OSAC 2021-S-0021 Forensic Autosomal STR DNA Statistical Analyses - General Protocol, Protocol Verification, and Case Record Requirements

Human Forensic Biology Subcommittee
Biology Scientific Area Committee
Organization of Scientific Area Committees (OSAC) for Forensic Science

Draft OSAC Proposed Standard

OSAC 2021-S-0021 Forensic Autosomal STR DNA Statistical Analyses - General Protocol, Protocol Verification, and Case Record Requirements

Prepared by
Human Forensic Biology Subcommittee
Version 1.0
July, 2021

Disclaimer:

This OSAC Proposed Standard was written by the Human Forensic Biology Subcommittee of the Organization of Scientific Area Committees (OSAC) for Forensic Science following a process that includes an open comment period. This Proposed Standard will be submitted to a standards developing organization and is subject to change.

There may be references in an OSAC Proposed Standard to other publications under development by OSAC. The information in the Proposed Standard, and underlying concepts and methodologies, may be used by the forensic-science community before the completion of such companion publications.

Any identification of commercial equipment, instruments, or materials in the Proposed Standard is not a recommendation or endorsement by the U.S. Government and does not imply that the equipment, instruments, or materials are necessarily the best available for the purpose.

To be placed on the OSAC Registry, certain types of standards first must be reviewed by a Scientific and Technical Review Panel (STRP). The STRP process is vital to OSAC's mission of generating and recognizing scientifically sound standards for producing and interpreting forensic science results. The STRP shall provide critical and knowledgeable reviews of draft standards or of proposed revisions of standards previously published by standards developing organizations

(SDOs) to ensure that the published methods that practitioners employ are scientifically valid, and the resulting claims are trustworthy.

The STRP panel will consist of an independent and diverse panel, including subject matter experts, human factors scientists, quality assurance personnel, and legal experts, which will be tasked with evaluating the proposed standard based on a comprehensive list of science-based criteria.

For more information about this important process, please visit our website at: https://www.nist.gov/topics/organization-scientific-area-committees-forensic-science/scientific-technical-review-panels

Foreword

Detailed and comprehensive protocols are needed to ensure that appropriate statistical calculations are performed consistently for evidentiary DNA profiles. These calculations are provided to aid in the assessment of an inclusion or positive association of a DNA profile with the profile of a known individual. Specific requirements for a laboratory's protocol for performing statistical analyses, its verification, and requirements for case record documentation are provided. These requirements include documentation of when statistical calculations shall be performed and when they are not required; descriptions of the statistical methods available for use in the laboratory and relevant supporting information for their use; the use of assumptions in the calculations; documentation of the data used and relevant information for the calculations performed; and documented verification and consistency of use of the protocol in the laboratory.

This standard addresses general requirements for calculations commonly performed in forensic DNA testing laboratories. These may include the likelihood ratio (LR), the random match probability (RMP), and the combined probability of inclusion/exclusion (CPI/CPE). This document applies to any manual calculations or software using fixed formulae and/or continuous or semi-continuous methods. This document applies to calculations resulting from the comparison of DNA profiles for identity testing (i.e., could the DNA have come from the same source?) as well as biological relationship testing (i.e., could the individuals be related?). While this standard applies directly to testing performed using the polymerase chain reaction (PCR) amplification of autosomal loci having short tandem repeats (STR), many of the general requirements may also apply to other types of DNA testing and analysis. Additional information regarding the application of and specific requirements for the various statistical calculation methods routinely used in forensic DNA testing laboratories may be found in Annex A and the Bibliography (Annex B).

This standard is to be used in conjunction with the FBI's *Quality Assurance Standards for Forensic DNA Testing Laboratories* [1] and the following ANSI/ASB Standards: (1) ANSI/ASB Standard 018, Validation Standards for Probabilistic Genotyping Systems, First Edition, 2020; (2) ANSI/ASB Standard 020, Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol, First Edition, 2018; (3) ANSI/ASB Standard 040, Standard for Forensic DNA Interpretation and Comparison Protocols, First Edition, 2019; *(4) ANSI/ASB Standard 41, Assigning Propositions for Likelihood Ratios in Forensic DNA Interpretations, First Edition, 2020 and (5) ANSI/ASB Standard 123, Routine Internal Evaluation of a Laboratory's Interpretation and Comparison Protocol* as well as any current or future standards or recommendations that provide guidance for the appropriate use of specific statistical calculation methods and software.

