Stephen G. Bunch,[1] Ph.D.

Consecutive Matching Striation Criteria: A General Critique

REFERENCE: Bunch SG. Consecutive matching striation criteria: A general critique. J Forensic Sci 2000:45(5):955–962.

ABSTRACT: In the forensic science of firearms and toolmark identification, examiners traditionally have drawn conclusions of identity from subjective criteria. This paper critically explores the general validity of one proposed objective-criteria regime—that of counting consecutive matching striations on fired bullets. Practical considerations are discussed, as well as theoretical ones, with both discussions viewed from the perspective of Bayesian logic. It is concluded that drawbacks exist for this particular objective-criteria regime, but that research and logical analysis should continue.

KEYWORDS: forensic science, firearms, toolmarks, objective criteria, consecutive matching striations, line counting, validity

Since Al Biasotti conducted his original identification-criteria research in the 1950s, the debate over the relative virtues of objective and subjective methods in forensic firearms identification—specifically over the virtues of counting consecutive matching striations on bullets—has blown hot and cold. Recently the debate has heated up, in part owing to the Supreme Court's decision in *Daubert* v. *Merrill Dow Pharmaceuticals, Inc*. This ruling places a premium on the demonstrated scientific validity of purported scientific testimony. An objective decision-making regime, which purportedly describes the counting of striations, appears more likely to successfully meet a *Daubert* challenge than does the subjective regime that currently prevails in the discipline. Thus, and in view of its increasing popularity, this paper sets out to critically examine consecutive matching striation (CMS) models, from the standpoint of both technical substance and the interpretation of results. The treatment of technical substance will be relatively brief, interpretive issues less so. But before wrestling with either, and because this paper is principally addressed to practicing firearm-toolmark examiners, we should first review the logic of probability as it relates to the interpretation of evidence in general and to forensic evidence and courtroom testimony in particular.

Bayesian Analysis

Classical, or frequentist, probability theory arose from the study of games of chance and the idea of randomness. In this framework, when during a long series of repetitive trials the relative frequency of an event approaches a fixed number, that number is termed the probability of the event (1). One tends to think of classical probability methods as dealing with aggregate data and closed systems, such as a game of chance. If we measure the height of 100 randomly selected 4th graders from a particular elementary school, we can rest assured that the calculated mean is a pretty good estimate of the mean of another group of 100 pupils from the same school. But what if we wish to estimate the height of a particular child who had been absent on measuring day? The frequentist would look at the frequency distribution of the data, do some quick calculations, and conclude that the probability is X that the interval between height A and height C contains the height, B, of the new pupil. Closed system. Very clean.

But the other pupils protest. They have additional, relevant information. They inform the frequentist that the pupil is a male, is the oldest in the class, drinks lots of milk, and has parents who are very tall. Unwilling to invoke older, more general axioms of probability, unwilling to introduce subjective elements, our frequentist is thus unwilling to logically evaluate this new information and may decline to venture any estimate at all.

As it happens, Bayes' Rule (named after Thomas Bayes, the 18th century clergyman who first proved it) is a logical theorem that shows us how to rationally incorporate new information into a probability, within either closed or open, real-world, systems. In this framework, probability is defined as the degree of belief in a proposition or event, and is always conditioned on particular assumptions. What Bayes and his successors did was to "rationalize" the concept of probability. One of the virtues of Bayesian analysis is that its practitioners can better analyze specific events, such as crimes. Bayes' Rule is thus:

$$P(A \mid B) = P(A)\,[P(B \mid A)/P(B)]$$

where

$P(A)$ = the probability of A, given (or assuming) no other information. This term is called the prior probability of A.
$P(B)$ = the probability of B, given no other information.
$P(A \mid B)$ = the probability of A, given (assuming) the truth of B.
$P(B \mid A)$ = the probability of B, given the truth of A.
$\overline{A}$ = not A

For example, if A is tails, and B is a fair coin, $P(A \mid B) = \frac{1}{2}$. For forensic purposes it's usually preferable to use the odds form of Bayes' Rule, which is thus

$$P(A \mid B)/\,P(\overline{A} \mid B) = [P(A)/\,P(\overline{A})][P(B \mid A)/\,P(B \mid \overline{A})]$$

where on the right side of the equation the first term in brackets is known as the prior odds, and the second term in brackets is known as the likelihood ratio (LR). The left side of the equation is known as the posterior odds, the odds we ultimately wish to know. Thus we

[1] Firearms-Toolmarks Unit, FBI Laboratory, Washington, DC.

Received 15 Dec. 1998; and in revised form 26 July 1999; accepted 25 Oct. 1999.

have the following: posterior odds = prior odds × likelihood ratio. In other words, we multiply the prior odds in favor of A, which are the odds before any new information, by the likelihood ratio, which contains the new information, to arrive at the posterior odds in favor of A, which incorporates the new information. For the courtroom, let's substitute G (guilty) for A, and E (evidence) for B. Now the equation reads as follows: the odds of guilty given a piece of evidence = the odds of guilty before the evidence is presented, times the likelihood ratio. The likelihood ratio is the probability of the evidence given the defendant is guilty, divided by the probability of the evidence given the defendant is not guilty. For firearms examiners, the evidence could be, say, five maximum-CMS on a bullet.

