Science and Justice 56 (2016) 402–410

Contents lists available at ScienceDirect

Science and Justice

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

Position paper

Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios

Duncan Taylor [a,b,*], Tacha Hicks [c], Christophe Champod [d]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice and Fondation pour la formation continue UNIL-EPFL, University of Lausanne, Lausanne-Dorigny, Switzerland
[d] Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, Lausanne-Dorigny, Switzerland

ARTICLE INFO

Keywords:
Sensitivity analysis
Bayesian networks
Likelihood ratio
Data
Source level propositions

ABSTRACT

Bayesian networks are being increasingly used to address complex questions of forensic interest. Like all probabilities, those that underlie the nodes within a network rely on structured data and knowledge. Obviously, the more structured data we have, the better. But, in real life, the numbers of experiments that can be carried out are limited. It is thus important to know if/when our knowledge is sufficient and when one needs to perform further experiments to be in a position to report the value of the observations made. To explore the impact of the amount of data that are available for assessing results, we have constructed Bayesian Networks and explored the sensitivity of the likelihood ratios to changes to the data that underlie each node. Bayesian networks are constructed and sensitivity analyses performed using freely available R libraries (gRain and BNlearn). We demonstrate how the analyses can be used to yield information about the robustness provided by the data used to inform the conditional probability table, and also how they can be used to direct further research for maximum effect. By maximum effect, we mean to contribute with the least investment to an increased robustness. In addition, the paper investigates the consequences of the sensitivity analysis to the discussion on how the evidence shall be reported for a given state of knowledge in terms of underpinning data.

1. Introduction

This paper aims to contribute to the ongoing debate regarding the relevancy of reporting the precision associated with a likelihood ratio (*LR*). The case of the determination of the nature of body fluids will be used to illustrate the argument and explore practical implications. For the sake of this introduction, we reduce the problem to a potential bloodstain where a stain is observed on the garment of a person of interest. One single presumptive test for human blood is carried out and gives a positive result. The test is known to have a false positive rate of 0.01 and a false negative rate of 0.1. If the propositions of prosecution and defence are respectively 'The stain is human blood' (H_p) and 'The stain is not human blood' (H_d), the *LR* associated with the positive result can be written as: Pr(a positive test result| H_p, I)/Pr(a positive test result| H_d, I) = 0.9/0.01. The information 'I' represents what is known, told and assumed, here the data associated with the test. In this case, our *LR* is assigned as 90. The typical questions that will be explored in this paper are:

- How sensitive is our *LR* of 90 to the data that underpin the rates of false positive and false negative?
- Should this sensitivity be reflected in the reporting of the *LR* by the introduction, for example, of a confidence (or credible) interval associated with our *LR*?

The argument we will try to convey is that, in the above case, there is no such thing as a "true value" for the likelihood ratio, and that the *LR* of 90 conveys in itself all that needs to be known about the weight to be assigned to the forensic results. However, we do not wish to imply that measuring the variability of likelihood ratios and their dependency on the data is useless for forensic scientists. It is useful to decide whether knowledge is sufficient for robust reporting. However, this is a different question (a question about data) and cannot be answered by giving the value of the results in the case at hand.

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.scijus.2016.06.010
1355-0306/

These questions are currently debated in the literature [1–3]. For example, Ali et al. [4], following their analysis of sampling variability in the training sets used in biometrics, suggest that "*a range of LRs should be reported which incorporates the sampling variability instead of reporting a single value of the LR.*" Sjerps et al. [2] advocate the need for a full transparency on the statistical analysis and not depriving the fact finder from any information that may help them to assess the trustworthy of the reported *LR*. They also refer to forensic publications introducing variability measures on the *LR* in areas such DNA, traces, drugs or speaker recognition. Taroni et al. [3] argue that the very nature of the *LR* encapsulates all required uncertainty and does not need any complementary measures. Our own thinking aligns with that of Taroni et al., who in [3], eloquently wrote that probability is a state of mind, and to present multiple values for a probability (such as a point estimate and a probability interval) is akin to having two different states of mind, and is hence logically flawed.

Body fluid attribution in forensic science is more complex than the above example because the types of fluids encountered in forensic science may be more varied than human blood. Moreover forensic observations can be multiple and dependent (e.g., visual observations, presumptive and confirmation tests). To deal with this complexity, we will use Bayesian networks (*BN*), a tool that can be used to graphically display the dependency relationships and interactions between different elements within a dataset. There have been numerous applications of *BN* within forensic science: quality control monitoring [5], preparation for legal challenges [6], complex pedigree evaluation [7], DNA profile mixture evaluation [8] helping to address activity level propositions [9], as well as numerous applications outside legal or forensic applications (refer to [10] for a review of Bayesian Networks). We direct the reader to [11] for explanations of the structure and terminology of *BN*s. Recently, the authors published a paper that used *BN* to combine DNA profiling results with the results of body fluid tests in order to help address propositions at the source level [12]. When assessing such results, the data used to inform probabilities can be limited to a few experiments. Because the conclusion of the forensic scientist depends on these data, it is important to know whether their knowledge is sufficient to ensure robust reporting, and when it is necessary to perform further research. This aim of the present contribution is to show how *BN*s can help us in this task. The basic structure of the *BN* network that achieved this can be seen in Fig. 1.

The definition of each node is:

Profile matches – This node has states 'yes' and 'no' and is the node that is instantiated when a DNA profile obtained from a recovered trace possesses the same alleles as the reference of the POI.
POI DNA present – This node has states 'yes' and 'no' and is the node that specifies whether the DNA of the POI is the source of the stain.
Hp/Hd – This node has states 'Hp' and 'Hd' and marries the DNA results with the results of tests for body fluid identification (in this particular network, the body fluid is blood, however it could be configured for questions regarding any body fluid by changing prior probabilities in the 'Nature of Stain' node; see appendix table A5).
Nature of stain – This node has states 'blood', 'semen', 'saliva', 'trace' and 'none'.
Quant – Each numerical category represents the concentration of DNA detected per mm^2 of sampled area. Categories are '0', '0 to 50', '50 to 500', '500 to 5000' and '5000+'.
Visual – This node has states 'red/brown', 'white/yellow' and 'none/other' to indicate the presence (or absence) of a visual stain.
HemaStix result – This node has states 'positive' and 'negative' that correspond to the test result.
HemaTrace result – This node has states 'positive' and 'negative'.

