Contents lists available at SciVerse ScienceDirect

Science and Justice

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

Technical note

Application of likelihood ratios for firearm and toolmark analysis

Stephen Bunch, Gerhard Wevers *

ESR, Mt Albert Science Centre, Hampstead Rd, Mt Albert, Private Bag 92021, Auckland, New Zealand

ABSTRACT

Historically firearm and toolmark examiners have rendered categorical or inconclusive opinions and eschewed probabilistic ones, especially in the United States. We suggest this practice may no longer be necessary or desirable, and outline an alternative approach that is within a comprehensive logical/Bayesian paradigm. Hypothetical forensic and non-forensic examples are provided for readers who are practicing firearm and toolmark examiners, and the strengths and weaknesses of both approaches are considered.

© 2013 Forensic Science Society. Published by Elsevier Ireland Ltd. All rights reserved.

1. Introduction

In the United States especially, firearm–toolmark examiners often provide in their scientific reports conclusions that are categorical— identifications (individualizations) and exclusions. These conclusions attempt to answer, for example, this question: Was this bullet fired from this gun? The answers given are often Yes or No, with a "reasonable degree of scientific certainty," or practical certainty. These results, though usually highly probative we believe, do not reach, and cannot reach, perfection (not even an exclusion of two bullets with fundamentally different rifling characteristics reaches absolute certainty, though it comes vanishingly close). Thus what is important to note here is that (1) these conclusions are in the form of posterior, or final odds, and (2) mathematically, posterior odds can never equal infinity (probability of 1).

Moreover, posterior odds cannot rationally exist without the assumption of prior odds, whether explicitly or implicitly. The unwitting assumption of prior odds is not so apparent with firearm and toolmark analysis, but it happens nonetheless with each categorical conclusion offered.

It has been a forceful argument of many experts in evidence evaluation for many years that forensic experts ideally should *not* assume prior odds and then produce posterior odds [1–3]. If a scientist ignores all non-scientific, contextual information in assigning a prior, then final probabilities could be very low and extremely misleading to attorneys and the court. But if she does account for outside, non-scientific information in the prior (that is, she is thinking of the probability as a juror would), then not only is she risking scientific bias, but she would be effectively usurping the rightful role of the court in assessing investigative and other information, and the court also would likely double count evidence contained in the scientist's prior. The forensic scientist should properly consider only outside information that could affect her laboratory observations (what one actually sees under the microscope) or non-case information that is scientific and background in nature (such as the technical literature, training and experience) and that bears on the interpretation of her results. There is no bright line crisply separating the proper from the improper use of contextual information by the forensic scientist. Judgment is involved. But ideally the scientist should leave the assignment of prior odds to the judge and/or jury.

There are at least three solutions to this less-than-ideal state of affairs. First, the discipline could continue with the current paradigm where practiced, but in the interest of intellectual honesty examiners could provide transparency in reports and testimony by (1) avoiding any suggestion of absolute certainty and indeed by providing words such as "practical certainty" to categorical conclusions, and by (2) disclosing the assumption of contextual information that is embedded in prior odds, which odds are implicit and logically necessary to effect these conclusions. It is unclear how the second of these would be accomplished without adding a detailed verbal explanation, and we view this solution as less than desirable.

As a second alternative, examiners could provide in the examination report a conclusion by using the complete Bayes' rule. That is, provide a final probability in quantitative terms. Both an LR and the prior odds must be calculated and assigned, respectively. Clearly, however, this

* Corresponding author. Tel.: +64 9 8153912; fax: +64 9 8496046.
E-mail address: gerhard.wevers@esr.cri.nz (G. Wevers).

runs counter to the principle that scientists should not be assigning prior odds. Another option would be for the scientist to leave the assignment of priors to the attorneys and the court, but provide guidance. That is, the scientist could provide a sensitivity table in her report, such as the one below, to demonstrate how changing the prior odds would change the final odds.

Prior odds	Likelihood ratio (LR)	Posterior odds
1 to 1000	5000	5 to 1
1 to 20	5000	250 to 1
1 to 5	5000	1000 to 1
1 to 1	5000	5000 to 1

This solution is actually similar to the third alternative below, with the exceptions that LRs must be quantitative and that they instruct the contributor and the court on how to reason to final odds given the examiner's LR conclusion.