To acquire a better intuitive feel for how Bayes' Rule can work in the real world, consider the following example, borrowed from the work of Bernard Robertson and G. A. Vignaux (2). Imagine a police officer has a portable breath analyzer in his patrol car. The analyzer is designed to provide two answers to the relevant question, Is this motorist over the legal limit for levels of blood alcohol? A positive answer (over the limit) is indicated on the analyzer by a red flashing light; a negative by a green light. No instrument is perfect, however; the analyzer can give false positive and false negative readings. That is, it is possible for the analyzer to flash red when the motorist is under the legal limit, and flash green when over the legal limit. For obvious reasons, the police department and the local district attorney wish to minimize false positive readings. Thus the police adjust the analyzer to give false positive readings (red) for only 5 out of 1000 innocent motorists (a .5% rate). Unfortunately this means the police department must accept a higher rate of false negative readings (green)—for our example, "exonerating" 50 out of 1000 guilty motorists (a 5% rate).

Armed with the above information, the question we now wish to answer is the following: What are the odds that a particular motorist is over the limit if the analyzer flashes red? From the odds form of Bayes' Rule, we first calculate the likelihood ratio, which is, again, the probability of obtaining the evidence (a red light) given the motorist is over the limit (guilty), divided by the probability of obtaining the evidence given the motorist is under the limit (not guilty). Clearly the numerator in this case is .95 (950 over 1000), the denominator .005 (5 over 1000). Dividing, we obtain a likelihood ratio of 190.

If we stop here, how do we interpret the likelihood ratio in this case? It simply means that, whatever the odds in favor of guilt before this test, we now must multiply those odds by 190 to arrive at our updated assessment of the odds in favor of guilt. The likelihood ratio measures the strength of the evidence, but it does not assess posterior odds by itself. For that we must have the prior odds in favor of guilt.

To understand how prior odds works, imagine that our police officer was randomly stopping motorists on a busy highway at 10:00 A.M. on a Wednesday. From his past experience conducting these random stops at this location and at this time, he estimates that about 1 of every 150 motorists will be over the legal limit for blood alcohol. Thus, for any particular motorist, he would plausibly estimate the prior odds in favor of guilt at 1 to 149 (or, 149 to 1 against). For a "red light" case, then, the posterior odds in favor of guilt are (1/149)(190) = 1.28. That is, under these circumstances, our police officer would rationally believe that the odds in favor of this motorist being legally drunk are 1.28 to 1. (Since for odds of A to B the equivalent probability is A/(A+B), in our case the probability of guilt is 1.28/(2.28), or 56%.)

Now let's assume our officer is posted at this same highway at 1:30 A.M. on New Year's Day. Let's also assume that instead of stopping motorists at random he is stopping only those who are "driving drunk." From past experience he estimates that about 4 of every 5 of these motorists will be legally drunk. Thus, under these circumstances, our police officer must rationally believe that the odds in favor of a particular red-light motorist being legally drunk are (4/1)(190) = 760, or 760 to 1. This probability is 99.87%.

Note two things about this example, however. First, our police officer chose not to factor in other information that a real officer might. For example, observing that a motorist was disheveled and holding a liquor bottle, with several more empties in the back seat, surely would cause a real officer to make a reassessment. If he knew enough about probability, he in fact would insert a second likelihood ratio into the odds form of Bayes' Rule, so long as he was certain that this new information was independent of the results of the analyzer. The new likelihood ratio would be the probability of observing this evidence given guilt, over the probability of observing this evidence given innocence. Mathematically, it is perfectly legitimate to insert this new likelihood ratio, as well as others, into the equation. Indeed, this is how a perfectly rational jury would combine new evidence with what has already been heard.

Second, there clearly is an element of subjectivity involved in assessing prior odds in real world cases. Another police officer may have used a different figure. Then too, imagine that you were a driver on the New Year's highway at 1:30 A.M., but that your erratic driving had been caused by sleepiness. You also knew that you had not consumed alcohol in the past two weeks. When you were stopped, you might have assessed the prior probability in favor of your drunkenness at about 1 in 100,000 (it is at least possible that you forgot when you last partook). Thus if the light flashed red, your assessment of posterior odds in favor of drunkenness was (1/99,999)(190) = .0019, or .0019 to 1, or 1 to 526 in favor of drunkenness, or 526 to 1 against drunkenness. Clearly this result would differ from that of the police officer.

Bayesian Analysis and Forensic Science in the Courtroom

At this point it must be observed that there is no rational or scientific ground for making claims of absolute certainty in any of the traditional identification sciences, which include fingerprint, document, firearms, toolmark, and shoe and tire-tread analysis. Case-specific conclusions of identity rest on a fundamental proposition, or hypothesis; namely, that no two fingerprints, bullets, etc., from different sources will appear sufficiently similar to induce a competent forensic examiner to posit a common source. But as any logician or philosopher of science would insist, no hypothesis can be proved absolutely. In this case, proof could be attained only if at the same instant every competent examiner compared every possible combination of prints or bullets, with no resultant errors—an impossible task, with the proof valid only for an instant. From this also emerges the important distinction between simple facts and hypotheses (or theories). I have observed directly that I, my family, friends, and relatives have eight fingers and two thumbs. This is a simple fact and is unchallenged. But then to assert that all human beings (or 99%, or 99.999%) have eight fingers and two thumbs is an act of inductive inference—it is to set forth a hypothesis.

Consequently, statements asserting identity often include the following: "with a reasonable degree of scientific certainty; practical certainty; moral certainty; beyond any credible doubt; a practical impossibility of dissimilar origin." So long as traditional, subjective forensic examinations are conducted, this kind of concluding terminology is acceptable (3). But when we consider more objective decision-making regimes, quantitative conclusions become possible, and therefore desirable—provided they are understood by examiner and jury alike.

