ASOCIACIÓN NACIONAL
DE
MÉDICOS FORENSES

Spanish Journal of Legal Medicine
Revista Española de Medicina Legal

www.elsevier.es/mlegal

REVIEW

DNA test evaluation in large-scale identification cases of missing persons☆

Lourdes Prieto[a], **Yarimar Ruiz**[b], **Elías Hernandis**[c,*], **Ángel Carracedo**[a]

[a] *Grupo de Medicina Xenómica, Instituto de Ciencias Forenses, Universidade de Santiago de Compostela, Santiago de Compostela, Spain*
[b] *Comité Internacional de la Cruz Roja, Mexico City, Mexico*
[c] *Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid, Spain*

Received 28 April 2021; accepted 16 August 2021
Available online 9 August 2022

KEYWORDS
Identification;
DNA;
Probability;
Disaster Victim Identification (DVI)

Abstract The evaluation of the DNA test in massive identification cases requires the use of Bayes' Theorem to estimate the probability of identification from *a priori* data together with probabilities obtained from the DNA test itself. To apply it, one needs to specify the prior probabilities of the hypotheses. An interdisciplinary team and an identification coordinator are key stakeholders in this process. The statistical approach can be complex but there exists validated non-commercial software, such as Familias, which aid in estimating the likelihood ratios of the DNA test for the given hypotheses. Next, the posterior probabilities in massive identification events can be estimated using the one to one, PM-driven, AM-driven, or Global approaches published recently by Kling et al. (2021) which are discussed in this article. The Identification Coordinator has a key role in formulating the hypotheses of the case, in establishing the prior probabilities, the identification threshold, and in consolidating the integrated identification report together with the multidisciplinary team through the reconciliation of the case.

PALABRAS CLAVE
Identificación;
ADN;
Probabilidad;
Identificación de

Valoración de la prueba de ADN en las identificaciones a gran escala de personas desaparecidas

Resumen La valoración de prueba de ADN en casos de identificación masiva exige el uso por los peritos del teorema de Bayes para estimar la probabilidad de identificación a partir de unos

☆ Please cite this article as: Prieto L, Ruiz Y, Hernandis E, Carracedo Á. Valoración de la prueba de ADN en las identificaciones a gran escala de personas desaparecidas. Revista Española de Medicina Legal. 2022. https://doi.org/10.1016/j.reml.2021.08.003

* Corresponding author.
E-mail address: elias@hernandis.me (E. Hernandis).

víctimas en desastres masivos (DVI)

datos *a priori* a los que suman las probabilidades proporcionadas por la prueba de ADN. Para aplicarlo hace falta por una parte especificar la probabilidad *a priori* de las hipótesis de identidad que se pueden plantear de modo que un equipo multidisciplinario y la figura de un coordinador de identificaciones son clave. El abordaje estadístico puede ser complejo pero existen programas validados no comerciales, como el software Familias que facilitan las estimas de las razones de verosimilitud de la prueba de ADN para las hipótesis que se establezcan. A continuación la probabilidad *a posteriori* en eventos de identificación a gran escala se puede estimar a través de las aproximaciones *one to one*, *PM-driven, AM-driven* y *Global approach* publicadas recientemente por Kling et al. (2021) y que son descritas en detalle en este artículo. El papel del Coordinador de Identificación es clave en la formulación de las hipótesis del caso, en el establecimiento de las probabilidades *a priori*, del umbral de identificación y en consolidar el reporte integrado de identificación junto al equipo multidisciplinario a través de la reconciliación del caso.

Introduction

DNA testing is a fundamental tool in mass victim identification;[1–3] however, it is not the only test to consider in these scenarios. Traditionally, the results of genetic analysis for forensic purposes have been accompanied by a statistical assessment to measure their significance. The likelihood ratio (LR) is used to compare 2 alternative hypotheses and is generally used to assess the results of genetic testing[4,5] without taking other information into account. Its value therefore depends only on genetic data (allele frequencies,[6] degree of compatibility between the profiles to be compared, quality of the genetic profiles, etc.).

However, 2 genetic profiles being compatible (e.g., they share one allele in each marker) does not mean that they have to come from a parent and their child. In a scenario involving multiple victims, a multitude of comparisons need to be made of genetic profiles from the deceased (postmortem [PM] profiles) with genetic profiles from relatives or personal belongings of missing persons (antemortem [AM]). In this massive context, apparent genetic matches can be found without there necessarily being a family relationship between the sample donors, especially if the genetic profiles are partial. Additional information is necessary to achieve identification, as a genetic match does not mean that we have directly identified a victim. In other scenarios where DNA testing is useful (e.g., in criminal investigations), it is the judicial authority who will incorporate the value of the LR with other non-genetic evidence to reach a final conclusion. But in mass identification cases, experts from different areas usually decide on the final result; reporting an LR value is not sufficient for that purpose, as it does not report the final likelihood of identification, the so-called posterior probability. This comes from Bayesian statistics. However, this theorem is not easy to apply in that it allows combining non-genetic information about the case with genetic information. Therefore, we need to establish what the prior probability of identification is before DNA analysis.