The authors' preference is to abandon conclusions in the form of posterior odds and shift to using a full likelihood approach (this also could be termed a probabilistic or logical approach). Importantly, an LR could be provided in either qualitative or quantitative terms. Consider how this solution could work in the following example.

2. A concrete example

In addition to Fig. 1, here are the relevant facts for this hypothetical case, from either the submittal letter or the bullet examination:

1. A submitted .22 caliber projectile is on the left of the photograph; a second .22 caliber projectile was recovered from another crime scene 300 miles distant and is on the right. Both bullets are typical in appearance and measurable features with bullets fired from .22 Long Rifle cartridges.
2. No firearm was submitted or found.
3. Fig. 1 displays the best microscopic agreement observed. Overall, the microscopic agreement between the two specimens was very poor, although a small amount of correspondence was observed around the bullets.
4. All class characteristics were in agreement. The rifling was 7 lands/grooves with a right twist. A search of the FBI's "General Rifling Characteristics" database returned no firearms in .22 caliber with these general rifling characteristics (GRCs).
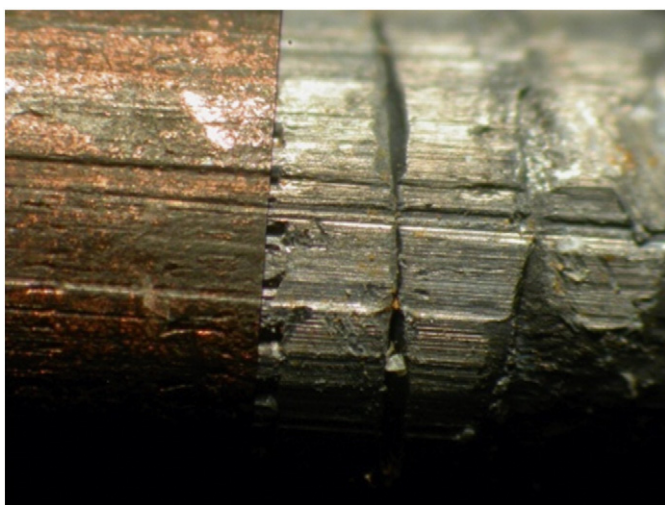

Fig. 1. Photograph of a comparison between two .22 caliber projectiles found 300 miles apart.

5. Also submitted to the laboratory was a videotape showing the defendant threatening to shoot a convenience store clerk with what appears to be a large-caliber revolver. The tape was submitted for the purpose of having the revolver's make and model identified as part of a second examination.

The contributor wishes to know if these two projectiles were fired from the same firearm. Assume that you are an examiner working this case. By using the traditional categorical model, an inconclusive result would be the norm. How would one (you) calculate a likelihood ratio in this case? First, let us note that when determining or calculating a likelihood ratio, the only matters typically under consideration are your observations and your scientific knowledge base. Thus for a microscopic comparison anything associated with the activity (the shooting), such as, the videotape and the differing locations where the bullets were recovered, should be ignored.

Let us also note again that a likelihood ratio needs not be quantitative [4]. For this case, the prosecution hypothesis is that the two bullets were fired from the same firearm. The defense hypothesis is that they were fired from different firearms. We will designate these propositions with symbols, SF and DF, respectively. The LR is thus the probability of your examination observations given the scientific background information and assuming that they were fired from the same firearm, divided by the probability of your examination observations given the background information and that they were fired from different firearms. This can be written symbolically as follows:

$$LR = \frac{P(obs \,|\, I, \ SF)}{P(obs \,|\, I, \ DF)}$$

where "I" represents the scientific background information (and is usually omitted from the expression for the sake of brevity). The vertical line (|) denotes a conditional, and simply means "assuming that," or "given that." Thus P(obs|I, SF) can be read as "the probability of the observations given that the bullets were fired from the same firearm." Happily, one of the features of the odds form of the Bayes' rule is that multiple LRs containing separate pieces of evidence (observations) that are independent (i.e., changing one does not change another) can simply be multiplied together to form a more updated, single LR for a given hypothesis and conditioning information. In our example, class-characteristic evidence is very likely independent of microscopic-agreement evidence. Changing one should not change the other, or only slightly, such that the assumption of independence is reasonable. In this case there are two pieces of evidence—the caliber and the GRC data/match on the one hand, and the degree of microscopic correspondence on the other. Thus we can assign two LRs, one for each, then multiply them together to obtain the overall strength of the bullet evidence.