Keywords: *statistics, statistical analysis, protocol, protocol verification, consistency, random match probability (RMP), combined probability of inclusion or exclusion (CPI/CPE), likelihood ratio (LR), probabilistic genotyping, DNA profile, DNA mixtures*

Table of Contents

Forensic Autosomal STR DNA Statistical Analyses - General Protocol, Protocol Verification, and Case Record Requirements

1 Scope

Forensic DNA testing requires that statistical calculations be performed on evidentiary DNA profiles that are established as relevant in the context of the case to aid in the assessment of an inclusion or positive association with a known individual. Calculations commonly used are the likelihood ratio (LR), random match probability (RMP), or combined probability of inclusion or exclusion (CPI/CPE). This standard provides general requirements for the laboratory protocol for performing statistical analyses, verification and consistency of use of the protocol, and documentation in the case record of all pertinent information regarding the statistical calculations. This standard applies directly to testing performed using the polymerase chain reaction (PCR) amplification of autosomal loci having short tandem repeats (STR); many of the general requirements may also apply to other types of DNA testing and analysis.

2 Normative References

There are no normative reference documents. Annex B, Bibliography, contains informative references.

3 Terms and Definitions

For purposes of this document, the following definitions apply.

3.1
combined probability of exclusion (CPE)
The probability that a randomly selected, unrelated individual would be excluded as a contributor to the mixture; produced by multiplying the probabilities of inclusion from each locus and subtracting the product from 1 (i.e., 1-CPI).

3.2
combined probability of inclusion (CPI)
The probability that a randomly selected, unrelated individual would be included as a possible contributor to a mixture; produced by multiplying the probabilities of inclusion from each locus.

3.3
conditioning
The act of assuming one or more pieces of information when assigning a conditional probability. The information might be the profile of an individual, or profiles of a set of individuals, who are assumed to have contributed DNA to the evidentiary item under a particular proposition, or it might simply be the assumption that a particular proposition is

true. Any events (or information) that have been used for conditioning are placed to the right of the conditioning bar in a conditional probability expression.

3.4
likelihood ratio (LR)
The ratio of two conditional probabilities of the same event under mutually exclusive hypotheses. The general formula is: LR= Pr (E|H_1, I)/Pr (E|H_2, I). For DNA testing, a statement of comparison of the probability of the evidence (E) (i.e., the DNA profile), given two competing hypotheses, inclusionary (H_1) or exclusionary (H_2) for an individual or specific sets of individuals, and in the context of relevant information (I). (Note: alternative nomenclature is provided in Annex A.)

3.5
probabilistic genotyping
The use of biological modeling (i.e., statistical modeling informed by biological data), statistical theory, computer algorithms, and/or probability distributions, to infer genotypes and/or calculate likelihood ratios.

3.6
proposition
A statement that is true or false, associated with the standpoint of one of the parties on a disputed issue of interest.

3.7
random match probability (RMP)
The probability of randomly selecting an unrelated individual from the population who could be a potential contributor to an evidentiary profile.

4 Requirements

Refer to Annex A, Information on Random Match Probability (RMP), Likelihood Ratio (LR) and Combined Probability of Inclusion or Exclusion (CPI/CPE), for additional information regarding the statistical values applicable to autosomal DNA testing and the following requirements.

4.1 The laboratory shall have and follow a protocol for performing statistical analyses that includes the following:

4.1.1 Descriptions of scenarios where statistical analyses must be performed and scenarios where statistical analyses are not required.

NOTE No statistical analysis is required for an exclusion determined manually.

NOTE Statistical analyses on the evidentiary DNA profile are not required, but may be performed, when a comparison has not been made to known reference data (e.g. to provide

important or relevant information for a particular case when no reference sample is available).

4.1.2 A requirement that any reported positive association of an evidentiary DNA profile to the DNA profile from a known individual be supported by a statistical analysis. The data from each locus used for comparison and for stating a positive association shall be included in the statistical calculation.

NOTE This does not apply to the inclusion of an individual whose DNA is reasonably expected to be present on the item of evidence based on how and from where the biological sample was collected, as defined by the laboratory protocol and/or as documented in the case record for a specific case scenario (e.g. swabbings of an area of an individual's body; clothing worn in close contact with the individual's body).