Large-scale identification processes occur as a consequence of mass disappearances related to armed conflicts, natural disasters, violations of human rights, or international humanitarian law, among others. These cases can overwhelm local forensic services, requiring procedures and practices to be adapted.[7–9] This adaptation ranges from capacity building and management of corpses and forensic information to hypotheses of identity, statistical approach, and assessment of DNA evidence.

Today, in many medico-legal systems that perform large-scale identifications of victims and missing persons, genetic experts still work in isolation from other forensic specialties, and continue to use default prior probability values. One of the most common errors is to apply a value of .5 probability (50%), classically applied in paternity tests, although of very debatable use in judicial evidence. In this article, we address the importance of understanding identification as a multidisciplinary task and the risks of not using adequate prior probability in mass identification cases.

Bayes theorem in cases of mass identification

To calculate an LR, at least 2 hypotheses about the facts need to be stated. The hypotheses must be mutually exclusive (if one is true, the other must be false), for example:

H_1 = The skeletal remains belong to a child of P and M.

H_2 = The skeletal remains do not belong to a child of P and M.

The LR measures the probability of having attained the genetic profiles that were obtained in the skeletal remains, and in P and M (whatever this result is; i.e., whether they share alleles or not) under the 2 hypotheses mentioned. And it is formulated as follows (1):

$$LR = \frac{P(E|H_1)}{P(E|H_2)} = \frac{\text{probability of obtaining the results of the genetic test assuming that H1 is true}}{\text{probability of obtaining the results of the genetic test assuming that H2 is true}} \qquad (1)$$

where E = evidence (the genetic result in the simples analysed), P = probability and the symbol "|" = assuming.

The results obtained in the analysis may support one or another hypothesis, depending mainly on whether or not the genetic profiles are compatible with each other and on the frequency of shared alleles. The LR can take values between 0 and infinity. Under the usual assumption of independence between markers, the LR for the genetic profile is obtained by multiplying the LRs of each marker. If the value is below 1, the genetic results are more supportive of the denominator hypothesis, usually the one that assumes no relatedness. If the LR value is 1, the genetic results do not support one hypothesis over the other, i.e., the genetic evidence is neutral. A clear example of this situation is when no results have been obtained for some genetic marker, in which case the LR for this marker is 1. If the LR value is greater than 1, the genetic results are more supportive of the numerator hypothesis, and more so the higher the value than 1.[10]

The LR value obtained in the DNA test can be incorporated with other non-genetic information to calculate the posterior probability of identification by applying Bayes' theorem. Its formulation is simple when there are only 2 hypotheses to test (2):

$$P(H_1|E) = \frac{P(E|H_1)\,P(H_1)}{P(E|H_1)\,P(H_1) + P(E|H_2)\,P(H_2)} \qquad (2)$$

$P(H_1 \mid E)$ is the posterior probability of identification, i.e., the probability that is ultimately of interest to the investigator or judicial authority hearing the case. $P(H_1)$ is the prior probability of hypothesis 1 and $P(H_2)$ is the prior probability of hypothesis 2. $P(E \mid H_1)$ and $P(E \mid H_2)$ were defined earlier in the LR. Therefore, the posterior probability is a combination of the LR and the prior probability, and answers what the probability of a hypothesis is given the data (the genetic and non-genetic evidence). Bayes' theorem can also be expressed with odds. Annex 1 of the supplementary material shows how the 2 forms of the theorem are related.

Returning to the form defined in (2) of the theorem, if $P(H_1)$ is .5; and therefore $P(H_2)$ will also be .5, the formula simplifies (3):

$$P(H_1|E) = \frac{P(E|H_1)}{P(E|H_1) + P(E|H_2)} \qquad (3)$$

And if we divide the numerator and denominator of (3) by $P(E \mid H2)$, we get the famous formula commonly used in paternity cases to calculate the posterior probability (4):

$$P(H_1|E) = \frac{LR_1}{LR_1 + 1} \qquad (4)$$

However, in cases of mass identification, we are confronted with many more hypotheses; e.g., in the case of mass grave finds we must test a set of hypotheses and not just 2 as in isolated identification cases. Kling et al.,[6] give the example of a simple scenario of a mass grave in which 3 bodies are found (V1, V2, and V3) and 3 families are looking for 3 missing persons (F1, F2, and F3). The hypotheses to be tested would be:

H_1 = V1 belongs to family F1.
H_2 = V2 belongs to family F1.
H_3 = V3 belongs to family F1.
H_4 = Another unknown victim belongs to family F1.

H_4 contemplates that the missing person that family F1 is looking for is not actually in the set of human remains investigated. In the case of families F2 and F3, the hypotheses would be the same as for F1, but replacing F1 by F2 and F3, respectively.