Within the work by Taylor et al. [12] a theoretical series of court questions were used to drive the work forward from a simple *BN* to one which could consider a complex mixed DNA profile scenario. The work in [12] provides a useful starting point for the evaluation of evidence when biological source is in question; however there is a common line of questioning in court to which the scenario could be extended.

Q: What are the sample sizes on which you are basing your calculations and is that big enough?

This question is founded in traditional frequentist thinking, where the scientist may want to consider all possible datasets that could have been obtained (given different experiments, or alternative data) but were not. This is then commonly referred to as 'sampling variation' and can be taken into account by producing a distribution of the *LR*s and reporting a confidence or probability interval. On the other hand, Bayesian inference makes probability statements posterior to the data, i.e. it is inferentially complete. Indeed, Bayesian inference provides the conditional probability distribution of the next observation given prior belief and all the data observed thus far. There is therefore no need to worry about the data that could exist but have not been yet obtained, as this is encapsulated in our probabilities.

Still, underlying the question is an important concept, whilst the *LR* being provided utilises the prior beliefs of the scientist and the available data, is this accumulated knowledge enough to provide a robust opinion? By robust in the forensic context, we mean a piece of information that has limited opportunity to mislead the court. In such cases, scientists, in order to decide whether or not the data available are sufficient to warrant a robust opinion, can explore the impact of the size of the dataset on their evaluation. Because *LR*s depend on the data used to inform probabilities, it goes without saying that using different data, will lead to different *LR*s. However, ideally, if the data sufficiently reflect the phenomenon we want to account for, sensitivity analyses should not lead to *LR*s that are

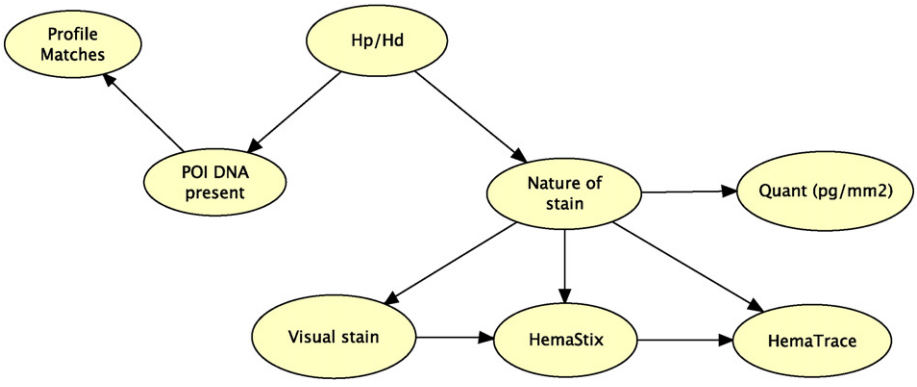

Fig. 1. Bayesian network for the presence of human blood in a sample and incorporating DNA profiling results [8].

of vastly different magnitude. This means that our evaluation will be considered robust, if the system, informed by a different set of experiments, leads to similar *LR* values. Thus, exploring the sensitivity of the *LR* to the underlying data allows us to explore whether our knowledge is sufficient to ensure robust conclusions. Sensitivity analysis is achieved by simulating cases under a range of datasets and exploring the impact of this on the output.

In this paper we will show, with the use of sensitivity analyses, how the scientist can decide whether or not data are sufficient or if more experiments would be needed to help with the issue at hand, and contribute to the debate regarding any precision associated with the *LR*.

Sensitivity analyses can also be used to investigate what the main factors that have a bearing on evaluation are. This is an extremely useful tool for laboratories who wish to use their limited resources to the greatest effect, and for researchers in planning their experimental design.

2. Sensitivity analyses

The probability tables for each node in the Bayesian network are given in [12], and we reproduce the counts in Appendix 1. The network presented in [12] was developed using the *BN* commercial software Hugin (www.hugin.com). The sensitivity of the *LR* to the data underlying the nodes (or parts of nodes) can be explored by resampling the counts from a Dirichlet distribution using software R [13] and the RHugin libraries v8.0 (http://rhugin.r-forge.r-project.org/). Alternatively, for those who do not have Hugin, the *BN* can be constructed and resampled using the freely available R libraries, gRain [14] and BNlearn [15]. We provide the R code for the latter as supplementary material.

We apply a Dirichlet$(1, \ldots,1)$ prior to calculate the posterior mean probability of each count for use in the *BN*. If state i of category k has $n_{i,k}$ observations then the posterior probabilities are calculated by:

$$p_{i,k} = \frac{n_{i,k}+1}{I+\sum_i n_{i,k}}$$ where I is the number of different states that exist in that category

To carry out sensitivity analyses we use an iterative process. For y iterations we redraw each observational count $n_{i,k}$ by:

$$n_{i,k,y} = gamma(n_{i,k}+1, 1) \quad \text{for each} \quad n_{i,k}$$

and then normalise to obtain posterior probabilities, $p_{i,k,y}$:

$$p_{i,k,y} = \frac{n_{i,k,y}}{\sum_i n_{i,k,y}}$$

This is carried out for each of the nodes based on observational data. We use the following scenario to demonstrate the process:

