# The use of statistics in legal proceedings

## A PRIMER FOR COURTS

THE ROYAL SOCIETY

RSE
The Royal Society of Edinburgh
KNOWLEDGE MADE USEFUL

*The use of statistics in legal proceedings: a primer for courts*
Issued: November 2020 DES6439
ISBN: 978-1-78252-486-1



To request additional copies of this document please contact:

The Royal Society
6 – 9 Carlton House Terrace
London SW1Y 5AG
**T** +44 20 7451 2571
**E** law@royalsociety.org
**W** royalsociety.org/science-and-law

This primer can be viewed online at
**royalsociety.org/science-and-law**

# Contents

# Science and the law primers
## Foreword

The judicial primers project is a unique collaboration between members of the judiciary, the Royal Society and the Royal Society of Edinburgh. The primers have been created under the direction of a Steering Group initially chaired by Lord Hughes of Ombersley who was succeeded by Lady Justice Rafferty DBE, and are designed to assist the judiciary when handling scientific evidence in the courtroom. They have been written by leading scientists and members of the judiciary, peer reviewed by practitioners and approved by the Councils of the Royal Society and the Royal Society of Edinburgh.

Each primer presents an easily understood, accurate position on the scientific topic in question, and considers the limitations of the science and the challenges associated with its application. The way scientific evidence is used can vary between jurisdictions, but the underpinning science and methodologies remain consistent. For this reason we trust these primers will prove helpful in many jurisdictions throughout the world and assist the judiciary in their understanding of scientific topics. The primers are not intended to replace expert scientific evidence; they are intended to help understand it and assess it, by providing a basic, and so far as possible uncontroversial, statement of the underlying science.

The production of this primer on the use of statistics in legal proceedings has been led by Professor Niamh Nic Daéid FRSE. We are most grateful to her, to the Executive Director of the Royal Society, Dr Julie Maxton CBE, the Chief Executive of the Royal Society of Edinburgh, Dr Rebekah Widdowfield, and the members of the Primers Steering Group, the Editorial Board and the Writing Group. Please see the back page for a full list of acknowledgements.

**Sir Venki Ramakrishnan**
**President of the Royal Society**

**Dame Anne Glover**
**President of the Royal Society of Edinburgh**

# Summary, introduction and scope

The aim of this primer is to provide assistance to the judiciary and legal professionals in understanding the principles of evaluating evidence (that has a statistical basis) presented in the courts. The primer is presented in two parts. The first part provides a general introduction to the use of statistical and probabilistic tools within legal processes with some examples presented, including some relating to evidence types commonly presented to the courts.

The second part consists of five appendices.

Appendices 1 – 3 provide specific information about how statistical and probabilistic tools may be used in assisting the delivery of evaluative opinions by forensic practitioners relating to common types of scientific evidence encountered primarily in criminal cases, for example trace evidence (eg fibres, glass, DNA) and impression evidence (eg footwear marks, toolmarks, fingerprints).

Appendices 4 and 5 relate to specific statistical methods and to their use in assessing statistical significance and relative risk. These areas generally have more relevance in civil proceedings.

This short guide cannot equip the judiciary and legal professionals with all the necessary skills required, but it should be useful for signposting where problems may arise and where external expertise may be needed.

# 1. What is statistical science?

Reasoning about data is increasingly recognised as an essential skill for modern life. Fact-finding and the assessment of expert evidence in court cases often requires an understanding of probability, statistics and numbers. Various different statistical and probabilistic tools can be used to address different questions relating to the context of individual cases presented to the courts and the choice of which tools to use will depend on the questions that need to be addressed. Standard types of questions which can be answered by statistical science may be categorised as follows:

- Descriptive statistics: eg What is the number of rapes reported in the country? How many drugs of a particular type are found in drug seizures?
- Inference from observed data to a larger population: eg Given the responses to the British Crime Survey, what is the estimated number of illegal drug users in the UK?
- Inference from observed data to a scientific conclusion: eg Did the exposure to the emissions from an incinerator raise the risk of birth defects[1]?
- Prediction: eg Given a set of characteristics, what is the chance that the accused will reoffend[2]?
- Evaluation: The evaluation of scientific findings in court uses probability as a measure of uncertainty. This is based upon the findings, associated data, expert knowledge, case-specific propositions and conditioning information[3].

All these situations are characterised by uncertainty, and probability theory provides the tools and language for handling and communicating uncertainty.

Statistical science has developed a wide range of powerful techniques for quantifying the impact of some sources of uncertainty, eg calculating margins of error from a survey, measuring the support for a proposition (also called a hypothesis) from observed data or assessing the probability of a future event. Other sources of uncertainty are not so easily quantified but can still be informally assessed and communicated, eg those arising from the reliability of survey respondents, the quality of scientific studies and the relevance of available and good quality datasets to the facts of a legal case.

Unavoidable uncertainty about the future is often termed chance, also known as aleatory uncertainty, and the assignment of probabilities to future events is familiar. Legal cases generally deal with uncertainty in the sense of lack of knowledge, also known as epistemic uncertainty. Fortunately, the theory of probability can still be applied in this context. Uncertainty of measurement can also arise and this, in general, can be characterised for objective measurements (eg how much of a controlled substance may be in an analysed sample) but is more challenging for more subjective measurements (eg in the examination of toolmarks).

### 1.1 Use of statistical science and types of evidence

Statistical science can be called upon to support expert knowledge when dealing with a variety of types of evidence and proceedings. These include:

- evaluation of DNA evidence[4];
- evaluation of trace evidence, eg fibres, glass, paint or firearms discharge residues (Appendix 2);
- evaluation of pattern-matching evidence, eg toolmarks, ballistics and fingerprint evidence (Appendix 3); and
- causation of illness or injury in a civil case, where it may be helpful to apply epidemiological research (the study of occurrence, aetiology, prognosis and treatment of illness in populations) to individual cases (Appendix 5).

The way in which statistical science may be used in a legal context is illustrated in Figure 1.

**FIGURE 1**

The process by which statistical science may be used in legal proceedings and in which relevant past data are used to draw conclusions about the facts of a current case.

1. Identification of an observation related to a specific item of evidence relevant to some aspect of the current case.

↓

2. Providing a list of appropriate potential propositions concerning the evidence, which may include prosecution allegations and any defence alternative assertions.

↓

3. Identifying resources containing relevant past data or creating new relevant datasets.

↓

4. Methods for using the accumulated information in a database (if available) to derive a numerical or verbal expression of the probative value of the observation related to the item of evidence with regard to the competing propositions.

↓

5. Communication of probative value of evidence.

### 1.2 Communication of the probative value when statistical science is used

When conclusions based on statistical science are drawn from data, it is crucial that the data and the reasoning supporting those conclusions are transparent. Under the term 'intelligent transparency' Baroness Onora O'Neill[5] has argued that the data and reasoning must be:

- accessible: ie easily available and not, for example, hidden behind a proprietary algorithm;
- understandable: to everyone involved in the case, including a jury;
- useable: they address current specific concerns; and
- assessable: where the 'working' is open to scrutiny by legal and other professionals.

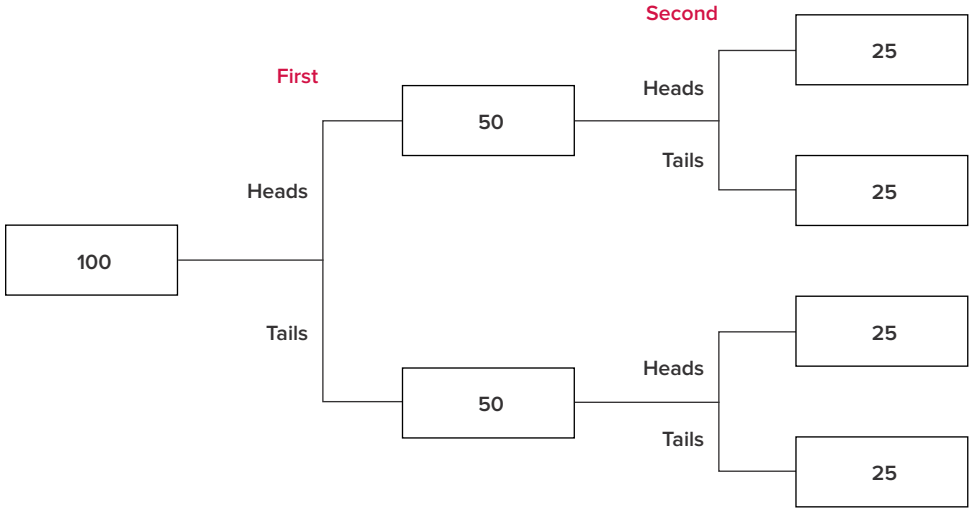
There is always uncertainty involved in statistics, particularly in assessing the relevance of available historical data to a current case. There may also be some extent of disagreement between different professionals, but this may not be of real substance. Statistical conclusions are only as reliable as the model (and data) from which they are derived.

Estimates are neither wholly right nor wholly wrong, conclusions are not mechanistic and sometimes the only database available is the experience of the professional. In such situations transparency is particularly necessary and the experience needs to be documented with emphasis on the relevance to the case in question. The professional judgement of the appropriate experts (expert knowledge) is inevitably involved in each stage of the process outlined in Figure 1. Probability is a conceptual device that helps us think and reason logically when faced with uncertainty about the occurrence of a questioned event in the past, present or the future.

Probability helps us think clearly and coherently about uncertain events.

**FIGURE 2**

Expected frequency tree when repeating a double-flip of a coin 100 times. We would expect the first flip to be heads in 50 of these experiments, and both flips to be heads in 25.




By way of an example to explain probability, we can use the idea of 'expected frequency' (such a frequentist approach is not appropriate in forensic inference where probability is conditional and personal but is used here to explain the notion of probability). When faced with a question concerning likely outcomes if a coin is flipped twice, you ask yourself: What would I expect to happen if I tried the experiment many times? Take the example that you repeated this double-flip experiment 100 times. As shown in Figure 2, in 25 out of these 100 repeats you would expect to get two heads. Therefore, the reasoning goes, the probability that on a particular attempt you would get two heads is 1 in 4, or ¼. Which, fortunately, is the correct answer. This probability can be equivalently expressed as a fraction (¼), a decimal (0.25), a percentage (25%), a proportion (1 out of 4) or betting odds (3 to 1 against).

Probability follows basic rules in which:

- probabilities of all possible events add to 1 (eg in Figure 2 there are four possible events, so their probabilities add to 1); the probability of not getting two heads is ¾ or 75%;
- probabilities are multiplied for sequences of events which are independent (eg the probability of two heads in a row is ½ × ½ = ¼); and
- probabilities are added when considering probabilities of separate (mutually exclusive) sets of events (eg the probability of getting two heads or two tails is ¼ + ¼ = ½).

It is a common misapprehension that probabilities can only be used for future events with some randomness. While it is true that an event has either happened or not, many statisticians will feel that it is reasonable to assign probabilities to our personal uncertainty about unknown facts, as the following example shows.

Suppose I have a coin and I ask you for your probability that it will come up heads. You answer “50:50”, or similar (50% or ½). Then I flip it, cover up the result before either of us sees it and again ask for your probability that it is heads. You may, after a pause, say “50:50”. Then I take a quick look at the coin, without showing you, and repeat the question. Again, if you are like most people, you eventually say “50:50”. This simple exercise reveals a major distinction between two types of uncertainty: what is known as aleatory uncertainty before I flip the coin – the ‘chance’ of an unpredictable event – and epistemic uncertainty after I flip the coin – an expression of our personal ignorance about an event that is fixed but unknown. In forensic science we are almost always concerned with epistemic uncertainty about the facts of a past situation.

Probabilities are commonly used to express epistemic uncertainty in legal settings. Section 2 provides more detail on the meaning of probability in such contexts, including Bayes’ theorem and the role of the likelihood ratio (LR).

# 2. Probability and the principles of evaluating scientific evidence

Specifically, in a legal context, probability can help fact-finders assess the impact of evidence on the truth or otherwise of a particular proposition. It has a well-documented history in academic legal literature[6]. Each item of evidence can be used to support one or more proposition(s). Evidence may, on occasion, point directly to incriminating or exculpating a suspect of a particular crime, but will more likely have probative value in discriminating between competing propositions for either the source of some material found in relation to a scene of the crime or an alleged activity connected to a crime.