In cases with multiple hypotheses, the application of Bayes' theorem to calculate the posterior probability is formulated as follows (5):

$$P(H_i|E) = \frac{P(E|H_i)\,P(H_i)}{\sum_{j=1}^{k} P(E|H_j)P(H_j)} \qquad (5)$$

Where $P(H_i)$ is the prior probability of identification for the hypothesis of interest and the summation $\sum_{j=1}^{k} P(E \mid H_j)P(H_j)$ indicates the sum of the hypotheses 1 to k. In our example, if the hypothesis of interest were H_1, the formula (5) would be expressed (6):

$$P(H_1|E) = \frac{P(E|H_1)\,P(H_1)}{P(E|H_1)\,P(H_1) + P(E|H_2)\,P(H_2) + P(E|H_3)P(H_3) + P(E|H_4)P(H_4)} \qquad (6)$$

We can simplify this formula (6) by dividing the numerator and denominator by the term $P(E \mid H_4)$. This will give us (7):

$$P(H_1|E) = \frac{LR_1 P(H_1)}{LR_1 P(H_1) + LR_2 P(H_2) + LR_3 P(H_3) + LR_4 P(H_4)} \qquad (7)$$

The value of LR_4 is 1, as it was obtained by dividing $P(E \mid H_4)$ / $P(E \mid H_4)$. And if in addition the value of the prior probability is equal for all hypotheses, we can simplify further. Suppose that in this case $P(H_1) = P(H_2) = P(H_3) = P(H_4) = 1/4$ (see next section on prior probabilities). We then obtain (8):

$$P(H_1|E) = \frac{LR_1}{LR_1 + LR_2 + LR_3 + LR_4} \qquad (8)$$

So far, we have seen that there is a large difference between applying Bayes' theorem in scenarios with only 2 hypotheses and in scenarios with k hypotheses. We will now focus on what happens in massive scenarios when Bayes' theorem is applied. For this, we refer again to Kling et al.,[6] who describe different approaches to setting the hypotheses. In this section, we only describe the different ways of approaching the problems, but in section 4, the reader will see how each of them is applied in more detail with a simple example. The different approaches include:

a) One-to-one: this approach is usually applied in non-mass scenarios, where there is a presumed identity, and a single *PM* profile is compared with a single pedigree. However, it is also applied in mass scenarios, when the *PM* profile is compared with all pedigrees, and finally only 2 hypotheses are considered when Bayes'. theorem is applied. The fact that only 2 hypotheses are considered does not imply that the value of the prior probability has to be .5; in section 4, the reader can see an *ad hoc* example. But we are now interested in the hypotheses:
   H_1 = Victim 1 is related to family F1.
   H_2 = Victim 1 is not related to family F1.
b) PM-driven: in this approach the hypotheses are set considering the victims, i.e., the question to be answered is: which family is related to this particular victim? This approach attempts to provide the most likely family for each victim, and does not consider that 2 different victims could be associated with the same family. The hypotheses would therefore be:
   H_1 = Victim 1 is related to family F1.
   H_2 = Victim 1 is related to family F2.
   …
   H_N =Victim 1 is related to family NF, where N is the number of families searching for their loved ones.
   H_{N+1} =Victim 1 is not related to any of the N families.
c) AM-driven*:* in this approach, the hypotheses are set considering the families, i.e., the question to be answered is: which victim is related to this particular family? It is not considered that the same victim could be associated with 2 different families. And the hypotheses are set:
   H_1 =Family F1 is related to V1.
   H_2 = Family F1 is related to V2.
   ….
   H_M = Family F1 is related to victim M, where M is the number of victims.
   H_{M+1} =Family F1 is not related to any of the M victims.

The one-to-one approach is the most usual due to its simplicity, but it is not without problems. By considering only 2 hypotheses, e.g., V1 is related to F1 vs. V1 is not related to F1; the possibility that victim 1 is related to F2, F3, etc. is not considered. In other words, although the comparison has been made with all profiles, in the statistical assessment, we are only taking into account one of the results of the comparison.

However, in the PM-driven and AM-driven approaches, all comparisons are taken into account in the statistical assessment of the results, as several hypotheses are considered at the same time. However, both approaches have the disadvantage that the statistical evaluation is performed sequentially, i.e., each hypothesis is considered one at a time in the numerator of the formula (5). This can have a great influence in cases where there is more than one missing person in a family. For example, if a father P and a son H are missing and only H's mother is available as reference sample, father P cannot be identified unless H's genetic profile is elevated to the status of reference profile. Therefore, the identification in this case is sequential, first identifying the son, and once that identification is certain, the father is identified.

The solution to this problem is offered by the authors themselves, through what is termed the *global approach*, although this approach requires a great computational effort. This method considers all the possible solutions that can be obtained in the mass identification project; e.g., no victim is identified, only 1, only 2, etc. An example of possible S solutions of a case using this global approach is shown in Fig. 1. The main advantage of the method is that it is not possible to assign the same victim to different families or to assign different victims to the same family in the same position in the pedigree, as the approach considers all the results at once and not sequentially as in the other 3 approaches. There being 2 missing persons within the same family is not a problem with the global approach, as in the example above where P and H were missing, it is not necessary to upgrade H's genetic profile to an AM profile, but both can be identified at the same time.