So what are your probabilities of observing .22 caliber and GRCs of 7/right on these bullets? Go to your background information. Let us assume your laboratory does not yet possess an internal caliber/GRC database from physical evidence submissions. But you do know that the FBI-GRC database returns zero guns that are .22 rimfire and .22 centerfire caliber and are also 7/right. Your search of sales literature and the internet also returns nothing. And you have never seen 7/right in your experience. Let us pretend more research finally uncovers one make of rifle in .22 Long Rifle caliber made in Latin America in the 1950s with rifling of 7/right, but that is all (again, this is a hypothetical example). Therefore, as 7/right is a rare rifling characteristic, you would assess the probability of observing 7/right rifling, assuming that the bullets were fired from different guns, as quite low. And conversely, as this rifling characteristic is rare, the probability of observing 7/right rifling, assuming the bullets were fired from the same gun, is extraordinarily high, virtually 1. Thus the ratio of 1 to "quite low" is quite high. So for now our class-characteristics LR is "quite high."

But we should refine this further. In New Zealand the Institute of Environment Science and Research Limited (ESR) has adopted the following verbal likelihood scale that we will use for illustration purposes:

Report language	Likelihood ratio
Extremely strong support against	<0.000001
Very strong support against	0.000001–0.001
Strong support against	0.001–0.01
Moderate support against	0.01–0.1
Slight support against	0.1–1
Neutral	1
Slight support for	1–10
Moderate support for	10–100
Strong support for	100–1000
Very strong support for	1000–1,000,000
Extremely strong support for	>1,000,000

From the caliber match, the GRC data and the GRC match, and from just perusing the verbal side of the scale, you might decide that, given the rarity of these GRCs, that the LR is approximately 1000 and that *strong support for* is the best single fit with your observations.

You next observe the overall microscopic agreement between the specimens. Again, it is very poor, but some correspondence does exist, although much less than would be expected for a known match comparison. As part of your overall knowledge base, you especially consider the possibility of a subclass marking from the technical literature and from your laboratory experience and training. You decide that due to the poor correspondence, the microscopic agreement alone—separate from the GRCs—has an LR greater than 1, but less than 10 and therefore provides only *slight support for* the proposition that the bullets were fired from the same gun. Combining these two pieces of evidence (the two LRs), you believe that the best fit is *very strong support for*. Note that this result is quite different from a traditional conclusion of inconclusive.

Most examiners would probably prefer the above qualitative over the quantitative approach for two reasons: (1) they are familiar and comfortable with present-day, traditional conclusions that also are reached via a qualitative process, and (2) there is no firm and accepted database for population frequencies of class characteristics for bullets, nor a probability model for microscopic marks that is nearly as well developed as those for, say, DNA evidence. Nevertheless, a quantitative approach is logical and justifiable as an option. Further, it is often easy to dismiss how much information already exists from extant research. Then too, it would be a relatively small matter for a laboratory to keep track of class-characteristic data on all submitted firearms, via their own internal database. The frequencies yielded would be directly relevant to calculating class-characteristic likelihood ratios. (Note that caliber and GRC frequencies from firearms submitted to a laboratory are arguably more suitable for class-characteristic LRs than are background population frequencies—which would be much more difficult or impossible to obtain in any case. The information used would be from a pool of potential crime guns, as opposed to a pool of all guns.) In actual practice, of course, an examiner could use both the qualitative and quantitative approaches and compare them to settle on a final conclusion.

In any event, let us proceed with the quantitative method for this example. Of all the firearms in the relevant geographic area, or better, from your laboratory's caliber/gauge/GRC database (and this would include shotguns), and also from your knowledge base, you estimate that about 1 in 800 fire .22 caliber projectiles with GRCs of 7/right. This yields a class-characteristics LR of 800 given the alternative hypothesis in this case. As for the micro-correspondence, you again take into account your knowledge base, including if you wish, the number and groupings of consecutive matching striations (CMS) that you observe. You decide that this evidence alone is weak; that perhaps 1 in 10 "innocent" guns might display this low level of

agreement (again, including shotguns). And the micro LR's numerator—the probability of observing this level of correspondence given they were fired in the same gun? You might judge this as anywhere from between zero and 1, depending on several factors that could affect what you observe, but for the sake of demonstration, let us use 0.5. So the LR for this micro evidence is $(0.5/0.1) = 5$, and thus the total, overall LR follows:

$$LR = \frac{800}{1} \times \frac{0.5}{0.1} = 4000.$$

This means that you believe that your observations were 4000 times more likely if the bullets were fired from the same gun as opposed to different guns. You consult the verbal scale; your verbal conclusion is that your observations provide *very strong support for* the proposition that the bullets were fired from the same gun as opposed to having been fired in different guns.

