability for the matching person, if innocent, and n the database size, both $1/p$, originating from a suspect-centered framework, and $1/(np)$, originating from a database-centered framework, arise as likelihood ratio. Both have been defended and both have been criticized in the literature.

We will clarify the situation by not introducing models and choices of prior probabilities until they are needed. This allows to derive the posterior odds in their most general form, which applies whenever we know that a single person among a list is not excluded as potential trace donor. We show that we need only three probabilities, that pertain to the observed match, to the database, and to the matching person respectively.

How these required probabilities behave in a given context, then, differs from one situation to another. This is understandable since database searches may be done under various circumstances. They may be carried out with or without a suspect already in mind and, depending on the operational procedures, one may or may not be informed about the personal details of the person who gives the match. We show how to evaluate the required probabilities in all such cases.

We will motivate why we believe that for some database searches, the $1/p$ likelihood ratio is more natural, whereas for others, $1/(np)$ seems the more sensible choice. This is not motivated by the mathematics: mathematically, the approaches are equivalent. It is motivated by considering which model best reflects the actual situation, taking into account what question was asked to begin with, and by the practical consideration of judging which likelihood ratio comes closer to the posterior odds based on the information available in the case.

This article is intended to be both a research and a review article, and we end with an in-depth discussion of various arguments that have been brought forward in favor or against either $1/p$ or $1/(np)$.

1. Introduction, history, and context

The evidential value of a unique match in a database search has been a source of considerable debate that reached its peak around 15 years ago [1–8]. The controversy circled around the question whether or not a unique match of a trace profile constitutes weaker or stronger evidence compared to the so called probable cause scenario in which only one suspect is typed and found to match. When we compare a database search to a probable cause scenario, there are two opposite effects: on the one hand the database search could give rise to coincidental matches, but on the other hand the search result excludes a number of potential candidates as donor of the crime trace. So on the one hand, it could be felt that some correction ought to be put in place to account for the fact that many comparisons are done, but on the other hand, it can also be argued that the database result gives more evidence against the matching individual than if he had been the only person that was compared. The fact that other persons can no longer be the source of the trace is also to some extent evidence against the matching suspect.

Initially, a report of the National Research Council [1] advised to circumvent this (perceived) problem by simply not taking the match into account as evidence in court, but to only use it to be able to define the matching individual to be a suspect, thereby getting back into the probable cause case. Additional DNA testing would then supply

* Corresponding author at: Netherlands Forensic Institute, The Netherlands
E-mail address: k.slooten@nfi.minvenj.nl (K. Slooten).

https://doi.org/10.1016/j.fsigen.2019.102229
Received 2 September 2019; Received in revised form 18 December 2019; Accepted 18 December 2019
Available online 31 December 2019
1872-4973/

evidence that can be subjected to a probabilistic assessment yielding a statistic (random match probability or likelihood ratio) to be used in court.

Clearly, this is not an optimal solution. For one thing, it could very well be the case that the same single match would also have been obtained if fewer loci had been used in the search. Therefore, exactly which loci can be used as evidence depends on the database search settings, even if all data are precisely the same. From a practical point of view, in some cases it means that no DNA evidence can be used, for example if there is no trace material left to analyze for further DNA typing.

Several publications subsequently appeared concerning the issue of database matches. In [9] the use of likelihood ratios was advocated because of their optimality in (frequentist) decision making, but a Bonferroni correction was proposed for the match probability. This correction, proposed to be $1-(1-p)^n$, gives the probability of having at least one match by chance in the database, and is equal to about np for $np \ll 1$.

Other authors, using likelihood ratios as well but subsequently applying Bayes rule to obtain probabilistic assessments on hypotheses concerned with the guilt or innocence of the suspect, claimed that a database search strengthens the DNA evidence against the suspect. For example, [10] argues against the hypothesis testing framework for the evaluation of DNA evidence in general, and for database searches note that "In a wide range of settings, the DNA evidence is slightly stronger when it is obtained after a search", noting that the intuition behind that result is the fact that the search, contrary to a probable cause case, excludes other individuals as trace donors. In [6] this point of view is further elaborated.

In 1996, a second report of the National Research Council in the United States [2] appeared, which aligned with the frequentist intuition. The report stated that "If the only reason that the person becomes a suspect is that his DNA profile turned up in a database, the calculations must be modified", where the calculation that was referred to is the calculation of the random match probability, or of the likelihood ratio equal to its inverse. It was recommended (Recommendation 5.1) to multiply the random match probability with the number n of persons in the database "to describe the impact of the DNA evidence under the hypothesis that the source of the evidence sample is someone in the database". Even though the hypothesis that someone in the database left the trace is explicitly mentioned, it is not immediately clear to us from the phrasing of the report whether the number $1/(np)$ was intended to play a role in a Bayesian analysis, or rather was intended as a frequentist instrument to correct for the number of hypotheses tested, such as the Bonferroni correction. The report also mentioned that this correction was proposed as suitable for databases that contain only a small fraction of the whole population and that, if this were not the case, a more complicated analysis would be required without going into the details of what such an analysis should then be like.

This recommendation was criticized by many, because it goes against the Bayesian analysis that had previously been provided. To strengthen those arguments, absurd conclusions were derived assuming the evidence weakens in larger databases. One of the arguments put forward indeed was that if the database grows to the full population, clearly a unique match must identify the trace donor with certainty, whilst the evidential value as per NRC-II (ignoring that the report had stated the correction to be applicable only for relatively small databases) has the opposite effect because of the division by n.

In 1999, the NRC-II recommendation gained some statistical support when Stockmarr [4] provided a rationale for the evidential value of $1/(np)$. He showed that this number can be obtained as the likelihood ratio of a unique match when considering the hypotheses that the donor of the trace profile is in the database versus its negation.

Now that both $1/p$ and $1/(np)$ both can appear as likelihood ratio, a controversy was born: which one is considered to be applicable? It was rapidly argued by various authors [3,7] that the choice of the hypotheses, and hence of the ensuing likelihood ratio was *mathematically* unimportant in the sense that they lead to the same posterior odds. This last fact is understandable, since after finding a unique match, the hypothesis that the matching person is the trace donor is equivalent with the hypothesis that the donor is in the database. However, even if the choice of hypotheses is mathematically unimportant, it is another issue which likelihood ratio a forensic laboratory should *report*. Various authors have argued that only the hypotheses against the identified suspect are admissible. For example, in [3] it is argued that "Stockmarr makes a fundamental logical error when he suggests that the court can replace these hypotheses by H_p and H_d [hypotheses on the database containing the trace donor] and still use the resulting likelihood ratio as if it were directly relevant to the case against Smith [the identified suspect]." Similar arguments were used in [11,12] and more recently in [13].

In many jurisdictions a forensic laboratory has the task to assess the strength of evidence, and to communicate this to the investigating authorities or to fact finders such as judges or juries. It is our impression that a likelihood ratio of $1/p$ is usually chosen. However, the German Stain Commission issued a recommendation in favor or reporting a likelihood ratio of $1/(np)$ in [14].

Meanwhile, the forensic databases have grown considerably, and millions of comparisons are routinely made with crime stain profiles. The difference between $1/p$ and $1/(np)$ can therefore be, depending on p, of significant importance for the subsequent process or trial. It is well known that likelihood ratios are often erroneously interpreted as odds on the hypotheses, a mistake known as the prosecutor's fallacy, and the implications of such a fallacy also depend on whether $1/p$ or $1/(np)$ is presented as likelihood ratio. It is therefore perhaps not a surprise that the debate is still ongoing. In fact not only the debate is still ongoing, but also more generally there is discussion as to which are the relevant probabilities that influence the evidential value or the prior/posterior odds on the hypotheses.

Recently there has been renewed attention to the evidential value of unique database matches. In [13] the authors claim to derive "the" likelihood ratio for unique database matches in a new way, and in [15] it is argued that the size of the offender population can be estimated and should play a role in computing posterior odds for a unique database match. Both the implied unambiguity of the likelihood ratio in [13] and the claimed direct relevance of the size of the offender population in [15] are, we believe, incorrect.

These misunderstandings persist for several reasons. One reason is that, depending on the specifics of the search, one's intuition leans more toward $1/p$ as most natural value for the evidential value or more towards $1/(np)$. Both values can be perceived as being natural. Second, in the literature most (if not all) explicit calculations have been made assuming some ad hoc model for the population and typically uniform prior distributions for the trace donor in the population and for the database as a subset from the population. Moreover, the match probability p is usually taken to be the same for all persons in the database, whereas in reality this does not have to be the case since there is often more genetic information for some database members than for others. The uniform prior has some appeal in the sense that it facilitates computations, but it often is not realistic and care must be taken not to take conclusions for the uniform prior case as conclusions for the general case.

We will show that in this particular situation the general case is actually the simplest one, and treating it in full generality makes it clear where the aforementioned problems appear and also how they can be approached. In the current paper, therefore, we first present a surprisingly simple formula for the posterior odds upon a unique match which is valid in *all* circumstances, irrespective of the question whether or not the suspect became suspect as a result of the match, or whether there was already interest in him or her. Although in most papers on this subject uniform prior odds are assumed, we deviate from this habit for the reasons given above. With our general formula, it becomes clear

exactly which quantities are important for the calculation of posterior odds and likelihood ratios, and which are not.

The title of this article refers to the controversy around using either likelihood ratio $1/p$ or $1/(np)$ in database unique hits (and of course also to a much harder problem, suitably with capital letters[1]). We will show that sometimes, but not always, either number can be used, if properly understood, but that what we believe is the most natural choice may depend on the circumstances and on the question that was originally asked to the forensic laboratory.