Fortunately, there is free software to help us use any of these approaches to the problem. The Familias[11] programme provides the necessary tools to define the hypotheses according to the AM-driven, PM-driven, and one-to-one approaches. Recently, Vigeland and Egeland[12] have developed a package for R called "dvir" to approach the problem from a global perspective (global approach, called joint approach in their publication).

The prior probability of identification

The prior probability of identification is the belief that an unidentified deceased person may belong to a certain family. It is a subjective probability, but can be based on certain data.

One of the most common mistakes when setting the value of the prior probability of identification in mass scenarios is to apply the same value used in paternity tests (50%). Usually, the expert geneticist does not have any non-genetic information from the paternity case to establish a prior probability value and, almost more out of tradition than accuracy, experts sometimes use the value of 50%.

It is believed that a probability of 50% is uninformative, as the probability of paternity is given the same value as the probability of non-paternity (100%-50% = 50%). However, this is not the case. By applying a prior probability value of 50%, we are assigning a 50% probability of paternity to the alleged father, and the remaining 50% is distributed among all the males in the population of interest, and therefore a higher probability is assigned to the alleged father than to each of the other possible fathers.[6] Also in cases of judicial paternity, it is the judge who should set the value of the prior probability, and in this sense, the expert is overstepping their role.[13]

It should also be noted that if we apply a prior probability of 50% and contrast 2 hypotheses (related vs. unrelated), in cases of mass identification, we only need a value of LR = 1000 to reach a posterior probability of 99.9%, and with an LR = 10 000 we will reach a value of 99.99%. These posterior probability values can lead a judge to determine the identification with high certainty, however, a value of LR = 1000 can be achieved even if the genetic match is not a true genetic match.

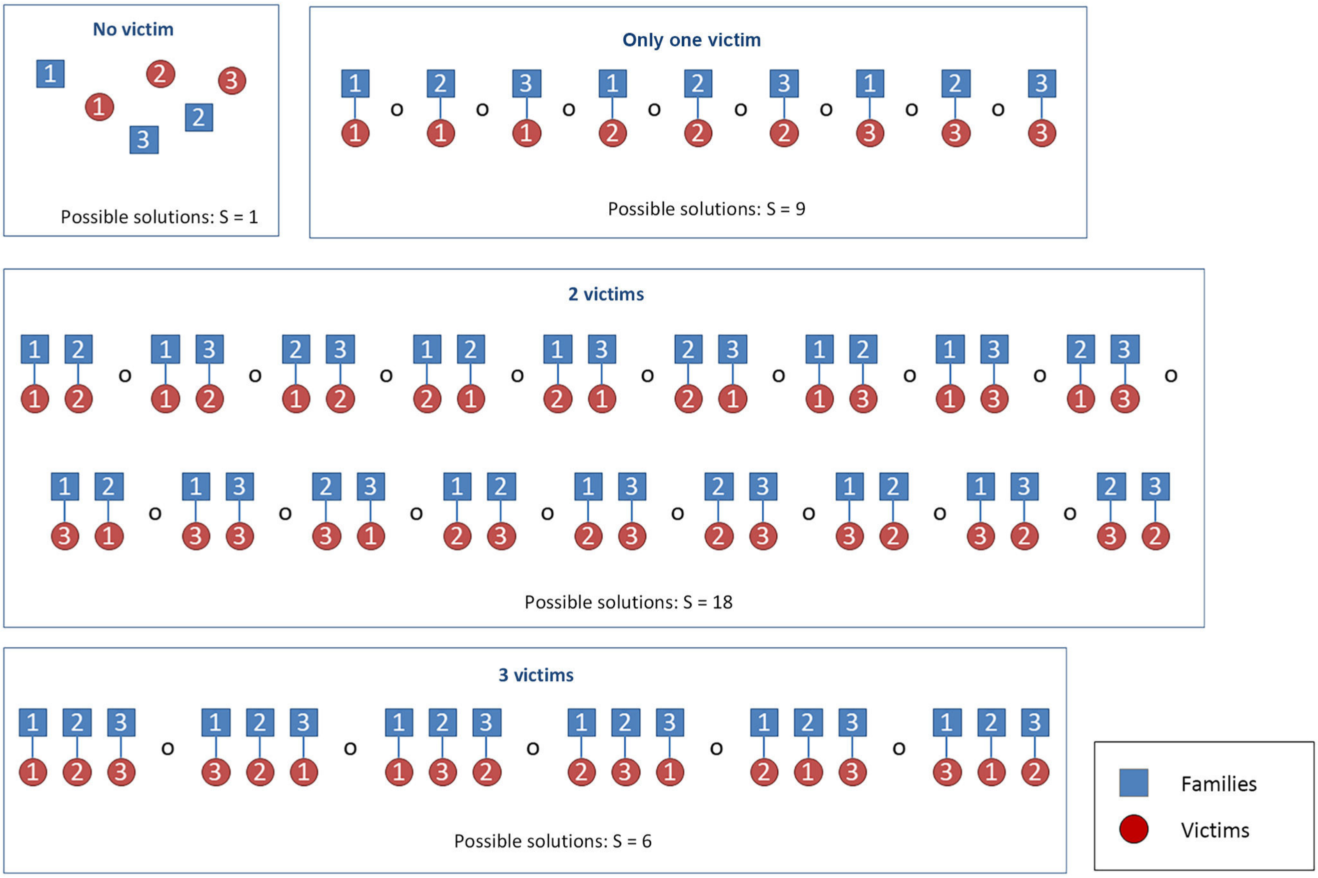
No victim
Possible solutions: S = 1
Only one victim
Possible solutions: S = 9
2 victims
Possible solutions: S = 18
3 victims
Possible solutions: S = 6
Families
Victims