Note here that if the scientist wishes, a "worst-case/best-case" bracket could inform the conclusion, over and above the quantitative LR range in the verbal scale. Doing so is attractive for those cases for which unreliable or no data exists, or for which the examiner has little confidence in the data. An examiner may believe that a single LR could be significantly off the mark, but at the same time strongly believe that the LR lies within certain limits. The procedure is analogous to providing quantitative distance brackets in gunshot residue examinations for muzzle-to-garment distance. In this case, one might strongly believe that the true population frequency of .22 caliber *and* 7/right in the relevant geographic area lies in the bracket of from 1 in 300 to 1 in 10,000. It is the same for the micro-correspondence in that it might lie between 1 in 8 and 1 in 35. These figures would correspond to total LRs of 1200 to 175,000—ignoring for simplicity sake any bracket for the micro-LR numerator—which in turn would correspond on the verbal scale to the language of *very strong support for*. Thus your report language could state that your observations provide *very strong support for* the proposition that the bullets were fired in the same gun as opposed to having been fired in different guns.

3. A second example

In addition to the above photomicrographs, here are many of the relevant facts for this hypothetical case, taken from the shotshell examinations:

1. Two 12-gauge shotguns were submitted: a break-open, hammerless single-barrelled shotgun, and a pump-action shotgun. Also submitted was a fired 12-gauge evidence shotshell (Fig. 2).

Fig 2. Single-barrelled test-fire on left; evidence shotshell on right.

2. The single-barrelled shotgun's firing pin was not free-floating, and protruded slightly into the breech area when the action was in an open position.

3. Whether visible in these reproductions or not, the single-barrelled test-fired shotshells featured fine, parallel breechface marks, along with the vertical mark on the primer at the 12:00 position. These marks were highly similar among all test-fires. The pump-action test-fired shotshells featured non-linear, "pebbly" marks on the primer, as did the submitted evidence shotshell. The vertical mark was absent from both pump-action test-fired shotshells and from the evidence shotshell (Fig. 3).

4. The firing pin impressions for the evidence and all test-fired specimens were highly similar in size and shape, with virtually no marks in their interiors. Thus, by using the Association of Firearm & Tool Mark Examiners (AFTE) Glossary definition of class characteristics, they (the firing pin impressions) were virtually the same for all specimens [6]. The two marks on the battery cup of the single-barrelled test-fired shotshells in the 7:00 position were inconsistent across test-fired shotshells. They were absent on the test shotshells prior to firing (Fig. 4).
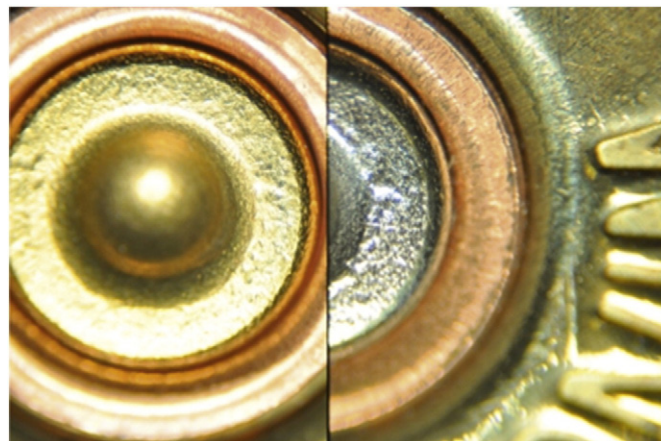
The contributor wishes to know if the submitted, fired evidence shotshell was fired from the single-barrelled shotgun. Again, assume that you are an examiner working in this case. By using the traditional model, many examiners might individualize the submitted specimen to the pump-action shotgun, given the small marks on the left and right sides of the primers. Some might exclude the evidence shotshell as having been fired in the single-barrelled shotgun. Others, on the ground of similar class characteristics, might render an inconclusive verdict. But how might you approach this case using likelihood ratios?