NOTE Statistical analyses are not required in support of a positive association between two sets of evidentiary data, but may be calculated and provided (e.g., DNA profiles in common between two blood stains of unknown origin found at two different crime scenes to aid in assessing the possibility they may be from the same individual).

4.1.3 A requirement that statistical analyses shall only be performed on loci deemed suitable for comparison based upon the laboratory's documented interpretation and comparison protocol (e.g., where stochastic phenomena such as allelic drop-out, allelic drop-in, or stutter are not explicitly accounted for in the statistical model being used). If the data at a locus have been deemed unsuitable for comparison, then no statistical value can be provided for that locus.

NOTE This requirement may not be applicable for some probabilistic genotyping software. This requirement is meant to eliminate the practice of omitting loci which do not exhibit the alleles of one or more individuals when compared to the known reference standard. Although such practice has been historically labelled as neutral or conservative, it typically is not, and can be especially problematic with interpretation methods that do not allow explicit modelling of allelic dropout or other stochastic phenomena.

4.1.4 A description of statistical analysis methods available for use in the laboratory, to include the following:

4.1.4.1 When statistical analyses are generated from manual calculations or software (e.g., RMP, CPI/CPE, and LR not from probabilistic genotyping software), provide all equations used in the calculations including the following.

4.1.4.1.1 For a homozygous genotype at a locus.

4.1.4.1.2 For a heterozygous genotype at a locus.

4.1.4.1.3 Where a theta (θ) correction factor(s) is used and provide the value of theta used in the calculation.

4.1.4.1.4 For the possible genotype combinations when data from more than one contributor (i.e., mixture) are present at a locus.

4.1.4.1.5 For combining genotype frequencies across multiple loci in a DNA profile.

4.1.4.1.6 For minimum allele frequencies, if used, for the population databases.

4.1.4.1.7 For biological relationships, if used.

4.1.4.2 For calculations generated using probabilistic genotyping software, provide the following.

4.1.4.2.1 References to the published literature, and any other relevant information (e.g., technical and/or user's manual), for the equations and the calculations used by the software for computing likelihood ratios.

4.1.4.2.2 The statistical basis for defining inclusion, exclusion, inconclusive and uninterpretable when those terms are used by the laboratory.

4.1.4.2.3 A requirement that when multiple persons of interest have likelihood ratios that support an association to a DNA mixture, within the capabilities of the approach used, an analysis shall be performed using proposition pairs that test whether the multiple persons of interest can be included together in the observed DNA profile. *(note: borrowed from LR Props document -- need to be sure this stays consistent with that document as moves through the process)*

4.1.4.2.4 A protocol regarding the use of replicate profile data, if performed by the laboratory.

4.1.4.3 A description of when each statistical method can be employed in the laboratory.

4.1.4.4 When multiple methods are available in the laboratory for calculating statistical values and more than one may be appropriately used for a particular case sample scenario and/or DNA profile per 4.1.4.3, then the protocol shall state which statistical analysis method shall be used and/or how to determine which method will be used. For example, the protocol may permit the use of RMP and LR calculations for single source DNA profiles; in this situation, the protocol shall clearly define which calculation should be used under which scenario to ensure reliability based on validation studies and consistency within the laboratory.

Similarly, a CPI/CPE, RMP and/or LR calculation may be appropriate for use for a mixed DNA profile; again, the protocol shall clearly define which calculation should be used. A

common scenario where this may be relevant is an assumed two-person contributor mixed DNA profile obtained from a vaginal, oral or breast swab where the DNA profile from the known female contributor is available and each of the approaches may be applicable.

4.1.5 The source of the population database(s) used in any statistical analyses.

4.1.6 Procedures describing when and how alternate databases and/or theta correction values shall be applied.

4.1.7 What types of assumptions can be made, when those assumptions can be made, and how they shall be incorporated into the statistical analysis. Such assumptions may include, but are not limited to, the number of contributors, the presence of possible artifacts (e.g., stutter) and/or stochastic effects, and the presence of assumed contributors. In addition, the protocol shall also define the use of conditioning information in propositions used to calculate likelihood ratios. The protocol shall provide information regarding the appropriate situation for the use of assumptions (and/or conditioning information used in the proposition for an LR) typically permitted in the laboratory that may impact the statistical analyses. Assumption(s) used that may impact the statistical analyses shall be documented in the case record as required by 4.3.5.

4.1.8 A description of the appropriate validated software and version number to be used for each type of statistical analysis.