For example, competing source-level propositions concerning a fragment of glass found on a suspect's clothing might include that the glass fragments came, or did not come, from a particular broken window. Competing activity-level propositions might be that the suspect broke or did not break the particular glass in question at a particular time.

## 2.1 What probability is not

Probability is not an inherent property of material or objects. For example, a pack of playing cards does not possess a specific probability for an ace being drawn from it. But you may say, given your current state of knowledge about the properties of the pack, "that the probability of drawing an ace from a pack of playing cards is 4 in 52, or 1 in 13". Such a probability is based on certain assumptions. It is assumed that, among other considerations, (i) it is a full pack of 52 cards; (ii) it is well shuffled; (iii) the cards are exactly the same shape, size and condition so as not to influence the physical act of a draw; and (iv) the person drawing the card will do so without any bias. The probability figure of 1/13 (or roughly 8%) is your own best assessment, based on your knowledge of the composition of a full pack of cards and based on the implicit assumptions outlined. It is a probability with which most of us would agree and it is one that would work well in gambling games, but it is not an inherent property of the aces in the pack or of any of the other cards in the pack. The implication here is that even in cases with well-structured chance devices, such as a deck of cards, if you have already been dealt a couple of aces, your probability for the next card being an ace has dropped considerably. Probability is personal: it is personal in the sense that it depends on the knowledge available to the person making the judgement and on the assumptions they make.

But that is not to say that personal probability is conjured up on a whim or a preference, or to suggest it is not based on acquired data. Where relevant data are available, it is expected that they will be taken into account in assigning a probability. For example, suppose that reliable information is available on the proportion of individuals in a target population that possess a particular observable feature, such as skin, hair or eye colour. Then, our assessment that a person drawn randomly from that population will show a particular feature of interest ought to be informed by the available knowledge about the composition of the population. When new testing systems are used, the frequencies of given traits in a population will not be widely known.

### 2.2 Personal probabilities

We make personal assignments of probability every day:

- What is the probability that I will miss the bus this morning if I have one more cup of coffee?
- What is the probability that I will be caught if I break into this property?

In such circumstances, the probabilities that we assign, albeit not mathematically evaluated or even verbalised in this way, will depend on our knowledge and understanding of the factors and risks involved. Such probabilities are also known as personal ‘degrees of belief’.

Some people may attempt to answer the above questions by thinking of past experiences in similar situations. For example, they may consider how many mornings in the past they have missed the bus when having one more cup of coffee, though this may give rise to many other questions, such as the extent to which today’s morning is comparable to previous experiences.

Statisticians and scientists refer to data on the proportion of times an event has occurred as a relative frequency. However, and most importantly, there are some situations for which relative frequencies cannot be meaningfully conceived. In legal cases, for example, the fact-finder must deal with singular, non-repeatable, one-off events for which the notion of relative frequency may not be helpful. This does not preclude the possibility that useful frequency data may be available (eg scientific data on the occurrence of genetic features or the prevalence of a disease) to help decide aspects of the case (examples of this type of use are in Appendices 4 and 5). Where such data are available and are relevant, they ought to be used in probability assignment as one source of information among others.

Experts assign personal probabilities based on their experience, knowledge and understanding of their type of expert evidence. However, a challenge with such probabilities is the potential influence of cognitive effects. The reliability of expert-assigned probabilities is determined by various factors, including:

- the extent and relevance of the expert's experience;
- the ability of the expert to compile and store systematically those experiences in their memory;
- the expert's ability to recall accurately the relevant data;
- the expert's ability to avoid and mitigate against bias while inputting expert knowledge; and
- calibration, ie measuring the extent to which those events assigned a probability of (say) 40% actually do have a relative frequency of occurrence close to 40%.

In general, the more that experts base their assignments of probability on relevant, shared and robust data, the greater is the trustworthiness of those assignments. The more they base their assignments on their recalled experience and knowledge and on their intuition, the more those assignments will be open to justified challenge.

### 2.3 Datasets containing relevant past observations

Evaluation of evidence using likelihood ratios (LRs) often involves the use of datasets and statistical assumptions. It is important that these datasets, assumptions and calculations are clearly stated and appropriate for the problem. Validation tests should be carried out to gauge the statistical assumptions and to ensure that the LR values presented in court are reliable. Datasets should be from relevant populations and should not bias the results of the analysis.

The Court of Appeal decided in Regina v Abadom[7] that an expert is entitled to draw upon material produced by others in the field in which his or her expertise lies, and indeed where any reliable data are available that bear upon the question the expert is addressing. It is part of the duty of the expert to take this into account. So, a crucial judgement concerns the reliability and relevance of available data. The dataset should have high intrinsic quality, reliability and high relevance to the question being addressed by the use of the data. A national dataset of informally collected examples of glass or footwear from people's homes may be of limited relevance to local criminal investigations but of value in informing background abundance of the items in question, while local datasets collected to address specific aspects of criminal incidents or of suspect populations may have high relevance. Ideally, there will be an appropriately large enough random sample from a population that matches agreed features of the case, but this is a high bar that is rarely achieved and means that expert judgement and full transparency are required to deal with such limitations.

### 2.4 Probative value expressed as a likelihood ratio

Technically, the LR is the probability of the evidence assuming that proposition A is true divided by the probability of the evidence assuming tha proposition B is true:

$$LR = \frac{\text{probability of the evidence, if A is true}}{\text{probability of the evidence, if B is true}}$$

LRs are typically attached to DNA evidence in which a 'match' of some degree is found between the suspect's DNA profile and the DNA profile derived from a trace found at the scene of a crime. The two competing hypotheses are that the DNA profile in the recovered trace material originates from the suspect or it originates from someone else, so that we can express the LR as:

$$LR = \frac{\text{probability of the DNA profile 'match', if the suspect left the trace}}{\text{probability of the DNA profile 'match', if the trace was left by someone else}}$$

The 'DNA evidence' is the suspect's DNA together with the DNA trace from the crime scene. For the specific situation when the trace contains plenty of DNA and it is deemed to have come from one person, the LR above can be written, after some mathematical operations and given some assumptions, as:

$$LR = \frac{1}{\text{random match probability}}$$

The random match probability is the probability of finding an evidence match if selecting at random from within a particular population. For example, in the context of a DNA sample[8], it is the probability of observing a DNA profile of an unknown person that is the same as the DNA profile from a crime scene stain (and assuming a particular population genetic model). Typical LRs for DNA evidence are in the millions or billions, although the exact values may be contested, such as when there are complications due to the traces containing a mix of DNA from multiple people. Further information is provided in Forensic DNA analysis: a primer for the courts[9].

Table 1 shows an example of a verbal scale used for communicating LRs (a similar example can be found in Willis[10]). In most cases (including DNA at an activity level where the activity that caused the DNA to be deposited is the issue) the LR will be based on a semi-quantitative (ie an order of magnitude) assessment and a verbal equivalent may be presented to the court. The reason that verbal expressions are defined numerically is to provide a consistency in their use rather than to translate available numbers into a common language. In those cases where a quantitative assignment is possible there is a strong argument for presenting the LR value to the court without a verbal qualifier but also an argument for avoiding the risk that lay persons (eg juries) may misunderstand conclusions stated in numbers as absolute measurements. An LR equal to 1 supports neither proposition preferentially.

The LR is not a specific measurement, but rather is the weight of evidence of the scientific findings in two competing scenarios (prosecution and defence). There will almost always be some natural variation in LRs depending on different assumptions and the quality and relevance of the datasets and what is known about the transfer, persistence, recovery and background abundance of the particular type of evidence under scrutiny. The value of the LR on the scale shown in Table 1 is preferably assigned based on robust data extracted from a relevant dataset.

With good quality and relevant data it may be possible to generate a numerical assessment using the LR relating to evidential support. However, often the available data are either poor or non-existent (particularly true for knowledge relating to the transfer of material and its persistence once transferred). In such cases, the expert forms a personal opinion based on domain knowledge of processes and on personal experience that can be disclosed and audited. In these situations, verbal expressions or orders of magnitude of the LR may be helpful and the basis of any statement of expert opinion formed this way must always be made clear.

**TABLE 1**

Example of verbal interpretations of likelihood ratios (LRs) – in this case for source-level propositions.

| Value of LR | Verbal equivalent | Supported proposition |
|---|---|---|
| <0.000001 | Extremely strong support | Different source |
| 0.0001-0.000001 | Very strong support | |
| 0.001 – 0.0001 | Strong support | |
| 0.01 – 0.001 | Moderately strong support | |
| 0.1 – 0.01 | Moderate support | |
| <1 – 0.1 | Limited (or weak) support | |
| About 1 | Neutral | Neither proposition is supported in preference to the other |
| >1 – 10 | Limited (weak) support | Same source |
| 10 – 100 | Moderate support | |
| 100 – 1,000 | Moderately strong support | |
| 1,000 – 10,000 | Strong support | |
| 10,000 – 1,000,000 | Very strong support | |
| >1,000,000 | Extremely strong support | |

## 2.5 Bayes' theorem and the likelihood ratio

Bayes' theorem provides a general rule for updating probabilities about a proposition in the light of new evidence. It says that:

$$\text{the posterior (final) odds for a proposition} = \begin{array}{l}\text{the LR x the prior}\\ \text{(initial) odds for the proposition}\end{array}$$

For example, suppose a hypothetical screening test for doping in sports is claimed to be '95% accurate', meaning that if an athlete is doping there is a 95% chance (probability 0.95; sensitivity) of obtaining a positive test result, and if the athlete is a non-doper there is a 95% chance (probability 0.95; specificity) of obtaining a negative test result. Such general performance characteristics have been determined through tests under controlled conditions, ie by applying the test in so-called ground truth cases, where it is known whether a tested person is doping or not.

Assuming that the odds of an athlete taking drugs prior to being subject to a screening test are 1 in 50 (1:50), then if an athlete tests positive what is the probability that they are truly doping?

The LR (explained in Section 2.4) is the probability of a positive test given the proposition that the athlete is doping (95%) divided by the probability of a positive test given the proposition that the athlete is not doping (5%, ie 1 - specificity). This ratio is 19 (LR = 0.95/0.05 = 19).

Bayes' theorem tells us that the posterior odds of the athlete having taken drugs can be computed by multiplying the prior odds of that proposition by the LR provided by the positive test. In this form, we have to work with odds not probability. Odds are related mathematically to probability and a very simple conversion can be used to give the value for probability where the odds of m:n correspond to the probability m/(m + n).

So, for the doping example,

- the prior odds for the proposition 'athlete is doping' versus 'athlete is not doping' are 1:50, which correspond to a probability of 1/(1 + 50) or a prior probability of approximately 0.02 (the actual value is 0.0196, which is equivalent to 1.96%);
- the LR is 0.95/0.05 = 19; and
- therefore, by Bayes' theorem, the posterior odds that the athlete is doping are (1:50) x 19 = 19:50, giving a posterior probability of doping of 19/(19 + 50) ≃ 0.28 or 28%.

So, even though drug testing could be claimed to be '95% accurate' (based on the sensitivity and specificity metric) this does not mean that, in the event of a positive result, there is a 95% chance that the athlete is doping. In this example, the probability that the athlete is doping, given a positive test result, is approximately 28%. The posterior odds that an athlete is doping crucially depend on the prior odds for the proposition 'athlete is doping' versus 'athlete is not doping' (in the example this was 1:50) prior to considering the result of the screening test (the LR result). This means that if conclusions are drawn from test results in isolation there could be misinterpretations of what is meant by the accuracy of the test. This could cause conclusions such as athletes being incorrectly accused of doping because they failed a drug test.

In practice, the Court of Appeal has ruled that Bayes' theorem should not be used by a jury to combine and weigh evidence[11], but LRs assessed by experts are permitted if they have a sound basis. Many real-world cases involve multiple and related items of evidence, making probabilistic inference much more complex and intricate than the illustrative doping example given above.

# 3. Issues with the potential for misunderstanding

### 3.1 Prosecutor's fallacy

The prosecutor's fallacy occurs when the probability of the evidence (matching DNA profile, glass fragment of the same refractive index as the fragment recovered from the target window, etc) given innocence (the random match probability) is incorrectly interpreted as the probability of innocence given the evidence. This is formally known as 'transposing the conditional', and is a clear breach of logic that becomes obvious when the correct statement 'If it's a dog, the chances are very high that it has four legs' is transposed to 'If an animal has four legs, the chances are very high that it's a dog'.