For the sake of brevity, let us consider only the quantitative approach, though realizing that the purely qualitative approach is perfectly legitimate. The prosecution hypothesis is that the evidence shotshell was fired in the single-barrelled shotgun. We have two plausible defense hypotheses: that the submitted evidence shotshell was fired in some unknown firearm, or that it was fired in the pump-action shotgun.

First let us do the math given that we are comparing the prosecution hypothesis to defense hypothesis #1, in that the shotshell was fired in some unknown firearm. From your knowledge base, and these examination observations, you might assign the following probabilities that comprise two LRs, one for the class characteristics and the other for the micro-correspondence:

P(class observations|shotshell fired from the single barrelled shotgun) = 0.95
P(class observations|shotshell fired from an unknown firearm) = 0.2

Fig. 3. Pump-action test-fire on left; evidence shotshell on right.

Fig. 4. Pump-action test-fire on left; evidence shotshell on right.

P(micro observations|shotshell fired from the single barrelled shotgun) = 1/2000, or 0.0005.
P(micro observations|shotshell fired from an unknown firearm) ≅ 1.
Thus the overall LR for comparing the single-barrelled shotgun to the unknown = (0.95/0.2) × (0.0005/1) ≅ 0.00238.

Since this is less than one, it weighs against the prosecution hypothesis in the LR's numerator and in favor of the defense hypothesis that is in the LR's denominator. (The reciprocal, 420.1, favors the defense hypothesis by this amount. That is, the evidence is 420 times more likely given that the shotshell was fired in an unknown shotgun as opposed to being fired in the single-barrelled shotgun.)

Consulting the verbal scale, our conclusion is that our observations provide *strong support against* the proposition that the shotshell was fired in the single-barrelled shotgun, as opposed to having been fired in some unknown shotgun. (This is equivalent to saying that the observations provide *strong support for* the proposition that the shotshell was fired in some unknown firearm as opposed to having been fired in the single-barrelled shotgun.)

Now let us do the computations given that we are comparing the prosecution hypothesis to defense hypothesis #2, that the shotshell was fired in the pump-action shotgun. You might assign the following probabilities:

P(class observations|shotshell fired from the single barrelled shotgun) = 0.95
P(class observations|shotshell fired from the pump action shotgun) = 0.95
P(micro observations|shotshell fired from the single barrelled shotgun) = 0.0005.
P(micro observations|shotshell fired from the pump action shotgun) = 0.9.
Thus the overall LR for comparing the single-barrelled shotgun to the pump-action shotgun = (0.95/0.95) × (0.0005/0.9) ≅ 0.00056.

Consulting the verbal scale, our conclusion is that our observations provide *very strong support against* the proposition that the shotshell was fired in the single-barrel shotgun as opposed to the pump-action shotgun. (This is equivalent to saying that the observations provide *very strong support for* the proposition that the shotshell was fired in the pump-action shotgun as opposed to the single-barrelled shotgun.)

Several observations should be made here. First, as mentioned, probabilities are personal. Examiners will vary in their assignment of them. For example, consider the 0.9 figure for the P(micro-obs|pump-action). Here one would normally tend toward higher probabilities if the micro-

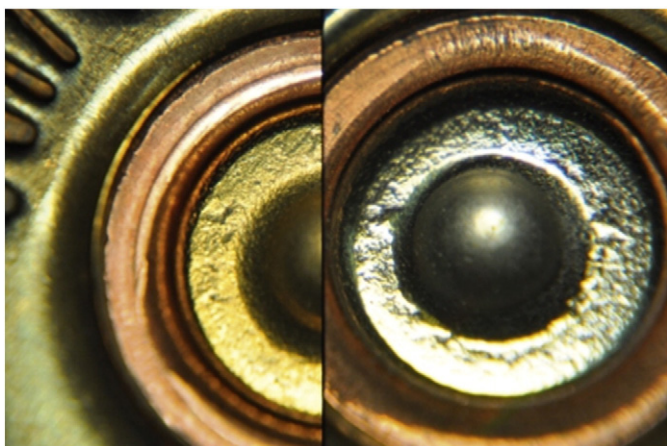correspondence were high, and tend toward lower probabilities as the correspondence decreases. However, if the guns were available for test-firing, information about how well they reproduce marks from shot-to-shot would be highly relevant, as would information about the time span between crime-shot and test-firing, information about tampering, etc. (It is our view that the latter type of information would be a legitimate scientific information to be included in the assignment of probabilities, for the reason that the examination observations would be affected.) Clearly, an expert's judgment is involved. Thus examiner "A" might assign a numerator value of 0.9 for this LR; examiner "B" may assign a value of 0.7 or 0.95, or what-have-you.