4.1.9 A description of when the variable input parameters should be modified and the appropriate values to be used for any parameter or input value that can be changed by the analyst in the software.

4.1.10 A requirement that statistical analyses be performed only at those loci common to both profiles (e.g., when one of the profiles used for comparison has data at fewer loci than the other profile in the comparison, as in a partial, incomplete profile or data from different multiplex kits) for non-probabilistic genotyping (e.g., manual) methods.

4.1.11 A requirement that a new statistical analysis must be performed when subsequent review of the profile data alters how it is used in the original statistical analysis.

4.1.12 A requirement that two or more conceptually different statistics shall not be combined. Specific examples include not multiplying a random match probability with either a combined probability of inclusion or a likelihood ratio, and not multiplying a combined probability of inclusion with a likelihood ratio.

4.1.13 Statements of any known limitations for the use of any formulae and/or software based on external or internal validation studies, and situations where profile data cannot be used for statistical calculations shall be clearly defined in the protocol. Some possible limitations include the number of contributors that may be assumed when using certain formula(e) or software, limitations established through the laboratory validation studies, functions that have not been validated by the laboratory, and when data are insufficient for using the statistical analysis method (e.g., the inability to use CPI/CPE calculations if there is a reasonable risk that data are missing from a locus).

4.2 The laboratory shall verify and document that the protocols for performing statistical analyses generate appropriate values and are performed consistently within the laboratory for all types of DNA profiles typically encountered by the laboratory.

4.2.1 Verification of the protocols shall be performed on single source and mixed DNA samples of known origin that are different from those used in the initial validation studies for the amplification kit and/or statistical analysis software or used to establish the statistical analysis protocol.

4.2.2 Verification of the statistical analysis protocol shall demonstrate that its use returns the same value within the laboratory for the same DNA profile when using procedures without an element of randomness (e.g., Popstats or non-probabilistic genotyping software).

4.2.3 Verification of the statistical calculations protocol shall demonstrate that its use with probabilistic genotyping software having an element of randomness results in consistent values between different runs with the same inputs, as defined by the laboratory based on validation studies for both true contributors and non-contributors.

4.2.4 Verification shall include a demonstration of consistency among analysts in the laboratory for the calculated statistical values using examples representative of the range of samples handled by the laboratory. The laboratory shall define the acceptable range of variability in the statistical values generated for use in the evaluation of the consistency within the laboratory.

4.2.5 Verification shall be performed on new, existing, and modified statistical interpretation protocols.

4.2.6 For verification of the Statistical Analyses protocol, the laboratory shall use data generated and processed under similar testing conditions to those routinely used by the laboratory. The data for all contributors to the DNA used in the verification shall be known and available for the assessment of the data and the proposed statistical analyses protocol. DNA data from different sets of contributors than used in the initial validation studies shall be used to verify the protocol. These supplemental data sets shall span the range of data anticipated to be interpreted by the laboratory.

4.2.7 The validation of the protocol shall be completed prior to implementation of the protocol for casework. Additional validation studies and/or protocol development shall be necessary if deficiencies in the protocol or inconsistencies within the laboratory are identified through this verification process.

4.2.8 Any subsequent modifications to any DNA testing or data interpretation protocol shall include an evaluation for its impact on DNA statistical calculations. These modifications shall be updated in the relevant protocol(s) addressing these requirements, as needed.

4.2.9 Methods, equations, software, etc. shall not be used for statistical calculations without the prerequisite validation, protocol development and verification of the protocol for accuracy and consistency.

4.3 The laboratory shall document the following in the case record for each statistical analysis performed.

4.3.1 The population database(s) used and the source(s) of the database(s).

4.3.2 The statistical analysis method(s) used, and, if applicable, the software program and version number used.

4.3.3 The theta correction factor value(s) used.

4.3.4 The genetic loci and data used for statistical calculations.

4.3.5 All assumptions made when performing the statistical analysis, including but not limited to number of contributors and/or assumed contributors, and in the case of paternity or kinship analysis, any alleged or assumed biological relationships.

4.3.6 All statistical analyses performed, including analyses performed using different assumptions and/or different propositions (e.g., conditioning on different DNA profiles), regardless of whether the statistical analysis is reported by the laboratory.

4.3.7 The actual value used by the analyst with each statistical analysis for any parameter or input value that can be changed in the software (e.g., random number seeds, number of Markov Chain Monte Carlo iterations, probability of drop-out and/or drop-in).