We mention this here so the reader will not be tempted to dismiss a probabilistic analysis on the grounds that (A) absolute proof of identity conclusions is possible, (B) the ability to successfully handle a hostile cross-examination trumps science and logic, and/or (C) a CMS model is an identification model, not a probabilistic model. (B) is a matter of opinion, but we would hope that good science and logic determine the content of expert testimony, not the structure of the justice system. I have just argued that (A) is false, and as we hope to show, (C) also is false.

In the meantime, some authorities on evidence interpretation in the courtroom have insisted that, if possible, expert witnesses should provide the jury with likelihood ratios, not posterior odds (4). The reasons for this recommendation are several. Most important perhaps, it is the province of the jury to decide the ultimate issue of guilt. If an expert witness is in effect testifying to the posterior odds of guilt, he is usurping the obligation of the jury. He is substituting his judgment for that of the jury. Related to this, recall the example of the breath analyzer. Assessing prior odds can be and often is a subjective process. Best to leave as much subjectivity as possible to the jury. Moreover, if a witness provides the jury with a likelihood ratio, the jury can—at least theoretically—combine this evidence with evidence offered by other witnesses, as already observed. But testimony on the ultimate issue (posterior odds) cannot logically be combined with other evidence.

Finally, for the expert witness to use a prior that truly reflects her belief would result in the double counting of evidence. A firearms examiner might estimate, from experience, that one-half the firearms submitted to her laboratory unit were used in the crime alleged, and thus that one-half the defendants were guilty (more about this logical leap shortly). Thus her prior odds that a bullet passed through a particular barrel for a new case could be 1 to 1. But these odds indirectly incorporate information (from the police investigation, for example) that the court has not commissioned the examiner to judge. The court has commissioned the jury to judge it; for the examiner to use these odds in calculating a posterior-odds conclusion would result in the jury counting some evidence twice.

To solve this essentially legal problem, the examiner must assume a "naive" prior. That is, she must assume no knowledge of laboratory submission history and no knowledge of the crime, except to the extent that it occurred in a certain place and involved a firearm (if only one was used). Thus the naive prior odds in favor of the guilt of a single defendant would be 1/(the number of firearms owners in the relevant geographic area, minus 1). Of course the question of what constitutes the relevant geographic area could be a source of endless debate. Obtaining good figures for the number of firearms owners also could prove difficult.

In contrast to posterior odds, the likelihood ratio has both logical and juridical appeal. It is a direct measure of the strength of any piece of evidence. Evidence that is strong will have a high LR, evidence that is weak will have a low LR. Evidence that is useless, of no value, will have a likelihood ratio of $1/1 = 1$. For example, if an expert testifies that the blood stain found at the crime scene was reddish in color, and that the defendant's blood is also reddish in color, the LR would be one—the probability of a reddish blood stain if the defendant is guilty is the same as the probability of a reddish blood stain if the defendant is not guilty, i.e., one. This evidence is useless.

[As an aside, it should also be mentioned here that Bayes' theorem as applied to forensic evidence is actually more flexible than it has been presented up to now. The jury, logically, must compare the complete prosecution hypothesis (presumably that the defendant is guilty) with the complete defense hypothesis. For the latter there might exist a variety of possibilities. The simple case is where the defense asserts the defendant is not guilty and that some other unknown person is guilty. Or the defense may assert that another named person is the guilty party, or that the guilty party is a member of a certain subgroup, either of which can affect the denominator of the likelihood ratio. Thus, for the LR, instead of speaking about the probability of the evidence given guilt and not guilt, we can speak of the probability of the evidence given the prosecution and defense hypotheses. Present CMS regimes can deal with only the simple case in which the defense asserts the bullet was fired from an unknown firearm with similar rifling characteristics.]

So, from the foregoing, is it possible to provide likelihood ratios in forensic firearms examination? The answer is maybe. For the moment, however, most ratios presented in court would be of debatable value, as we'll see.

CMS Decision Criteria

Biasotti was the first—in print—to suggest the possibility of counting consecutive matching striations (or "lines") on bullets for use as a criterion for identifying a particular firearm as having fired a particular bullet (5). Further empirical work was undertaken at the California Criminalistics Institute—although this body of research remains largely unpublished—and recently technical articles have appeared that invoked the CMS-count approach (6,7). The general procedure is first to fire numerous bullets through many firearms of the same make and model. Keeping good records, the researcher then microscopically compares specimens known to have been fired from the same barrel and compares specimens known to have been fired from different barrels, counting how many striations match well in multiple "runs" of CMS. Biasotti found no .38 Special caliber lead bullets fired from different Smith & Wesson revolver barrels that displayed more than 3 CMS, and no metal-jacketed bullets that displayed more than 4 CMS. His clear suggestion was that, with further validation and breadth of research, examiners could reach conclusions of identity for firearms when CMS counts exceeded fixed threshold values. (A Bayesian approach was not taken. The analytical and interpretational errors that this avoidance sometimes can give rise to are discussed later.)

In the simple max-CMS model discussed here, by definition, the only CMS run on a bullet that matters is the one—or more—featuring the maximum CMS count. After the data are collected, a histogram can be plotted showing how the max-CMS values vary with the probability of finding them under the comparison microscope. Two data sets can be plotted, one for bullets fired from the same barrel (SG for same-gun) and one for those fired from different barrels (DG for different-gun). A hypothetical histogram is shown in Fig. 1.