But this variance is perfectly legitimate and simply reflects reality. Fully informed and competent experts can, and will, reasonably disagree in their judgments, whether radiologists or civil engineers; nor is this different from traditional practice. It is well understood and accepted that for any specific case, some examiners may conclude Yes (or No), while others offer inconclusive conclusions. Moreover, the New Zealand experience suggests that there is a good degree of consistency in verbal conclusions across examiners (inter-examiner reliability/precision), and this performance parameter is capable of a more meaningful measurement for scaled conclusions, via validity and proficiency tests, than it is for the Yes/No/Maybe responses.

Second, notice how the LRs–and thus the strength of the evidence–change depending on the hypotheses compared. Comparing one possibility to many possibilities is not equivalent to comparing one possibility to another single possibility.

Third, all known scientifically relevant information can be invoked when assigning probabilities in the LRs. This includes empirical and theoretical studies on the probability of random striation matches, research on subclass marks, and laboratory error rates from validity and proficiency tests. Error rates are a source of much debate, and clearly past error rates should not be imported directly into a specific examination. Nevertheless, the possibility of a false-positive laboratory error from specimen mix-ups, for example, is real. Though admittedly not a purely scientific consideration per se, the probability of an error in a report or in testimony obviously bears on the administration of justice. And the probability of laboratory error for a specific case easily can be many orders of magnitude greater than, say, a DNA random match probability. One of the advantages of the likelihood approach is that, if the scientist wishes, and if it is in accord with laboratory policy, the potentiality for laboratory error can be incorporated directly into the LRs.

4. Laboratory error

For example, let us assume that in a specific cartridge-case examination for which your initial micro-correspondence LR very strongly supports the prosecution hypothesis of same-gun (SG), your probability for the microscopic observations assuming a different gun (DG) is 1 in 2000 (essentially your probability of a random- or a subclass-match, where "match" represents the quality and quantity of micro-correspondence observed in this comparison). This figure of 0.0005 would be the denominator of a micro-correspondence LR. But now you also wish to consider the possibility of a false-positive (FP) laboratory/practitioner error, specifically a false-positive error from a specimen mix-up between evidence and test-fired specimens. Given your lab's history, given past false-positive errors in your lab stemming from specimen mix-ups, your lab's review policies, and given the nature of the corrective actions taken subsequently, and of course your own mix-up experience, you assign a false-positive probability due to mix-up of 1 in 800 (0.00125) *for this specific case*. The two events, a coincidental or subclass match on the one hand, and a false positive match due to a mix-up on the other, are mutually exclusive, i.e., they cannot happen at the same time. Thus the two probabilities are simply added: the LR's denominator would now be $0.0005 + 0.00125 = 0.00175$, and the micro LR, assuming for simplicity a numerator of 1, would be

approximately 571, not 2000. In symbols this would be expressed as follows:

$$\frac{P(\text{micro observations}|SG)}{P(\text{micro observations or FP}|DG)} = \frac{1}{0.0005 + 0.00125} \approx 571.$$

Thus the inclusion of the false-positive potentiality from an evidence/control mix-up has changed the microscopic LR from 2000 to 571, both favoring the prosecution hypothesis.

To avoid double counting the potential lab error, the class-characteristics LR would remain the same: let us assign its value as 50 in this example. Therefore the overall post-error LR would be $571 \times 50 = 28{,}550$. From the verbal scale this result would still be classified as very strong evidence in favor of the prosecution hypothesis.

Of course laboratory/practitioner errors of the false-negative type also can occur. These can be included for cases in which initial LRs support a negative, e.g., support a defense hypothesis of DG. Though we include no sample calculation here, for these cases the probability of an error would be incorporated into an LR's numerator instead of a denominator.

5. Conclusion: the pros and cons of the likelihood approach compared with the current categorical approach

The main flaw and weakness of the dominant paradigm have been largely set forth in this paper already: categorical conclusions take the structure of posterior odds, which ideally should not be provided in reports and testimony. By their very nature, posterior odds in the firearm–toolmark discipline incorporate non-scientific contextual information that is contained in the prior odds, which odds thus are inappropriate for the forensic scientist to assume. Moreover, presently most firearm and toolmark examiners are unaware that they are assuming prior odds, and thus are unaware that they are incorporating contextual information in their conclusions.