### 3.2 Defence attorney's fallacy

The defence attorney's fallacy occurs when it is reported how many people with the matching characteristics are likely to be found in a defined population (eg the population of the UK). This assumes that the perpetrator is part of some arbitrarily large population and that there is no other information available, so that everyone is equally likely to be the perpetrator. Under these assumptions, it is deduced that there is a small probability that the suspect is the perpetrator. It is a fallacy to infer from this that the evidence is weak.

### 3.3 Combining evidence

A standard statistical approach is to include all evidence in a single calculation, each item weighted by its relative strength expressed by LRs. This technique is used, for example, in email spam filters and in Case study 1.

**CASE STUDY 1**

## An archaeological case

On Saturday 25 August 2012, archaeologists began an excavation for Richard III's remains by digging in a car park in Leicester. Within a few hours they found their first skeleton and the question was whether this was Richard III. Table 2 shows the specific items of evidence and their likelihood ratios (LRs) regarding the propositions that the skeleton was that of Richard III. These LRs were, as far as possible, based on sound statistical evidence, but there was inevitably some uncertainty so that conservative values were assigned, and verbally interpreted here (because of their uncertainty and therefore qualitative nature) using the terms in Table 1.

Probability theory permits, given certain assumptions about the evidence, the multiplication of these LRs to give a final number that represents 'extremely strong' evidence to support the proposition that the skeleton was, rather than was not, that of Richard III. Of course, the final assignation of the skeleton would not be based on the LR alone but would involve other evidence as well.

**TABLE 2**

Likelihood ratios (LRs) assessed for items of evidence found on a skeleton recovered in Leicester.

| Evidence | LR (conservative assignment) | Verbal equivalent |
|---|---|---|
| Radiocarbon dating CE 1456 – 1530 | 1.8 | Weak support |
| Age and sex of skeleton | 5.3 | Weak support |
| Scoliosis | 212 | Moderately strong support |
| Post-mortem wounds | 42 | Moderate support |
| Mitochondrial DNA match | 478 | Moderately strong support |
| Y chromosome not matching | 0.16 | Weak evidence against |
| Combined evidence | 6.5 million | Extremely strong support |

Such an approach is used when combining multiple aspects of a forensic scientist's findings, eg when combining individual peaks within a DNA profile. Some contrast this with the way in which the law approaches fact-finding in civil cases: 'If a legal rule requires a fact to be proved (a "fact in issue"), a judge or jury must decide whether or not it happened. There is no room for a finding that it might have happened. The law operates a binary system in which the only values are 0 and 1. The fact either happened or it did not[12].'

In statistical science, however, the uncertainty regarding facts is carried through any chain of reasoning and influences trust in the final conclusions. The expert can properly ask what the level of probability or uncertainty is. However, courts are perfectly used to bringing into the calculation of a primary conclusion uncertain disputed facts along the way and without necessarily resolving each uncertainty. Sometimes evidence going to disputed contributory facts, when combined with other evidence going to a different disputed contributory fact, may enable a conclusion to be reached safely on the principal fact in issue. Likewise, juries are commonly directed that they do not need to resolve every dispute in the evidence, so long as they are satisfied beyond reasonable doubt of the guilt of the accused. Some disputed facts can safely be left unresolved and scientific findings will generally have a degree of uncertainty rather than a definite value.

### 3.4 Coincidences and rare events

Intuition is notoriously poor at assessing how 'surprising' an event is. Just because an event is exceedingly rare for a particular person (eg winning the lottery at odds of 45 million to 1 against) does not necessarily mean it is a surprising event to occur to someone (because of the large number of tickets sold). When three major plane crashes occurred within an eight-day period in 2014, it appeared to many to be beyond coincidence. But there is around a 60% probability that such a 'cluster' will happen at some point over a ten-year period.

### 3.5 Interpretation of 'beyond reasonable doubt' and 'balance of probabilities'

These might appear to be expressions of probability for either criminal guilt or on behalf of one of the sides in a civil case, respectively. In fact, they relate to the strength of the evidence required by the legal system, the prior odds provided for, or against, a particular explanation of events and the relative merits and losses associated with accurate and erroneous consequences of a decision.

# 4. The role of expert witnesses and what should be expected from them

The guidelines for expert witnesses in England and Wales[13] provide a useful reminder that expert witnesses have a duty to give independent, impartial and unbiased evidence, and not to stray outside their area of expertise. Crucially, any reference to data that have been collected by others has to be justified, and their limitations communicated clearly. Similar considerations apply in civil cases[14].

The Forensic Science Regulator has the responsibility for reporting standards used by forensic experts and reports must be structured to provide an understanding of the probative value of the evidence. Crucially, legal professionals should be able to recognise complex and non-standard situations in which an expert in probability, forensic inference or statistical reasoning may need to be consulted.

Communication of probability is fraught with the potential for misunderstanding. One only has to look at the classic problem of the prosecutor's fallacy to see what may happen when such misunderstanding occurs. Judgments from the Court of Appeal (in England and Wales) relating to expert evidence also reveal a variation in the way in which the Court understands and accepts or rejects probabilistic evidence. In some ways, this misunderstanding reflects a basic difference in the manner by which people reason and discuss issues. Some prefer a narrative style of argument while others prefer a numerical approach. This difference does not necessarily correlate with whether you are a scientist or a lawyer.

It is important also to recognise with whom the expert is attempting to communicate. In a criminal court it will be with the magistrates or the jury; in civil litigation it will be with the judge. The onus is on the expert to use simple and accurate language in the communication of their evidence so that they are understood. It is equally important that the lawyers ask the experts questions such that the full nature of the evidence together with its strengths and limitations are exposed to the recipients (magistrates, judge, jury).

The Forensic Science Regulator is developing a quality standard for forensic practitioners on the development of evaluative opinions. This will set out a standardised, transparent approach to the expression of opinions for forensic science disciplines, based on case-specific propositions. Where it is not possible for numerical values to be assigned to probabilities, a revised verbal scale, with fewer categories than presented in Table 1, is being considered. The standard is also expected to require each expert to set out the specific data and/or experience and expertise on which their evaluation is based.

# 5. Conclusions and the future

As forensic tests become more sensitive, and the amount and complexity of scientific evidence increases, there will be a need for more sophisticated models and statistics to obtain meaningful inferences and interpretation of the evidence given the specific circumstances of a case. This includes the generation of ground truth datasets (where the provenance of the data is known) relating to trace evidence and pattern evidence as well as an understanding of the transfer, persistence, recovery and background abundance of materials in general and in case-specific circumstances. The development of concepts of 'calibrated knowledge' and their use in decision-making in respect to the generation of propositions and the evaluation of these propositions within the context of cases are highly relevant. There is an increasing use of LR calculations for pattern evidence where the LR is obtained by dividing two probability assignments. Both require judgement on the part of the expert based on a corpus of knowledge that can be divided into two broad categories:

[S] Knowledge derived from robust systematic studies, ideally published, where the relevant features have been measured and studied statistically.

[E] Knowledge derived from personal experience, ie the expert's training and professional experience in their forensic specialism.

Published scientific data are used wherever possible as a basis for these assessments. If relevant published data are not available, then data from unpublished sources or ad hoc experiments may be used as long as they have been peer reviewed and documented on file. Knowledge such as personal experience in similar cases and peer consultations may be used provided that the practitioner can justify their use and demonstrate their basis[15]. In addition to the nature of the knowledge invoked, it is critically important that the expert discloses transparently the nature, provenance, extent and relevancy of the knowledge used to inform their LR. Transparency is paramount to ensure scrutiny and ultimately to allow courts to assess credibility, ie how LRs were derived and their robustness. Because LRs may be based on different experts' knowledge, there may be a legitimate and understandable difference in opinion between two experts.

These developments will require improvements to the content of expert reports, to the competence of counsel to understand such reports and interrogate experts and to the ability of the judiciary to handle such evidence appropriately. These improvements will require actions by many stakeholders, including the Forensic Science Regulator to set standards for reporting, the academic and professional bodies to agree that the education of legal professionals should include dealing with scientific evidence, and training related to scientific evidence and its evaluation within the judicial colleges.

The lack of a common language among experts, lawyers, judges and lay people about what is meant by probability and statistics and how these concepts are used by experts to provide answers to specific case-related questions remains challenging. This can lead to misunderstandings and confusion.

There are also gaps in data and knowledge related to many types of evidence, including how materials transfer between people and between people and surfaces. Similarly, data on the persistence of materials once transferred and on background abundance of materials are sparse. This requires a greater reliance on the expert's knowledge and understanding of evidence and applying this to specific case circumstances.

In some circumstances where data are well known and well defined (eg repetitive measurements made by a scientific instrument) significance testing (Appendix 4) can be undertaken to provide a fundamental tool in uncovering relevant information about the data and what inferences can be made. An example may be the measurement of uncertainty or error relating to the determination of alcohol or drugs in a sample. Concepts such as causation and relative risk (Appendix 5) can be explored with the help of statistical methods; however, the significance of such associations remains primarily a matter of expert judgement.

This primer forms only a basic introduction to the evaluation of evidence based on statistical and probabilistic reasoning.

# Appendix 1: The use of probability

### Probability in a legal context

With the exception of evidence that is eliminative, scientific findings rarely provide evidence that is conclusive for a particular questioned event. The probative value of scientific findings needs to be assessed and taken into account by the fact-finders when considering whether a contested event in the past, for example whether the defendant committed the offence, is true. The notion of probability applies not only to the ultimate question but also to intermediate questions, such as: What is the probability that the defendant grabbed the victim's (complainer in Scotland) clothing at the relevant time?

In criminal cases, the fact-finder must believe beyond reasonable doubt and be sure that the event occurred. There have been attempts to define what 'beyond reasonable doubt' and to 'be sure' mean in numerical, probabilistic terms but there is no agreement on, or even a strong drive to adopt, a particular number.

In contrast, the fact-finder in civil cases relies on the concept of balance of probabilities. One common understanding is that the fact-finder must form a judgement, based on the strength of the evidence, as to whether their belief for the plaintiff's contention is greater than or less than 50%. The notion of probability may also be invoked in court when expert evidence is being adduced.

As an example, suppose the court is being presented with evidence of a particular bloodstain pattern in a case of assault. The expert testifies that it is 'highly probable' the bloodstain pattern would be observed on the defendant's shoes if the defendant had, as alleged, kicked the victim. The notion here is of 'observations' and expert expectations of observing material/analytical results, etc if a proposition were true rather than to discuss the 'cause' of the bloodstain pattern. Discussion of 'cause' requires the generation of explanations after the observations have been made.

To avoid possible confusion, the expert should explain to the court that 'highly probable' is a description of their expectation that this particular pattern of bloodstaining on the defendant's shoes would have been observed if the defendant had, as alleged, kicked the victim. But, for balance, it is necessary also to ask: What is the probability that this pattern of bloodstaining would have been observed if the defendant had not kicked the victim? The answer to this second question may be, for example, 'very low'.

These assignments are the expert's probability judgements made after taking into account the results of any experimentation aligned to the case circumstances, the body of documented knowledge and data in the specialism, as well as their own experience in the field. There are no such things as the 'right' or 'correct' probabilities in this case.

### Expert explanations, or expressions of possibility

There is a form of expert opinion that does not involve assessment of probability and is classified[16] as an explanation (possibility). Phrases that an expert will use to express an explanation include 'consistent with', 'could have' and 'cannot exclude'. For example, the expert may explain that "the finding of a fibre on the clothing of a suspect which is indistinguishable from fibres of the complainer's clothing could be the result of secondary rather than primary transfer". Such explanations, however, are generally unhelpful for fact-finders when deciding on the truth of a contested event because merely presenting explanations bears the potential of suggesting that the expert is opining directly on the probability of the competing versions of the event rather than on the value of the findings within the framework of circumstances of the alleged activities.

Explanations may offer assistance during the investigative proceedings but are limited insofar as there is no assessment of the probative value of the findings. The explanations may not be an exhaustive list of possibilities and there is no assessment of how probable each explanation may be, rendering them generally not useful for decision-making. However, while exploring alternative explanations (causes) for the evidence is a perfectly valid procedure before a trial, questions about alternative explanations may be posed later in court by defence counsel to dilute the force of the principal conclusions.