4.3.8 Case-specific scenarios where calculations are not needed shall be documented in the case record.

Annex A (informative)

Information on Random Match Probability (RMP), Likelihood Ratio (LR) and Combined Probability of Inclusion or Exclusion (CPI/CPE)

Additional information regarding the three major types of statistical values calculated for forensic STR DNA profiles is provided below. It should be noted that for Random Match Probability (RMP), Likelihood Ratio (LR) and Combined Probability of Inclusion/Exclusion (CPI/CPE):

1) These three terms refer only to statistical values and their respective calculations.
2) The use of all three statistical calculation methods requires prior independent interpretation of the STR DNA profile, which includes (but is not limited to) determination of the alleles and loci suitable for comparison, the risk of allele drop-out and drop-in at each locus and across the profile, and the assumed number of contributors. None of these are interpretation methods and play no direct role in the interpretation of the DNA profile.
3) A single statistical calculation method must be used across all loci that are suitable for comparison in a given profile; it is not permissible to combine any of these different statistical calculations for a single profile per Requirement 4.1.12.
4) There may be situations where the insufficiency of data and/or the inability to perform a statistical calculation for a profile or for one or more loci within a profile precludes that profile or locus, respectively, from being used for comparison purposes, causing that profile or locus to be reported as unsuitable for comparison, and thus inconclusive.
5) The calculated values are estimates and will vary depending on the allele frequency database used, the quality of the DNA profile, the number of loci having data, the model and formula(e) used and many other variables that impact the calculations.

Random Match Probability (RMP)

Some of the key features and use of the Random Match Probability statistical calculation method for STR DNA single source and mixture profiles are provided here:

1) The RMP may be used for single source profiles and for some mixtures.
    a) For mixtures, the RMP may be calculated for one contributor to a mixture, a subset of contributors, or for the combined genotypes of all contributors. Within a mixture:
        i) May be used for single source profiles that may be resolved (e.g., single major or minor contributor; deduced single contributor when using the genotypes from one or more assumed contributors in the determination of possible genotypes).
        ii) May be used for multiple contributor profiles by considering the combinations of possible genotypes at a locus (e.g., two contributor profiles) by summing the probabilities for all genotypes included at the locus.
            (1) Has sometimes been referred to as modified RMP or restricted RMP.

b) The assumed number of contributors to the DNA mixture shall be assessed along with the genotypes from any assumed contributor(s) to limit, or restrict, the possible genotypes at a locus that are then used for the calculation of the RMP.
c) It may be practical to limit the RMP calculation to profiles, or the portion of a profile, with a defined maximum number of contributors.
d) The RMP can be used for profiles where stochastic effects may be present.

2) The equations using Recommendation 4.1 of the NRC II (1996)[4] for RMP calculations are:
a) $p^2 + p(1-p)\theta$ for homozygous loci, where p is the frequency of allele P at a single locus and $\theta = 0.01$ (for most populations in the United States) or 0.03 (for some isolated populations).
b) 2pq for heterozygous loci, where p is the frequency of allele P at a single locus and q is the frequency of allele Q at the same locus.
c) For single alleles at a locus for which the second allele cannot be determined (e.g., due to possible allele drop-out or allele masking at a possible shared allele), one of the three following equations may be used: a) 2p; b) $2p-p^2$ or c) $p^2 + 2p(1-p)$, where p is the frequency of the single obligate allele P.
d) The product rule is used to calculate the RMP across multiple loci.
e) Equations using Recommendation 4.2 of the NRC II (1996) may also be used. These equations provide corrections for both homozygous and heterozygous profiles.

3) The RMP can be approximated by the estimated frequency of occurrence for a given genotype or set of genotypes, in a particular reference population, that make up the profile of a DNA contributor among random unrelated individuals. It is commonly reported as 1 in the number of individuals by inverting the resulting frequency after applying the product rule across all loci.

4) The RMP is calculated for the genotypes of the single source or mixed evidentiary DNA profile independently of (and even prior to) comparison to the profile from any known individual (other than assumed contributors) since the calculation is based on the evidence data alone.
a) It is necessary to calculate different RMP values for a DNA mixture when different profiles can be resolved [e.g., one RMP for the major contributor(s), and one RMP for the minor contributor(s)].
b) If a subset of loci are used to calculate the RMP, then the selection of loci used should be determined independently of (and even prior to) comparison to any reference profile.