Consequently, it is also our belief that the traditional model is far more prone to a contextual bias. Consider the following four statements:

1. Given the potentiality for contextual bias, no responsible forensic scientist should include information that is completely external to an examination when forming an evaluative opinion; only the observations before him and the relevant background scientific information should be brought to bear. The outside information is for the court to consider, not the scientist.
2. Firearm and toolmark examiners, indeed examiners in all of the individualizing sciences, often offer categorical conclusions. And these take the form of posterior odds;
3. Posterior odds require prior odds. In actual practice, examiners who individualize unwittingly assume a primitive or naive prior of 1:1 or thereabouts, which contains non-scientific contextual information. Thus their categorical conclusion also necessarily draws on non-scientific contextual information.
4. When seeking an answer of posterior odds, Bayes' rule provides that new, relevant information of all types be accounted for in either the prior odds or via an additional likelihood ratio having to do with this new information. In this way posterior odds are updated, as we have seen.

But this set of statements is clearly self-contradictory. Interestingly, the experiments conducted by Itiel Dror and his colleagues with fingerprint examiners illustrate this contradiction [6,7]. The examiners often changed–updated–their conclusions given new, outside information. This was a quite logical step in and of itself, in accordance with the Bayes' rule. But it violated the principle that forensic

scientists should never incorporate non-scientific information external to the examination process.

Another weakness of the traditional paradigm is the all-or-nothing nature of the conclusions. Imagine a bullet comparison in which very strong micro correspondence slowly disappears. At the beginning of this thought experiment the correspondence would result in a traditional individualization. But as the correspondence gradually disappears, the examiner reaches a decision point where she "falls off the cliff" and renders an inconclusive decision. The result is that probative information is lost, information that in general can go either toward exoneration or incrimination. Another example of lost information is in the present treatment of caliber and GRC data for bullets. It is used as a filter, a screening device to be applied before proceeding onward. If further microscopic examination yields little or nothing, the caliber and GRC evidence are then lost in an inconclusive outcome. The rationale for this is sometimes given that there exist many firearms bearing those GRCs, along with that caliber, so that this information should not be considered. This reasoning is faulty. To place this reasoning in sharp relief, we know that habitual cigarette smoking increases the probability of lung cancer by many times. And yet perhaps we have all heard the "Uncle Henry" fallacy: "My Uncle Henry lived to 95 and smoked liked a chimney all his life without a speck of lung cancer. Obviously people smoke and do not get lung cancer. I do not put any stock in all that risk nonsense."

True, there are no formal databases containing caliber and GRC frequencies (though each laboratory could relatively easily construct them). But examiners know much about relative caliber/gauge and GRC frequencies from their laboratory experience, from the FBI's GRC database, from doing research and from general familiarity with firearms. Their knowledge is not perfect, but none is. Even DNA databases are imperfect representations of reality. Forensic scientists should be paid to use their expert judgment, and all judgment is imperfect [4]. Note here that part of using expert judgment may in some cases lie in the use of best-case/worst-case analyses, as presented earlier. If the boundaries of the LR-bracket from such an analysis are above and below a value of one, then no conclusion or a weak conclusion is rational and expected. If the bracket boundaries are reasonable and either less than one or greater than one, then it is our view that forensic scientists should seriously consider reporting these findings in the interests of justice and sound public policy and be willing to discuss their reasoning and assumptions in testimony and perhaps in reports. Once again, however, expert judgment is involved in deciding this question.

Related to this, much of the data that firearm examiners observe is continuous in nature. There are no discrete levels attached to the quantity and quality of micro correspondence. And yet the conclusions offered in the traditional paradigm are bluntly discrete. Clearly many individualization and exclusion conclusions are stronger than others. The traditional approach has no means to accommodate this fact.

Of course, lost information also occurs at the opposite end of the spectrum from individualizations. The generally accepted policy and protocol is to offer exclusions only with clearly incompatible class characteristics, with some exceptions [8]. But this exclusion criterion appears to be more stringent than the criteria for individualizations. The latter have to do with the degree of microscopic correspondence observed. The criterion for exclusions largely ignores incompatible microscopic correspondence (in many or most cases) for the reason that it is conceivable that the firearm or tool could have changed or been tampered with from the time of the crime to the submission of the specimen to the laboratory, thus changing the microscopic marks. But just as with cigarette smoking or GRCs, a possibility of tampering, for example, does not logically justify ignoring probabilities and expert judgment about them. A likelihood approach not only creates a greater space for examiner judgment and a range of conclusions at the negative end of the spectrum, it also allows the examiner to apply the same or near-same standard across the spectrum: class characteristics can

tend toward the positive or negative; likewise, microscopic correspondence (or lack thereof) also can tend toward the positive or negative.