### The basis for assigning probabilities

All probability assignments can be viewed as being conditioned on some form of pre-existing information. That information could include relevant details of the circumstances of the case in question, actuarial data, technical data, expert knowledge, results of competency testing, etc. It is important for all those people who use and rely on probability assignments to understand the notion of conditioning.

It is important for the people who make probability assignments to declare the conditioning factors that have influenced those assignments – and it is important for the recipients of expert information to probe the foundations of those declarations. It is also vital that the fact-finder is made fully aware of those influences. The type and quantity of conditioning information taken into account will vary depending on, first, the role of the person assigning the probability within the fact-finding process and, second, the question being asked.

For some roles and some types of question, the basis for assigning a probability may be straightforward, but for others assigning a probability may be problematic because of a lack of knowledge or a lack of relevant data or because it is not within the competence of the person being asked the question. Case studies 2 and 3 provide two examples in a criminal context to illustrate the sources of expert probability assignments.

**CASE STUDY 2**

## A DNA case

The fact-finder is presented with expert evidence of matching DNA profiles extracted from a sample from a defendant and from a bloodstain left at the scene of a crime. A question for the fact-finder would be: Is the DNA extracted and analysed from the bloodstain that of the defendant? And the ultimate question would be: Is the defendant guilty of an offence? It is the fact-finder's role to answer the two questions. Whether they do so probabilistically is entirely their choice. It is not the expert's role to answer the two questions, however tempting it may be to answer the first question. What the expert can do is to provide the fact-finder with their expert, justified view on the probability of observing the DNA evidence, ie DNA profile from the bloodstain and the defendant's DNA profile, under two competing propositions. If we look at the first question, the pair of competing propositions would be:

- Hp (the prosecution proposition): The DNA originated from the defendant.
- Hd (the defence proposition): The DNA originated from someone unrelated to the defendant.

The expert can help the fact-finder by providing their probability of obtaining the DNA evidence if the DNA had originated from the defendant and, alternatively, if it had originated from someone unrelated to the defendant.

The first probability assignment, ie the probability of obtaining the DNA evidence if the DNA from the scene had come from the defendant, would be based on the expert's knowledge of the reliability of the process of DNA profiling in producing the true profile from a stain. The expert should base their assessment on whatever relevant data there may be. Turning to the alternative proposition, additional considerations for assigning a probability of the DNA evidence come into play. Under the proposition that the DNA had originated from someone unrelated to the defendant, the probability of obtaining a match with the defendant depends additionally on the proportion of individuals who have this DNA type in the relevant population.

There are databases of DNA profiles from samples of people from various ethnic groupings. These databases can be consulted to assess the rarity of the (matching) DNA profile. That statistic can then be used as a basis, among other considerations (such as genetic relationships among individuals), for assigning a probability of obtaining a match IF the bloodstaining had originated from someone unrelated to the defendant. What the expert brings to that assignment is knowledge and understanding of the impact of relatedness among people, in the form of a population-genetic model, and of choosing the most relevant database(s) for the case in question. It is not just a simple question of using a frequency of occurrence as a probability: a more subtle treatment is required.

**CASE STUDY 3**

## A firearm discharge residue case

A person has been accused of discharging a firearm during the commission of a robbery. The firearm was not recovered but spent cartridge cases were found at the scene. Swabs were taken from the hands of the defendant about five hours after the incident. Subsequent chemical analysis of the swabs revealed the presence of a small amount of firearm discharge residue (FDR, also called a gunshot residue, GSR) that had the same qualitative composition as the reference FDR recovered from the cartridge cases.

The fact-finder's questions would include: Did the FDR on the defendant's hand swabs come from the gun that fired the cartridges? And: What is the probability that the defendant fired the gun, given matching FDR had been found? The expert's role is to offer probabilities for observing the particular FDR, given the truth of the prosecution proposition and defence alternative that flow from the questions facing the fact-finder. Looking at the issue of whether the defendant fired the gun, the two competing propositions would be:

- Hp: The defendant fired the gun (at the relevant time).
- Hd: The defendant did not fire the gun (it was some other person).

Note that, for a sensible definition of the alternative proposition, it is relevant to enquire, if possible, about what the defendant says. There may be situations in which the defendant asserts that he was a bystander, or that he provided first aid to the victim. These details are relevant to help specify the alternative proposition and task-relevant conditioning information. Defining the alternative proposition as the simple negation of the prosecution's proposition is usually not sufficient. The expert should be able to offer probabilities for obtaining the FDR evidence given the truth of these propositions.

The task to be addressed is that the propositions are contested – we do not know which of the competing propositions is true. Did the defendant or someone else fire the gun? The principles of inductive logic (where the conclusion may be probable based upon the evidence presented) dictate that, to assess the probability that an uncertain proposition is true, the probabilities of the evidence under the competing propositions need to be considered (see Section 2.5 on doping). So, how does the expert assign such probabilities?

### Under the first proposition

The defendant fired the gun (at the relevant time): the expert would rely on whatever is known generally about transfer and persistence of FDR and apply that knowledge to assign a probability for obtaining the evidence under this proposition. The expert may consider undertaking experiments to replicate as far as possible the conditions of the incident, to provide more data and knowledge. The probability that they assign will be their best assessment of the probability of obtaining the evidence. Other experts may disagree with the assignment but, with transparent communication and explanation, the expert should be able to demonstrate and justify how they arrived at a particular probability.

### Under the alternative proposition

The defendant did not fire the gun (at a relevant time, it was some other person): the expert would use whatever survey or other data there may be to help them consider the probability of obtaining the evidence under this proposition, and given the task-relevant information. Again, this will be the expert's considered view on that probability. Generally, the body of knowledge upon which the expert relies should be available for auditing and disclosure.

**How do we know if the expert's probability assignments are reliable?**

The answer to that lies, first, in the expert being able to explain in a transparent way the basis of their assignment and, second, in the expert revealing the results of any relevant calibration of their past opinions through, for example, competency testing. The expert should be able to defend their opinion upon challenge. Through transparency and effective explanation on the part of the expert, the fact-finder should be able to understand the basis of the declared probabilities and take a view on how to incorporate the expert's opinion in the fact-finder's own reasoning process.

**How may experts use probability to assist fact-finders in their decision-making?**

It is important for all participants in the justice process to understand and be clear about several key issues and questions in the use of probability. These are:

- The essential distinction between probabilities for evidence (usually the expert's observations and analytical results) and probabilities for propositions (ie the facts in issue).
- Whether probabilities for propositions are being assigned before or after expert evidence is presented. If before the consideration of expert evidence, these probabilities are called prior probabilities; if after the presentation of expert evidence, they are called posterior probabilities (see Section 2.5 on the doping example).
- Who is best placed, in terms of their roles in the legal process, to provide these different types of probability?
- What is the information that has conditioned (or influenced) the assignment of probabilities to the propositions, and is it relevant to the task?

Generally, the probabilities for obtaining expert observations, conditioned on the truth or otherwise of the proposition in question, are in the domain of the experts. The expert should have sufficient data and the knowledge and understanding of the evidence to assign defensible, informed probabilities. The expert should be able to convey and explain these probabilities to the fact-finder to help them deliberate on the truth or falsehood of the proposition in question. It is the fact-finder who has received other, non-expert evidence in a case and who is therefore in the best position to take a view on the truth of the proposition in question. However, this is not a hard-and-fast rule and there will be some situations in which the expert can provide informed probabilities for the truth of the proposition.

### What are the limitations of using probability?

Perhaps the main limitation is the lack of a common understanding among experts, lawyers, judges and lay people of the notion of probability and the extent to which it can be applied. Until there is a shared understanding and a common language about probability, and an agreement on how best to express probability, then misunderstandings and confusion will occur, resulting in expert evidence being valued inappropriately or perhaps ruled as inadmissible when it might be helpful and valid. Even among experts, there is misunderstanding about, and a variation in, the adoption of probability. In the field of DNA profiling, practitioners use probabilistic software for the logical interpretation of complex DNA mixtures. Without such software, the interpretation of such mixtures would be very difficult. Discussion of the potential and limitations of such software is provided in Forensic DNA analysis: a primer for the courts[17]. In other fields, however, practitioners are only just beginning to explore probabilistic thinking. The Royal Statistical Society and the Inns of Court College of Advocates have together produced guidance on the use of probability[18] that provides a good starting point.

Another limitation is the lack of relevant data to inform probabilities in some areas of expertise. In some areas, such as in textile fibres and 'touch' DNA, there is an extensive body of research and survey data on which to draw. In other areas, such as toolmarks or ballistics, there is only limited published research on important considerations relevant for assessing probative value. In the absence of reliable, informative and structured data, the expert must rely on their knowledge and understanding of the evidence type, provided that the basis of such opinion is documented, can be audited and is disclosed. It is in such areas particularly that evidence of the reliability of the expert's opinions would be highly desirable.

# Appendix 2: Evaluation of trace evidence

Trace evidence refers to any material that is transferred between persons or objects during contact and is commonly recovered in connection with an alleged crime. The term is often used to refer to the collection of materials frequently encountered by forensic scientists such as glass, paint, fibres, firearm discharge residue and DNA.

In forensic casework, the objective is often to compare known material and questioned material. Either can originate from the suspect or from the scene. Examples include glass from a window compared with glass fragments recovered from a suspect or fibres from a suspect's jumper compared with recovered fibres from a victim. The material might be compared at the source level (eg are the fragments from the same window or not?) or at the activity level (eg did the suspect break the window or not?). Source level is rarely sufficient to assist the questions relevant to the case. Glass recovered from a suspect that is indistinguishable from a broken window is of little value without an assessment of how probable such a finding is if the suspect broke the window versus that the suspect had nothing to do with the breaking.

In order to tackle the question of source, forensic laboratory analysis is typically undertaken to measure different features or characteristics of the known and questioned material. To interpret the weight of the evidence (or likelihood ratio (LR)) when comparing material at the source-level usually requires evaluating the similarity of the features of the two sets of material (how closely they match each other based on the analysis undertaken) and the rarity of the observed features. The rarer the features observed, the stronger the evidence may be. See Appendix 3 for more details on source-level comparisons.

Analysing trace evidence at the source level alone, without reference to the activities associated with the evidence, can be misleading. It is often the case that the most relevant questions are related to the activities which may have led to the trace materials being transferred. In order to obtain results that are helpful to address these questions and that are not misleading, other factors need to be considered in addition to the source-level questions. These other factors include the probabilities of transfer, persistence and recovery of the material in the context of the alleged activities. Statistical approaches which only assess the similarity and rarity of the materials can miss factors which affect relevance within the context of the circumstances of a given case – this can have a major impact on the evidential weight of the findings.

Trace evidence must be viewed in the context of the case. As described in Section 2, at least two competing propositions should be addressed. In addressing these for a specific case, the expert considers how probable the findings are in each of these competing propositions. The result is presented in the form of an LR. This highlights that the results do not have a stand-alone value; rather, their value is dependent on the proposition being addressed or the questions to be answered.

The issues to be considered to address activity-level propositions are similar for all trace materials but the factors affecting the issues vary from one material to another. To assess how probable it is to find matching materials (glass, fibres, etc) if a particular action took place, data on transfer, persistence and recovery are needed as well as data on how common the materials under consideration are in the environment. Levels of background abundance are also needed when considering the findings if the alternative is true, ie that the activity did not take place or was not carried out by the suspect. An example is knowing the background abundance of groups of glass fragments which would be found on clothing in a given population. Sometimes, where insufficient data are available, expert judgement or personal opinion are used to assign the required probabilities. This should be made explicit in the report. Ideally ad hoc experiments should be carried out in the absence of data.

### Assumptions and data

When evaluating LRs for trace evidence, statistical assumptions are needed both at the source level and at the activity level. For example, it may be assumed that the distribution of measurements follows a particular statistical model or that two events that are alleged to have occurred are statistically independent of one another. These assumptions will depend on the type of evidence and the competing propositions being considered. For some evidence types and propositions, the statistical assumptions are well understood and validated. For others the methods are less well developed. When analysing evidence at the activity level, propositions can be very case dependent. As a result, the statistical approaches and datasets used to evaluate the evidence may be based on the personal judgement of the forensic expert.

