Forensic Science International: Synergy 2 (2020) 701–702

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

Commentary on: I. Dror, N Scurich "(Mis)use of scientific measurements in forensic science" Forensic Science International: Synergy 2020 https://doi.org/10.1016/j.fsisyn.2020.08.006

Keywords:
Error rates
Daubert
Forensic science
Inconclusive
Expert decision marking
Study design

Dear Editor

The article describes instances in which the authors believe inconclusive calls have been misused or ignored in determining estimated error rates. They propose two general study designs intended to determine when inconclusive calls should and should not be regarded as errors. In one proposal, this would be done *a priori* by a panel of expert examiners; in the other, it would be done *a posteriori* by the majority of calls on a given set by participating examiners. We find the proposals outlined in this paper to be objectionable and flawed, and believe that they could undermine efforts to clearly characterize the quality of forensic examination.

The difficulty that arises from the authors' logic comes from the apparent insistence that every call be regarded as either "correct" or an "error" (and the pejorative implication of the word "error" makes this even more difficult). For either true positives or true negatives, there are three possible outcomes: identification, inconclusive, and elimination. An insistence that these be somehow partitioned into "correct" and "error" calls is overly simplistic. In these examiner error rate studies, only one ground truth exists: the fact that a comparison is a true same source (i.e. know match) or true different source (i.e. known non-match). The authors argue that each set can be additionally characterized as having sufficient encoded information to support definitive conclusions (identification or elimination) or not (inconclusive). If one can arrive at this additional "truth" of sufficient encoded information, then any inconclusive call would be regarded as "error" in samples with sufficient information. Likewise, a definitive call on a specimen that lacks sufficient encoded information would also be an "error". However, ground truth regarding the sufficiency of encoded information is not available.

The authors propose two solutions for determining whether sufficient information is available. The first relies on the judgement of a panel of experts (*a priori*) and the second relies on the majority of participants' responses (*a posteriori*) as the arbiters of this new "ground truth". This is logically problematic for two reasons. First it substitutes the concept of reproducibility for the concept of accuracy. That is, it considers as "correct" (an accuracy concept) calls that represent agreement among multiple examiners (the essence of reproducibility). As an example of how this is flawed, consider a scenario where the panel of experts judge a known same source set to lack sufficient data, and thus under the authors' proposal, this set *should* be called inconclusive. Any definitive answer, including one that agrees with the ground truth of same source (i.e. identification), would be scored as an error. Upon return of test results, a majority of participants report "identifications" for this set. Under the authors' proposed paradigm, these test participants' answers would be erroneous, despite the majority being factually correct when considering the fundamental ground truth: these are same-source comparisons. The second problem these proposals would introduce is the circular logic of assessing examiner accuracy via a process that assumes examiner accuracy. To further illustrate this problem, consider actual data from a recently presented firearms examiner error rate study [1]. In this study, one test set has a ground truth of being a known same source comparison. Twenty-six (26) examiners provided conclusions on this set, with 10 identifications, 16 inconclusives, and 0 eliminations. Using the authors' proposal of using test returns to arbitrate the correctness of these answers is problematic. It is illogical to judge the 10 examiners who reached a conclusion of identification as erroneous when the ground truth comports with their conclusion. Additionally, the portion of 10 identifications vs 16 inconclusives does not provide a clear consensus.

We must deal with the fact that ground truth always has two categories but (at least) three distinct and meaningful calls can be made in each case. Insisting this can be artificially reduced to two categories in an effort to match ground truth and fit with traditional diagnostic testing error rate metrics is a dangerous mischaracterization of the forensic examination process. The word "error" suggests a dichotomy, which simply isn't sufficient to characterize accepted ranges of forensic pattern comparison conclusions. We propose alternative indices that would provide meaningful performance metrics.

As an example, consider a study with sets of both true same source and true different source comparisons. When analyzing the different source comparisons, participants would report the

DOIs of original article: https://doi.org/10.1016/j.fsisyn.2020.10.005, https://doi.org/10.1016/j.fsisyn.2020.08.006.

https://doi.org/10.1016/j.fsisyn.2020.10.004

number of identification calls (ID), elimination calls (E) and inconclusive calls (I). Then one might report:

- ID/(ID+E+I) = “incorrect identification rate”
- E/(ID+E+I) = “correct elimination rate” (i.e. specificity)
- I/(ID+E+I) = “different source inconclusive rate”

Conversely, when analyzing the true same source comparisons, participants would again report the numbers of calls made in each of the three categories. Then one could report:

- E/(ID+E+I) = “incorrect elimination rate”
- ID/(ID+E+I) = “correct identification rate” (i.e. sensitivity)
- I/(ID+E+I) = “same source inconclusive rate”

Traditional diagnostic error rate metrics appear to have caused confusion because “error” suggests a binary test result. This results in a semantics problem because “error rate” can’t be interpreted as 1-“correct rate”. The solution to this is to eliminate the semantics problem, understand that there are two ground truth states and three meaningful call categories, and report statistics that fully and honestly describe the situation. This cannot be accomplished when there is an insistence that everything be regarded as either “error” or “correct”.

Declaration of competing interest

No conflict of interest.

Reference

1 R. Lilien, T. Weller, Results of the 2018 3D Virtual Comparison Microscopy Error Rate Study (VCMERS) AFTE 2019 Seminar, May 30 2019.

Todd J. Weller, MS*
Weller Forensics, LLC, PO Box 106, Burlingame, CA, 94011, USA

Max D. Morris, PhD
Iowa St University Department of Statistics, 2438 Osborn Dr Ames, IA, 50011, USA
E-mail address: mmorris@iastate.edu.

* Corresponding author. Weller Forensics, LLC, PO Box 106, Burlingame, CA, 94011, USA.
E-mail address: toddweller@wellerforensics.com (T.J. Weller).

6 October 2020