> *An assault has taken place and there has potentially been a transfer of blood between the offender and the victim (both of whom have reportedly been injured during the assault). The suspect's shirt is examined and a red/brown stain identified that gives a positive result to a presumptive test for blood (in this case the Hemastix test). A 2 mm by 2 mm cutting is taken and submitted for DNA profiling. Quantification determined that from the sample taken of the stain 5 ng of DNA was present and a portion of this is taken and subjected to PCR to generate a STR profile. Upon doing so a single source profile is obtained that corresponds to the victim in the case. The suspect does not deny being in contact with the victim, but denies that the stain is the victim's blood. The results of the presumptive test and of DNA analysis are assessed given the following propositions: The DNA comes from the victim's blood versus the DNA comes from the victim but is not blood (i.e., it is either semen, saliva, or trace – trace refers to DNA that is not associated with any of the above defined body fluids).*[1]

We therefore have the following information:

- Visual appearance of stain – Red/Brown
- Quantification – 1.2 ng/mm^2 (in 500–5000 pg/mm^2)
- HemaStix result – positive
- DNA profile result – Presence of victim's DNA has been conceded

The issue revolves at to whether or not the stain is blood (as opposed to another biological fluid than can lead to a DNA profile) and not the source of the obtained DNA profile.

In this scenario, the sensitivity of the *LR* is investigated in the light of the actual data (underlying counts) in the following nodes; Quant, Visual and HemaStix. Note that in the Quant category 'none' we have not allowed for any sensitivity as this category is defined as an absence of human biological material and hence should always provide a quant of 0. 10,000 iterations of the observational counts were made following to the procedure outlined above. It provides the distribution of *LRs* shown in Fig. 2 (on a $\log_{10}$ scale), for which the 50%, 5% and 1% quantiles *LR*s are 160, 40 and 30 respectively. This distribution shows the impact of the data on our evaluation. We wish to stress here that there is no "true" value for a *LR* so this distribution does not show the uncertainty of the *LR*, but how robust (or sensitive) it is depending on the data used. Each data point on the distribution only represents a *LR* computed under the conditioning of different data.

3. Using sensitivity analyses to direct research

The benefits of sensitivity analyses are two-fold. Firstly, it demonstrates if the basis of our opinion is robust and what the impact would have been if another practitioner had chosen different casefiles from which to collect data. Here, the results always favour the first proposition over the alternative. The degree of support, depending on the data used, varies between 50 and 1000, with most data giving a *LR* in the order of 100. The sensitivity analysis allows the appreciation of the impact of the data on our *LR*s.

The second benefit is that the *BN* can be probed in order to determine which nodes have the most impact on evaluation. This can direct us to which experiments would provide the most benefit if we were to collect more data to inform the conditional probability tables of the *BN*.

3.1. Identifying the source of variability

Consider the different *LRs* shown in Fig. 2, and imagine that the laboratory, which has compiled these data, wishes to increase the robustness associated with their assigned *LR*s, by acquiring more knowledge. Limited resourcing means that choices must be made as to where the laboratories research efforts are best spent. It may also be the case that addition of further observation into some areas will yield virtually no decrease in variability from case to case. In this instance it would be very useful to know from where the case to case variation is coming. To do this, the counts in each node can be sampled,

[1] Note that the *LR* obtained will depend critically on the relative prior probabilities specified for each body fluid under the defence proposition. In this instance, within the framework of circumstances, it would be odd for the defence to suggest semen is a viable alternative to blood with the same prior probability as saliva or trace. In our example we do use equal prior probabilities (see appendix table A5), however we could easily use prior probabilities such as 0.1, 0.4, 0.5 (for semen, saliva or trace respectively). Such a shift in prior beliefs changes nothing about the basic concepts we discuss in this paper it would just change the numerical value of our *LR*s. Finally, because of the introduction of prior probabilities to weigh initially the three options under the alternative, some authors may prefer to call this likelihood ratio (*LR*) a Bayes factor (*BF*). We took the option to keep *LR* all throughout.

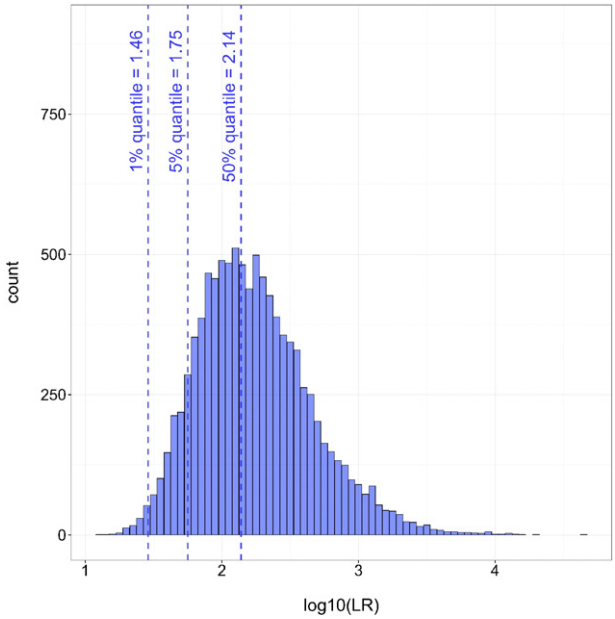

Fig. 2. Different likelihood ratios obtained from the BN shown in Fig. 1, depending on the data on which our evaluation is based. This shows how sensitive our evaluation is to the data used. Put another way, this shows the impact of the data on our evaluation.

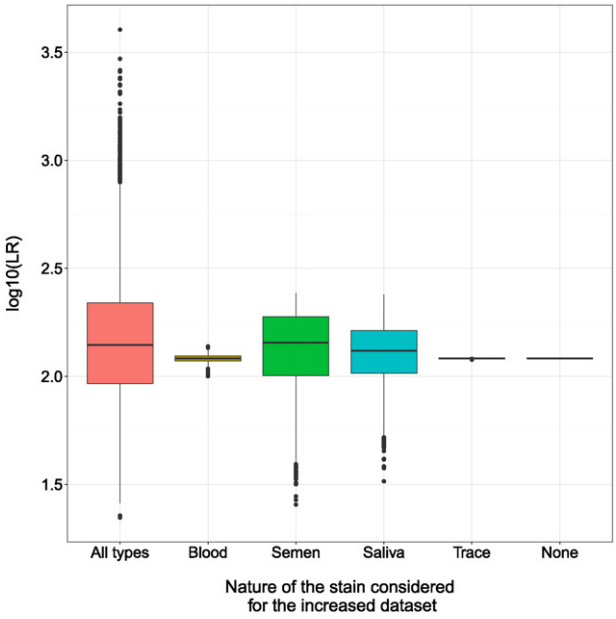

Fig. 4. Different LRs obtained when resampling the visual node (All types), or just single stain category within the visual node. This shows how sensitive our LR is to these data.

one node at a time leaving all others unchanged. For the network shown in Fig. 1 we carry out such an analysis to demonstrate how sensitive our likelihood ratios are to the data used to inform our probabilities (see Fig. 3).