Where any statistical assumptions or datasets have been used to evaluate evidence, these should be clearly explained and justified in the case report. It is important that checks have been carried out to test whether the statistical assumptions used are appropriate for the evidence type and the propositions being assessed. One way of doing this is to test the statistical approach on an existing dataset where the ground truth is known and to assess the proportion of times that the LR gives a misleading or incorrect result.

For example, for a source-level comparison this would mean evaluating the proportion of times that the LR is greater than 1 when the two sets of material are from different sources and the proportion of times that the LR is less than 1 when the two sets of material are from the same source. Both proportions should be small in a model that fits the evidence type for which it is being used.

### Fibres

Fibres are shed from surfaces of various materials such as clothing, carpets and car seats. They vary greatly in composition and colour. Studies have found that fibres which are common, such as blue wool, have not been detected on surfaces in high numbers except in areas where a known source has been in contact. Hence, fibres can be very useful in reconstructing the activities that occurred during the contact between textiles. The tendency to shed fibres is governed by many factors, including the looseness of the weave, the size of the fibres and the age of the garment. This is easy to visualise when we consider the difference between the shedding of a new carpet and that of one that has been in place for some time.

Whether fibres transfer or not depends on the shedability and on the type of contact. Information on both factors is needed to assess the range of fibres likely to be transferred. A smooth shell suit will not be expected to yield transferred fibres even if the contact is prolonged while a woollen jumper will give rise to transferred fibres with limited contact. Little peer reviewed published literature exists in relation to the shedability of fibres. These considerations highlight why case context is important in assigning the probabilities of the findings in competing scenarios. Fibres are best considered in the totality of the case and it is rarely useful to consider only their sources or the presence of a single fibre.

Transfers of fibres can be by direct contact (eg from the suspect's clothes to the victim's clothes) or by indirect contact (eg fibres transferred from the victim's clothes to the suspect's clothes via an intermediary object); the latter is called secondary transfer. It is not possible for the expert to opine on whether the transfer of fibres was primary or secondary transfer, or whether or not the transfer occurred during the alleged activity. However, fibre experts are well placed to assess the probability of particular findings in either scenario.

### Example 1: competing sources of fibres

A sexual assault is reported as having taken place in the bedroom of a residence. This is denied by the suspect, who alleges that consensual sex occurred in the living room and that he never entered the bedroom. Fibres are an ideal evidence type to help distinguish between these two scenarios. For example, a large number of fibres indistinguishable from the bedcover would be expected to be recovered from the suspect's socks and fibres indistinguishable from the socks and the bedcover would be expected to be recovered from the victim if the assault took place in the bedroom.

A large number of fibres indistinguishable from the bedcover would not be expected to be recovered from the suspect's socks given the alternative scenario. In this case, the LR (ie the weight to be assigned to the findings) would involve the relative consideration of these probabilities and the pair of competing propositions would be:

- Hp: The assault took place in the bedroom.
- Hd: Consensual sex took place in the living room.

Given the above assignments, if a large number of matching fibres were obtained, we would expect a value of the LR above 1, supporting the proposition that the assault occurred in the bedroom. The actual values for the probabilities will depend on factors such as the tendency of the socks and the bedcover to shed fibres, the time between the incident and the seizure and examination of the items, the frequency of occurrence of fibre types in given situations and the statistical assumptions used to link these factors together. The number of recovered indistinguishable fibres from the bedcover are relevant because it is possible that a small number of bedcover fibres may be present in the living room. Much of this becomes a matter of professional judgement and experience.

### Example 2: missing fibre types

In an alleged assault, the victim (complainer in Scotland) is wearing a green T-shirt and the suspect a red football supporter's jersey. The jersey is found to consist of a range of coloured polyester and cotton fibres. Shedding tests show that the fibres shed in roughly equal proportions. The red polyester fibres recovered from the victim's T-shirt are found to be indistinguishable from the suspect's jersey but no cotton fibres are recovered. No green fibres matching the victim's T-shirt are recovered from the suspect's jersey. Here we have an example of the findings not fitting the expectations in the context of the case. The pair of competing propositions would be:

- Hp: The suspect grappled with the victim at a given time.
- Hd: The suspect never had any contact with the victim.

With the information above, the expert will inform that the findings are unlikely given Hp but more expected given Hd. The LR will be less than 1, ie supporting Hd.

This example is designed to highlight the necessity of considering the findings in competing scenarios and considering the context of the case. The matching red fibres viewed in isolation are misleading no matter how robust the analytical tests applied or how rare are the fibre types.

## Glass

When a window is broken a large number of fragments fall back in the direction of the blow. Thus, a person delivering the blow is expected to have small fragments in their hair or on the surface of their clothing depending on how close they were to the window as it was breaking, the height of the window, the type of glass and the activities that followed[19]. One of the main tests used to examine glass fragments at the source level consists of measuring the refractive index, which varies both within a pane of glass and between sources of glass. The refractive indices of the glass fragments from the window are measured and glass fragments from the suspect are also analysed and put into matching and non-matching groups of glass if glass of more than one refractive index is present[20]. However, assessing the closeness of the 'match' between glass recovered from clothing and the window glass does not provide sufficient information to evaluate propositions concerning whether or not the suspect broke the window. Even when additional analytical tests are applied, this activity-related question is not answered. Other information is required relating to, for example, how glass fragments are transferred and retained following the breaking of glass objects, or how prevalent

glass fragments are on surfaces not directly connected to a recent break. Glass fragments can be exchanged following a large range of daily activities.

To evaluate the glass fragment results, information is needed on how probable the findings are if the suspect broke the window(s) against how probable the findings are if the suspect had nothing to do with breaking the window(s). To address the first probability, two possible ways that the glass fragments can arise must be considered – either fragments were transferred when the window was broken and non-matching glass, if present, was already on the clothing or no glass was transferred from the window and all the glass on the clothing, both matching and non-matching, was already there. To assess the probability of the findings (ie glass matching the window found on the clothing), if the suspect had nothing to do with breaking the glass, we need information on the probability of finding glass on innocent members of the population. This information is critical, and it is useful to consider a population as close to the suspect as possible. One well-known dataset considers glass on the clothing of persons who come to the attention of the police rather than the general population.

### Firearm discharge residue (FDR)

FDR is a combination of small particles produced when a gun is fired following the explosion in the barrel. This evidence type is complicated, consisting of a non-uniform population of particles which includes so-called 'unique' particles containing lead, barium and antimony. However, to be considered FDR, other characteristic particles also need to be present. Accepting this means that a single particle is not FDR.

Different ammunitions give rise to different residues, but a very high number have similar compositions providing little discrimination. Even when the ammunition type is known, it is considered good practice to compare residue from the discharged cartridge case, barrel of the gun or the bullet hole – the known material(s) – and the residue recovered from the hands or clothing of the suspect – the questioned material(s). This is because variation in the proportions of particles can occur. It is common to see reports from forensic scientists in which the number of particles recovered is factually reported, or a statement that FDR was detected. Such statements can often be accompanied by a disclaimer that the findings are 'consistent with' the suspect being close to a person firing the gun or touching a surface with firearm residue on it. Similarly, a negative finding is explained away by loss or time delay. The evaluation of the meaning of such findings in the context of the case is generally left out of the report.

However, as with other trace evidence types, the more relevant question is whether or not the suspect undertook an activity which could result in the scientific findings. In the case of FDR, it is more meaningful to address whether the findings are more or less likely if the person fired or did not fire the gun in the circumstances of the case. These assessments will go beyond statements of consistency but should be qualified in terms of probability.

Example

In a given set of circumstances, the expert may indicate that the probability of finding FDR on a person who fired a gun is high if the person is sampled soon after the gun had been fired. That expectation will be balanced against the probability of the finding if the suspect did not fire the gun. The latter probability is low as surveys on members of the general population show few instances of FDR. Depending on the circumstances, finding FDR on a person's hands is expected to provide support for firing of a gun rather than not firing a gun, ie the LR will be greater than 1.

To inform these probabilities and assign a meaningful LR, information is needed on the type of weapon and length of time between the alleged firing and sampling. Ideally, tests should be carried out under the conditions of the known circumstances of the case. Data regarding the presence of FDR as a background in a given population (of individuals or objects) is also required to assess how prevalent the material may be considering activities other than discharging a firearm. At the current time the understanding of the transfer, persistence, recovery and background abundance of FDR is limited.

### Drugs on banknotes

Banknotes seized from people who have been found guilty of drug crime on average have higher levels of drug contamination than banknotes found in general circulation. Different analytical techniques can be used to obtain measurements of drug traces on banknotes and can result in different measurements. Hence it is important that comparisons are made using datasets obtained using the same analytical technique. For some drugs (eg cocaine, which is found on most banknotes) the measurements of drug found on the set of seized banknotes are related to the quantities of drug on the notes. For other drugs (such as heroin) the measurements might simply be the presence or absence of the drug on each banknote.

The strength of this evidence in relation to the following propositions can be evaluated using an LR and the pair of competing propositions would be:

- Hp: The banknotes are associated with a person involved in drug crime.
- Hd: The banknotes are from general circulation.

Statistical models can be used to evaluate the LR; for some examples, see Wilson *et al*.[21]. The assumptions behind the statistical models must be checked and the models validated. Selecting suitable databases for these statistical models can be a challenge[22] as there may be both regional variations and variations over time associated with particular drug use behaviours in different areas. Having relevant localised ground truth data (analysed samples from known locations over different known time periods) is essential to assess whether this is the case. It is important that the datasets are consistent with the propositions. For example, if the propositions are specific to a particular drug, then the dataset should also be specific to that drug over the appropriate time period. It is also difficult to obtain a dataset of banknotes 'associated with a person involved in drug crime', which can make it difficult to estimate the probability of the findings under Hp. It is therefore key that the statistical assumptions supporting data selection are described and justified.

Many of the general guidelines described in the four scenarios presented above are applicable to other types of evidence. The same can be said of the challenges posed by the limited availability of ground truth databases and lack of experimental studies for estimating transfer, persistence and recovery probabilities for most types of trace evidence. Going forward, another challenge will be to keep up with emerging materials such as glass for smartphones and organic ammunition for guns. Successful quantification of the evidential value of such trace evidence and understanding of the limitations of the methods developed will depend on the availability of reliable databases and experiments simulating case scenarios for the different types of trace evidence. It will also rely on appropriate probabilistic and statistical models, which will need to be tested and validated for each type of trace evidence and for different sets of competing propositions.

# Appendix 3: Evaluation of impression evidence

The purpose of this appendix is to explain the type and significance of the conclusions reached by forensic experts dealing with impression evidence. It will also explain the basis (statistical or otherwise) upon which these conclusions are formed.

### What is 'impression evidence'?

The term 'impression evidence' is used here to refer to the field of forensic examination in which an expert is asked to compare, primarily using visual methods, items of disputed source with items of known source. These items are called 'impressions' because they are marks or signs left by a person or an object following physical contact with a surface. The field covers a large number of forensic specialisms, each having its own terminology when referring to these impressions (Table 3). In this appendix, the generic term of 'impression' will be used when the argument holds for all types of impressions. When making examples in a given forensic specialism, its specific terminology will be used.

Questioned impressions are of disputed sources (a person or an object). They generally have been left unintentionally and recovered in association with the investigation of some criminal activity (eg a fingermark in blood recovered from a crime scene). Known impressions are of undisputed sources (person or object) and have been obtained under known and controlled conditions (eg reference fingerprints from a known individual taken in police custody).

### What is the purpose of the forensic examination?

For all these specialisms, the purpose of the forensic examination is to help assess whether or not questioned impression(s) originated from the source(s) that produced the reference impressions. The issue can typically be phrased as follows:

- Were the (finger-/palm-) marks recovered at the scene left by this individual or by some other unknown person?
- Were the (footwear-/tool-) marks recovered at the scene made by this object (shoe or tool) or another unknown shoe or tool?
- Was the bullet recovered from a body fired by the seized firearm or by some other unknown firearm?