Excluding class evidence for the moment, to calculate a likelihood ratio from the data set is straightforward. For any max-CMS, simply divide the SG probability by the DG probability (see Table 1 below). Though these data are hypothetical, note that in this simple model they yield LRs greater-than-one only for max-CMS of three or more. A dash indicates no CMS observations were effected.

The LR evidence for a comparison with a max-CMS count of six could be presented in court using, for example, either of two statements:

(1) This result is 110 times more likely if the bullet passed through this barrel than through an unknown barrel with similar rifling characteristics. Or,

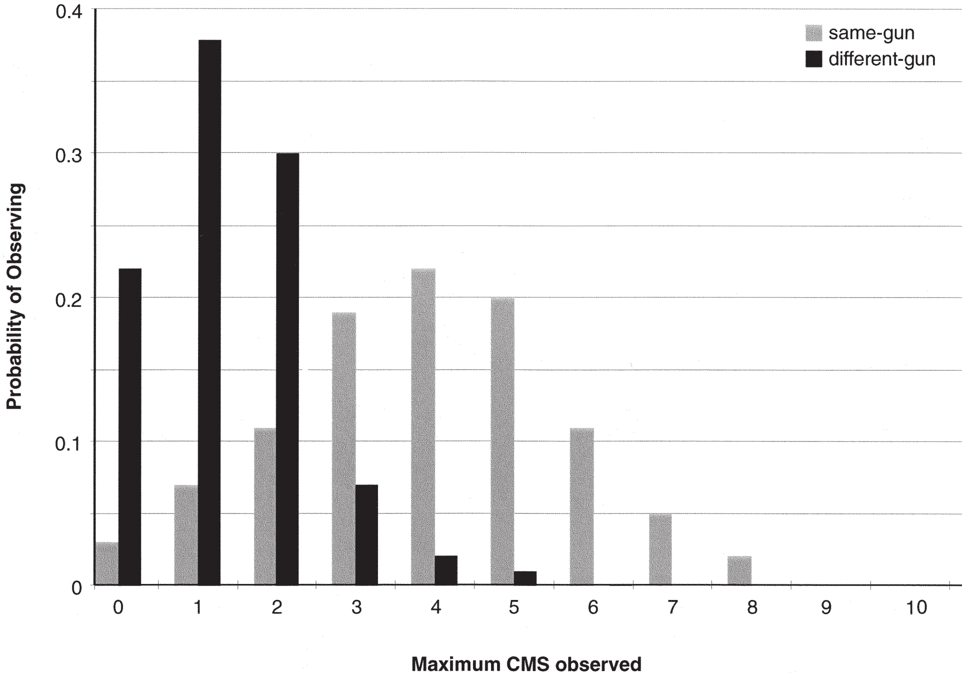

FIG. 1—*Histogram using hypothetical data.*

TABLE 1—*Hypothetical max-CMS probabilities and LRs.*

max-CMS	(1) = Pr|SG	(2) = Pr|DG	LR = (1)/(2)
0	.030	.220	.136
1	.070	.379	.185
2	.110	.300	.367
3	.190	.070	2.71
4	.220	.020	11.0
5	.200	.010	20.0
6	.110	.001	110
7	.050	—	—
8	.020	—	—
9	—	—	—

(2) Whatever the jury's present judgment of the odds in favor of the defendant's guilt, multiply those odds by 110, assuming that only the defendant could have discharged the firearm.

(The qualifying phrase in #2 points up a complication. The jury decides the issue of guilty/not guilty. Yet the firearms examiner is providing a likelihood ratio that directly relates only to the odds that a bullet passed through a particular barrel. To link the posterior odds of guilt to the provided LR requires either that the qualifying phrase above be included; or that the jury intuitively discount the LR as additional evidence is presented bearing on the odds that the defendant was/was not the shooter; or that the jury rationally discount the LR. Though beyond the scope of this paper, and beyond the direct concern of the firearms examiner, rational discounting theoretically can be achieved by considering additional, intervening hypotheses and by invoking the law of total probability (8).) It also should be mentioned that the numerical LR testimony can be supplemented by verbal add-ons if the examiner wishes. With LRs between, say, 100 to 1000 for example, the examiner could state that the results constitute weak, moderately strong, strong, or very strong (pick one) evidence.

Provided the data set incorporates a sufficiently large body of data from quality research, curve fitting is unnecessary, perhaps even undesirable from a Bayesian perspective. But it is unavoidable for the higher CMS regions where empirical data do not exist. For our hypothetical example, if seven max-CMS are observed the examiner clearly has no denominator with which to calculate the LR. She must rely on extrapolation of the data—hence the need for using a mathematical probability distribution. This would be both rational and defensible provided the extrapolation remained within reasonable limits. For the kind of discrete data discussed here, the Poisson distribution seems appropriate:

$$P(\varkappa) \mid (\ldots) = e^{-\lambda}\, \lambda^{\varkappa}/\varkappa!,$$

where $\varkappa$ is the max-CMS count and λ is the weighted average max-CMS count from actual data. For our hypothetical data the weighted averages are 3.91 and 1.32 for the SG and DG data, respectively. The results are presented in Table 2 below.