Fig. 1 Global approach. Scenario with 3 victims and 3 families of 3 missing persons. All the results that can be obtained (S) are shown, taking into account that it is possible that no match can be obtained: 1. No genetic match. 2. Only one match (F1 with V1, or F2 with V1, …, or F3 with V3). 3. Two genetic matches (F1 with V1 and F2 with V2, or F1 with V1 and F3 with V2, …, or F2 with V3 and F3 with V2). 4. Three genetic matches (F1 with V1, F2 with V2 and F3 with V3, …, or F1 with V3, F2 with V1 and F3 with V2). 5. The total number of possible S solutions S is 1 + 9 + 18 + 6 = 34. 6. Translated from Klingt al.[6]

As we mention above, a match between genetic profiles is not synonymous with identification. Let us imagine that in the analysis of a mass event, we find a match between the *PM* genetic profile obtained from skeletal remains and the *AM* genetic profile of a 10-year-old child looking for his father. It could happen that this compatibility is false, especially if there are many individuals involved or the *PM* genetic profiles are partial. If the anthropologist determines that the skeletal remains in question come from a sub-adult individual, the match can clearly be given as false, as the prior probability in this case would have a value of 0, since the age of the individuals precludes their having a parent–child relationship. Unfortunately, it is not possible to reach such clear conclusions in all cases.

Traditionally, mass scenarios have been classified as closed or open. In closed cases, the number of missing people is usually quantifiable, and their identity is known. The typical example is a plane crash where there is a list of passengers, although this may not be 100% reliable. In open cases, the actual number of missing persons may be unknown, i.e., their estimation is more uncertain. In real life, however, it is more common to find mixed cases, i.e., cases where the identity of some of the missing persons is known, but not all of them.[14]

If there is no other information, the number of missing persons is usually taken into account to establish the prior probability of identification.[13] Thus, in the case mentioned in the previous section, where the number of missing persons was 3, the prior probability could be set at 1/4 (1/[n° missing persons + 1]), in order to assume that the person we are looking for may not be in the grave under investigation. This approach assumes that all victims have the same prior probability of identification (flat priors). And the probability of non-identification for each individual would be 3/4. The reader can find examples in the exercises by the Spanish and Portuguese Speaking Working Group of the International Society for Forensic Genetics (GHEP-ISFG).[15,16]

The prior probability of identification can be redefined by other experts such as anthropologists, taking into account other characteristics such as sex or age. If in the case of the 3 missing persons, 2 are women, and the human remains to be identified still present physical characteristics that make their female sex visible, one could then redefine the above prior probability to 1/3 (1/[n° missing women + 1]).

However, we should not forget that there can also be uncertainty in determining the age or sex of human remains, especially if we are faced with a scenario of incomplete skeletal remains. For example, age estimation in *PM* remains may not be very accurate depending on the age of the individual, as occurs when the morphology of the pubic symphysis is considered in individuals over 40 years of age.[17]

Other characteristics such as tattoos, marks, prostheses can be more difficult to quantify. The statements of witnesses who may have been present at the events can also be considered, but they are also difficult to quantify and may even be unreliable. [18] In Budowle et al.,[19] the reader can find a discussion of the problems that can arise when 100% credibility is given to witness statements, other information (location of the grave, PM interval, demographic variables) or when potential dependence between the different variables used to establish the prior probability of identification has not been taken into account.

To define the prior probability, Gill etal.,[20] looked at the work conducted on the dental pieces of the skeletal remains that were later identified as belonging to the Romanov family. An example of how to apply Bayesian thinking to identification can also be found in King et al.[21] The reader will see that sometimes non-genetic information is difficult to quantify. The frequency of various characteristics or conditions is not precisely known, and therefore the prior probability value is difficult to estimate.[22]

Therefore, in conclusion, the estimation of prior probability considering the number of missing persons is usually conservative and defensible. Estimates can be made even when the exact number is not known. For example, in the terrorist attack on the World Trade Centre in 2001, a prior probability of 1/3000 was used because the number of missing persons was estimated at slightly less than 3000.[23] This estimate (estimated number of missing persons) is the most appropriate estimate in large-scale identification cases, for DNA laboratories to report the list of possible identifications that can be corroborated or ruled out after including non-genetic information.