We have set up the paper as a hybrid between an overview paper and a research paper. We aim both to give a full account of the main arguments brought forward in the database discussion, as well as to move the debate forward by presenting a treatment that is as general as possible, and apply that general framework to resolve all controversies.

In the next section we derive and discuss the basic formula. After that, we apply it to various special cases in Section 3, namely to the so called cold case situation, the targeted search situation, and the probable cause case. In Section 4 we discuss the various likelihood ratios that may be most appropriate to report. Which likelihood ratio is to be reported depends on the precise circumstances, and we will give a number of examples. In Section 5 we investigate the situation in which we receive less information than the full identity of the unique match. In Section 6 we treat a number of arguments that have been put forward in the course of the controversy, especially those where people claim that the $1/p$ or $1/(np)$ likelihood ratio would be wrong. We will see that the claims that have been made do not always hold up, and point out where they go wrong. We end with some conclusions in Section 7.

2. The likelihood ratio for a single non-exclusion

We start by considering the most general situation, because this will allow us to derive all further results as special cases, and because we believe it already sheds a lot of light on the problem. Consider a population $\mathcal{P}$ of individuals, and a DNA database $\mathcal{D} = \{d_1, \ldots, d_n\}$ of n DNA profiles of members of $\mathcal{P}$. The database is typically not homogeneous, in the sense that different loci may be typed for the different d_i: the amount of genetic information in d_i varies. The reason for this is that historically the DNA typing technology has advanced to give genetic information on more and more loci. Therefore, older profiles typed with previous technology may have data on fewer loci. Moreover, when DNA profiling is done sometimes not all loci are successfully typed, so that for some profiles data on a few loci are missing.

Now suppose we have a DNA profile g_C of an unknown person $C \in \mathcal{P}$. The C stands for Criminal, and we can, if we want, think of g_C as a profile found at the scene of a crime. We are going to consider the case where some specific d_i is the only profile in $\mathcal{D}$ that is not proven to be different from (and in that sense, matches with) g_C. In particular we are interested in the evidential value of such a unique match. The profiles g_C and d_i are said to match if they are the same on all *common* loci that are typed for both profiles. In order to evaluate this evidential value, we need the probability of such a match to happen by chance. For person i this is the probability that a randomly chosen person matches with g_C on all common loci of d_i and g_C. This probability is called the *random match probability* (RMP) of person i. The fact that $\mathcal{D}$ is not homogeneous implies that the match probabilities of the persons in $\mathcal{D}$ are not all the same. Indeed, the RMP for person i depends on the trace profile g_C and on the set of loci typed for d_i.

We denote by p_i the RMP of person i, noting that p_i depends on both d_i (since this tells us the loci for person i) and g_C (since this tells us the loci of C's profile, as well as the profile we need to have a match with). The administrator of the database may have full knowledge about the profiles and, therefore, also about all random match probabilities. We put ourselves in the position of the investigating authorities, who do not have that knowledge.

The investigating authorities ask the database administrator whether or not the profile g_C has a match in $\mathcal{D}$. We (the investigating authorities) are interested in the situation in which it is reported to us that there is a unique match with an identified person i_0 in the database, an event denoted by E_{i_0}. How does the occurrence of E_{i_0} affect the probability that $C = i_0$?

One might perhaps think that in order to answer this question, we need a lot of additional modeling. After all, we are not aware of the nature of the comparisons with d_j for $j \neq i_0$, i.e., we have no knowledge about the random match probabilities of the members of $\mathcal{D}\setminus\{i_0\}$. However, it suffices to have a prior probability $P(C = i_0)$, to know the random match probability p_{i_0} of individual i_0, and to have assessed the probability $P(C \in \mathcal{D})$ that the database contains C, since the posterior odds of $C = i_0$ versus $C \neq i_0$ are given by

$$\frac{P(C = i_0|E_{i_0})}{P(C \neq i_0|E_{i_0})} = \frac{1}{p_{i_0}} \frac{P(C = i_0)}{P(C \notin \mathcal{D})}. \tag{2.1}$$

We will prove this formula in Appendix A, where we also further detail the mathematical model.

It is customary to denote the matching individual by S, standing for Suspect, since providing a match with g_C of course usually leads to suspicion of being C. Once the unique match is there and the index i_0 of the matching person is revealed, S is defined, and we can speak about the (prior) probability $P(S = C)$ that the uniquely matching person is C. We denote by E_S the event that there is a unique match with the so found S. Hence we can rewrite (2.1) as

$$\frac{P(C = S|E_S)}{P(C \neq S|E_S)} = \frac{1}{p_S} \frac{P(C = S)}{P(C \notin \mathcal{D})}, \tag{2.2}$$

and it is this form which we typically use. Note that a version of this formula already appeared in [3], in a more specialized situation.

Upon a unique match with an identified person S, the quantities $P(C = S)$, $P(C \notin \mathcal{D})$ and p_S can be defined, and the right hand side of (2.2) can be computed. The first two probabilities are prior probabilities of C being either S or not belonging to $\mathcal{D}$, and with the prior of C in hand these can be computed. The quantity p_S is just the random match probability of the found S, as explained above.

Before we continue, we take a moment to reflect on the three probabilities that determine the posterior odds on $C = S$. First of all, we need the random match probability p_S of S. This makes perfect sense: if the probability that S would be indistinguishable from C if S and C were different persons becomes smaller, the probability that $C = S$ becomes larger, all other circumstances being equal. We also observe that the *only* random match probability of relevance is the one with S. How likely it was beforehand that other members of $\mathcal{D}$ would be excluded is not important anymore, once S is the only non-excluded individual. We only need to know that all others are excluded, nothing else. Our notation, allowing for varying random match probabilities, makes this explicit, and the fact is easy to miss if we would assume that all random match probabilities would be the same.

Second, we need the prior probability that $C \in \mathcal{D}$, and this also makes sense. Indeed, the better the database is suited for our purpose, the larger the probability that if we find a single match, it is with the right person, all other circumstances being equal. Finally, we need the prior probability that $C = S$, and this is understandable as well: the stronger the case against S without the database result, the stronger the case will be with the database result, again of course assuming all other circumstance being equal.

We next draw attention to the fact that (2.2) is a completely general expression, applying whenever there is just a single person in the database who cannot be excluded as trace donor. In particular it is valid for database searches without having a prior specific suspicion against any database member, or for a search where there is already interest in the matching individual beforehand, or for a probable cause case where

[1] https://en.wikipedia.org/wiki/P_versus_NP_problem.

there is no one tested except S, which simply means that $\mathcal{D} = \{S\}$. The estimation or assignment of the relevant three probabilities on the right hand side in (2.2) depends on the further context and will be different for each of the three scenarios that we just mentioned.

We stress the fact that some quantities that could perhaps be thought to be relevant for the posterior odds on $C = S$ do not explicitly enter into (2.2), such as the size of the (general or offender population) or the size of the database. These quantities will therefore only be interesting for us if we employ a model in which they are needed to estimate any of the probabilities in the right hand side of (2.2), and in any case they are only of indirect relevance. Once the probabilities in (2.2) are known, the posterior odds are known as well, so in a different (offender) population with the same pS, $P(C = S)$ and $P(C \in \mathcal{D})$, they would be the same even if $\mathcal{D}$ itself were different.

The posterior odds in (2.2) are not in the form of a likelihood ratio times prior odds for a hypothesis versus its negation, but in several ways we can rewrite it as such a product. First of all we may write

$$\frac{P(C = S|E_S)}{P(C \neq S|E_S)} = \frac{1}{p_S}\frac{P(C = S)}{P(C \notin \mathcal{D})}\frac{P(C \neq S)}{P(C \neq S)} = \frac{1}{p_S}\frac{1}{P(C \notin \mathcal{D}|C \neq S)}\frac{P(C = S)}{P(C \neq S)}. \tag{2.3}$$

Hence the likelihood ratio corresponding to $C = S$ versus $C \neq S$ is given by

$$\frac{1}{p_S}\frac{1}{P(C \notin \mathcal{D}|C \neq S)}, \tag{2.4}$$

which is at least equal to $1/p_S$.

On the other hand, when we condition on E_S, the hypotheses $C \in \mathcal{D}$ and $S = C$ are equivalent. Hence $P(C \in \mathcal{D}|E_S) = P(C = S|E_S)$ and we therefore also have

$$\frac{P(C \in \mathcal{D}|E_S)}{P(C \notin \mathcal{D}|E_S)} = \frac{1}{p_S}\frac{P(C = S)}{P(C \notin \mathcal{D})}\frac{P(C \in \mathcal{D})}{P(C \in \mathcal{D})} = \frac{1}{p_S}P(C = S|C \in \mathcal{D})\frac{P(C \in \mathcal{D})}{P(C \notin \mathcal{D})}. \tag{2.5}$$

Hence the likelihood ratio of $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$ is equal to

$$\frac{1}{p_S}P(C = S|C \in \mathcal{D}), \tag{2.6}$$

and this quantity is *at most* $1/p_S$.

In both formulations, we see that the likelihood ratio reduces to just $1/p_S$ in case $\mathcal{D} = \{S\}$, which corresponds to the classical 'probable cause' situation in which only the profile of a suspect S is compared with that of C. The hypotheses $C = S$ and $C \in \mathcal{D}$ are then equivalent. In case of a uniform prior of C on $\mathcal{D}$, (2.6) reduces to $1/(np_S)$, that is, we retrieve Stockmarr's (cf. [4]) expression.