Concerning the technical substance of a Biasotti-style CMS counting regime, doubtless it offers numerous theoretical and practical benefits. It is inherently more scientific than the subjective regime currently used by the vast majority of examiners and thus perhaps more likely to successfully pass as a scientific theory or technique at a *Daubert* hearing. In regards to the testability and error rate guidelines stemming from this ruling, certainly the CMS regime is testable and with far more research could be deeply tested. True, conclusions resting on a solid CMS regime would have no error rate as such, since it's a probability model, and there can exist problems such as evidence and evidence/test-fire mixups, the counting of CMS that prove too subjective, and faulty research. Still, these can be checked, as in DNA regimes, with collective simulations or tests that compare LRs for known-non-matches to LRs for known matches. Using pristine bullets, LRs for known matches should be relatively high, non-matches uniformly low.

TABLE 2—*Poisson distributions and LRs for hypothetical max-CMS data.*

max-CMS	(1) = Pr(x)|SG	(2) = Pr(x)|DG	LR = (1)/(2)
0	.020	.267	.075
1	.078	.353	.221
2	.153	.233	.657
3	.200	.102	1.96
4	.195	.034	5.74
5	.153	.0089	17.2
6	.099	.0020	49.5
7	.056	.00037	151
8	.027	.000061	443
9	.0118	.000009	1311
10	.0046	.0000012	3833

A well researched CMS regime could result in greater confidence being placed in examiner conclusions, inasmuch as they largely would rest on the validity of published research rather than on examiner appeals in court to the trustworthiness of subjective results, or on appeals to the results of relatively few proficiency tests. Other benefits: photomicrographs possibly could be used by other examiners to review conclusions; broad research on CMS counting, which included all types of barrels, would obviate the need for examiners to invoke any theory of toolmark uniqueness (any so-called subclass marks on bullets would be insinuated into the research data at a frequency roughly similar to that in the real-world barrel population, and thus be insinuated into the LRs); with the use of LRs, there would be none of the "falling off the cliff" that sometimes exists when examiners draw a bright line between conclusions of identity and conclusions of inconclusive, i.e., lesser degrees of probative striation evidence would not be effectively ignored; and, finally, there could be fewer professional risks to individual examiners. Leaving aside for the moment the question of variation in counting CMS and the language of the law, fixed decision rules imply that an examiner would be testifying to the objective results of her examination, not to her subjective opinion. If the results are mistaken, the fault must lie either with the validation research or with quality control in the laboratory. Only the latter could implicate the examiner (provided she wasn't a key contributor to the validation research).

Then there is the issue of bias. A weakness of subjective examinations is what one would presume is their greater vulnerability to various sources of possible bias. One source, for example, is the written account of the crime that the contributing agency usually submits with the evidence. Therein a suspect is often implicated. No "control" suspects are mentioned, nor are control "evidence" samples submitted. In one experiment, a group of hair examiner trainees was divided in half; subgroup A was given hair samples provided in the usual fashion, from the crime scene and from one suspect; subgroup B was provided samples from the crime scene and from five possible suspects. In actuality the crime scene hair matched none of the suspects' hair, but 30.8% of those in subgroup A concluded they did match the suspect. Only 3.8% of those in subgroup B concluded they matched a suspect (9).

Hair matching, like present-day firearm identification, in the final analysis is subjective. Had these students been armed with clear-cut objective matching criteria, and in drawing their conclusions instructed not to deviate from these criteria without good reason, the results of subgroup A would doubtless have been closer to those of subgroup B. Objective criteria allow the examiner to better insulate himself from all sorts of outside factors that conceivably can influence results.

Practical Difficulties

Nevertheless, benefits notwithstanding, there clearly exist practical difficulties with a CMS regime. Some critics of counting CMS argue that it oversimplifies reality, and they point to the subtleties of pattern recognition that defy quantification and objectification. True, counting striations simplifies, but to a scientist simplicity is a virtue. Astrophysicists calculate the motions of heavenly bodies by hypothetically reducing an irregular mass with size and shape to a point mass with no size or shape. And their predictions are remarkably accurate. Moreover, the human brain often discerns "obvious" patterns from meaningless jumble (e.g., faces and objects in clouds, canals on Mars, the man-in-the-moon, and "trends" in the random, successive changes in common stock prices). The issue really reduces to what magnitude of LRs one can expect from a CMS regime. If typical LRs are in the neighborhood of 10 to 100, then the critics would perhaps be right. The strength of this kind of evidence is relatively low.

Next comes the question of subjectivity in counting striations. Interestingly, Evett and Williams conducted a test in England and Wales in which fingerprint examiners were given 10 sets of latent and ink prints to compare. One set was from different persons but with the prints modified to show many points of similarity. The examiners were asked to conclude if the prints in each set were from the same source, and if they concluded they were, to determine also the number of points of similarity. Not a single examiner misidentified the severest test set, the modified set. But counting points of similarity proved highly subjective, examiner counts varying from 10 to 40 for one set, 8 to 26 for another, and 14 to 56 for a third (10). Thus if a 16 point or greater standard were used, some examiners would have reported the results as inconclusive, while most would have been happy to go into court with a positive identification.

Does this degree of subjectivity in counting "objective" points of similarity also hold for qualified firearm-toolmark examiners when counting CMS? With consistent, national training, individual judgments on the quality and quantity of striations should converge; but they will never be unanimous. This simply means that examiners would sometimes report different LRs for the same evidence bullet. This is not so bad as it might first appear. It is merely the analogue of examiners, using traditional methods, drawing two different conclusions about the same bullet: identification or inconclusive. Differing LRs simply reflect the fact that even objective regimes can contain subjective elements.