Fig. 3 shows our different *LRs* from Fig. 2 by resampling all nodes, seen on the far left, and then the *LRs* obtained when just resampling one node at a time; Visual, Quant and HemaStix. From these likelihood ratios it can be seen that the Visual and HemaStix nodes contribute most to our different *LRs*. These results can be further decomposed into the categories within the nodes to determine which has more impact on the observed variability.

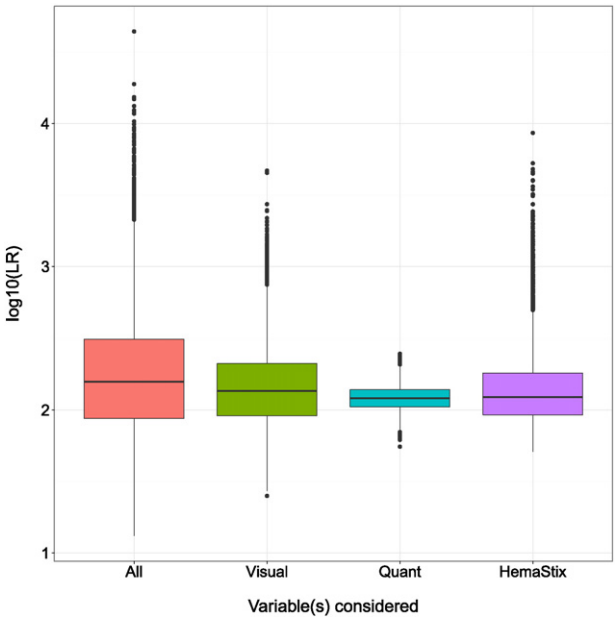

Fig. 3. Different likelihood ratios obtained when resampling all nodes, and respectively just single nodes in the BN. Our LRs are most sensitive to the nodes 'Visual' and 'HemaStix'.

Fig. 4 shows the sensitivity of our *LRs* depending on the visual node (from Fig. 3) and the different *LRs* obtained when sampling just the counts for the visual node for each stain type, one at a time.

3.2. Addressing the source of variability

Having identified the areas of the *BN* that contribute to the evaluation, the forensic scientist may now wish to know how much further experimental work is required to reduce the variability. This can be achieved by artificially increasing the size of the counts of interest (without affecting the relative proportions) and increasing them to see when the variability reduces to the point at which is desired.

We recreate the 'Visual' category by increasing the counts of all subcategories (in proportion to what they currently are). We then resample the counts in the entire *BN* as seen in the 'All' category of Fig. 3. The results of this experiment are shown in Fig. 5.

Note that as the sample size increases in Fig. 5 there is little change in variability but a definite trend to increase the magnitude of our *LR*. The means of our LRs in the six sensitivity experiments shown in Fig. 5 are 180, 270, 320, 360, 390 and 480. The reason for this is the application of the sample size multiplier to the raw count, which is then resampled from a Dirichlet distribution with a Dir(1, ...,1) prior. If the counts were resampled from $gamma(n_{i,k},\ 1)$ rather than $gamma(n_{i,k}+1,\ 1)$ distributions then the mean would be approximately the same for all sample sizes. The application of a Dirichlet(1, …,1) is also the reason why the variation does not decrease with increasing sample size. This is because a proportional sample size increase when the raw count of zero, will still produce a count of zero and will update the Dir(1,1, …,1) prior to a posterior count of 1 (regardless of the size of the inflation). This resampling of the count of 1, then limits the range of the simulated *LRs*. To demonstrate this concept Fig. 6 shows the same analysis as conducted in Fig. 5, but resampling counts from $gamma(n_{i,k},\ 1)$. Fig. 6 exhibits a roughly constant mean and a reduced variance with increased data size, as expected.

With the results from Fig. 5, the scientist now knows that by just a four-fold increase in the number of observations in the visual appearance of semen and saliva stains they would update their *LR* by a factor two (assuming that additional stains show the same trend as the

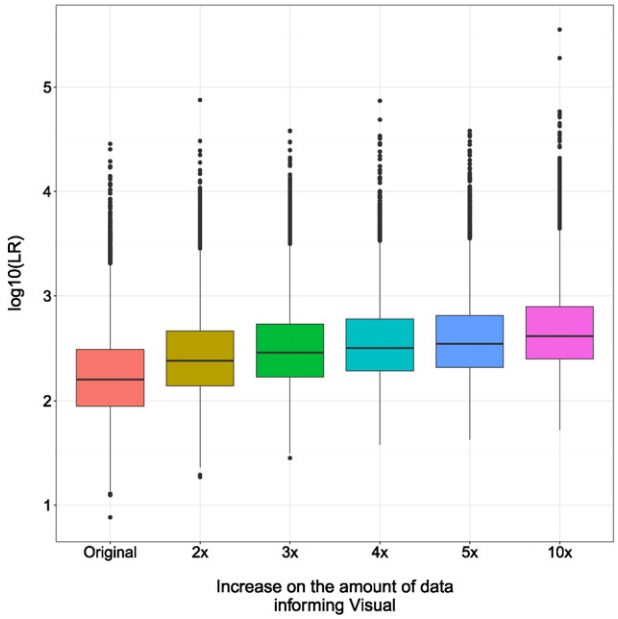

Fig. 5. Different LRs obtained when increasing sample size in the all categories in the visual node by a factor of 2, 3, 4, 5 or 10.