**TABLE 3**

Some forensic specialisms and relevant terminology.

| Forensic specialism (as colloquially referred to) | Questioned impression | Known impressions |
|---|---|---|
| Fingerprint examination | Fingermark<br>Palmmark | Finger prints<br>Palm prints |
| Barefoot examination (with no visible, friction-ridge, skin impressions) | Barefoot mark | Barefoot prints |
| Footwear mark examination | Footwear mark | Known shoes and their associated reference prints (reference prints taken in two dimensions or impressions in three dimensions) |
| Toolmark examination | Toolmark | Known tools and their associated reference impressions made on a soft surface |
| Firearm examination | Marks on a questioned fired bullet<br>Marks on a questioned fired cartridge case | Known bullets fired by a weapon<br>Known cartridges fired by a weapon |

It is important to stress that the issue of origin (or source) in all of these specialisms is generally associated with an implied activity made by a person or carried out using an object. It is these implied activities that led to the production of the questioned impressions (Table 4).

These activities are often not referred to specifically in the expert's report but remain implied. Indeed, experts will not systematically envisage all conceivable possibilities for a questioned impression to be produced but will consider the most reasonable activity arising from the case circumstances. For example, experts, unless instructed otherwise, will not account for fanciful scenarios such as:

- a fingermark not being left on the surface by a living hand but using a forged dummy finger obtained from an individual; and
- a footwear mark in the snow being the result of the landing of a shoe after being discarded from a car.

**TABLE 4**

Implied activities resulting in the impression.

| Forensic specialism | Implied activity |
|---|---|
| Fingerprint examination | The individual handled an object or touched a surface with bare hands. |
| Barefoot examination | The individual walked barefoot on a surface. |
| Footwear mark examination | A person wearing these shoes walked on the floor. |
| Toolmark examination | A person using this tool forced the safe, the door or the window. |
| Firearm examination | A person with that firearm fired a cartridge that led to a bullet and a cartridge case. |

### How is the comparison work carried out?

For all of the above forensic specialisms, the forensic examination of a given questioned impression starts with its analysis, where the examiner assesses the relevant visible features, gathers as much information as possible from the questioned impression alone and assesses if it can be used for comparison purposes. Questioned impressions deemed of no value will not be compared any further. The term 'no value' will typically be assigned to impressions showing insufficient detail to allow a meaningful comparison. Only impressions that are declared 'of value' will be compared (side by side or by superimposition) with potential sources in the form of known impressions. The potential sources have been generated either through police investigations (eg an arrest or a seizure) or by searching a database holding known impressions, attributed to individuals, that have been previously put on record. The comparison stage consists of deciding if the features of the questioned impression observed in analysis are in sufficiently close agreement or disagreement with the submitted known impressions.

Relevant features are specific to the forensic specialism and are detailed in Table 5. They can be shared by many (such as the manufacturing size of a tool shared by all tools produced of that size) or by a few (such as the acquired damage in the form of cuts on the outsole of a shoe). In other words, the discriminative power of relevant features varies depending on the specific type of feature considered. Furthermore, depending on their size and on how the impression is produced (types of surfaces, residue on the surface, movement, materials, etc), features may not be reproduced in the impressions and even when features are made by the same source (eg a shoe) impressions are never identical. Hence, features in disagreement may be found between impressions, despite sharing the same source. This is because the respective impressions have been subject to distortion, movement, superimposition or background noise or have changed appearance over time. On the other hand, while findings in agreement should be observed when the impressions were produced by a common source, they may also be found when they were produced by different sources. Matches between different sources are known as adventitious matches because another source has produced, by chance, the same level of agreement.

**TABLE 5**

Relevant features of different evidence types.

| Forensic specialism | Relevant features (not exhaustive) |
|---|---|
| Fingerprint examination | The general flow of the friction-ridge skin (papillary lines), often classified in general patterns such as arches, loops and whorls.<br>The ridge endings and bifurcations (referred to as minutiae or Galton's details) made by the papillary lines and the combination thereof.<br>The marks left by scars or other damage.<br>The specific shapes of edges of the papillary lines and the pores present in them. |
| Barefoot examination | The size (length and width) of the foot from heel to toes and the relative size and position of the toe impressions. |
| Footwear mark examination | The overall manufacturing design of the outsole (the geometric elements of the design and how they are arranged relative to each other).<br>The size of the outsole as specified by the manufacturer.<br>The general level of wear of the outsole at the time of the impressions.<br>The acquired features shown in the impressions in the form of cuts, removal of material or damage to the outsole. |
| Toolmark examination | The width and size of the tool and its shape as given by the manufacturer.<br>The acquired defects on the surface, removal of material and small imperfections of the surface due to its usage. |
| Firearm examination | The calibre, number, widths and twist of the lands and grooves of the barrel through which a bullet was fired.<br>The relative positions of the firing pin, extractor and ejector coming in contact with a cartridge case.<br>The striated marks, due to usage, left on the impressions of lands and grooves on a fired bullet.<br>The breech face impression left on the back of the cartridge when fired. |

The results of all these observations on both the known-source and the questioned impression (in agreement or disagreement) are what will be called the comparison findings. These findings will then be evaluated with regards to the proposition of common source against the proposition of different sources. This evaluation stage encapsulates the interpretation of the findings and generation of associated conclusions. In the UK the evaluation is carried out holistically by the expert using their knowledge and experience. It is more infrequent that such an evaluation is undertaken using a likelihood ratio (LR) approach. However, such approaches are encouraged by forensic practitioners and are used elsewhere, particularly in continental Europe. Finally, it is customary for each examination followed by a conclusion to be reviewed independently by a second examiner. This is called the verification stage. The four above stages (analysis, comparison, evaluation and verification) are generally referred to by the acronym ACE-V. In the UK, it is common to use this approach for the comparison of fingerprint evidence but ACE-V is not necessarily used for other types of evidence involving the comparison of visual patterns.

**What conclusions do experts reach following a comparison?**

In all specialisms, it will be customary for the expert, when the comparison findings allow, to reach a conclusion of either an 'identification' or an 'exclusion'. An identification is declared when the examiner is subjectively convinced that the mark could not have been made by any other source, even if all other possible sources have not been examined. This means that the expert takes a decision on the issue of the source of the impression.

The term 'identification' refers to the decision of the expert that the questioned and known impressions originated from the same source. It is a categorical opinion, and should not be misconstrued as being a factual certainty. No forensic examination covered in this appendix can claim to factually demonstrate the source of a questioned impression. A decision of 'identification' is not a fact; it is the opinion of an expert based on their measurements, observations and experience and it is a statement of an expert's probability that the impression was made by different sources other than the questioned source is so small that it is negligible.

The opinion of an 'identification' should not be taken as stating, for example:

- that the impression was associated with a specific individual or object to the exclusion of all others in the world; or
- that it is absolutely certain (or with a 100% certainty) that a specific individual or an object is the source of the questioned impressions; or
- that it is the result of the comparison of all impressions in the world's population.

The term 'exclusion' refers to the decision of the expert that the questioned and known impressions did not originate from the same source; in other words, they have been produced by different sources. When a decision of 'identification' or 'exclusion' cannot be reached, the expert may either, depending on the forensic specialism:

- indicate that the comparison is 'inconclusive'; such a conclusion could be more than the expression of neutral findings but does not provide 'certainty' in the eyes of the examiner; or
- indicate the strength of support the findings will bring to the question of the source; that strength will be qualified either verbally or numerically.

Table 6 shows these options as practised in a few of the forensic specialisms considered. Resorting to 'inconclusive' as opposed to a more nuanced range of possibilities expressing graded degrees of support depends on the habit of the expert or on institutional/organisational policy. For example, while many fingerprint or firearm experts will only ever report 'inconclusive' in instances where they are unable to either 'identify' or 'exclude', other experts in those fields will qualify their 'inconclusive' conclusions.

### How do experts decide on an identification or an exclusion?

A decision on identification is taken by the expert when, in their opinion, there are insufficient features in disagreement to conclude that the impressions came from different sources and that the observed features (between questioned and known impressions) are in sufficient correspondence such that the expert would not expect to see them repeated if that impression came from a different source. As indicated before, these conclusions are not facts and the expert must concede that there is some probability, however small, that their decision might be erroneous.

A conclusion of 'inconclusive' means that no categorical decisions ('identification' or 'exclusion') have been made and that no meaningful assessment of its value is given to the court. This may be due to various reasons, among them the low quality of the impressions, the failure of the known impressions to fully represent the source, or comparison findings that are not judged sufficiently discriminating to make a categorical decision. However, the findings in such an 'inconclusive' case may still provide support one way or the other towards the issue of source. In such cases, depending on the forensic specialism, the expert may provide guidance as to the weight to be assigned to these comparison findings in favour of the proposition of common source or of different sources.

This can be done by expressing the degree of support that the findings provide in favour or against these propositions using an LR based on a qualitative assessment for impression evidence and expressed using a scale such as that provided in Table 1.

**TABLE 6**

Some of the range of conclusions reached.

| Conclusions | Meaning of conclusions |
|---|---|
| Exclusion | Exclusion |
| Inconclusive | The findings provide (qualified) support for the questioned impression originating from another source rather than the source under examination.<br>The findings do not provide support either for the source under examination or for another source.<br>The findings provide (qualified) support for the questioned impression originating from the source under examination rather than another source. |
| Identification | Identification |

For example, in a case involving a footwear mark where the expert noted a correspondence between a mark and a sole in terms of overall design, size, general level of wear and the presence of three cuts located in the heel area, and there is no significant discrepancy, they may state:

- the comparison findings, in my opinion, provide extremely strong support for the view that the mark has been left by that sole as opposed to some other sole; or
- the comparison findings are more than 1,000 times more likely to be observed if the mark has been left by that sole rather than if the mark has been left by some other sole.

If the findings had been different, for example if the general level of wear on both the mark and the sole were different and not easily reconcilable (with potential additional wear over the time period between the recovery of the mark and its seizure), the expert may express (reversing the propositions):

- that the comparison findings, in my opinion, provide moderate support for the view that the mark has been left by some other sole; or
- that the comparison findings are between 10 and 100 times more likely to be observed if the mark has been left by some other sole.

The LR is always assigned numerically first and then, if so chosen, translated into a verbal expression. In most cases involving impression evidence, the LR will be expressed by an order of magnitude (10, 100, 1,000, 10,000, 1,000,000, etc). If this methodology is used then an LR should be the prerequisite required to reach any conclusion, including an identification or exclusion decision. An LR is obtained by dividing two probability assignments. Both require judgement on the part of the expert based on a corpus of knowledge that can be divided into two broad categories:

[S] Knowledge derived from robust systematic studies, ideally published, where the relevant features have been measured and studied statistically.

[E] Knowledge derived from personal experience, ie the expert's training and professional experience in the forensic specialism.

Because LRs may be based on different experts' knowledge, there may be a legitimate and understandable difference in opinion between two experts. Table 7 gives, for a few specialisms, examples of the type of knowledge used by experts.

**TABLE 7**

Types of knowledge used in some forensic specialisms.

| Forensic specialism | Nature of the knowledge used by the examiner to assign a likelihood ratio (LR) |
|---|---|
| Fingerprint examination | [S] Relative frequencies of fingerprint general patterns or of types of minutiae observed on fingerprints. LR computations obtained from a statistical model.<br>[E] Knowledge on the distortion marks may show in comparison with their corresponding prints owing, for example, to the elasticity of the skin or the movements of the hand when grasping an object. Personal experience derived from the systematic observations of fingerprints from different individuals. |
| Footwear mark examination | [S] Studies on the relative frequencies of the general designs of the outsoles and their sizes in the selected population. Data showing how wear develops on used outsoles.<br>[E] Knowledge of the variations in size between a mark and its corresponding shoe owing to its deposition on the floor. Knowledge on the occurrence and persistence of inclusions such as stones caught between polyurethane structures of the outsole. |
| Firearm examination | [S] Studies associated with the systematic search of matching features occurring on bullets fired by different firearms. Data associated with the evolution of striated features on bullets due to their successive firings in a given barrel.<br>[E] Knowledge of how features produced by the manufacturing process can be distinguished from features acquired through the use of a firearm. |
| Barefoot examination | [S] Studies on the variability observed between features measured on barefoot impressions from different individuals.<br>[E] Knowledge of the variations between barefoot impressions from the same donor. The variations can be due to the donor, the walked surface or the residue left on the surface (eg blood). |

Before an expert decides on an identification or an exclusion, they assess the comparison findings in the form of an LR. The expert decides an identification when the LR is sufficiently high in their opinion to safely take that decision bearing in mind all other relevant evidence. Conversely, the expert decides an exclusion when the LR is sufficiently low for them to decide accordingly. What represents an LR as being sufficiently high or sufficiently low cannot be precisely defined and has an element of subjective judgement.