Likelihood Ratio (LR)

Some of the key features and use of the Likelihood Ratio method for STR DNA single source and mixture profiles are provided here:

1) The LR may be used for single source profiles and for some mixtures.
a) A binary LR (non-probabilistic LR) cannot be used for profiles where allele drop-out and/or drop-in may have occurred.

b) A probabilistic LR can be used for profiles where allele drop-out and/or drop-in may have occurred.

2) The LR is a ratio of probabilities of observing the evidence (i.e., DNA profile obtained) under opposing propositions. It is NOT a measure of frequency or a probability.

3) The general equation for the LR is:

$$LR = \frac{\Pr\left(E \mid H_p, I\right)}{\Pr\left(E \mid H_d, I\right)}$$

where Pr = Probability, E = Evidence, H_p = Proposition of the prosecution, H_d = Proposition of the defense, and I = relevant Information in formulating the propositions and assigning the probabilities. Propositions may be referred to as prosecution/defense propositions, proposition 1/proposition 2, prosecution/alternate propositions, inclusionary propositions/exclusionary propositions or other terms that communicate the propositions are different from one another.

a) A proposition represents the set of contributor(s), known and unknown, who may have contributed to the observed DNA profile. There is no requirement that a particular proposition is true.

b) The propositions shall depend on case information and the claims (or reasonably assumed claims) of each of the parties. The propositions may be changed at the request of either party.

c) Propositions must be mutually exclusive. At least one element of the proposition must be different so that they may not both be true at the same time (e.g., Proposition 1 states the Person of Interest (POI) is the source of the DNA and Proposition 2 states a random, unrelated person in the population is the source of the DNA), or the value of the LR will equal 1.

d) A particular contributor genotype may be known or assumed to be a contributor in a proposition.

i) A conditioning profile is a profile that is assumed to be present in both propositions.

ii) A conditioning profile may be a profile assumed to be present due to the collection and/or origin of the evidence item (e.g., intimate sample) or it might simply be the assumption that a particular profile is present under a given set of propositions.

e) Consideration must be given to calculating a separate LR for each included contributor as well as an LR for the contributors together per requirement 4.1.4.2.3. This may prevent a major contributor from having undue influence on the weight of the evidence for a minor contributor. Conditioning profiles may be useful in this scenario.

f) For a binary LR calculation, the weight given to a plausible genotype is 1 and the weight given to an implausible genotype is 0 (hence the name “binary”).

g) For a probabilistic LR calculation, the weight given to a genotype can vary between 0 and 1.

h) The weights of the same genotypes may differ for different propositions in the probabilistic LR calculation.

4) An LR is reported as a ratio of the probabilities of the evidence given the propositions, and not as a ratio of the probabilities of the propositions. For example, appropriate statements include: "The evidence is LR times more likely to be observed if Proposition 1 is true rather than if Proposition 2 is true" or "It is LR times more likely that the DNA profile would be observed if Proposition 1 is true rather than if Proposition 2 is true".
   a) It is reported as an LR; it is NOT reported as 1 in X number of individuals.
      i) For a single source profile, often the LR and RMP values are numerically the reciprocal of each other; however, they answer fundamentally different questions.
5) A given LR is only for the propositions stated under the relevant information. If the propositions change or if the relevant information changes, then a new LR must be calculated.
   a) The value of the LR will change when the data and/or propositions change.
      i) LRs generated under the same set of propositions using probabilistic genotyping software with an element of randomness will generally vary within an expected limited range.
6) A probabilistic LR calculation can return a value less than one (or negative logLR), which communicates that more weight of evidence is given to the defense or alternative proposition.
7) A probabilistic or a binary LR calculation can return a value of one (or logLR of 0), which communicates that equal weight of evidence is given to both propositions. Neither proposition is supported over the other.
8) A probabilistic or a binary LR calculation can return a value greater than one (or positive logLR), which communicates that more weight of evidence is given to the prosecution proposition.