A more serious problem: obtaining a truly valid and usable CMS regime would necessitate a large-scale research program involving numerous varieties of bullets and barrels, tens of thousands of test firings, and possibly careful mathematical curve fitting. The need to analyze multiple runs of CMS only complicates matters, but without this kind of analysis much information is lost. Indeed, the firearm-toolmark community could commit a long-term mistake by underestimating the scope of the research required to truly validate a CMS regime. As the history of DNA validation has revealed, hard scientific research and the attendant statistical analyses can elicit equally hard questions that scrutinize every level of analysis and interpretation. Have disinterested "control" examiners been used in the research? Are the groupings of research barrels representative of the current barrel population? Can we conclusively show that they are, or why they need not be? Can an examiner clearly show that, say, the Poisson function provides a better fit to the data than an alternative? Should the fitted curves or the actual probability data be used across-the-board for LR calculations? Why? The list goes on, but the foregoing questions are the simple ones. The point is, if CMS criteria are widely adopted, we can rest assured that tough questions eventually will be raised in court, possibly by high-powered scientists and statisticians hired by opposing counsel.

The final practical difficulty involves explaining and defending in the courtroom conclusions resting on a CMS regime. Examiners schooled in subjective methods may fail to understand or appreciate the research and the logic of interpreting this kind of evidence. Thus they may find it difficult to explain them to judge and jury. The variation in examiner LRs discussed above, while logically harmless, also could prove difficult to explain and therefore be legally harmful. It can be done; DNA examiners successfully wrestle with these difficulties regularly. But if firearm examiners wrestle with them less successfully, it could be a blow to the profession and to the administration of justice.

Theoretical Problems

If the foregoing practical problems are important, the more theoretical problems associated with a CMS regime also must be examined. To begin, when examining bullets the examiner looks for evidence in two categories. The first involves the rifling impressions left on a bullet by the barrel. In number, direction of twist, and widths, the impressions on test fired bullets must match those on the evidence bullets. This is class evidence; it can greatly restrict the number of suspect barrels. The second, of course, involves the agreement of microscopic, striated marks within the rifling impressions, the very type of evidence we have been discussing. A likelihood ratio can theoretically be calculated for each—and since they are independent, multiplied together—but one for class evidence would require knowing the proportion of firearms within the relevant population that share the class characteristics of the evidence firearm. Let's say a homicide occurs in Denver, and the firearm population of Colorado is taken as the relevant population of firearms. What proportion of these match in rifling characteristics the suspected weapon? We don't know. No database exists, and at the moment it seems unlikely that one could be constructed. Manufacturers are sometimes reluctant to release past production figures, let alone current ones. And one cannot establish with confidence what happens to firearms once in the hands of the consumer. Estimates can perhaps be made, which would be better than nothing, but these would be rough and endlessly challenged in court. And without the class characteristics LR, the strength of an examiner's testimony would be diminished, and in particular cases could be immensely diminished. This is not a weakness in the CMS regime per se, but it does logically flow from the kind of probabilistic analysis that a CMS regime requires.

A more serious weakness in CMS counting, however, is that evidence bullets are not fired in new, clean barrels, as would be the research bullets. Real barrels change over time. A bullet fired from a new, clean barrel might, under the microscope, appear nothing like one fired later from the same barrel. To the point, a barrel's interior can become worn, corroded, or the object of willful tampering in the span of time between the crime and the recovery of test fired bullets. One can argue that this merely results in a lower max-CMS count, which reduces the LR, which in turn favors the defense. But a little thought and a study of Fig. 2 shows this is not necessarily the case.

This plot—and continuing with the earlier hypothetical data—shows the kind of shift needed if we assume a barrel has changed from the time the evidence bullet was fired to the time the test bullets were fired (CG = changed-gun). That is to say, the SG probability data must shift downward to some position between the original SG and DG probability data. Why? Because we are less likely to find, say, five, or six, or any particular number of max-CMS on bullets from the same barrel, if that barrel changed between firings. The SG data, derived from barrels that changed very little, no longer apply. Meanwhile, the DG probability data would shift little if at all. The probabilities for chance striation matches