3. Evaluation of the probabilities in particular cases

From now on we assume that the identity of S is known, so that (2.2) is our basic expression, valid in all circumstances where only a single individual S in a set $\mathcal{D}$ turns out not to be excluded as candidate for being C. For the formal derivation of this result, it was irrelevant how large the database is, whether or not there was a suspicion against S, why the search was conducted, or in which order evidence has been gathered.

None of these additional aspects are needed to derive (2.2) and therefore they will not change (2.2) algebraically. But in order to assess the required probabilities, of course different situations can lead to different *numerical* evaluations of these relevant probabilities, and therefore also to different posterior odds.

To illustrate this, we next treat some special cases in more detail. For database searches, we distinguish between various different scenarios. In the first one we assume the search is carried out without any other relevant information about C other than the obtained profile, that is, there is no additional evidence against any database member. We call this a *cold case search*. After the search we are informed that there is a match with a profile in the database, with or without knowing the identity (i.e., the index i_0) of the donor of the matching profile.

A second type of search, which we call a *targeted search*, arises if a suspect S has been identified and that suspect happens to already be in the database, say $S = i$. We then carry out the search to confirm this suspicion. In that case, a single match with another person would have surprised us much more than if we indeed obtain E_i, a single match with the already identified suspect.

The classical *probable cause* situation is the one where the identified suspect S is the only person whose profile is compared to that of C. Mathematically, this corresponds to a targeted database search in a database consisting only of S.

3.1. Cold case search

First, we assume that we have carried out a cold case search, meaning that we have done the database search because the identity of C is unknown and we believe that C might be one of the members of $\mathcal{D}$. In this case, defining a prior $P(C = S)$ from scratch seems hard and therefore (2.3) is hard to evaluate directly. However, if we take the route (2.5) then within $\mathcal{D}$ things are easier: without any further information about the individuals in the database, the only option is to choose $P(C = S|C \in \mathcal{D}) = 1/n$. Therefore (2.5) becomes

$$\frac{P(C = S|E_S)}{P(C \neq S|E_S)} = \frac{P(C \in \mathcal{D}|E_S)}{P(C \notin \mathcal{D}|E_S)} = \frac{1}{np_S}\frac{P(C \in \mathcal{D})}{P(C \notin \mathcal{D})}.$$

It remains to provide a numerical assessment of $P(C \in \mathcal{D})$. One way to do so is to let $P(C \in \mathcal{D})$ be equal to the proportion of traces that have been previously searched with and have given rise to a match in the database. In doing so, one implicitly assumes that all previous matches were with the true donor of the trace, and that the traces that were searched with in the past form a sufficiently representative sample to be useful for an estimate of $P(C \in \mathcal{D})$. This assumption is not entirely unproblematic. If we take the type of crime into account, the estimate for $P(C \in \mathcal{D})$ may change depending on whether the case is, for example, a burglary case, a homicide or a sexual assault case. Furthermore, one may argue that the trace donor C need not be the actual offender. This, however, is also be possible for previous searches; the probability $P(C \in \mathcal{D})$ therefore applies to C as trace donor and not to C as offender. For an in-depth discussion on these issues we refer to [16].

Bearing these cautions in mind, it is not uncommon for databases to be sufficiently large as to have odds $P(C \in \mathcal{D})/P(C \notin \mathcal{D})$ that are within one order of magnitude of being even. If that is the case, the posterior odds are of the same order of magnitude as the likelihood ratio, and we can then say that the odds on the match being with the trace donor are within one order of magnitude of $1/(np_S)$. If, for example, $P(C \in \mathcal{D}) = P(C \notin \mathcal{D}) = 0.5$, $n = 10^6$ and $p_S = 10^{-9}$, the odds are 1000:1 that the match is with the actual trace donor. Of course, when the specifics of the crime and of the uncovered suspect are brought into consideration, these odds will need to be further updated. If, for example, it turns out the match is with a person yet to be born when the crime was committed, they will be reduced to zero. But this cannot happen very often, since there will be a thousand true matches for every coincidental one for these n and p_S.

In case the match is indeed with the trace donor, and the trace donor is the actual offender, further evidence can potentially be uncovered which will raise the odds from 1000:1 to a larger number. When further non-genetic evidence I is found and taken into account, the result (2.2) still applies, but all probabilities need to be conditioned on I. This has no effect on the match probability p_S but now $P(C = S \mid I) > P(C = S)$. For $\mathcal{D}$, since additional evidence against one of its members S has been found, the probability that $\mathcal{D}$ contains C cannot decrease and hence we have $P(C \notin \mathcal{D}|I) \leq P(C \notin \mathcal{D})$. Putting this together we see that the posterior odds on $C = S$ increase, reflecting the

strengthening of the case against S due to the new evidence I.

3.2. Targeted search

The preceding discussion brings us naturally to the targeted search case. In this case, evidence against S is found before the database search is done. Since there is no temporal order for probabilities, we must arrive at the same posterior odds regardless of whether S is identified via the database cold case search and further evidence is subsequently found, or when this happens in the reverse order. If we take into account the additional evidence before we process the evidence E_S, we will no longer have $P(C = S|C \in \mathcal{D}) = 1/n$, but a much larger value, approaching $P(C = S|C \in \mathcal{D}) \approx 1$ as more and more evidence against S is uncovered. In that case, S was – before the database search – pretty much the only plausible candidate for C, which in turn means that $P(C = S) \approx P(C \in \mathcal{D})$, making the hypotheses $C = S$ and $C \in \mathcal{D}$ much closer to being equivalent then in the cold case.

In terms of (2.3) and (2.5), both terms $P(C = S|C \in \mathcal{D})$ and $P(C \notin \mathcal{D}|C \neq S)$ are close to 1, so that the likelihood ratio is close to $1/p_S$, regardless of whether we start out with hypotheses about S (in which case the likelihood ratio is larger than $1/p_S$) or about $\mathcal{D}$ (in which case it is smaller). The exclusions that the database search has provided are, in other words, essentially irrelevant since we already believed that S was by far the most plausible candidate for being C before carrying out the search. Learning that the other database members, who we already believed not to be C, are indeed not C, then has only very little impact.

3.3. Probable cause

Now we arrive naturally at the probable cause case, which we can think of in various ways. We can set $\mathcal{D} = \{S\}$ so that no other comparisons have been done other than between S and C, who turned out to have matching profiles. Alternatively, we can think of a database in which all individuals apart from S were already excluded prior to the search, that is, $P(C = S|C \in \mathcal{D}) = P(C \notin \mathcal{D}|C \neq S) = 1$. The latter formulation is nothing but an extreme case of the targeted search case which we discussed above. Regardless of how we think about it, the hypotheses $C = S$ and $C \in \mathcal{D}$ are then equivalent prior to learning E_S, so that (2.3) and (2.5) coincide. The likelihood ratio in favor of $C = S$ (or in favor of $C \in \mathcal{D}$, which is now the same hypothesis) is then exactly equal to $1/p_S$.

3.4. Casework

Of course, a case is not going to be confined to one of these three categories once and for all, but at any given point in time we will have information that makes us regard the case as most similar to one of the three types above. For example, a cold case search may be initially carried out, after which evidence against S is found. When that evidence is taken into account, we are in the same situation as for a targeted search. Conversely, a suspect may be identified via other means than the database, and be the only one that is compared to the trace profile. If, subsequently, a database search is carried out and no further matches are found, this is also equivalent to a targeted search. If, on the other hand, the evidence leading to the identification of S as suspect turns out to be erroneous and is dismissed, we could also come close to a situation best described as a cold case search, because there is no evidence any more distinguishing S from the other database members other than the matching profile.

4. Which likelihood ratio?

So far, we have seen that the evaluation of the posterior odds on $C = S$ is, at least mathematically, straightforward via (2.2). As the expression (2.2) also makes clear, the relevant hypotheses following the search result are that $C = S$ or that $C \in \mathcal{D}$. Writing these posterior odds as a likelihood ratio times prior odds on a hypothesis versus its negation is possible in two ways, but both are a little artificial. Either, we obtain the likelihood ratio for the hypotheses $C \in D$ versus $C \notin \mathcal{D}$, which reflect the initial questions (at least, in a cold case search) but not the question that has come up following the search, namely whether $C = S$ or $C \neq S$. Or, we work with $C = S$ versus $C \neq S$ throughout, which does not reflect that prior to the search we were not especially interested in S and which gives by construction strong evidence in the form of a large likelihood ratio. Mathematically, there is no harm in either approach but we do not think they express the situation better than (2.2) does.

One way out is to abandon the likelihood ratio approach and to directly focus on the matter of interest via (2.2). However, forensic laboratories are often asked (e.g., cf. [17]) to provide an assessment of the evidence in terms of a likelihood ratio which makes it inevitable to choose one. As is clear from (2.4) and (2.6), two candidates emerge here: the suspect centered likelihood ratio $1/(p_S P(C \notin \mathcal{D}|S \neq C)) \geq 1/p_S$, or the database centered likelihood ratio $P(S = C|C \in \mathcal{D})/p_S \leq 1/p_S$. How large the difference between these are depends on the situation, hence we will discuss the cold case search and targeted search once more. (The latter includes the probable cause case as a special case, as remarked above.)

4.1. Cold case search

In this case, suppose that we do not know anything about C and assume uniform prior odds for C within $\mathcal{D}$. Then the discrepancy between the likelihood ratios is large: it is either at least $1/p_S$, or equal to $1/(np_S)$. Which one should we prefer? Obviously, from a mathematical point of view there is no problem: both lead to the same posterior odds, each within their own context. So, another question emerges: what is the most natural context here?