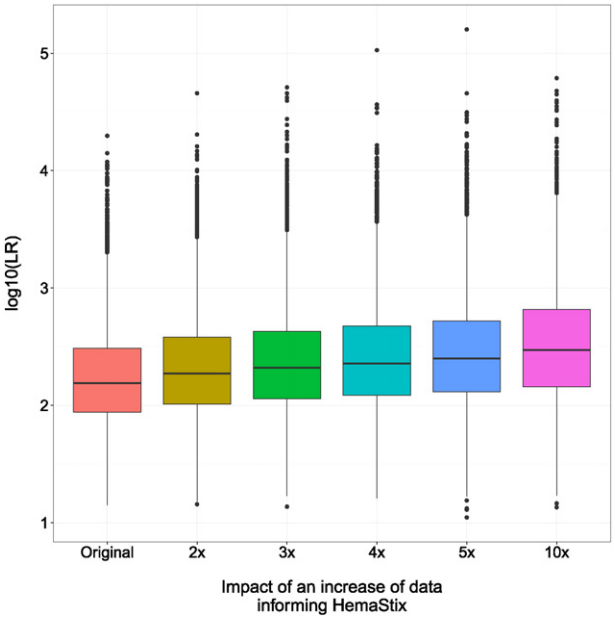

Fig. 7. Different LRs obtained when increasing sample size in the all categories in the HemaStix node by a factor of 2, 3, 4, 5 or 10.

currently examined data). This relatively small amount of additional work is likely to yield the greatest benefit to the scenario being considered in this paper. Still further investigation could be carried out (not shown here) on the effect of an increase in data for a specific component of the node. For example a similar investigation as shown in Fig. 5 could be carried out by exploring an increase in data within the visual node for just the semen and saliva categories (which, as can be seen in Fig. 4, contribute most significantly to the final *LR*s).

We also carry out the same process of increasing the counts (retaining the observed proportions) for the HemaStix node. The results of this increase can be seen in Fig. 7.

We can observe that when everything is kept proportional, there is also some gain to be had by increasing the data counts associated with HemaStix. The extent however is limited, and this is due to the data underpinning node 'visual' which induces much of the variability. Only a joint acquisition will bring more substantial reduction in the variability as shown in Fig. 8.

Fig. 8 shows that the most effective use of resources would be to increase by a small amount (e.g. by a factor of five) the observations in the HemaStix node then switch to a similar increase in the observations in the Visual node, rather than continuing to build on the HemaStix observations. This can be seen most clearly by the fact that the increase in both the visual and HemaStix counts by a factor of five, outperforms an increase in only either the visual (Fig. 5) or HemaStix (Fig. 7) by a larger factor, say 10.

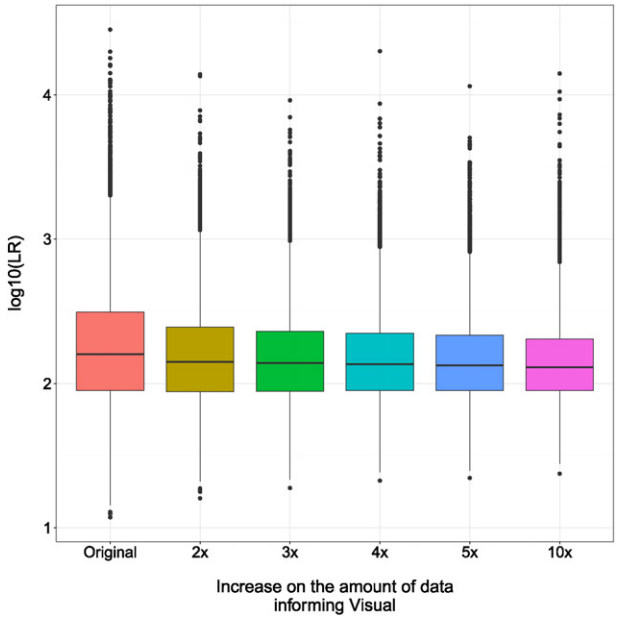

Fig. 6. LRs obtained when increasing sample size in the all categories in the visual node by a factor of 2, 3, 4, 5 or 10, but sampling unadjusted raw counts.

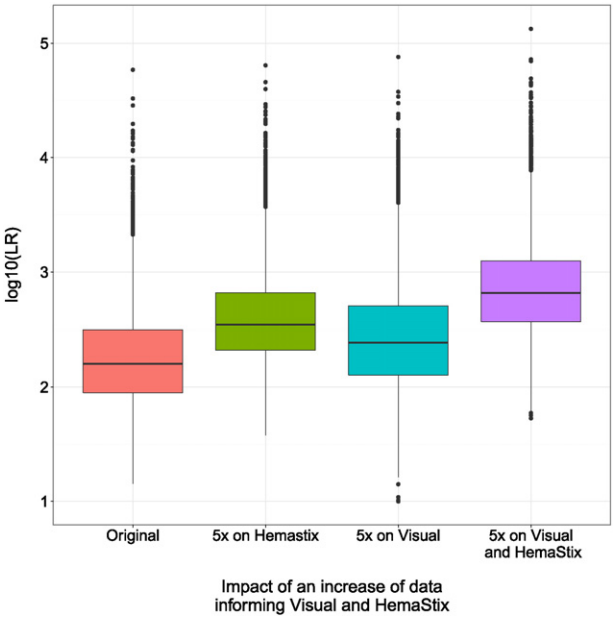

Fig. 8. Sensitivity of the LR to an increase in sample size for all categories in the HemaStix, the Visual nodes and then both together by a factor of 5.