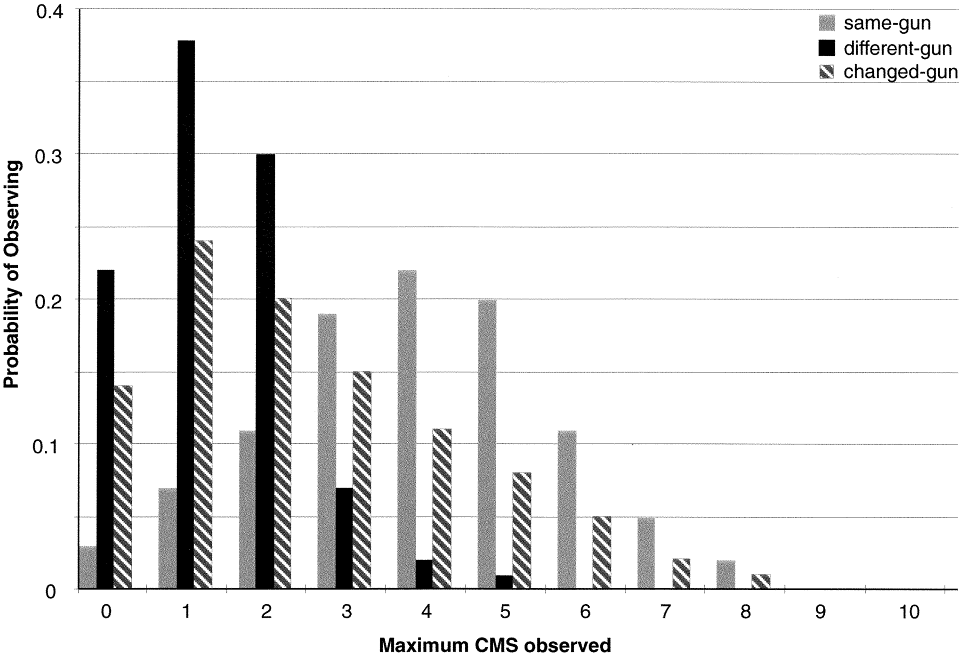

FIG. 2–*Histogram using hypothetical data, including changed-gun data.*

should not change significantly merely because a single barrel changes within the population of barrels.

To illustrate the problem, first consider scenario 1: an examiner is provided with an evidence pistol seized from the suspect immediately after the crime. Also provided was a pristine, metal-jacketed evidence bullet recovered from the victim's body. The barrel appears new and relatively clean, and we'll assume no changes in the barrel have occurred from corrosion. Test bullets are collected and compared to the evidence bullet, and the examiner finds six max-CMS on a land impression. He next consults the appropriate research data for this variety of bullet and barrel and finds the probabilities of finding six max-CMS are .110 and .001 for the SG and DG, respectively. Thus the calculated likelihood ratio is .110/.001 = 110. Since the research conditions approximated the actual conditions for barrel and evidence bullet—though in actuality this can never be strictly known—this ratio is well supported and realistic.

Now consider scenario 2: all the conditions are the same as above except for the barrel bore, which has changed somewhat. Note that the probability of now finding four max-CMS using the CG data is approximately equal to the probability of finding six max-CMS in scenario 1, i.e., .110. Of course, in actuality we cannot know the actual appearance of a CG histogram other than that it will fall between the original SG and DG probability data. There is no way to research it adequately, no way to know how much an evidence barrel has changed, and thus no way to know how much the SG data should shift when a barrel changes. All the examiner can do is use the original SG data derived from clean, relatively unchanged research barrels. Thus, at four max-CMS, the probabilities for scenario 2 are .220 and .020, which yields a likelihood ratio of 11.0. At first glance this figure seems reasonable, but the problem is that if we had used the hypothetical CG data which actually fit these circumstances, then the examiner's calculated LR would be .110/.020 = 5.5. That is to say, by using the SG data instead of the CG data (again, the CG data represent reality but are knowable in detail only in a simulation such as this), the examiner, in this context, has overstated the strength of the evidence by a factor of 2.

(As an aside, it appears that any future automated "identification" system would suffer from the same malady. Present-day automated systems score and rank comparisons, but the machines do not attempt to effect identifications or determine LRs. After conducting validation research involving same-guns and different-guns, a system conceivably could calculate LRs indirectly from a scoring algorithm. But the changed-barrel problem would remain.)

How serious is this theoretical weakness? On the one hand, it could be regarded as relatively benign. LRs in the neighborhood of 5.5 and 11 would be judged by the courts as weak evidence, and most actual cases probably would involve differences of similar magnitude. For practical legal purposes, differences of this kind often would be ignored. Moreover, the LR of 11 would in fact rest on the best information available to the examiner using a CMS model, and no set of information is perfect and complete, expecially in the context of the legal process.

Conversely, however, one can argue that a LR that is too-high by a factor of 2 is not immaterial, that when all the evidence is accounted for in a juror's mind, the "doubling" of the LR could mean the difference between a vote for conviction and acquittal. Then too, is it acceptable for an examiner to present in court a LR that is known to be very likely higher—and not randomly higher or lower—than would be the case given perfect knowledge, while at the same time he is unable to provide a solid estimate of how much higher? It is difficult to conceive how the magnitude of a barrel's changes could be reliably known to an investigator or examiner, or how the magnitude of change could be reliably measured. A more sophisticated multiple-run model is no more removed from this difficulty.

Research to Date

But what about the CMS research that already has been conducted? Is it useful? An honest answer is that it is only marginally so. Not only does the concept appear to fall short in the ways just presented, but the research-to-date suffers from further weaknesses. First, as already observed, the existing research findings are directly relevant for only particular barrel manufacturing methods, barrel lengths, barrel hardnesses, bullet hardnesses, and bullet surface materials. With much additional research, some of these variables may turn out irrelevant, but good scientific practice demands that all relevant variables be accounted for, and only research will reveal which variables are irrelevant, or at least of lesser importance. So far there has been a paucity of published, empirical validity research since Biasotti's 1959 article, and thus for a case with differing circumstances, drawing conclusions from the limited existing data is unjustified.

The remaining faults are interpretational. One of these is an unfortunate tendency, or at least suggestion, to latch onto a fixed "decision number." That is, if in a certain study no more than five CMS were found for known non-matches, then it is tempting to conclude that, for the research conditions, any CMS-count above five constitutes an identification. This is incomplete and misleading. It is akin to a proclamation (fictional) in the 1970s by nuclear engineers that a serious nuclear reactor accident had never happened and therefore wouldn't happen. The mathematical, different-gun probability distribution must approach the CMS axis asymptotically. It will never reach zero. The CMS model is most properly termed a probability model, not an identification model, and the selection of an appropriate mathematical probability distribution is necessary before making inferences where no actual probability data exist.