Initially, the question addressed was whether $C \in \mathcal{D}$. A forensic laboratory may receive a generic question to produce a DNA profile from a crime stain sample and compare the resulting profile with the database $\mathcal{D}$. Up to that point, no one in the database stands out. But when the match result E_S is obtained, attention shifts to S and the ultimate issue for a court is whether $C = S$ or not. Hence the context changes along the way. There is, therefore, no obvious choice from the contextual point of view either. Either $1/(np_S)$ is reported reflecting the original question that has been asked, or $1/p_S$ is reported reflecting the fact that the case now revolves about S.

Is there then, perhaps, a *practical* reason to choose for either likelihood ratio? We believe that this may indeed be the case, and that $1/(np_S)$ then has a practical advantage. We next explain why. First of all, we remark that probabilistic assessments are difficult to convey to judges and juries and that intuition may lead to wrong conclusions. A common pitfall is to understand likelihood ratios as posterior odds, a mistake commonly referred to as the prosecutor's fallacy (or base rate neglect). Those who do not make such mistakes will be able to both interpret the likelihood ratio $1/p_S$ (for initial hypotheses whether $S = C$ or not) or $1/(np_S)$ (for initial hypotheses whether $C \in \mathcal{D}$ or not) and make the correct inference on the posterior odds. However, since the posterior odds in the cold case search case based on the information known at the time of issuing the report are much closer to $1/(np_S)$ than to $1/p_S$, the harm done by a prosecutor's fallacy is much less if $1/(np_S)$ is reported in such a case than if $1/p_S$ is reported. Such a report could for example be as follows.

> A request has been received to generate a DNA profile from item X, and use that profile to search for its trace donor in the database $\mathcal{D}$. From item X a DNA profile has been generated and compared with the DNA profiles of the database $\mathcal{D}$. The findings are that a single match with database member S has been found and that the probability for S to produce a match, if not the trace donor, is p_S (e.g., $p_S = 10^{-9}$). At the time of the search the database contained n

profiles (e.g., $n = 1{,}000{,}000$). Furthermore the laboratory is not aware of any non-DNA information pertaining to specific database members as to the possibility that they are more plausible candidates for being the trace donor than others. From this it is concluded that the obtained database search result is $1/(np_S)$ (1000) times more likely if the database contains the trace donor than if the database does not contain the trace donor. Furthermore S is now the only person in $\mathcal{D}$ who can be the trace donor. Averaged over all cases, the database contains about a proportion α (e.g., 40%) of the trace donors that are searched in it. This figure would lead to odds of $(1/(np_S) \times (\alpha/(1-\alpha))$ to one (667:1) in favor of S being the trace donor. Further information pertaining to the case and to S can be used to revise the odds on S being the trace donor to a larger or smaller number.

The advantage of this approach is that it describes an evaluation of all involved probabilities, and that a prosecutor's fallacy is less likely to occur. A practical disadvantage is that if the number n changes rapidly over time. Another concern is in order to know whether $1/(np_S)$ as a likelihood ratio applies one has to know whether a uniform prior is applicable to the database or not. It seems to us that in most cases, using a uniform prior where there is more information available, will be in the interest of the matching individual, since for the n and p_S used in practice, most of the matches are with the true trace donors.

Another concern is that laboratories may not be expected or permitted to give more than only likelihood ratios. In that case, one could adapt the above report so as to leave out the estimate of the odds on $C \in \mathcal{D}$.

4.2. Targeted search

In this case a search has been done where S was already, prior to the search, a plausible candidate for being the trace donor C. We assume that the forensic laboratory has been requested to generate a DNA profile from item X and compare it with the DNA profiles in database $\mathcal{D}$, and in particular to the DNA profile of S. In that case $P(C = S|C \in \mathcal{D}) \neq 1/n$, and $1/(np_S)$ is not the likelihood ratio for $C \in \mathcal{D}$. In this case, both the purpose of the search and of the ensuing further investigation are to investigate whether $C = S$ and then it is natural to report also the corresponding likelihood ratio for these hypotheses, which is $1/p_S \times 1/P(C \notin \mathcal{D}|S \neq C) \geq 1/p_S$. Thus, a report could be phrased along the following lines:

> A request has been received to generate a DNA profile from item X, and use that profile to search for its trace donor in the database $\mathcal{D}$ with particular attention to individual S. From item X a DNA profile has been generated and compared with the DNA profiles of the persons in $\mathcal{D}$. The findings are that a single match with database member S has been found and that the probability for S to produce a match, if not the trace donor, is p_S. This provides evidence that S is indeed the trace donor. Based on the information described up to now, the match with S increases the odds on S being the trace donor by at least a factor $1/p_S$, compared to what they were prior to the search result. Further information pertaining to the case and S will allow to revise these odds to a larger or smaller number.

Of course it is also possible that a request is done to carry out a targeted search, leading to a match with a different individual S' than the expected one S. In that case, we are back into a cold case search case since we must then use that the database members other than S cannot be distinguished from each other than by their DNA profiles based on the available information.

5. No personal information on the person yielding the unique match

So far, we have assumed that we get to know who the match is with, in the sense that the index i_0 is revealed so that it is known which individual in the population S is. In this section we discuss two situations in which we have less information because we are not told who the match is with. First, we discuss the case in which we only get to know the matching profile, without knowing whose profile it is, but with possible extra information about S being in a certain subset of $\mathcal{D}$. After that we investigate what happens when we only know that there is a unique match, and not the random match probability of that profile, or anything else.

5.1. Partial information about S

There are situations in which we do not obtain full information about the unique match S, but instead only (apart from the RMP of the matching profile) that S is contained in a certain given subset $\mathcal{D}'$ of $\mathcal{D}$. As an example where this may happen, in searches between different jurisdictions, the database that has been searched sometimes only returns the matching profile but no further personal details about the person whose profile that is. So, the queried database only confirms that it holds a unique non-excluded individual and gives the databased profile of that person. In that case we do not know the identity of S, but only that S belongs to the collection of individuals who have the reported RMP. What can we say in this situation about the probability that the uniquely matching individual is C?

In this example $\mathcal{D}'$ is defined in terms of the RMP itself, but it helps to consider the more general situation in which the RMP of the match is known, together with the fact that S is contained in *some* given subset $\mathcal{D}'$ of $\mathcal{D}$.

Let us still denote the matching person by S, but note that S is unknown, so we cannot use the unconditional probability $P(C = S)$. We can, however, speak about the probability that $C = S$ given that there is a unique match. The random match probability p_S is still available.

We write $E_{\mathcal{D}'}$ for the event that there is a unique match in $\mathcal{D}$ with a person S only to be known in $\mathcal{D}'$. We claim that

$$\frac{P(C = S|E_{\mathcal{D}}')}{P(C \neq S|E_{\mathcal{D}}')} = \frac{1}{n'p_S}\frac{P(C \in \mathcal{D}')}{P(C \notin \mathcal{D})}. \tag{5.1}$$

This expression generalizes (2.2), which is obtained by taking $\mathcal{D}' = \{S\}$. In the general formulation, $P(C = S)$ is replaced by $P(C \in \mathcal{D}')/n'$, which is rather intuitive. Indeed, knowing $\mathcal{D}'$ has allowed us to narrow down the matching individual from someone in $\mathcal{D}$ to someone in $\mathcal{D}'$, but we have no information as to who the matching person is in $\mathcal{D}'$, hence the division by n'. Also note that we only use prior information that is available to us. We prove this formula in Appendix A.

Taking $\mathcal{D}' = \mathcal{D}$ in (5.1) yields

$$\frac{P(C = S|E_{\mathcal{D}})}{P(C \neq S|E_{\mathcal{D}})} = \frac{1}{np_S}\frac{P(C \in \mathcal{D})}{P(C \notin \mathcal{D})}. \tag{5.2}$$

Since on the left hand side we can replace $C = S$ by $C \in \mathcal{D}$ and $C \neq S$ by $C \notin \mathcal{D}$, this is a natural situation in which the likelihood ratio is equal to $1/(np_S)$.

More generally, in case $P(C \in \mathcal{D}') = \frac{n'}{n}P(C \in \mathcal{D})$, (5.1) reduces to

$$\frac{P(C = S|E_{\mathcal{D}}')}{P(C \neq S|E_{\mathcal{D}}')} = \frac{1}{np_S}\frac{P(C \in \mathcal{D})}{P(C \notin \mathcal{D})}, \tag{5.3}$$

and hence $\mathcal{D}'$ is then irrelevant. In particular we do not need to know the size n' of $\mathcal{D}'$. For example, when the prior for C on $\mathcal{D}$ is uniform, this will be the case.

5.2. Only existence of unique match is known

Suppose that we only get to know that there is a unique match with some person S, without any further information such as p_S or the identity of the matching person. Can we still say anything meaningful about the posterior probability that the matching person is C?

A little reflection shows that in order to say anything meaningful about this probability, we need more information. It is intuitively clear that in the absence of any individual information, we will need the random match probabilities p_i of all database profiles. Assuming that we know these, we again write $P(C = S \mid E)$ for the probability that the unique match is with the actual donor.

We claim that in odds form, we get

$$\frac{P(C = S|E)}{P(C \neq S|E)} = \frac{P(C \in \mathcal{D}|E)}{P(C \notin \mathcal{D}|E)} = \frac{\sum_{i=1}^{n} \frac{P(C=i)}{1-p_i}}{P(C \in \mathcal{D}) \sum_{j=1}^{n} \frac{p_j}{1-p_j}} \frac{P(C \in \mathcal{D})}{P(C \notin \mathcal{D})}. \tag{5.4}$$

We prove this formula in the Appendix. Note the difference between (5.1) and (5.4). In the former, we know (and use) the random match probability p_S of the unique match. All other random match probabilities are irrelevant. In (5.4) we know nothing about the matching persons, and all random match probabilities of the whole database are relevant. Also note that this time, we have no other option but using prior information about the full database, and this is precisely what happens in (5.4).