### Use of a likelihood ratio approach in practice in the UK

The explicit use of the LR approach for the interpretation of impression evidence is not yet the norm in the UK in all specialisms involving impression evidence but may be used for some evidence types (such as footwear comparison). The LR should be used as a measure of support for hypotheses in regards to identification or exclusion of a potential source in the context of impression evidence. This re-emphasises that the decision reached by the expert is an expression of personal belief and not a statement of fact. For example, if an expert when concluding a fingerprint identification ultimately decided for themselves that, given the LR, the identification can be made and that the risk of an erroneous identification is, for them and in the case circumstances, acceptable, then this may be incorrect. Information not known to the expert (eg that the defendant was in prison at the time of the alleged burglary where marks have been left) may provide a completely different perspective on the case, requiring the expert to revise their opinion. An opinion of 'identification' in these circumstances goes beyond the assessment of the comparison of findings and requires that the expert takes into account additional elements not of a scientific nature (eg the broad number of individuals or objects that could be the source of the impression). In these forensic specialisms, where the court allows the expert to consider non-scientific evidence, the court defers to the expert on the issue of the source of an impression. That delegation of the source decision has been the general practice in the UK. However, if courts decide not to allow experts to go that far, then the LR would be an appropriate way to express the strength of the expert's findings.

# Appendix 4: Statistical significance

Significance testing is a fundamental tool for scientific discovery that is widely used in medical and other scientific literature as a basis for making claims. The logic of the procedure is somewhat analogous to a criminal case. The steps for a significance test are as follows:

- A null hypothesis is presented. This is generally the proposition that the discovery is false, eg a pharmaceutical has no beneficial effect or the Higgs boson does not exist. This null hypothesis is set up as a default assumption and is only rejected if there is sufficiently convincing evidence.
- A test statistic is chosen, for which large values would tend to cast doubt on the null hypothesis. For example, the average observed benefit in patients within a control group compared with the average observed benefit in patients given a particular drug.
- Data are collected and the observed value for the test statistic is calculated.
- The probability of getting the observed value (or a more extreme value) of the test statistic given that the null hypothesis is assumed to be true is calculated. This is known as the P-value.
- If the P-value is very small, then the null hypothesis is rejected. The definition of 'small' depends on the stringency required: a standard threshold for declaring statistical significance (that a difference between the tested data and the expected value is a real difference and not just due to chance alone) is to find a P-value of less than 0.05 (1 in 20).

  For example, new pharmaceuticals generally require at least two independent clinical trials with P-values less than 0.05 to be able to claim they are effective.

  To claim the existence of the Higgs boson, physicists required a P-value of less than 1 in 3.5 million.

The P-value is essentially a measure of the incompatibility between the observed data and a pre-specified hypothesis: if the P-value is very small, either the null hypothesis is true and a very surprising event has occurred or the null hypothesis is false. Many problems can arise in the use and interpretation of statistical significance testing.

- The P-value is the probability of extreme evidence, given that the null hypothesis is true, but it is often interpreted as the probability that the null hypothesis is true, given the evidence. This is an example of the prosecutor's fallacy.
- If the null hypothesis is not rejected, it does not mean it is true. This is similar to someone who is not found guilty in an English court: they are found 'not guilty' rather than 'innocent'.
- With a large dataset, a statistically significant result may not necessarily be of any practical significance, eg the difference in cure rates between a new and a standard drug might be only 1%.
- If an exploratory analysis of data without a pre-existing hypothesis suggests a hypothesis, an independent set of data should be used to test that hypothesis. This is known as a confirmatory analysis.
- It is poor scientific practice to conduct multiple tests and only report the most significant – this is very likely to be a false discovery and will give biased estimates.

# Appendix 5: Causation and relative risk

Statistical methods cannot establish proof of a causal relationship in an association. The causal significance of an association is a matter of judgement that goes beyond any statement of statistical probability. To judge or evaluate the causal significance of the association between the attribute or agent and the disease, or effect upon health, a number of criteria must be utilised, none of which is an all-sufficient basis for judgment.

Relative risk (RR) estimates are used with increasing frequency in toxic tort/delict litigation as evidence for a causal link between the putative toxic exposure and the personal injury sustained by the claimant. The simplistic phrase 'doubling the risk' is unhelpful, because there is rarely a single risk with no variation by age and sex, and rarely only one estimate, let alone a very precise estimate of a particular RR.

An RR, or risk ratio, measures the size of the effect of a given risk factor on disease rates in specific populations. It describes the proportional increase in the probability of an event occurring in a group exposed to some condition, as measured from a baseline probability of an event occurring in a comparison group that has not been exposed to the condition. For example, men who smoke 15 – 24 cigarettes per day have an RR of lung cancer compared with never-smokers of about 26. The RR of lung cancer for regular drinkers of more than four glasses of wine per day compared with those who drink a glass of wine per week is about 3.2, averaged across smoking habits. After adjusting for smoking habits and other factors, the RR is 1.4[23].

The RR of lung cancer with heavy drinking is more than halved by adjusting for smoking; this illustrates the importance of having a careful definition of a causal hypothesis and not drawing conclusions from a single estimate. Estimates from a range of epidemiological studies are needed. A medical statistician or epidemiologist will usually consider how study results relate to the viewpoints identified by Sir Austin Bradford Hill for the assessment of causality[24].

The Bradford Hill criteria apply to general scientific conclusions for populations. But we may also be interested in individual cases, say in civil litigation where courts need to decide whether a particular exposure (say the asbestos encountered in a job) caused a negative outcome in a specific person (say John Smith's lung cancer). It can never be established with absolute certainty that the asbestos was the cause of the cancer, since it cannot be proved that the cancer would not have occurred without the exposure. But some courts have accepted that, on the 'balance of probabilities', a direct causal link has been established if the RR associated with the exposure is greater than two[25]. But why two?

Presumably the reasoning behind this conclusion is as follows[26]:

- Suppose that, in the normal run of things, out of 1,000 men like John Smith, 10 would get lung cancer. If asbestos more than doubles the risk, then if these 1,000 men had been exposed to asbestos, perhaps 25 would develop lung cancer.
- Of those exposed to asbestos who go on to develop lung cancer, fewer than half would have got lung cancer if they had not been exposed to the asbestos.
- So more than half of the lung cancers in this group will have been caused by the asbestos.
- Since John Smith is one of a group who was diagnosed with lung cancer, then on the balance of probabilities his lung cancer was caused by the asbestos.

These viewpoints are a common-sense approach to systematically evaluating a range of evidence. This section discusses the main concepts required to understand research findings. Discussion of the link between legal and scientific questions can be found in Dawid *et al.*[27] and Hutton[28].

### Rates, ratios and risks

- Rate or risk rate is the number of new cases identified out of the population under study. If 6 out of 100 babies are born with Down's syndrome, the risk is 6%. The time scale is not specified.
- An incidence rate is the number of new cases identified in a given time period divided by the total life time lived in that period by the population under study. If 6 new cases of lung cancer are identified in a year, and 1,000 people were monitored during that year, the incidence rate is 6 per 1,000 person-years.

- RRs are used to compare incidence rates in two groups (hence rate ratio), usually an exposed group compared with a non-exposed group. Given a lung cancer incidence rate of 6 per 1,000 person-years in chimney sweeps, and 2 per 1,000 person-years in the general population, the RR for chimney sweeps is 6/2 = 3.
- Excess risk is the difference between incidence rates in two groups. The excess risk of lung cancer for chimney sweeps is 6 - 2 = 4 per 1,000 person-years.

### Epidemiology, basic principles and study types

Epidemiological methods are used to investigate and estimate the frequency of medical conditions, the risk factors for those conditions, the effects of interventions and causes of ill-health and recovery.

### Case reports or series

Medical symptoms and conditions which appear to be unusual or novel are typically reported by describing a single case of a person with the conditions, or two or three cases. The role of these reports is both to alert colleagues to look out for similar cases and to encourage explanations and further investigation.

### Retrospective case-control studies

Case-control studies are efficient for investigating rare or newly emerging conditions or diseases, or those with a long gap between exposure and outcome, and allow multiple risk factors to be considered[29]. A study begins with finding a list of cases, such as people with throat cancer. Controls are selected to be people who are similar to the cases, for example in age, sex and socio-economic status, but without the disease. Exposure of both groups to factors such as radiation, as well as aspects of health such as weight and smoking, is assessed using medical or occupational records. Participants might be asked to recall exposure. Clear diagnostic criteria for the disease, and an equally thorough search for possible exposures among cases and controls, are important. By design, case-control studies investigate a single disease or condition. The main disadvantage of case-control studies is that they are open to several biases[30]. Bias can arise both through the quality of routine records and because the people with the disease are likely to take more effort to remember risk factors.

### Retrospective and prospective cohort studies

Cohort studies are observational studies which follow a group, or cohort, of subjects over time. The more explicit name 'longitudinal cohort study' stresses the time aspect. Cohort studies in health research aim to assess the possible factors in the development of disease or disability. The cohort must be explicitly defined, with inclusion criteria (eg employment in the civil service in 1967) and exclusion criteria (eg pre-existing cardiac disease). Basic characteristics of the cohort, such as age, sex, social status and current exposure to a range of factors are recorded.

A prospective cohort study actively enrols the defined cohort and collects baseline information. Subsequent health status is observed through follow-up. A retrospective cohort study, or historical cohort study, uses data from past records. In studies of health issues arising from occupational exposure, data from routine employment medical examinations and health and safety records can be used. A retrospective cohort study might provide information relatively quickly.

If the retrospective cohort continues to be followed after the first phase, the study includes both retrospective and prospective data. Cohort studies can provide information on a range of factors affecting a range of health states. It is easier to ensure consistency of measurement or recording of exposure factors and diagnosis in prospective cohorts than in retrospective cohorts or case-control studies. It is important that possible exposure or risk factors are clearly defined and that consistent effort is made to obtain data from the whole cohort. However, cohort studies are typically more expensive because they are larger. The results of a prospective cohort study can only be observed after some time, potentially a long time for slowly developing conditions.

### Registry data and databases

A disease or case register is a database which is intended to include all cases of a procedure, condition or disease in an identified population. Registers are used for clinical and service purposes as well as for epidemiological research. Successful registers need clear aims, appropriate data collection and validation systems, as well as regular analysis and feedback to interested parties. These requirements indicate that a multidisciplinary team, stable funding and relevant leadership are important. As disease registers contain medical information, both social attitudes to confidentiality and the jurisdiction will influence data collection and completeness[31]; Nordic registers have high coverage rates as consent is not required[32].

Clinical databases, such as a list of children referred to a specialist hospital, are not regarded as a disease register, because a relevant population cannot be identified. Parents might travel across regions or countries to seek help[33]. If cases are collected from a service which focuses on provision for those with cognitive deficits, accurate clinical diagnoses and reliable information on people with normal or good cognitive ability will not be available[34, 35].

### Randomised controlled trials

Patients who have agreed to enter the trial are allocated at random to receive one of the trial treatments. The role of randomisation is to minimise bias arising from both patients or doctors who have strong beliefs about what the response to treatment ought to be. Randomisation also balances factors such as age, underlying health and compliance with treatment across the groups in combination with allocation concealment (blinding). With effective randomisation, the probability of differences between the trial groups in responses and adverse effects arising by chance can be calculated. A potential disadvantage of randomised controlled trials is that only a selected population is included.

### Systematic review

The Cochrane Collaboration[36] defines a systematic review as a review of a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise all relevant research, and to collect and analyse data from the studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies, depending on the quality and quantity of information. Some systematic reviews can only provide descriptions of the main features of the included studies. A summary of the evidence based on the data for each patient in each study is generally regarded as the optimal approach, but this is acknowledged often to be impractical owing to constraints of confidentially and time. Systematic reviews are contrasted with narrative or expert reviews, which are based only on research that is known by, easily available to or acceptable to the reviewers.

### Bias

In epidemiology and statistics, bias typically refers to estimates which systematically misrepresent the quantity of interest. As a simple example, if a lecturer asks the 10 students, out of a class of 100 students, who have come to all his lectures whether his lectures are worth attending, the answers given cannot be assumed to represent the view of the whole class.