Combined Probability of Inclusion (CPI) and Combined Probability of Exclusion (CPE)

Some of the key features and use of the Combined Probability of Inclusion (CPI) and Combined Probability of Exclusion (CPE) statistical calculation method for mixed STR DNA profiles are provided here:

1) Also referred to as Random Man Not Excluded (RMNE).
2) Only appropriate use is to provide statistical calculations for a limited subset of mixed DNA profiles.
   a) Most applicable for use with profiles generated from the amplification of sufficiently high amounts of DNA such that stochastic effects, if present, are negligible, and have no impact on the interpretation and ability to generate statistical frequency calculations.
   b) Generally, most applicable for use with DNA profiles from two-person DNA mixtures or three person mixtures having two major contributors, where the CPI/CPE is calculated only for the two major contributors.

c) Rarely suitable for use with mixtures of three or more contributors, particularly when amplified with high sensitivity kits using recommended procedures, with the possible exception of when two distinguishable major contributor profiles are present. It can only be used with mixtures of three or more contributors when high levels of DNA are observed and no contributor is reasonably expected to have dropped out at any locus.

d) Commonly used for indistinguishable mixed DNA profiles (i.e., unable to associate alleles into genotypes for the contributors due to similarities in peak heights and the inability to assume the genotypes of one of the contributors).

3) Shall ONLY be used for profiles where there is very high confidence that all alleles, and thus all genotypes, for all contributors are present at each of the loci with data available for interpretation and comparison where there is no reason to expect that allele drop-out might have occurred. (See 2b above.)

a) Data from loci with one or more alleles below the stochastic threshold shall not be used for comparison or for calculating the CPI/CPE (with the one exception stated in (e) below).

b) The assumed number of contributors to the DNA mixture using the entire DNA profile shall be assessed and then used for evaluating the prospect that all genotypes from all contributors are present at each locus.

c) This determination shall occur prior to comparison of the DNA profile data to the profile from any known contributor (i.e., independently of any knowledge of data from other profiles).

d) If all alleles at a locus are above the stochastic threshold, but there are only a limited number of alleles as compared to the maximum expected allele count based on the assumed number of contributors (1-2 alleles in 2 person mixtures; 1-4 alleles in 3 person mixtures), then the possibility that drop-out has occurred shall be considered and the CPI/CPE calculation shall not be used if there is some reasonable possibility that drop-out explains the paucity of alleles. Peak heights at other loci and total peak height values at each locus shall be taken into account when assessing the data and the possibility of drop-out.

i) When the alleles from at least one contributor are below the stochastic threshold at multiple loci, it is reasonable to assume that the alleles for that individual will be below the stochastic threshold at all loci based on the mixture ratio of the contributors' DNA; thus, CPI/CPE cannot be used for this profile, even for the one or few loci with all alleles above the stochastic threshold as it is more likely that alleles are missing than the assumption that all alleles are present.

ii) If one or more alleles are missing from a locus, the CPI/CPE value resulting from the use of the existing alleles would underestimate the proportion of possible contributors as compared to the calculation using all of the alleles from all of the contributors. That is, the value calculated would give the appearance of the profile being rarer than it really is. Such a figure would be more prejudicial against the defendant. It is not generally accepted practice for rarer values to be

reported or presented in testimony when providing a statistical frequency for an individual who cannot be excluded as a possible contributor.

e) Loci with one or more peaks below an RFU-defined stochastic threshold may be used in the CPI calculation ONLY if the total number of alleles present at each locus is consistent with all alleles being present for the assumed number of contributors (e.g., six alleles are present at a locus and the assumption of three total contributors is used).

4) The equations for CPI/CPE calculation are:

a) Probability of inclusion for a locus = (the sum of allele frequencies)^2 = (P_A + P_B + P_C + … + P_N)^2, where P_A, P_B, P_C and P_N are the frequencies of alleles A, B, C and N, respectively, observed at the locus, where it is assumed that all alleles from all contributors to the DNA mixture are present, based on the data observed and the assumed number of contributors to the DNA profile.

i) The value at each locus is the cumulative frequency of all possible heterozygous and homozygous genotypes.

ii) For profiles where the maximum allele count is observed based on the assumed number of contributors to the DNA mixture, the CPI/CPE calculation would still incorporate the frequencies of homozygous genotypes included at that locus, however, individuals with homozygous genotypes could be excluded definitively from that locus during interpretation and comparison based on the assumed number of contributors.

b) The Combined Probability of Inclusion (CPI) is the product (i.e., multiplied together) of each of the probabilities of inclusion calculated from each locus used in the interpretation.

c) CPE = (1 – CPI); other equations are available in the publications referenced in Annex B, Bibliography.