4. Reporting considerations

From the above analysis we ought to ask '*what is the impact of these various simulated cases on the reporting practice?*'. In the case example we have used so far the following questions would be legitimate:

- Should the forensic scientist report a likelihood ratio plus a description of the distribution of *LR*s obtained (refer to Fig. 2)?
- If the forensic scientists report a single value for the *LR*, should they report some bound[2] of that distribution to 'err on the safe side', or should they report the mean or the median?

We are of the opinion that sensitivity analysis can be useful for scientists to decide whether or not they should perform more experiments or if they have sufficient knowledge to report. But, we believe that it is their duty to decide the value of their *LR* as we will explain below.

Our proposal is to distinguish the technical or validation work carried by the forensic scientist to understand and appreciate the uncertainties associated with the findings and the reporting principles he/she will apply to convey the results in a statement or in court. We have no doubts regarding the benefits of sensitivity analyses for the first task. It forces us to have a critical look at the available dataset, understand their limitations and the impact of these limitations on our assessment. When it comes to reporting, we need to strive to inform in a fair but useful manner. Fig. 2 (or at least some summary of it) should no doubt find its place in the case file and be amenable to review by any forensic scientist. We have doubt that the results of sensitivity analyses will help in a statement that can be used in court. The key question may be phrased as "in the forensic process, who assigns the *LR*?". We believe that this task should fall within the remit of the scientist. In this case, considering the variation observed, we could indicate a strength of the findings by using a description that covers an order of magnitude, for example "*the LR is of the order of hundreds*". Quoting a quantile confuses ideas of Bayesian and frequentist thinking. Claiming that the *LR* is the median (157.9) may convey a level of precision that the data would not allow, and we can imagine that in instances where large ranges for the *LR*s are obtained, quoting only the median may appear misrepresentative of the evidence. The order of magnitude we decide to quote is an informed decision based on the analysis expressed with a view to help the criminal justice process, without any intention to withdraw any information but with the idea of giving an assessment of the strength of the observations that can be placed in the hands of the fact finder.

The above situation still leaves unclear what value should be reported if a large *LR* range is observed or if an exact computed value is directly requested. This can be particularly problematic if lower parts of this range crosses unity (a $\log_{10}(LR)$ of zero). To illustrate the issue, we will use an example based on the *BN* used so far, but informing the conditional probability tables differently. Let us assume that for the nodes HemaStix, HemaTrace, Quant and Visual, we have no preference between any of the states, but in one case with a very limited number of observations (1 for each state corresponding to the prior counts) and in the second case with 99 observations acquired on each state. The difference between the two cases is that we build on a very reduced number of data points for the first and a large number of observations for the second. However, the relative proportions between each state remain the same for both cases. If we simulate the same scenario as in Fig. 2, we obtain the distributions shown in Fig. 9. The 1% and 99% quantiles are shown for both distributions.

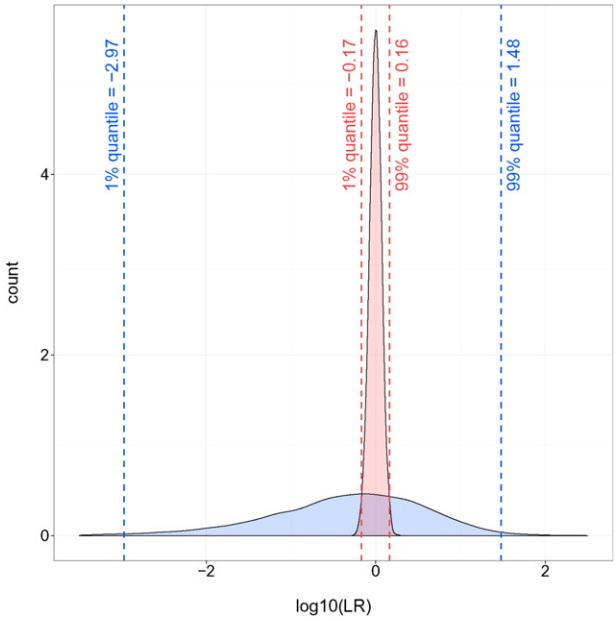

Fig. 9. Different LRs computed when all states in CPTs are equally likely but each informed initially by 1 count (density in blue) and by 99 observations (density in pink).

Given the fact that we have not assigned any preference to any states in the conditional probability tables (CPTs), it is not surprising to see that the mode of each distribution corresponds to a *LR* of 1 (the results neither support prosecution's or defence's proposition and are thus neutral), but the range of values obtained differs substantially between the two cases.

If we were to report the first case, a *LR* of 1 would impose itself, as it best reflects our state of knowledge, even if that knowledge is sparse. Taking a lower bound such as the 1% quantile ($\log_{10}(LR) = -2.97$) to "err" on the defence side would obviously be inadequate. A *LR* of 1 rightfully and fairly describes our belief regarding the various states in the CPTs. In the second case, the *LR* remains the same, but our trust in it is increased thanks to the amount of knowledge that is used in the *BN*. Reporting 1% quantile ($\log_{10}(LR) = -0.17$) is not helpful either as the best decision to take is to report a *LR* of 1. The dilemma that the scientist faces is to avoid using ranges of *LR* or quantiles that do not reflect appropriately their belief while conveying that the same reported *LR* may be more robust in the second case than in the first.

The essential difference between the two cases is that in the first the *LR* of 1 reflects upon a "state of ignorance" where we chose to assign equal opportunity to each state whereas the second case leads to an uninformative system but based on substantial knowledge that the states are equally likely. Pragmatically speaking, the first case tells us that it would be wise to invest on data acquisition regarding the test HemaStix, the second is a call to stop using the test for the purpose at hand. Put it differently, the first case is dominated by the chosen flat priors, whereas the second is driven by the data used to inform the CPTs. When it comes to the weight to be assigned to the observations, both cases are equivalent ($LR = 1$), but we ought to search for an additional piece of communication to highlight the difference between the two cases. Discussing the quality and quantity of data on which our evaluation is based is, we believe, a useful point to report: as the robustness of our *LR* will generally depend on the data that are available.