### Reporting standard

In order to assess the quality of the research described in a published article, sufficient information is required. When medical statisticians began to assess the quality of medical research publications, one difficulty was the lack of information provided regarding the design and analysis of studies[37, 38, 39]. Since 2010, a series of statements and guidelines, with accompanying checklists, have been published to facilitate understanding of the study and assessment of the validity of results and conclusions. These can be found on the website of the Equator (Enhancing the Quality and Transparency of Health Research) Network[40].

### Causality and the Bradford Hill criteria

Suppose there is a qualifying association: a large and statistically significant association between a possible causal agent and a particular pathology. If this observed association cannot be easily explained by confounding and bias, further consideration of cause and effect is worthwhile. Confounding describes the situation when a risk factor other than the exposure in question is associated with both the exposure and the outcome.
If the risk factor can be precisely measured, the effect of confounding can be estimated and the association between the exposure and outcome adjusted for the confounder. Higher alcohol consumption is associated with higher rates of lung cancer, but also with higher rates of smoking. In order to separate the effects of alcohol consumption and tobacco smoking on lung cancer, accurate records of alcohol and tobacco consumption are required.

Often a confounder cannot be precisely estimated and there may be various factors to consider as outlined in the Bradford Hill criteria illustrated in Regina v Abadom, 1983[41].

### Strength

'The strength of the association is expressed as a comparison between a standard or unexposed population and a population exposed to the putative causal agent. The RR of lung cancer for smokers is higher compared to never-smokers'[42]. 'Strength' is not well defined, although an RR of 5 would generally lead to further investigation. However, the baseline rate of the condition should also be taken into account. Headlines in 2000 which reported a doubling of risk of deep vein thrombosis associated with oral contraceptives, without reporting the absolute risk of 2 per 10,000 users per year, resulted in many women abruptly stopping taking their pills. Doubling the risk increases the absolute rate to 4 per 10,000 users.

### Consistency

'If an association is repeatedly observed by different people, in different places, circumstances and times, it is more reasonable to conclude that the association is not due to error, or imprecise definition, or a false positive statistical result'[43]. Further, the association should be observed in studies with a high methodological standard.

### Specificity

'Consideration should also be given to whether particular diseases only occur among workers within particular occupations. This is a supporting feature in some cases, but in other cases one agent might give rise to a range of reasons for death'[44]. The best example of a simple specific causal agent is thalidomide: the congenital deformity known as phocomelia is seen almost exclusively in the population of individuals exposed to thalidomide during gestation.

### Temporality

'This requires causal factors to be present before the disease'[45].

### Dose-response curve, or biological gradient

'If the frequency of a disease increases as consumption or exposure to a factor increases, this supports a causal association, for example, increasing levels of smoking associated with increased frequency of lung disease supports the hypothesis that smoking causes lung disease'[46].

### Plausibility

Biological plausibility that an effect is causal for an outcome to occur must be based on scientific reasoning or data, not just prior beliefs. Laboratory experiments might be possible, especially if the outcome effect can be modified by an appropriate experimental regime. However, extrapolation from animal experiments to humans is not straightforward. In the development of drugs, randomised experiments are required. For side effects with long-term treatments, or industrial exposures, other study designs are used.

### Coherence

'A cause and effect interpretation should not seriously conflict with generally known facts of the development and biology of the disease'[47].

### Experiment

'Sometimes evidence from laboratory or field experiments might be available'[48].

### Analogy

Bradford Hill commented: "In some circumstances it would be fair to judge by analogy. With the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy." This criterion has a limited role. If the criteria are not met, one cannot conclude that there is not a causal association. The conclusion is that there might be direct causal explanation, or an indirect explanation, or even that the association arose from some aspects of data collection or analysis. Competing explanations should be considered that might include unmeasured confounding factors or alternative factors which have an association of similar strength to the putative causal factor[49].

# References


1. Straight Statistics. 2009 Question marks over Corby judgement. See https://straightstatistics.fullfact.org/article/question-marks-over-corby-judgement (accessed 26 May 2020).
2. The Law Society Gazette. 2018 Police chief explains 'justice by algorithm' tool. See https://www.lawgazette.co.uk/news/police-chief-explains-justice-by-algorithm-tool-/5067033.article (accessed 26 May 2020).
3. Willis W. 2015 ENFSI guideline for evaluative reporting in forensic science. European Network of Forensic Science Institutes. See http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf (accessed 26 May 2020).
4. Royal Society and Royal Society of Edinburgh. 2017 Forensic DNA analysis: a primer for the courts. See https://royalsociety.org/-/media/about-us/programmes/science-and-law/royal-society-forensic-dna-analysis-primer-for-courts.pdf (accessed 24 August 2020).
5. BBC Reith Lectures. 2002 A question of trust. See http://www.bbc.co.uk/radio4/reith2002/lectures.shtml (accessed 26 May 2020).
6. Anderson T, Schum D, Twining W. 2005 Analysis of evidence (2nd edn, Law in Context). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511610585.
7. Regina v Abadom. 1983 1 Weekly Law Reports 126.
8. The Council of the Inns of Court and the Royal Statistical Society. 2017 Statistics and probability for advocates: understanding the use of statistical evidence in courts and tribunals. See https://www.statsref.com/ICCA-RSS-guide.pdf (accessed 26 May 2020).
9. *Op. cit.* note 4.
10. *Op. cit.* note 3.
11. Regina v Adams. 1996 England and Wales Court of Appeal (Criminal Division) 222.
12. Lord Hoffmann. 2009 Re B (Children) 1 AC 11.
13. Part 19, the Criminal Procedure Rules and Criminal Practice Direction 2015 (as amended 2018, 2019). See https://www.justice.gov.uk/courts/procedure-rules/criminal/docs/2015/crim-proc-rules-2015-part-19.pdf (accessed 26 May 2020).
14. Part 35, Civil Procedure Rules 1998 (as amended 2019). See https://www.justice.gov.uk/courts/procedure-rules/civil/rules/part35 (accessed 26 May 2020).
15. *Op. cit.* note 3.
16. Jackson G, Aitken C, Roberts P. 2015 Case assessment and interpretation of expert evidence – guidance for judges, lawyers, forensic scientists and expert witnesses. London: Royal Statistical Society.

17. *Op. cit.* note 4.

18. *Op. cit.* note 8.

19. Curran J, Hicks T, Buckleton J. 2000 Forensic interpretation of glass evidence. Boca Raton, FL: CRC Press LLC.

20. *Ibid.*

21. Wilson A, Aitken C, Sleeman R, *et al.* 2014 The evaluation of evidence relating to traces of cocaine on banknotes. Forensic Science International, 236, 67 – 76.

22. Aitken C, Wilson A, Sleeman E, *et al.* 2017 Distribution of cocaine on banknotes in general circulation in England and Wales. Forensic Science International, 270, 261 – 266.

23. Bagnardi V, Randi G, Lubin J, *et al.* 2009 Alcohol consumption and lung cancer risk in the Environment and Genetics in Lung Cancer Etiology (EAGLE) study. American Journal of Epidemiology, 171, 36 – 44.

24. Hill A. 1965 The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58(5), 295 – 300.

25. Spiegelhalter D. 2019 The art of statistics: learning from data. London: Pelican Books.

26. *Ibid.*

27. Dawid A, Faigman D, Fienberg S. 2014 Fitting science into legal contexts: assessing effects of causes or causes of effects? Sociological Methods and Research, 43, 359 – 421.

28. Hutton J. 2018 Expert evidence: civil law, epidemiology and data quality. Law, Probability and Risk. 17(2), 101 – 110.

29. Altman D. 1991 Practical statistics for medical research. London, Chapman and Hall, 93 – 96.

30. Sackett D. 1979 Bias in analytical research. Journal of Chronic Diseases, 32, 51 – 63.

31. Strauss D, Shavelle R. 1998 Life expectancy of adults with cerebral palsy. Developmental Medicine & Child Neurology, 40, 369 – 375.

32. Ludvigsson J, Häberg S, Knudsen G, *et al.* 2015 Ethical aspects of registry-based research in the Nordic countries. Clinical Epidemiology, 7, 491 – 508.

33. Hutton J. 2015 Weighing privacy against effective epidemiology. Developmental Medicine & Child Neurology, 57, 595 – 596.

34. Hutton J. 2006 Cerebral palsy life expectancy. Clinics in Perinatology, 33, 545 – 555.

35. Hutton J, Eccles M, Grimshaw J. 2008 Ethical issues in implementation research: a discussion of the problems in achieving informed consent. Implementation Science, 3, 52.

36. Cochrane UK. See http://uk.cochrane.org/ (accessed 26 May 2020).

37. *Op. cit.* note 35.

38. *Op. cit.* note 36.

39. Gore S, Altman D. 1982 Statistics in practice. London: British Medical Association.

40. Equator Network. See https://www.equator-network.org (accessed 26 May 2020).

41. *Op. cit.* note 7.

42. *Op. cit.* note 8.

43. *Op. cit.* note 8.

44. *Op. cit.* note 8.

45. *Op. cit.* note 8.

46. *Op. cit.* note 8.

47. *Op. cit.* note 8.

48. *Op. cit.* note 8.

49. *Op. cit.* note 8.

The members of the groups involved in producing this primer are listed below. The members acted in an individual and not organisational capacity and declared any conflicts of interest. They contributed on the basis of their own expertise and good judgement. The Royal Society and the Royal Society of Edinburgh gratefully acknowledge their contribution.

**Primer lead**

Professor Niamh Nic Daéid FRSE

**Editorial board**

Professor Colin Aitken

Professor Sheila Bird OBE FRSE

Sheriff Lorna Drummond QC

Lord Kitchin

Professor Niamh Nic Daéid FRSE

Professor Sir Bernard Silverman FRS

Professor Sir David Spiegelhalter FRS

**Writing group**

Associate Professor Alex Biedermann

Professor Christophe Champod

Professor Jane Hutton

Professor Graham Jackson

Lord Kitchin

Dr Tereza Neocleous

Professor Sir David Spiegelhalter FRS

Dr Sheila Willis

Dr Amy Wilson

**Primer steering group**

Dame Anne Rafferty DBE

Lord Hughes of Ombersley

Professor Dame Sue Black DBE FRSE

Sir Charles Godfray CBE FRS

Lord Justice Peter Jackson

Dame Ottoline Leyser DBE FRS

Dr Julie Maxton CBE

Dame Angela McLean DBE FRS

Professor Niamh Nic Daéid FRSE

Sir Muir Russell KCB FRSE

Professor Sarah Skerratt

Lord Turnbull

Mr Justice Wall

**Acknowledgements**

This project would also not have been possible without contributions and support from a range of individuals. In particular we wish to thank:

The Rt Hon the Lord Burnett of Maldon, Lord Chief Justice of England and Wales

The Rt Hon Lord Carloway, Lord President of the Court of Session and Lord Justice General

Sir Venkatraman Ramakrishnan, President of the Royal Society

Dame Ann Glover, President of the Royal Society of Edinburgh

We also wish to thank the Royal Statistical Society for contributing their time and expertise to this project.

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, technology, engineering, mathematics and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities are:

- Promoting excellence in science
- Supporting international collaboration
- Demonstrating the importance of science to everyone

**For further information**
The Royal Society
6 – 9 Carlton House Terrace
London SW1Y 5AG

**T** +44 20 7451 2571
**E** law@royalsociety.org
**W** royalsociety.org/science-and-law

Registered Charity No 207043

The Royal Society of Edinburgh (RSE), Scotland's National Academy, is a leading educational charity which operates on an independent and non-party-political basis to provide public benefit throughout Scotland. Established by Royal Charter in 1783 by key proponents of the Scottish Enlightenment, the RSE now has around 1600 Fellows from a wide range of disciplines. The work of the RSE includes awarding research funding, leading on major inquiries, informing public policy and delivering events across Scotland to inspire knowledge and learning.

**For further information**
The Royal Society of Edinburgh
22 – 26 George Street
Edinburgh EH2 2PQ

**T** +44 131 240 5000
**E** info@theRSE.org.uk
**W** rse.org.uk

Scottish Charity No SC000470


ISBN: 978-1-78252-486-1
Issued: November 2020 DES6439