5) The CPI value is an approximation of the proportion of random individuals unrelated to a true contributor in the mixture who would be expected to be included as possible contributors to the DNA mixture from the random population. It is commonly reported as 1 in X number of individuals.

a) The CPE is an approximation of the proportion of random individuals unrelated to a true contributor in the mixture who would be excluded as contributors to the DNA mixture from the random population. This value may be reported as Y out of X individuals, but is sometimes reported as a percentage.

b) The CPI/CPE calculation is not appropriate for use when a non-contributing individual related to a true contributor to the DNA mixture cannot be excluded as a possible contributor to the DNA mixture.

c) The CPI/CPE value is appropriate for use when related individuals are contributors to the DNA mixture.

6) The CPI/CPE is calculated for the mixed DNA profile independently of (and even prior to) comparison of the profile from any known individual since the calculation is based on the questioned profile alone;

a) Only one CPI/CPE frequency can be calculated for one mixed DNA profile.
b) A CPI/CPE calculation is based on the questioned profile alone. It should never be based on the profile of an individual who cannot be excluded as a contributor.

Additional information regarding CPI/CPE calculations and uses is available in publications referenced in Annex B, Bibliography.

Annex B (informative)

Bibliography

1] Quality Assurance Standards for Forensic DNA Testing Laboratories, http://media.wix.com/ugd/4344b0_4a22824ce56f43d4b1a4d2486409f95d.pdf

2] Buckleton, John, Bright, Jo-Ann., and Taylor, Duncan A. (Eds.). *Forensic DNA Evidence Interpretation.* 2nd edition, CRC Press, 2016.

3] Balding, David J. and Steele, Christopher D. *Weight-of-Evidence for Forensic DNA Profiles*. 2nd edition, Wiley, 2015.

4] National Research Council (NRC II). *The Evaluation of Forensic DNA Evidence*. National Academy Press, 1996. (Available online at: https://www.nap.edu/read/5141/).

5] Bieber, Frederick. R., Buckleton, John S., Budowle, Bruce, Butler, John M., and Coble, Michael D. "Evaluation of Forensic DNA Mixture Evidence: Protocol for Evaluation, Interpretation, and Statistical Calculations Using the Combined Probability of Inclusion." *BMC Genet.*, vol. 7, no. 1, 2016, pp .125.

6] Curran, James M. and Buckleton, John. "Inclusion Probabilities And Dropout." *J. For. Sci.,* vol. 55, no. 5, 2010, pp. 1171-1173.

7] Buckleton, John and Curran, James M. "A Discussion of the Merits Of Random Man Not Excluded and Likelihood Ratios." *For. Sci. Int: Gen.*, vol. 2, no. 4, 2008, pp. 343-348.

8] Bille, Todd W., Bright, Jo-Ann and Buckleton, John. "Application of Random Match Probability Calculations to Mixed STR Profiles." *J. For. Sci.*, vol. 58, no. 2, 2013, pp. 474-85

9] Gittelson, Simone., Kalafut Timothy, Myers, Stephen, Taylor, Duncan, Hicks Tacha, Taroni, Franco, Evett, Ian W., Bright, Jo-Ann, and Buckleton, John. "A Practical Guide for the Formulation of Propositions in the Bayesian Approach to DNA Evidence Interpretation in an Adversarial Environment." *J. For. Sci.* vol. 61, no. 1, 2016, pp. 186-195.

10] SWGDAM 2017 Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Laboratories. https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_50e2749756a242528e6285a5bb478f4c.pdf (October 2020)

ANSI/ASB Standards

11] ANSI/ASB Standard 018, Standard for Validation of Probabilistic Genotyping Systems, First Edition, 2020 http://www.asbstandardsboard.org/wp-content/uploads/2020/07/018_Std_e1.pdf (October 2020)

12] ANSI/ASB Standard 020, Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol,

First Edition, 2018 https://asb.aafs.org/wp-content/uploads/2018/09/020_Std_e1.pdf (October 2020)

13] ANSI/ASB Standard 040, Standard for Forensic DNA Interpretation and Comparison Protocols, First Edition, 2019 http://www.asbstandardsboard.org/wp-content/uploads/2019/10/Std_040_e1.pdf (October 2020)

14] ANSI/ASB Standard 041, Assigning Propositions for Likelihood Ratios in Forensic DNA Interpretations, First Edition, 20xx, (currently in progress at ASB - link to be added when available)

15] ANSI/ASB Standard 123, Routine Internal Evaluation of a Laboratory's Interpretation and Comparison Protocol, First Edition, 20xx, (currently in progress at ASB - link to be added when available)