Calculating the posterior probability: an example and different approaches

In this section, we return to the simple example of Kling et al.[6] where we have a scenario with 3 victims (V1, V2, and V3) and 3 families (F1, F2, and F3) each searching for a missing person (MP1, MP2, and MP3). Based on the heading on the value of the prior probability, we will say that for each victim V_i, the value of this probability is 1/4, thus considering the possibility that the victim we are looking for is not in the investigated scenario.

Let us assume that after the genetic analysis and comparing the profiles we obtain the results described in Table 1. The LR value has been calculated taking into account each identity hypothesis against the hypothesis that considers the absence of family relationship (in our example, H_4). Strictly speaking, the LR value should not be 0 if we consider the possibility of mutation, but we approximate it to 0 to simplify the calculations.

Table 1 Practical assumption of a scenario with 3 victims and 3 missing persons. The results are shown after comparing the postmortem genetic profiles with the antemortem genetic profiles

Victim	Family	LR
V1	F1	10^6
V1	F2	10
V1	F3	0
V2	F1	500
V2	F2	0
V2	F3	0
V3	F1	0
V3	F2	0
V3	F3	0

We now show how posterior probabilities are calculated with the one-to-one, PM-driven, and AM-driven approaches.

One-to-one

With this method we only consider 2 hypotheses, which in view of the results would be:

H_1 = V1 is related to family F1.

H_2 = V1 is not related to family F1.

The results obtained in the other comparisons, i.e., other candidate matches. The posterior probability would be calculated according to the formula (2), considering that if the prior probability of identification has a value of 1/4, the probability of non-identification will be 1-1/4 = 3/4. But for simplicity, as we already have LR values, we can use the formula (7) with only 2 hypotheses:

$$P(H_1|E) = \frac{LR_1\ P(H_1)}{LR_1\ P(H_1) + LR_2\ P(H_2)} = \frac{10^6 * 1/4}{\left(10^6 * 1/4\right) + (1 * 3/4)}$$
$$= .999997$$

In Annex 2 of the complementary material, you can see the same example, but applying Bayes' theorem in the form of a bet.

PM-driven

In this approach, a victim is compared with each of the families. In our case we would say, e.g., for V1:

H_1 = V1 is related to family F1.

H_2 = V1 is related to family F2.

H_3 = V1 is related to family F3.

H_4 = V1 is not related to any of the 3 families.

The posterior probability would be calculated according to the formula (5), but we can use the simplification,[8] as the prior probability values are equal. This gives us:

$$P(H_1|E) = \frac{LR_1}{LR_1 + LR_2 + LR_3 + LR_4} = \frac{10^6}{10^6 + 10 + 0 + 1}$$
$$= .999989$$

Annex 3 of the complementary material shows all the posterior probabilities for each hypothesis and for each victim.

AM-driven

In this approach, a family is compared to each of the victims. The hypotheses in our example would be, for F1:

H_1 = Family F1 is related to V1.

H_2 = Family F1 is related to V2.

H_3 = Family F1 is related to V3.

H_4 = Family F1 is related to another unknown victim.

The posterior probability would be calculated according to the formula (5), but we can use the simplification,[8] as the prior probability values are the same. This gives us:

$$P(H_1|E) = \frac{LR_1}{LR_1 + LR_2 + LR_3 + LR_4} = \frac{10^6}{10^6 + 500 + 0 + 1} = .9995$$

Annex 3 of the complementary material shows all the posterior probabilities for each hypothesis and for each family.

Global approach

As shown in Fig. 1, in this case all possible solutions are considered at the same time: no victim is identified, only 1 victim is identified, 2 are identified, and 3 are identified. In our example, there are a total of 34 possible solutions, i.e., 34 possible.

In our example, there are a total of 34 possible solutions; i.e., 34 different hypotheses. This give us:

H_1 = V1, V2, and V3 are neither MP1, MP2, nor MP3.

H_2 = V1 is MP1, V2 and V3 are neither MP1, MP2, nor MP3.

H_3 = V2 is MP1, V1 and V3 are neither MP1, MP2, nor MP3.

...

H_{11} = V1 is MP1 and V2 is MP2, V3 is neither MP1, MP2, nor MP3.

...

H_{29} = V1 is MP1, V2 is MP2 and V3 is MP3.

...

H_{34} = V1 is MP3, V2 is MP2 and V3 is MP1.

Table S3 in Annex 3 of the supplementary material shows all these possibilities, respecting the values of the LR used in the other approaches and adding values of LR = 0 for the cases that were not considered in the sequential approaches.

Applying the simplification of the formula (8), from this table, the posterior probabilities of individual identification can be calculated, taking into account each hypothesis involving a particular victim. For example, for V1, the numerator considers all LRs in which MP1 appears in column V1 of the Table. The denominator is the sum of all RLs. This gives us:

$$P(V1 = MP1|E) = \frac{LR_2 + LR_{11} + LR_{12} + LR_{17} + LR_{18} + LR_{29} + LR_{30}}{LR_1 + \ldots + LR_{34}} = \frac{10^6 + 0 + 0 + 0 + 0 + 0 + 0}{1 + 10^6 + 10 + 0 + 500 + 0 + \ldots + 0} = .999489261$$

Table S4 in Annex 3 of the supplementary material shows all individual posterior probabilities for each victim.