In the case where all the p_i's are the same and equal to p, then we do actually know the random match probability so we are back to the situation in Section 5.1. Indeed, in that case the likelihood ratio in (5.4) reduces to

$$\frac{\sum_{i=1}^{n} \frac{P(C=i \mid C \in \mathcal{D})}{1-p_i}}{\sum_{j=1}^{n} \frac{p_j}{1-p_j}} = \frac{\sum_{i=1}^{n} \frac{P(C=i \mid C \in \mathcal{D})}{1-p}}{\sum_{j=1}^{n} \frac{p}{1-p}} = \frac{1}{np}, \tag{5.5}$$

so this is another situation in which $1/(np)$ is the correct likelihood ratio.

6. An analysis of some controversies in the literature

The interpretation of database searches yielding a single match has seen fierce debates [4,6–8,15,18], where especially the likelihood ratio $1/(np)$ has often been ridiculed. In this section we take a closer look at the various arguments that in this debate have been proposed in favor or against the use of one of the likelihood ratios, or in favor or against the use of the various hypotheses of interest. Authors sometimes provide direct arguments why a specific choice should be used, but more often it is the case that arguments are provided against other choices. We will consider the arguments one by one. In every instance that a general rejection of one type of likelihood ratio or hypothesis is suggested, we will argue that this is unjustified.

We distinguish between a strong *case* and strong *evidence*. A case always refers to posterior odds or probabilities, while evidence always refers to a likelihood ratio. We should keep in mind though that typically various different likelihood ratios are possible, so that we should be careful whenever we speak about evidence. In this section we assume that all members of the database have RMP equal to p. This is mainly for simplicity, since it will make it easier to explain why the various arguments against either likelihood ratio are incorrect. Also, we will see that the discussion of the controversies leads to certain facts about unique database matches that we find interesting enough to discuss, and which have not appeared in print before, as far as we are aware.

6.1. Against 1/(np): a large database should give strong evidence

One of the first arguments against the use of the likelihood ratio $1/(np)$ was that it would imply that the larger the database is, the weaker the evidence against the suspect, and that this must be absurd. In the extreme case where the whole population would be in $\mathcal{D}$, the evidence would be the weakest possible (still according to adversaries of the $1/(np)$ likelihood ratio) while it clearly provides the strongest evidence possible.

How should we evaluate this argument? First of all, this argument can only be put forward when the $1/(np)$ likelihood ratio applies. As we have seen, this is so in a cold case search case when the prior for C on $\mathcal{D}$ is uniform (see the discussion after (2.6)), and in the case where we do not know the identity of S. Here we consider the latter case, and we use the prior and posterior odds as expressed in (5.1) and (5.2). Below, in the discussion of the so called cunning defense lawyer argument, we will also discuss what happens when S is identified and C is uniform on $\mathcal{D}$.

Suppose now that the whole population is in the database. In that case $P(C \in \mathcal{D}) = 1$ and $P(C \notin \mathcal{D}) = 0$ which means that the prior odds for these hypotheses are infinite. It then follows that the posterior odds are infinite as well. Rather than giving rise to absurdities, this leads to a posterior probability on $C = S$ equal to 1, as it should. Indeed, if S is the only match and the full population is in $\mathcal{D}$, then the posterior probability that $C = S$ must be 1.

How can we reconcile this with the fact that the likelihood ratio is very small in this case, namely equal to $1/(np)$? We can better see what happens when we assume that *nearly* everyone is in $\mathcal{D}$ so that $P(C \notin \mathcal{D})$ is small but not zero. In that case, the prior odds on $C \in \mathcal{D}$ are very large, but the likelihood ratio $1/(np)$ may even be smaller than one. Note that np is the expected number of matches in the database, assuming that $C \notin \mathcal{D}$ and that this match probability applies to all database members. If the number of matches found (being one) is smaller than the expected number np of such matches, then indeed this result provides evidence *against* $C \in \mathcal{D}$; this is only reasonable. Hence the posterior probability on $C \in \mathcal{D}$ decreases, but since it is now totally concentrated on the uniquely matching individual S, the posterior odds on the match being with the trace donor are very high. Of course, np could in principle also be very large but in a situation with a unique match this will not happen often, since it would correspond to a trace occurring much less frequently in $\mathcal{D}$ than predicted by p.

We conclude that there is no contradiction in the proper use of the likelihood ratio $1/(np)$, even when the database is very large.

6.2. Against 1/(np): growing database

A second argument put forward against the $1/(np)$ likelihood ratio, or the use of hypotheses $C \in \mathcal{D}$, used hypothetically growing databases. It is clear that if a match with S is found in the database, and there are no further matches found at a later point in time when the database has grown, the additional exclusion of new individuals can only increase the posterior odds that S is the trace donor. It may seem that this provides an argument against using $1/(np)$.

In order to purely assess the effect of the growth of the database, we disregard the identity of S. Consider a subset $\mathcal{D}' \subset \mathcal{D}$. For example $\mathcal{D}'$ may be the database at some point in time, and $\mathcal{D}$ the database at a later moment when new persons have been added. Suppose we first find a unique match in $\mathcal{D}'$ and after that, we find no further matches in $\mathcal{D}$. Alternatively, we first find the match with S in $\mathcal{D}$, and then find out that $S \in \mathcal{D}'$. At the end of this procedure, we have a unique match in $\mathcal{D}$, but now we also know that this unique match is in fact in $\mathcal{D}'$. This knowledge changes the situation compared to the earlier analysis with $\mathcal{D}'$ and $\mathcal{D}$, as we can verify using (5.1), which is the expression for the posterior odds in precisely this situation. Indeed, we learn from (5.1) that

$$\begin{aligned} \frac{P(C = S|E_{\mathcal{D}}')}{P(C \neq S|E_{\mathcal{D}}')} &= \frac{1}{n'p} \frac{P(C \in \mathcal{D}')}{P(C \notin \mathcal{D})} \\ &\geq \frac{1}{n'p} \frac{P(C \in \mathcal{D}')}{P(C \notin \mathcal{D}')}, \end{aligned}$$

since $P(C \in \mathcal{D}) \geq P(C \in \mathcal{D}')$. Thus, in complete generality, the additional exclusions in $\mathcal{D} \setminus \mathcal{D}'$ provide further evidence for $C = S$, as is to be expected.

So far, we have considered that we first learn that S provides a

single match in $\mathcal{D}'$, and then the database grows into $\mathcal{D}$ and no additional matches are found. Next, we consider a different situation: we assume that we know that S provides a unique match in the larger database $\mathcal{D}$, and then we learn that $S \in \mathcal{D}'$ for some subset $\mathcal{D}' \subset \mathcal{D}$. How does that change the odds on $C = S$?

Learning that $S \in \mathcal{D}'$ changes the posterior odds on $C = S$ from (5.2) into (5.1). The ratio between the new odds, when we have learned that $S \in \mathcal{D}'$, and the odds when we only knew that $S \in \mathcal{D}$, is, by using (5.1) and (5.2),

$$\frac{P(C \in D')/n'}{P(C \in \mathcal{D})/n}. \tag{6.1}$$

If $\mathcal{D}'$ is a subset of $\mathcal{D}$ such that the average prior $P(C \in \mathcal{D}')/n'$ is larger than the corresponding average prior for $\mathcal{D}$, then the knowledge that $S \in \mathcal{D}'$ increases the probability that $C = S$, and the opposite is of course also possible.

In case the prior for C is uniform on the population, we have that $P(C \in D)$ is equal to the proportion of the population that is in the database, in which case we get unity in (6.1), expressing that it neither strengthens nor weakens the case against S if we learn that $S \in \mathcal{D}'$, whatever $\mathcal{D}'$ is.

6.3. Matches in a larger database are not necessarily more likely to be with the trace donor

It might be perceived that a larger database is better in the sense that a higher proportion of the unique matches is with the actual trace donors. This, however, is not true in general. We can already see this from (5.2). Indeed, consider two databases $\mathcal{D}'$ and $\mathcal{D}$ of sizes $n' < n$ respectively. Let us write $f_{\mathcal{D}}$ for $P(C \in \mathcal{D})/P(C \notin \mathcal{D})$ and define $f_{\mathcal{D}}{}'$ similarly. Then the ratio between the posterior odds for a unique match in $\mathcal{D}'$ to be with the actual trace donor, and those for a unique match in $\mathcal{D}$ to be with the actual trace donors, is

$$\frac{f_{\mathcal{D}}{}'/n'}{f_{\mathcal{D}}/n}. \tag{6.2}$$

The above expression holds for any two databases, it is not necessary to assume that one is a subset of the other. Let us now assume that $\mathcal{D}' \subset \mathcal{D}$ to see how the odds on the match being with the trace donor may change from matches in $\mathcal{D}'$ to matches in $\mathcal{D}$.

We see that in that case, if the odds on $C \in \mathcal{D}$ compared to those on $C \in \mathcal{D}'$ grow more than the sizes of the databases, then in $\mathcal{D}$ a larger proportion of the matches is with the actual trace donors. If the prior for C on $\mathcal{D}$ is uniform, this is the case since then $f_{\mathcal{D}}{}' = n'/(N - n')$ and $f_{\mathcal{D}} = n/(N - n)$ so (6.2) evaluates to $(N - n)/(N - n')$, which is smaller than one if $n' < n$.

Thus, if we assume uniform prior probabilities for C then a unique match in a larger database is always more likely to be with C than a unique match in a smaller database.