The authors of the present paper are still exploring and debating various options on how to convey the sensitivity that the scientist is

[2] Typical practice would be to report a bound that minimised the reported discrimination power, i.e. a bound that is closest to a likelihood ratio of one. This of course becomes complicated by factors such as which side of one is considered 'conservative' i.e. for example if comparison of a suspect to a DNA profile from a weapon yields a likelihood ratio that favours the suspect's inclusion then a bound that drew the reported value closer to one would be considered conservative, however if the likelihood ratio favoured the suspects exclusion then perhaps a bound that drew the reported value closer to zero, and away from one, would be the conservative choice. Also complicating the choice of the reported bound is when the distribution of likelihood ratios crosses the value of one.

prepared to associate with his/her *LR*. To do so, we would like to consider a case, where most *LR*s support the prosecution proposition and where the range observed following the simulations is large but with some *LR*s below 1 (say for that 99% of *LR*s fall between 0.5 and 580), as shown in Fig. 10.

We shall view the result in Fig. 10 in light of the question posed at the beginning of this paper:

> *Q: What are the sample sizes on which you are basing your calculations and is that big enough?*

Our view is that, given the sensitivity of the *LR* to the data, a decision would be made by the scientists as to whether they feel there is enough data to report the value of the results. If, the sensitivity of the *LR*s is too great for some nodes where data paucity exists, the scientist would convey the message that the weight to be assigned to the findings cannot be robustly assessed and that more data are required to help in this case. No likelihood ratio would be quoted, thus the case would not be evaluated. If scientists are satisfied that the findings can be robustly assessed then they could answer the query with:

> *A: The LR I have calculated is 3. This is the value which most appropriately expresses my view of the value of the findings, given each of the two propositions. The probability assignments I have used in my calculation are based on observational data. I have considered the effect of the sensitivity of the LR to the data used for probability assignments. This procedure allows me to assess the robustness of the LR value I have given. After undertaking such a sensitivity analysis, I am satisfied that the knowledge on which my evaluation is based is sufficient. In my opinion, the number that best reflects the value of the evidence is 3.*

To implement such a practice, guidelines will need to be developed within the laboratory, based on the properties of the sensitivity analysis, in order to define what type of results can be reported. If the court requires further information then the results of sensitivity analyses could be provided. However, every effort would need to be made to ensure the results of these analyses are not interpreted as a, classic frequentist, account for sampling variation from which some lower bound could be reported.

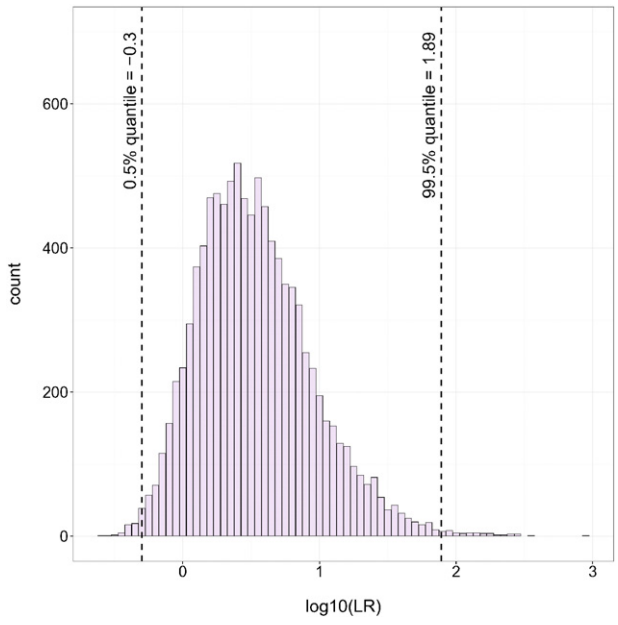

Fig. 10. Different LRs for a hypothetical case where 99% of LRs fall between 0.5 ($\log_{10} = -0.3$) and 580 ($\log_{10} = 1.89$), with a LR median of 3.

5. Conclusion

Bayesian networks are invaluable tools that can be used to bring together elements of a dataset that have complex interactions and dependencies. Once constructed, they can be used to answer questions of interest and help to judge the robustness of our evaluation and, in particular, whether the knowledge is judged sufficient or if more experiments should be performed. Sensitivity analyses are a method for evaluating how sensitive the *LR* is to changes in values that underlie conditional probability tables. Typically this is done by varying the values in the probability tables over some sensible range and graphing the associated change in *LRs*. In this work we demonstrate the sensitivity of the *LRs* when considering multiple nodes simultaneously, which is a slightly different treatment of the data, and allows comparisons of sensitivity between nodes to occur more easily. In the method we describe here if there are no observations within a category of a node table (for example the results of HemaStix testing on semen stains that appear Red/Brown in Table A3) then this naturally results in the sensitivity of the *LR* to this node. The key point is that it is not our belief (i.e., our probability) that is more robust, but the basis of this belief (i.e., the amount of knowledge, the data).

An additional benefit to performing the sensitivity analyses we describe on *BN* is that the results can be used to direct further work so that maximum effect is obtained from the resources dedicated to the task. We demonstrate here the path that an analyst can take to better understand their network and decide on a path of work.

We hope that our analysis also contributes, through this concrete example, to the discussion of whether to measure and report the variability of our likelihood ratios. It led us to argue that sensitivity of the system speaks well for the validation of the technique, but not for the weight to be assigned to the results. We understand that scientists wish to be transparent and indicate that, although their knowledge is imperfect, it is sufficient to warranty robust reporting. In that sense, as we have shown above, it is useful to explore how sensitive our likelihood ratios are to data, and what is their impact on the assessed value. This enables also to answer questions that we may face in Court on whether or not the data we have used are sufficient. But, whether we think that the data are sufficient and what is the value of our results are two different questions that deserve a different answer. Exploring the variability of our LRs given the data is useful to decide whether we ought to report in a case (i.e., to decide if the basis of our evaluation is robust, that is if the data are sufficient). But, as Taroni et al. [3] we are of the opinion that the Court should be given the scientist's *LR*. This single value (generally an order of magnitude will suffice) takes into account the data and the knowledge that we have: it encapsulates all our uncertainty.