Role of the identification coordinator in the analysis of large-scale cases

While in criminalistics or paternity cases, it is the judge who determines the prior probability, in large-scale disappearances the identification coordinator should play a key role in formulating the hypothesis of the case, establishing the prior probabilities, the identification threshold, and consolidating the integrated identification report together with the multidisciplinary team through the reconciliation of the case. The Public Prosecutor's Office does not act in all countries when a disappearance occurs. In countries where disappearances are continuous over time (e.g., Mexico), or in an armed conflict of long duration (e.g., Colombia), competent authorities called National Search Commissions have been established, which play a crucial role in the investigation of cases, and therefore in the establishment of prior probabilities, taking into account at all times the suggestions of the multidisciplinary team led by the identification coordinator.

The threshold for reporting a genetic match should be set in a way that strikes a balance between maximising identifications and minimising false identifications.[24] Genetic laboratories usually set posterior probability thresholds of 99.90%, 99.95%, or 99.99% for reporting a particular genetic match to the identification coordinator. A value of 99.95% posterior probability means that they are willing to assume that one in 2000 genetic matches is false.[22] Similarly, with a value of 99.90% it is assumed that 1 in 1000 matches will be false; and for 99.99%, there is expected to be 1 false match in 10 000. In exceptional cases, it may not be necessary to reach the posterior probability threshold to report a match of genetic data to the identification coordinator. For example, imagine a totally closed event in which the presumed identity of 4 individuals located in a grave is known, although it is unknown to whom each of the bodies belongs. If only mitochondrial DNA data has been obtained (therefore with moderate LR values), but the 4 haplotypes are totally different and each of them matches each of the 4 families searching for their loved one, this information that can lead to the identification of the individuals should be made available to the coordinator.

Discussions led by the identification coordinator on the reconciliation of a case are an essential step during the identification process where all information pertaining to the unidentified deceased person (PM) is compared and evaluated with all information pertaining to the missing person (AM). This is to be able to: i) reach a formal identification, or ii) determine the steps required to achieve an identification.

In this regard, before the reconciliation meeting takes place (and especially in cases of skeletal remains), the identification coordinator, once they receive a preliminary match of the case from the forensic genetics specialists, then corroborates the AM and contextual information checking consistency with the hypothesis of identity. Likewise, all specialists involved in the phases of the PM analysis process must confirm that all efforts, site visits, investigations, and collections of case information and other diligences have been completed, and then proceed to prepare the case reconciliation plan.

In addition to the archaeological, anthropological, odontological, and medical information that is part of the PM information, the following genetic information is discussed in reconciliation meetings for cases associated with large-scale disappearances: minimum number of individuals per DNA (in mixed cases), analysed samples that reached a profile suitable for hereditary–biological kinship analysis and/or for intra-skeletal association, biological reference samples belonging to relatives used in the analysis, obtained likelihood ratio (LR), prior probability used during the case and its basis, posterior probability obtained after the analysis, available family tree of the missing person that was not used in the comparison and assessment of the presence of other missing persons in the same family tree.

If consistency is observed after evaluation and comparison of PM data obtained from the investigations corresponding to each of the specialties (including genetics) in relation to the AM information (including the circumstances of the disappearance and other investigative information), and no unexplained discrepancies are found, then the identification is concluded. If, however, discrepancies are found in the reconciled information, additional recommendations are made (e.g., confirmation with additional reference samples or more markers, additional investigative information, additional anthropological analysis, etc.).

Conclusions and recommendations of forensic genetic analysis in large-scale identification cases

- Large-scale identification processes require procedures to be adapted including adjustments in information management. Expert geneticists need specialist training in forensic statistics and appropriate software handling.
- Part of these DNA procedures include posing multiple hypotheses of identity as well as the estimation of prior probability based on context information. It is recommended that in the absence of further context information, the prior probability can be estimated from the number of missing persons associated with an event.
- Bayes' theorem is one of the tools that provides a comprehensive analysis in large-scale identifications, because multiple hypotheses can be integrated, including the possibility that the victim being sought is not in the investigated scenario. Applying this theorem, both genetic and non-genetic information can be considered. The posterior probability in large-scale identification events can be estimated using the one-to-one, PM-driven, AM-driven, and global approach.
- For the correct statistical approach, there are free technological tools such as the Familias software that works in conjunction with packages developed in the R programming language.
- In large-scale identification events, the identification coordinator must play a key role in formulating the hypotheses of the case, establishing the prior probabilities, the identification threshold, and consolidating the integrated identification report together with the multidisciplinary team through the reconciliation of the case.

Funding

None.

Conflict of interests

The authors have no conflict of interests to declare.

Acknowledgements

We would like to thank Thore Egeland for the valuable comments received during the preparation of this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.remle.2021.08.002.