In general, it is possible that (6.2) is larger than one, which means that the posterior odds on a unique match to be with the trace donor are smaller in the larger database than those for matches in a subset of it.

As an example, let us now assume that $P(C \in \mathcal{D}) = \sqrt{n/N}$, where N is the size of the population and n the size of $\mathcal{D}$. Then a small fraction of the population in the database corresponds to a relatively large $P(C \in \mathcal{D})$. For instance, if $n/N = 0.01$ we have $P(C \in \mathcal{D}) = 0.1$. This is not an unnatural property at all, since the proportion of traces that yield a match may be much larger than the fraction of the population in the database.

We write $x = n/N$, so that $P(C \in \mathcal{D}) = x^{1/2}$. From (5.2) we have that the posterior odds on a match in $\mathcal{D}$, for size $x = n/N$, being with the trace donor are given by

$$\frac{1}{Np}\frac{x^{-1/2}}{1 - x^{1/2}}. \tag{6.3}$$

It is easy to see now that these posterior odds are minimal for $x = 1/4$: they *decrease* between $x = 0$ and $x = 1/4$, and then increase towards infinity as $x \to 1$. As n increases, so does $P(C \in \mathcal{D})$, as well as the possibility for adventitious matches. Both of these effects influence the posterior odds in case of a single match. For a uniformly sampled database, the increase of $P(C \in \mathcal{D})$ always outweighs the increased opportunity for adventitious matches, but we now see that this need not be so in general.

Thus, it is not generally true that a larger database is better in the sense that in a larger database, there may be a smaller proportion of the unique matches with the trace donor than in a smaller database. Formula (6.2) tells us that it may be the case that the posterior odds on the match being with the trace donor are *smaller* or *larger* or *the same* in an expanded database $\mathcal{D}$ than in a smaller $\mathcal{D}'$. If we learn that we have a match in $\mathcal{D}$, the odds on it being with the trace donor can be smaller, larger or identical to those that we get if we learn that we have a match in $\mathcal{D}'$ without knowing about $\mathcal{D} \setminus \mathcal{D}'$.

6.4. Against 1/(np): Cunning defense lawyer

A further argument that has been put forward against the $1/(np)$ rule is at first sight quite convincing [6]. Imagine that a match with a suspect has been obtained, outside the database. Then the suspect would be well advised by a cunning defense lawyer to insist a database search be carried out. Indeed, if no additional matches are found, the failure to find additional matches will substantially weaken the evidence against his client, since the likelihood ratio is now $1/(np)$ instead of $1/p$. On the other hand, if additional matches are found, then this is even better news for the suspect.

Convincing as this may sound, this argument against the $1/(np)$ rule is not correct. Clearly it is impossible that the *case* against the suspect weakens irrespective of the outcome of a database search, and indeed this is not what follows from our analysis, be it based on the suspect driven hypothesis $S = C$ or on $C \in \mathcal{D}$. We next explain this in detail.

First of all, we note that the $1/(np)$ rule is only valid under specific circumstances. With an identified S, we can only apply it when the prior of C is uniform on $\mathcal{D}$, see the discussion following (2.5) and (2.6). But in the situation described above, with a suspect already identified before carrying out a database search, uniform priors are not realistic. Every further tested person would have to have the same prior probability to be C as the suspect S, something which is clearly impossible to realize given the fact that S had been suspect before the search has been carried out. In other words, this argument against the use of the $1/(np)$ rule is flawed from the very start, since the assumptions are not fulfilled.

It is interesting and illuminating, though, to see what our analysis has to say about the cunning defense lawyer argument if we make the (unrealistic) assumption that the $1/(np)$ likelihood ratio applies. We already mentioned above that further exclusions cannot make the case against an already identified suspect weaker. This can indeed also be shown in the current situation, this time using either (2.3) or (2.5). In (2.3), we do not use the $1/(np)$ rule, but it is illuminating to see what is going on there. The likelihood ratio in (2.3) is $1/p \times 1/P(C \notin \mathcal{D} | S \neq C)$. If first S alone is compared with the trace we have, at that point, that $\mathcal{D} = \{S\}$ so that the likelihood ratio is $1/p$. If we then compare the profile of C with a database $\mathcal{D}'$ and let $\mathcal{D} = \mathcal{D}' \cup \{S\}$, the odds are further updated with a factor $1/P(C \notin \mathcal{D} | S \neq C)$, which is always in favor of $S = C$ (possibly neutral in case $P(C \in \mathcal{D}') = 0$, a situation unlikely to be encountered but not outside the scope of (2.3)). Thus, it is certainly not true that the case against S becomes weaker as a result of not finding any further matches.

We can also argue this from (2.5), where the likelihood ratio is $1/(np)$, assuming (as we have to in order to apply the $1/(np)$ rule) a uniform prior for C on $\mathcal{D}$. This runs as follows. Again we first have the situation that S is the only investigated person, so that $\mathcal{D} = \{S\}$ in (2.5). The posterior odds are then equal to

$$\frac{1}{p}\frac{\alpha}{1-\alpha}, \tag{6.4}$$

where $\alpha := P(C = S)$ is the prior for S.

Next we investigate $n-1$ further individuals, none of which matches. Note that in order to comply to the requirement that all investigated persons have the same prior α as S, we must have that $n\alpha \leq 1$. Now we can apply (2.5) again, this time with $\mathcal{D}$ defined as the full set of n tested persons. The posterior odds in (2.5) are now equal to

$$\frac{1}{np}\frac{n\alpha}{1-n\alpha} = \frac{1}{p}\frac{\alpha}{1-n\alpha}. \tag{6.5}$$

Since clearly (6.5) is at least as large as (6.4), we again conclude that the case against S cannot have weakened upon finding a number of further exclusions. Note that when $n\alpha = 1$, the posterior odds are infinite, hence the posterior probability that $S = C$ is 1. This is perfectly reasonable since in that case the database contains all persons with positive prior, and they are all excluded, except S. Hence S must be C.

We conclude that there is no way in which the cunning defense lawyer argument can be made correct. Either the argument does not apply because the assumptions are not fulfilled, or, when they are, the conclusion of the argument is incorrect. The cunning defense lawyer is, after all, not as cunning as it seemed.

6.5. Against 1/p: data driven hypotheses

Not only the likelihood ratio of $1/(np)$ has been challenged, also the $1/p$ likelihood ratio has received serious criticism. Probably the main argument against using (2.3) and the corresponding likelihood ratio of (at least) $1/p$ is that the hypothesis $S = C$ is *data driven* in the sense that it is only formulated *after* observing the match with S [4]. Without this match, there would perhaps be no reason whatsoever to consider this hypothesis and this seems unfair towards the suspect S. From this point of view the division by n in the $1/(np)$ rule is supposed to compensate for the data drive nature of the hypothesis. Is this criticism justified?

Although there is some truth in the idea that a data driven hypothesis must be compensated for, it is not the division by n that takes care of this, but instead the prior odds of $S = C$ versus $S \neq C$. When we compare (2.3) to (2.5), the prior in the former is smaller than in the latter, and this, simply, compensates for the fact that $S = C$ is a data driven hypothesis. When using a data driven hypothesis, the likelihood ratio tends to be large, but this is compensated by small prior odds. We conclude that there is no principle problem in using such data driven hypotheses, as long as the interplay between likelihood ratio and priors is understood; see also [8] for more details on this phenomenon. Hence this argument against the $1/p$ rule is unsound.

6.6. Relevance of (offender) population size

In [15] it is argued that the size of the offender population plays a role in the computation of posterior odds. In their setup they set $P(S = C|C \in \mathcal{D}) = 1/n$, so that the posterior odds in (2.5) reduce to

$$\frac{1}{np}\frac{P(C \in \mathcal{D})}{P(C \notin \mathcal{D})}. \tag{6.6}$$

These posterior odds involve only one more ingredient apart from n and p, namely the prior probability $P(C \in \mathcal{D})$ for the database to contain the offender. The paper [15] attaches importance to the size of the active criminal population, but that is a quantity absent from (6.6). Hence, we would be interested in the size of the criminal population (whatever this may be) *only if* this size would be essential for the estimation of $P(C \in \mathcal{D})$.

But it is not. In fact, the authors in [15] estimate the size N of this criminal population by assuming that $\mathcal{D}$ is a random sample from the active criminal population, and then estimate the size of that population from the estimate $P(C \in \mathcal{D})$. They use a mark-and-recapture framework to estimate the size of the active criminal population N as $N = n/P(C \in \mathcal{D})$, and then plug this in in the formula for the posterior odds.

However, this analysis is redundant because all that is needed to arrive at the posterior odds is an estimate of $P(C \in \mathcal{D})$ which is the very starting point of their procedure to estimate the size of the offender population. In terms of a criminal population, $P(C \in \mathcal{D})$ may be thought of as the coverage of the population by the database. How the coverage relates to the size of that population is another issue, immaterial for the present problem.

6.7. Inadmissibility of hypotheses about $C \in \mathcal{D}$

Several authors (e.g., [13], [6]) have claimed that, even though the posterior odds (2.3) and (2.5) both provide the same answer, the only relevant hypothesis is whether $S = C$ or not, since the trial is concerned with S and not with $\mathcal{D}$ as a whole. We fail to see however, how providing the likelihood ratio for $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$, which allows equally well to arrive at a correct evaluation of the posterior probabilities the court is interested in, should be banned from being reported. A court is indeed not concerned with the collective guilt or innocence of all database members, but in case of a single match, this collective guilt reduces only to S, so these viewpoints coincide. Moreover, in case the identity of S is not (yet) known, the most natural way to proceed is by using these hypotheses. The prior probability $P(C = S)$ is not available in case S is not known, and the analogue of (5.1) or variants thereof should be used. This formula is in terms of the hypotheses $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$. In case of a targeted search where there is already an existing suspicion, the $1/(np)$ likelihood ratio does not apply anymore to the hypotheses $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$. In that case, both frameworks can still be used, and the difference between them is then far less pronounced, both leading to a likelihood ratio not differing much from $1/p_S$.