Conflict of interest

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

Points of view in this document are those of the authors and do not necessarily represent the official position or policies of their organisations.

Appendix 1. Probability tables for nodes in BN shown in Fig. 1

Table A1
Count for cells in table for node 'Quant'.

	Stain nature	Blood	Semen	Saliva	Trace	None
pg/mm^2	0	0	0	0	0	1
	0 to 50	15	3	28	315	0
	50 to 500	28	4	11	43	0
	500 to 5000	39	9	19	2	0
	5000 +	4	14	0	2	0
	Total	86	30	58	362	1

Table A2
Count for cells in table for node 'Visual'.

	Stain nature	Blood	Semen	Saliva	Trace	None
Visual stain	Red/brown	120	0	1	0	25
	White/yellow	0	16	0	2	6
	None/other	30	15	47	371	30
	Total	150	31	48	373	61

Table A3
Count for cells in table for node 'Hemastix'.

	Stain nature	Blood			Semen		
	Visual	Red/brown	White/yellow	None/other	Red/brown	White/yellow	None/other
Hemastix	Positive	96	0	19	0	0	0
	Negative	2	0	5	0	3	8
	Total	98	0	24	0	3	8

	Stain nature	Saliva			Trace			None		
	Visual	Red/brown	White/yellow	None/other	Red/brown	White/yellow	None/other	Red/brown	White/yellow	None/other
Hemastix	Positive	0	0	0	0	0	0	10	1	0
	Negative	1	0	11	0	0	11	14	0	13
	Total	1	0	11	0	0	11	24	1	13

Table A4
Count for cells in table for node 'HemaTrace'.

	Stain nature	Blood		Semen		Saliva		Trace		None	
	Hemastix	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Hematrace	Positive	46	0	11	5	1	0	0	0	0	0
	Negative	1	2	1	1	1	9	0	10	5	12
	Total	47	2	12	6	2	9	0	10	5	12

Table A5
Probability table for node 'Biological Material Present' showing prior probabilities. Note that under Hd, only biological fluids other than blood are considered because the defence had agreed to the source of the DNA, hence to the presence of a biological fluid).

	Hp/Hd	Hp	Hd
Stain nature	Blood	1	0
	Semen	0	1/3
	Saliva	0	1/3
	Trace	0	1/3
	None	0	0

Table A6
Probability table for node 'POI DNA present' showing prior probabilities.

	Hp/Hd	Hp	Hd
POI DNA present	Yes	1	1
	No	0	0

Table A7
Probability table for node 'Profile matches' showing prior probabilities. LR_{SS}^{-1} is the inverse of the Sub-Source likelihood ratio.

	POI DNA present	YES	NO
Profile matches	Yes	1	LR_{SS}^{-1}
	No	0	$1-LR_{SS}^{-1}$

Table A8
Probability table for node 'Hp/Hd' showing prior probabilities.

Hp/Hd	Hp	0.5
	Hd	0.5

References

[1] A. Nordgaard, Comment on 'Dismissal of the illusion of uncertainty on the assessment of a likelihood ratio' by Taroni F., Bozza S., Biederman A. and Aitken C. Law Probab. Risk 15 (2016) 17–22.
[2] M.J. Sjerps, I. Alberink, A. Bolck, R.D. Stoel, P. Vergeer, J.H. van Zanten, Uncertainty and LR: to integrate or not to integrate, that's the question, Law Probab. Risk 15 (2016) 23–29.
[3] F. Taroni, S. Bozza, A. Biedermann, C. Aitken, Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio, Law Probab. Risk 15 (2016) 1–16.
[4] T. Ali, L. Spreeuwers, R. Veldhuis, D. Meuwly, Sampling variability in forensic likelihood-ratio computation: a simulation study, Sci. Justice 55 (2015) 499–508.
[5] Q.A. Le, G. Strylewicz, J.N. Doctor, Detecting blood laboratory errors using a Bayesian network: an evaluation on liver enzyme tests, Med. Decis. Mak. 31 (2011) 325–337.
[6] W. Edwards, Influence diagrams, Bayesian imperialism, and the Collins case: an appeal to reason, Cardozo Law Rev. 13 (1991) 1025–1074.
[7] A.P. Dawid, J. Mortera, V. Pascali, D. van Boxel, Probabilistic expert systems for forensic inference from genetic markers, Scand. J. Stat. 29 (2002) 577–595.
[8] J. Mortera, A.P. Dawid, S.L. Lauritzen, Probabilistic expert system for DNA mixture profiling, Theor. Popul. Biol. 63 (2003) 191–205.
[9] I.W. Evett, P.D. Gill, G. Jackson, J. Whitaker, C. Champod, Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks, J. Forensic Sci. 47 (2002) 520–530.
[10] A. Biedermann, F. Taroni, Bayesian networks for evaluating forensic DNA profiling evidence: a review and guide to literature, Forensic Sci. Int. Genet. 6 (2012) 147–157.
[11] F. Taroni, C. Aitken, P. Garbolino, A. Biedermann, Bayesian Networks and Probabilistic Inference in Forensic Science, John Wiley & Sons, Ltd., Chichester, 2006.
[12] D. Taylor, D. Abarno, C. Champod, T. Hicks, Evaluating forensic biology results given source level propositions, Forensic Sci. Int. Genet. 21 (2016) 54–67.
[13] M. Plummer, Bayesian Graphical Models Using MCMC, RJAGS, 2012.
[14] S. Højsgaard, Graphical independence networks with the gRain package for R, J. Stat. Softw. 46 (2012) 1–26.
[15] M. Scutari, Learning Bayesian networks with the bnlearn R package, J. Stat. Softw. 35 (2010) 1–22.