References

1. Vallejo G, Martín P, Alonso A. La identificación genética en grandes catástrofes. In: Barbería, editor. Catástrofes: identificación de víctimas y otros aspectos médico-forenses. Elsevier; 2015.
2. Vullo C. Identificación de desaparecidos en gran escala: grandes catástrofes (DVI) y desapariciones en contextos de crisis humanitaria (MPI). In: Crespillo Márquez M, Barrio Caballero P, editors. Genética forense. Del laboratorio a los tribunales. Madrid: Díaz de Santos; 2019. p. 405–24.
3. Vallejo G, Alonso A. La identificación genética en grandes catástrofes: avances científicos y normativos en España. Rev Esp Med Legal. 2009;35(1):19–27.
4. Prieto L, Carracedo A. El valor de la prueba de ADN. In: Crespillo M, Barrio P, editors. Genética Forense: del laboratorio a los tribunales. Díaz de Santos; 2019. p. 445–70.
5. Carracedo A, Prieto L. Más allá del efecto CSI: claves para una buena comunicación de la genética forense. Mètode Sci Stud J.. 2018;97:57–63.
6. Kling D, Egeland T, Tillmar A, Prieto L. Mass identifications. Statistical methods in forensic genetics, ISBN 9780128184233.1st ed. Elsevier, Acedemis Press; 2021.
7. Real Decreto. 32/2009 d1depeqsaePndaMfydPCescvm. Boletín Oficial del Estado. [Online]. 2009 [cited 2021 junio 12. Available from: https://www.boe.es/boe/dias/2009/02/06/pdfs/BOE-A-2009-2029.pdf.
8. Interpol. Disaster Victim Identification (DVI). [Online] cited. junio 12. Available from: https://www.interpol.int/How-we-work/Forensics/Disaster-Victim-Identification-DVI2021.
9. Morgan O, Tidball-Binz M, van Alphen D. La gestión de cadáveres en situaciones de desastre: Guía práctica para equipos de respuesta. [Online] cited 2021 junio 12. Available from:: https://www.icrc.org/es/doc/assets/files/other/icrc-003-0880.pdf2009.
10. Egeland T. Properties of Likelihood Ratios. [Online]. 2017 [cited 2021 abril 19. Available from: https://familias.name/mty/Day1-Part3-TE-LR-mty.pdf.
11. Kling D, Tillmar AO, Egeland T. Familias 3 - Extensions and new functionality. Forensic Sci Int Genet. 2014 Nov;13:121–7.
12. Vigeland MD, Egeland T. Joint DNA-based disaster victim identification. Res Square. 2021 enviado; DOI:10.21203/rs.3.rs-296414/v1.
13. Prinz M, Carracedo A, Mayr WR, Morling N, Parsons TJ, Sajantila A, et al. DNA Commission of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role

of forensic genetics for disaster victim identification (DVI). Forensic Sci Int Genet. 2007;1:3—12.

14. Comité Internacional de la Cruz Roja (CICR). Directrices para el uso de la genética forense en investigaciones sobre derechos humanos y derecho internacional humanitario. Ginebra: CICR. https://www.icrc.org/es/publication/directrices-uso-genetica-forense-investigaciones-ddhh-dih; 2021.
15. Vullo C, Romero M, Catelli L, Šakić M, Saragoni V, Jimenez MJ, et al. GHEP-ISFG collaborative simulated exercise for DVI/MPI: Lessons learned about large-scale profile database comparisons. Forensic Sci Int Genet. 2016;21:45—53.
16. Vullo C, Catelli L, Ibarra A, Papaioannou A, Álvarez J, López-Parra A, et al. Second GHEP-ISFG exercise for DVI: "DNA-led" victim's identification in a simulated air crash. Forensic Sci Int Genet. 2021;53, 102527.
17. Brooks S, Suchey JM. Skeletal age determination based on the os pubis: a comparison of the Acsádi-Nemeskéri and Suchey-Brooks methods. Hum Evol. 1990;5:227—38.
18. Scheck B, Neufeld P, Dwyer J. Actual Innocence. New York: New American Library; 2000.
19. Budowle B, Ge Y, Chakraborty R, Gill-King H. Use of prior odds for missing persons identifications. Investig Genet. 2011;2:15.
20. Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, et al. Identification of the remains of the Romanov family by DNA analysis. Nat Genet. 1994;6(2):130—5.
21. King TE, Fortes GG, Balaresque P, Thomas MG, Balding D, Maisano Delser P, et al. Identification of the remains of King Richard III. Nat Commun. 2014;5:5631.
22. Parsons T, Huel R. DNA and missing persons identification: practice, progress and perspectives.In: Amorim A, Budowle B, editors. Handbook of Forensic Genetics. London: World Scientific; 2017. p. 337—76.
23. Biesecker LG, Bailey-Wilson JE, Ballantyne J, Baum H, Bieber FR, Brenner C, et al. DNA identifications after the 9/11 World Trade Center attack. Science. 2005 Nov;310(5751):1122—3.
24. Brenner CH, Weir BS. Issues and strategies in the DNA identification of World Trade Center victims. Theor Popul Biol. 2003;63:173—8.