Even if dismissed by some statisticians, in the legal community, the relevance of the hypothesis $C \in \mathcal{D}$ and the matter of how to assign a probability to it, has not gone unnoticed. We agree with [16] when they write (p. 1451) that "[…] to apply Bayes' rule, the probability that the database contains the source of the forensic DNA, assessed prior to any consideration of whether an individual in the database actually matches, becomes a crucial input in determining the (posterior) likelihood that a particular matching defendant is the source of the forensic DNA."

We conclude that there is no reason to deem the hypotheses $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$ inadmissible.

6.8. A frequentist interpretation

The original motivation to believe that the value of the evidence decreases when n comparisons are done came from a frequentist framework, making a correction for multiple testing. The original $1/(np)$-rule was motivated from this point of view. However, we have seen in our expression for the posterior odds (2.2) that it is only p_S that we need, which is the probability for the matching individual to match by chance if innocent. Hence there is no need to think of p as being relevant for all database comparisons. Even if sometimes it would be appropriate to think of p as constant for some applications in real databases, we believe it could be detrimental for understanding the problem to act as if p is always constant. If that assumption is needed to reach a conclusion, it cannot be a conclusion pertaining to the general case. The same is of course true for conclusions only valid assuming a specific prior probability for C or $\mathcal{D}$.

With this in mind we mention the paper [19] that aims to reconcile the frequentist and Bayesian points of view. Using uniform priors for C and for $\mathcal{D}$ on a population of size N, and assuming p to be constant on the database, the authors derive expressions for two different "p-values". The first one is obtained by deriving the probability to find a single but coincidental match in the database. The second one is

obtained by conditioning on there being a single match, computing the conditional probability that this match is not with the trace donor. The authors observe that the latter p-value is actually nothing but the posterior probability on $C \neq S$, and they conclude that with the conditional frequentist approach the Bayesian and frequentist answers coincide. They write that "quantification of the evidence based on frequentist (P-values) and Bayesian (posterior probabilities) points of view coincide" in that case. They go on to remark that "Simultaneously we show that the unconditional case corresponds (approximately) to the np rule and we argue that this lack of conditioning is an argument against using the np rule".

We disagree with all of these comments. We first remark that the agreement they reach between frequentist and Bayesian quantities is by construction, since they simply *define* the p-value as the Bayesian posterior probability. Secondly, p-values do not represent a quantification of evidence against a hypothesis, and the strength of the evidence in the Bayesian framework is not represented by the posterior probabilities on the hypotheses but by the likelihood ratio.

The first p-value they derive is $P(E_S, C \notin \mathcal{D})$, the probability that the trace donor is not in $\mathcal{D}$ and that nonetheless there is a single match. This can be written as $P(E_S|C \notin \mathcal{D})P(C \notin \mathcal{D})$. To take account of the fact that a match has been observed, the second p-value they propose is $P(C \notin \mathcal{D}|E_S)$. None of these have much to do with likelihood ratios. When the authors remark that the unconditional approach gives a result approximately equal to np, they refer to $P(E_S|C \notin \mathcal{D})P(C \notin \mathcal{D})$. If p is constant on the population then the first term of this expression can be reasonably approximated by np if $np << 1$. But $np << 1$ does not in any way imply that $P(C \notin \mathcal{D}) \approx 1$ which would be needed to arrive at the conclusion that the unconditional p-value is approximately np. It is the case in their framework where they assume $n << N$ and that $P(C \in \mathcal{D}) = n/N$ but this is not an assumption one needs to make. The authors seem to conclude that, because a Bayesian quantity (the $1/(np)$ likelihood ratio) is, under special circumstances, somewhat similar to a result obtained with a suboptimal frequentist approach, that gives an argument against that likelihood ratio. But our comments above indicate that this argument is not convincing.

7. Discussion and conclusions

We have first derived a general expression for the posterior odds for the situation of a unique match in a database search in (2.2). This expression is valid for database searches without having a prior specific suspicion against any database member, or for a search where there is already interest in S beforehand, or for a probable cause case where there is no one tested except S, which simply means that $\mathcal{D} = \{S\}$. We do *not* make a uniform prior assumption, since this assumption (1) is often not realistic, and (2) obscures the picture of which quantities play a role and which do not.

By making the expression for the posterior odds as general as possible, we see that only three ingredients affect the posterior directly, namely

(i) The random match probability p_S of the suspect S. This random match probability need not be the same for all people, hence the subscript S. The random match probabilities of other individuals are unimportant. For the persons who are excluded, it does not matter how they were excluded or by which probability this happens. Of course it is necessary that the fact that they are excluded is not disputed.

(ii) The prior probability $P(S = C)$ that the suspects is the criminal (or the donor of the trace) C. In case the identity of S is known, then this identity may carry information and we need not necessarily have a uniform prior. If the identity is not known, then we obtain the same results as for a uniform prior.

(iii) The probability $P(C \notin \mathcal{D})$ that the database does not contain the criminal C. This reflects that a database which is more suited for the purpose (having a larger $P(C \in \mathcal{D})$) yields a higher probability that the matches it produces are with the correct persons.

All other quantities, such as the database size, database coverage, population size, offender population size and what not, only affect the poster odds indirectly when they come into a model that gives us the three required probabilities.

The estimation or assignment of these depends on the further context, and will be different for the three scenarios that we just mentioned. They may be numerically different in different situations, but it is important to conclude that there is no principle difference between a cold case database search, a targeted search, or a probable cause case. They are all covered by our analysis.

As far as the choice of the set of hypotheses is concerned (and along with that choice, the choice of the relevant likelihood ratio), we have argued that a universal best choice does not exist, and that the actual choice one makes should depend on the original question asked and further context. In [13] it is claimed that only $S = C$ versus $S \neq C$ is relevant, but we disagree. In fact, the $1/(np)$ likelihood ratio (if we assume uniform priors) following from the pair of hypotheses $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$ is far less dangerous than using $1/p$, in the sense that a possible wrong interpretation as posterior odds will be not so harmful since the prior odds are typically of order 1.

We have also shown that arguments against the $1/(np)$ rule are unsound, since these arguments assume that one should always use this rule, even when the uniform prior assumption that underlies it is not applicable. For instance, when there is already suspicion against a suspect S and after this a database search is carried out, the uniform prior assumption is not valid, and hence the $1/(np)$ likelihood ratio is simply not applicable. We have furthermore shown that it may very well be the case that single matches from an enlarged database have a smaller probability to be matches with the actual trace donor, falsifying the idea that matches in a larger databases always lead to stronger cases against the identified suspect than smaller databases.

Our conclusions show that it is best to express the odds and likelihood ratios in their most general form, and by doing so, all fallacies and controversies disappear. The evaluation of the required probabilities depends on the model that is thought to best reflect the circumstances of the case, and depending on those, the most natural way to express the likelihood ratio may lean towards $1/p$ or towards $1/(np)$. The latter however arises only in specific circumstances. Although we realize that previous authors have had similar expectations, we hope that with the results of the current paper, the database controversy can now be considered to be resolved.

Appendix A

In this appendix we detail the mathematics involved, and provide the proofs of (2.1), (5.1), and (5.4). In the text of this article, we considered g_C as non-random, but in the formal setup it is the realization of a random variable on which we condition. So, we let G_C be a random profile, representing the profile found at the crime scene, assuming it is donated by the criminal C. For each of the persons $i = 1, \ldots, n$, we let D_i be the random profiles in the database. The loci which are typed for person i are non-random.

For each i we can write

$$G_C = (F_i, H_i),$$

where F_i denotes the profile of G_C restricted to the loci that are typed both for G_C and for D_i, and where H_i denotes the profile on the remaining loci.

Similarly we can write

$D_i = (K_i, L_i)$,

where K_i is the profile of D_i restricted to the loci that are shared with G_C, and L_i the profile on the remaining loci.

We can write the event E_i as

$E_i = \{K_i = F_i \text{ and } K_j \neq F_j \text{ for all } j \neq i,\ G_C = g_C\}$.

A.1 Proof of (2.1)

The left hand side of (2.1) is equal to

$$\frac{P(C = i|E_i)}{P(C \notin \mathcal{D}|E_i)}, \tag{A.1}$$

since conditioned on E_i, the events $C \neq i$ and $C \notin \mathcal{D}$ are equivalent. It is, therefore, enough to show that the likelihood ratio of the evidence E_i for the hypotheses $C = i$ versus $C \notin \mathcal{D}$) is equal to $1/p_i$. Thus we compute

$$\begin{aligned}\frac{P(E_i|C = i)}{P(E_i|C \notin \mathcal{D})} &= \frac{P(K_i = F_i \text{ and } K_j \neq F_j \text{ for all } j \neq i,\ G_C = g_C|C = i)}{P(K_i = F_i \text{ and } K_j \neq F_j \text{ for all} j \neq i,\ G_C = g_C|C \notin \mathcal{D})}\\ &= \frac{P(K_j \neq F_j \text{ for all } j \neq i,\ G_C = g_C|C = i)}{P(K_j \neq F_j \text{ for all } j \neq i,\ G_C = g_C|C \notin \mathcal{D})}\times\\ &\quad\times \frac{P(K_i = F_i|K_j \neq F_j \text{ for all } j \neq i,\ G_C = g_C,\ C = i)}{P(K_i = F_i|K_j \neq F_j \text{ for all } j \neq i,\ G_C = g_C,\ C \notin \mathcal{D})}\\ &= \frac{1}{P(K_i = F_i|K_j \neq F_j \text{ for all } j \neq i,\ G_C = g_C,\ C \notin \mathcal{D})}.\end{aligned}$$

So far, we have not made any assumption about independence. If we assume (as we did in the text of this article) that all profiles are independent of each other, then the last expression reduces to $1/P(K_i = f_i)$, that is, the probability that the profiles D_i and G_C agree on the overlapping loci, conditioned on $G_C = g_c = (f_i, h_i)$. This proves (2.1).