A second interpretational fault is sometimes ignoring the significance of the same-gun histogram (or probability distribution). It may seem quite powerful to testify that the probability of finding, say, eight max-CMS on a given bullet, assuming it was fired from a different gun, is only 1 in 1000. But what if the probability of finding eight max-CMS on a given bullet, assuming it was fired from the suspect's gun (same-gun data), is also 1 in 1000? In a hypothetical case such as this, the CMS evidence is useless, in no sense relevant to guilt or innocence. The likelihood ratio would be 1. A same-gun probability of 1 in 500? This would constitute very weak evidence, with the LR = 2. To take a non-firearms example, imagine that a daughter accuses her father of molesting her during childhood, a charge based on recovered memory techniques. At the trial the therapist testifies that the daughter currently has dreams of her father molesting her as a child. He further insists that research shows that only 1 in 200 women (these are fictitious figures) who *were* never molested by their fathers have such dreams. But what he neglects to mention is that only 1 in 120 women who were molested by their fathers have such dreams. Thus the likelihood ratio is 1.67—very weak evidence in favor of childhood molestation.

Finally, invoking only the different-gun histogram (the denominator of the LR) invites the examiner, the judge, and the jury to reason illogically about the evidence. This occurs when one transposes the conditional, i.e., when the probability of the evidence assuming guilt is mistakenly thought to be, and presented as, the probability of guilt assuming the evidence (again treating guilt as tantamount to the evidence bullet being fired from the suspect barrel). The probability of a dorsal fin, given a shark, is not equal to the probability of a shark, given a dorsal fin. Trivial examples are obvious, but real world ones are not always so. Probability models for bullets have been theorized that suggest the probability of chance striation matches, given the bullets are from different barrels (11). One might estimate, for example, that the probability is one in 100,000 that a bullet from one barrel will accidently match a bullet from another barrel. In symbols

$$P(E \mid \text{not guilty}) = P(\text{striation match} \mid \text{different barrel}) = 1/100{,}000.$$

The error would occur if an examiner testified that, given the striation match, the probability that the evidence bullet originated from a non-suspect barrel is 1 in 100,000. In symbols, the examiner is asserting that $P(\text{different barrel} \mid \text{striation match}) = 1/100{,}000$. This is an easy trap to fall into, but clearly the 1/100,000 figure would belong in the denominator of a likelihood ratio.

Conclusion

It is arguably unfair to draw harsh conclusions about a CMS regime without subjecting its dominant rival—the traditional, subjective regime—to an equally critical examination. Nevertheless, and for now setting aside the practical difficulties, it appears that the inability of this probability model to deal rigorously with barrel changes is a weakness worthy of note, the seriousness of which is debatable (it's quite possible that further research and hard thinking could resolve the issue satisfactorily). Indeed, some questions do arise regarding the scientific status of present day subjective examinations; but with measures such as professional certification and rigorous validation/proficiency testing, the traditional, subjective examination regime can strengthen its scientific grounding. Whether CMS or objective-automated regimes eventually supplants it remains to be seen, and of course, research and logical analysis should continue, even accelerate. At least for the moment, however, the benefit of the doubt should go to the traditional methods.

Acknowledgments

I would like to thank Paul Schrecker and Douglas Murphy of the FBI Laboratory's Firearms-Toolmarks Unit for their general comments and suggestions. I also would like to thank Sandy Holliday of the FBI Laboratory's Special Projects Section for her assistance in the preparation of the figures. Any errors or omissions are solely the responsibility of the author.

References

1. Moore DS. Statistics: Concepts and controversies. 2nd ed. New York: W. H. Freeman & Co., 1985.
2. Robertson B, Vignaux GA. Interpreting evidence: Evaluating forensic science in the courtroom. Chichester: John Wiley & Sons, 1995.
3. Stoney DA. What made us ever think we could individualize using statistics? J Forensic Sci Soc 1991;31(2):197–9.
4. Aitken CGG. Statistics and the evaluation of evidence for forensic scientists. West Sussex: John Wiley & Sons, 1995, and Robertson and Vignaux, Interpreting evidence.
5. Biasotti AA. A statistical study of the individual characteristics of fired bullets. J Forensic Sci 1959;4:34–50.
6. Biasotti AA, Murdock J. Firearms and toolmark identification: The scientific basis of firearms and toolmark identification, Section 23–2.0. In: Faigman L, Kaye DH, Saks MJ, Sanders J, editors. Modern scientific evidence: The law and science of expert testimony. St. Paul: West Publishing, 1997;(2):144–150.
7. Miller J, McLean M. Criteria for identification of toolmarks. AFTE Journal 1998 Winter;30(1):15–43.
8. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A hierarchy of propositions: Deciding which level to address in casework. Science and Justice 1998;38(4):231–239; and Evett IW, Lambert JA, Buckleton JS, editors. A bayesian approach to interpreting footwear marks in forensic casework. Science and Justice 1998;38(4):241–7.
9. Jonakait RN. Forensic science: The need for regulation. Harvard Law Technol 1991;109(4):161–3.
10. Evett IW, Williams RL. A review of the sixteen points standard in England and Wales. J Forensic Ident 1996;46(1):49–73.
11. Heard BJ. Handbook of firearms and ballistics. Chichester: John Wiley & Sons, 1997;136–9.

Additional information and reprint requests:
Stephen G. Bunch
Firearms-Toolmarks Unit
FBI Laboratory
Washington, DC 20535