Then too, the examiner using the traditional paradigm must decide when class characteristics are truly "incompatible." Again, this kind of incompatibility is not a discrete phenomenon but a continuous one. Judgment is called for in many cases (7/right vs. 5/left on bullets is obviously easy. Differing firing pin impressions on cartridge cases often are not). But a discrete cutoff, as with individualizations, means that once again, "falling off the cliff" can and does occur.

We argue that there are no scientific or logical advantages to the traditional approach. Only deficits. Nevertheless, tradition does offer some practical advantages. First, there is the ease of use by examiners, and the ease of understanding (though this ease of understanding is often illusory when the Bayes' rule and prior odds are introduced as a necessity for complete understanding). Moreover, judges and lawyers currently understand the notion of final probability better than they understand the meaning of likelihood ratios.

These practical advantages are hardly lethal to the likelihood approach, however. The now-disbanded British Forensic Science Service from 2000 to 2006 provided lectures in Continuing Education Seminars for judges in order for them to better understand the entire likelihood approach. This educational program was received positively [4].

The pros and cons for the likelihood approach are more or less the mirror images of those for the traditional approach. It is scientifically and logically sound. There is no usurping the role of the court/jury; there is no falling-off-the-cliff problem; evidence is not lost. All scientific information is capable of being smoothly and logically accounted for, including CMS data and laboratory/examiner error possibilities. There is less risk of a contextual bias. The likelihood approach should improve the overall administration of justice.

But initially at least, not all judges and juries may feel comfortable with likelihoods; perhaps much more telling, nor would examiners steeped in traditional thinking. Training for examiners would require updating. And persistent training workshops for jurists also would prove helpful, as in the British experience. Still, for some time there could be resistance in the courts, at least to some degree, and certainly resistance would come from many established examiners. But at the end of the day, as in Britain and other countries, we believe that the logical force of the likelihood approach would win converts and prevail on a practical basis as well. The experience in New Zealand has been quite positive overall.

Moreover, adoption of this approach would improve the scientific and professional status of both the science and of the firearm–toolmark practitioner, respectively. Conclusions would flow more logically and coincide far more closely with the strength of the observations/evidence. There would be no need to explain in trials or scientific validity hearings how the term *individualization* fails to mean certainty. The likelihood approach logically would do away with the misleading all-or-nothing debates over, for example, subclass marks, for this possibility could be incorporated into LRs.

For the firearm–toolmark examiner, assertions that they are mere technicians would be harder to make stick. The hallmark of a technician is the lack of theoretical knowledge of a subject and the unquestioning, rote compliance with written procedures. Training and understanding of the likelihood approach would move the needle farther toward a scientist and away from a technician. To properly execute this method, the examiner going to court must have a deeper theoretical understanding of the science and the significance of findings. In sum, it is our recommendation that examiners worldwide–especially in the United States–begin moving toward the likelihood approach and toward the standardization of verbal scales and training.

References

[1] B. Robertson, G.A. Vignaux, Interpreting Evidence: Evaluating Forensic Science in the Courtroom, John Wiley & Sons, Chichester, 1995, pp. 20–21, (219–220).

[2] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, Chichester, 2004, pp. 105–112.

[3] Guest Editorial, Expressing evaluative opinions: a position statement, Science & Justice 51 (2011) 1–2.

[4] C.E.H. Berger, J. Buckleton, C. Champod, I.W. Evett, G. Jackson, Evidence evaluation: a response to the Court of Appeal judgment in R v T, Science & Justice 51 (2011) 43–49.

[5] Association of Firearm & Tool Mark Examiners (AFTE) Glossary, 5th ed., version 5.070207, 2007, pp. 172.

[6] I.E. Dror, D. Charlton, A.E. Peron, Contextual information renders experts vulnerable to making erroneous identifications, Forensic Science International 156 (2006) 74–78.

[7] I.E. Dror, D. Charlton, Why experts make errors, Journal of Forensic Identification 56 (2006) 600–616.

[8] Guideline on exclusions, www.swggun.org.