However, we note that formally, p_i must be calculated conditional on the event that the profile $G_C = g_C$ has been observed already, and on the information about the K_j for $j \neq i$. It would be rather tedious to incorporate the latter information, and in any case the effect would be essentially vanishing. Conditioning on $G_C = g_C$ is usually incorporated with the standard θ-correction. This means that RMP corresponding to person i is equal to

$P(K_i = f_i|G_C = (f_i, h_i),\ C \neq i)$,

and this expression then replaces p_i in (2.1). Similar remarks apply to the remaining proofs as well.

A.2 Proof of (5.4)

Recall that E_i is the event that we have a single match with individual $i \in \mathcal{D}$. From the odds in (A.1) we conclude that

$$P(C = i|E_i) = \frac{P(C = i)}{P(C = i) + p_i P(C \notin \mathcal{D})}. \tag{A.2}$$

This probability depends on p_i. This is only natural, since if we learn that a match has occurred with a smaller p_i we will be more confident in $C = i$ than if p_i were larger.

Next we write

$$P(C = S|E) = \sum_{i=1}^{n} P(C = i|E_i)P(E_i|E).$$

Writing

$$\Pi_i := \prod_{j \neq i} (1 - p_j)$$

we have

$P(E_i) = P(C = i)\Pi_i + P(C \notin \mathcal{D})p_i\Pi_i$.

Hence

$$\begin{aligned}P(E_i|E) &= \frac{P(E_i)}{P(E)} = \frac{P(E_i)}{\sum_{j=1}^{n} P(E_j)}\\ &= \frac{P(C = i)\Pi_i + P(C \notin \mathcal{D})p_i\Pi_i}{\sum_{j=1}^{n} (P(C = j)\Pi_j + P(C \notin \mathcal{D})p_j\Pi_j)}\\ &= \frac{\frac{P(C=i)}{1-p_i} + P(C \notin \mathcal{D})\frac{p_i}{1-p_i}}{\sum_{j=1}^{n}\left(\frac{P(C=j)}{1-p_j} + P(C \notin \mathcal{D})\frac{p_j}{1-p_j}\right)},\end{aligned}$$

assuming that the $p_i < 1$ for all i. Therefore, using (A.2) we find

$$P(C = S|E) = \sum_{i=1}^{n} \left(\frac{P(C=i)}{P(C=i) + p_i P(C \notin \mathcal{D})} \times \frac{\frac{P(C=i)}{1-p_i} + P(C \notin \mathcal{D})\frac{p_i}{1-p_i}}{\sum_{j=1}^{n} \left(\frac{P(C=j)}{1-p_j} + P(C \notin \mathcal{D})\frac{p_j}{1-p_j} \right)} \right)$$
$$= \frac{\sum_{i=1}^{n} \frac{P(C=i)}{1-p_i}}{\sum_{j=1}^{n} \frac{P(C=j)}{1-p_j} + P(C \notin \mathcal{D}) \sum_{j=1}^{n} \frac{p_j}{1-p_j}},$$

from which (5.4) follows.

A.3 Proof of (5.1)

The event $E'_{\mathcal{D}}$ can be decomposed as follows: (1) the event, denoted $E'_{\mathcal{D}}{}'$, that in $\mathcal{D}'$ there is a unique match, and (2) the event, denoted by M, that there are no matches outside $\mathcal{D}'$.

As for the first component, it follows from (5.5), applied to $\mathcal{D}'$ and p_S, that

$$\frac{P(C = S|E'_{\mathcal{D}}{}')}{P(C \neq S|E'_{\mathcal{D}}{}')} = \frac{P(C \in \mathcal{D}'|E'_{\mathcal{D}}{}')}{P(C \notin \mathcal{D}'|E'_{\mathcal{D}}{}')} = \frac{1}{n' p_S} \frac{P(C \in \mathcal{D}')}{P(C \notin \mathcal{D}')}. \tag{A.3}$$

Hence the likelihood ratio of the evidence $E'_{\mathcal{D}}{}'$ for $C \in \mathcal{D}'$ versus $C \notin \mathcal{D}'$ is $1/(n'p_S)$.

We now have

$$\frac{P(C = S|E_{\mathcal{D}}{}')}{P(C \neq S|E_{\mathcal{D}}{}')} = \frac{P(C \in \mathcal{D}|E_{\mathcal{D}}{}')}{P(C \notin \mathcal{D}|E_{\mathcal{D}}{}')} = \frac{P(C \in \mathcal{D}|E'_{\mathcal{D}}{}' \cap M)}{P(C \notin \mathcal{D}|E'_{\mathcal{D}}{}' \cap M)}$$
$$= \frac{P(E'_{\mathcal{D}}{}'|C \in \mathcal{D}')}{P(E'_{\mathcal{D}}{}'|C \notin \mathcal{D}')} \times \frac{P(M|C \in \mathcal{D}', E'_{\mathcal{D}}{}')}{P(M|C \notin \mathcal{D}', E'_{\mathcal{D}}{}')} \times \frac{P(C \in \mathcal{D}')}{P(C \notin \mathcal{D}')}.$$

The first fraction is the likelihood ratio from (A.3) and is equal to $1/(n'p_S)$. As for the second term, we first note that the conditioning on $E'_{\mathcal{D}}{}'$ can be deleted. If $C \in \mathcal{D}'$, then M occurs with some probability π, depending on the composition of $\mathcal{D} \setminus \mathcal{D}'$. If $C \notin \mathcal{D}'$, this event has probability $\pi P(C \notin \mathcal{D}|C \notin \mathcal{D}')$: the only way to have no further matches is if $C \notin \mathcal{D}$ at all, and in that case we have probability π again.

Hence we conclude that

$$\frac{P(C = S|E_{\mathcal{D}}{}')}{P(C \neq S|E_{\mathcal{D}}{}')} = \frac{1}{P(C \notin \mathcal{D}|C \notin \mathcal{D}')} \frac{1}{n' p_S} \frac{P(C \in \mathcal{D}')}{P(C \notin \mathcal{D}')}$$
$$= \frac{1}{n' p_S} \frac{P(C \in \mathcal{D}')}{P(C \notin \mathcal{D})}.$$

Note. The opinions expressed in this article, e.g., on reports for matches in databases, are those of the authors and do not necessarily reflect practices or opinions of institutes the authors have affiliations with.

References

[1] N. Research Council, DNA Technology in Forensic Science, National Academy Press, Washington DC, 1992.
[2] N. Research Council, The Evaluation of Forensic DNA Evidence, National Academy Press, 1996.
[3] A. Dawid, Comment on Stockmarr's "likelihood ratios for evaluating DNA evidence when the suspect is found through a database search", Biometrics (2001).
[4] A. Stockmarr, Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search, Biometrics (1999) 671–677.
[5] D. Balding, The DNA database controversy, Biometrics (2002).
[6] D. Balding, P. Donnelly, Evaluating DNA profile evidence when the suspect is found through a database search, J. Forensic Sci. 41 (1996) 603–607.
[7] R. Meester, M. Sjerps, The evidential value in the DNA database search controversy and the two-stain problem, Biometrics 59 (2003) 727–732.
[8] R. Meester, M. Sjerps, Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence, Law Probab. Risk (2004).
[9] A. Collins, N. Morton, Likelihood ratios for DNA identification, Proc. Natl. Acad. Sci. U.S.A. 91 (1994) 6007–6011.
[10] D. Balding, P. Donnelly, Inferring identity from DNA profile evidence, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 11741–11745.
[11] I. Evett, L. Foreman, B. Weir, Letter to the editor, Biometrics 56 (2000) 1274–1275.
[12] B. Weir, The second National Research Council report on forensic DNA evidence, Am. J. Human Genet. 59 (1996) 497–500.
[13] C. Berger, P. Vergeer, J. Buckleton, A more straightforward derivation of the LR for a database search, Forensic Sci. Int.: Genet. (2015).
[14] P. Schneider, et al., Allgemeine Empfehlungen der Spurenkommission zur statistischen Bewertung von DNA-Datenbank Treffern, Rechtsmedizin 20 (2010) 111–115.
[15] J. Wixted, N. Christenfeld, J. Rouder, Calculating the posterior odds from a single-match DNA database search, Law Probab. Risk (2019).
[16] I. Ayres, B. Nalebuff, The rule of probabilities: a practical approach for applying Bayes' rule to the analysis of DNA evidence, Stanford Law Rev. 67 (2015) 1447–1503.
[17] ENFSI Guidelines for Evaluative Reporting in Forensic Science, (2015).
[18] A. Biedermann, S. Gittelson, F. Taroni, Recent misconceptions about the 'database search problem': a probabilistic analysis using Bayesian networks, Forensic Sci. Int. 212 (2011) 51–60.
[19] G. Storvik, T. Egeland, The DNA database search controversy revisited: bridging the Bayesian-frequentist gap, Biometrics 63 (2007) 922–925.