Title: A response to Marquis et al (2017) What is the error margin of your signature analysis?

Author: Geoffrey Stewart Morrison

PII: S0379-0738(18)30104-X
DOI: https://doi.org/10.1016/j.forsciint.2018.03.009
Reference: FSI 9199

To appear in: *FSI*

Author: Kaye Ballentyne

PII: S0379-0738(18)30104-X
DOI: https://doi.org/10.1016/j.forsciint.2018.03.009
Reference: FSI 9199

To appear in: *FSI*

Author: Patrick Henry Geoghegan

PII: S0379-0738(18)30104-X
DOI: https://doi.org/10.1016/j.forsciint.2018.03.009
Reference: FSI 9199

To appear in: *FSI*

Received date: 1-3-2018

Please cite this article as: Patrick Henry Geoghegan, A response to Marquis et al (2017) What is the error margin of your signature analysis?, Forensic Science International https://doi.org/10.1016/j.forsciint.2018.03.009

A response to Marquis et al (2017) What is the error margin of your signature analysis?

Dr Geoffrey Stewart Morrison [*]

Associate Professor of Forensic Speech Science, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom

Director and Forensic Consultant, Forensic Evaluation Ltd, Birmingham, England, United Kingdom

Dr Kaye Ballentyne

Senior Research and Development Officer, Office of the Chief Forensic Scientist, Victoria Police Forensic Services Department, Melbourne, Victoria, Australia

Dr Patrick Henry Geoghegan

Lecturer in Biomedical Engineering, School of Life and Health Sciences, Aston University, Birmingham, England, United Kingdom

[*] Corresponding author. *E-mail address:* geoff-morrison@forensic-evaluation.net (G.S. Morrison).

Highlights

- A court requested an "error margin", but Marquis et al (2017) recommended:
- A method based on subjective judgement and not subjected to empirical validation.
- If a subjective-judgement-based method is used, we argue that it should be:
- Empirically calibrated.
- Empirically validated under conditions reflecting those of the case.

Abstract

Marquis et al (2017) [What is the error margin of your signature analysis? *Forensic Science International*, 281, e1–e8] ostensibly presents a model of how to respond to a request from a court to state an "error margin" for a conclusion from a forensic analysis. We interpret the court's request as an explicit request for meaningful empirical validation to be conducted and the results reported. Marquis et al (2017), however, recommends a method based entirely on subjective judgement and does not subject it to any empirical validation. We believe that much resistance to the adoption of the likelihood ratio framework is not to the idea of assessing the relative probabilities (or likelihoods) of the evidence under prosecution and defence hypotheses *per se*, but to what is perceived to be unwarranted subjective assignment of those probabilities. In order to maximize transparency, replicability, and resistance to cognitive bias, we recommend the use of methods based on relevant data, quantitative measurements, and statistical models. If the method is based on subjective judgement, the output should be empirically calibrated. Irrespective of the basis of the method, its implementation should be empirically validated under conditions reflecting those of the case at hand.

Keywords: Likelihood ratio; Subjective judgement; Empirical calibration; Empirical validation

Dear Editor:

A case report [1] recently published in *Forensic Science International* ostensibly presents a model of how to respond to a request from a court to:

> 'Please state, preferably in %, what is the degree of certainty of your conclusions [regarding whether a questioned signature was written by a particular known writer]. If, even under the best case scenario there remains an unavoidable error margin for this analysis according to the state of scientific and technical knowledge, please state what that error margin is (preferably in %).' ([1] p. e1)

We agree with the response in [1] that the appropriate way for a forensic practitioner to express their conclusions is as a numeric likelihood ratio value (which cannot be expressed as a percentage). We disagree, however, with the recommendations in [1] as to how a forensic practitioner should generate the likelihood ratio value that they report. The method described in [1] used "the knowledge and the experience of the examiners" (p. e3) to generate a value for the numerator of the likelihood ratio, and used their "knowledge, experience and training" (p. e4) to generate a value for the denominator. The likelihood ratio value reported was based entirely on the practitioners' uncalibrated and untested subjective judgement.

The comments we make in the present letter are with respect to the abovementioned method in general, irrespective of the branch of forensic science in which it is applied, and irrespective of legal jurisdiction. Nothing we write is intended to be specific to forensic examination of signatures in particular. Nor is anything we write intended to be specific to the Swiss judicial system in particular.

Although we prefer methods based on relevant data, quantitative measurements, and statistical models because they are transparent, replicable, and more resistant to cognitive bias, we do not object to the use of subjective judgement *per se*. If subjective judgement is used, however, procedures must be adopted to reduce the potential for cognitive bias, the likelihood ratio value generated should be empirically calibrated, and the implementation of the whole method must be empirically validated under conditions reflecting those of the case.[1] For supporting arguments see [3]–[11], for specific proposals on how to empirically calibrate human judgements see [12]–[16] (the procedure recommended in [17] can also be considered a procedure for generating empirically calibrated likelihood ratios [18]), and for descriptions of validation metrics for use within the likelihood ratio framework see [19] and [20].

Data used for training and calibrating an implementation of a method based on quantitative measurements and statistical models, or used for calibrating an implementation of a method based on

[1] By "implementation" of a method, we mean the method as used by the particular practitioner, see [2] §4.2.8–4.2.11, 4.3.2. By validation of the "whole" method we mean that the entire set of tools and procedures used to generate a likelihood ratio should be tested as a single unified system. Separately validating components of the system would not suffice, see [2] §3.3.1–3.3.2.

subjective judgement, must be sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions of the known and questioned specimens in the case that the likelihood ratio value generated is a reasonable answer to the question posed by the competing hypotheses specified for the case (e.g., the hypothesis that the questioned specimen came from the known source versus the hypothesis that the questioned specimen came from some other source selected at random from a specified population). Likewise, data used for empirically testing the performance of an implementation of a method, irrespective of whether is it is based on quantitative measurements and statistical models or on subjective judgement, must be sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions of the known and questioned specimens in the case that the test results provide a meaningful indication of how well the method works under the conditions of the case. If such data cannot be obtained, then meaningful empirical validation cannot be conducted, and the court lacks the information necessary for deciding on the extent to which it can trust a forensic practitioner's conclusion. Hence, with respect to admissibility, empirical demonstration of a sufficient level of performance under conditions pertinent to the case at hand is required by United States Federal Rule of Evidence 702[2] and the criteria set out in the *Daubert* trilogy of Supreme Court rulings,[3] and is also recommended by section 19A of the England & Wales Criminal Practice Directions[4] (see [3] and [21]). That obtaining case-relevant data for testing (and for training and calibration) may be financially costly and time consuming ([1] p. e3) does not absolve forensic practitioners from the requirement to conduct empirical validation under casework conditions. As President Obama's Council of Advisors on Science and Technology stated:

> neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. Similarly, an expert's expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For forensic feature-comparison methods, establishing foundational validity based on empirical evidence is thus a *sine qua non*. Nothing can substitute for it. ([17] p. 6, emphasis in original)

Our interpretation of the court's request for an "error margin" to be stated ([1] p. e1), is that this was an explicit request for meaningful empirical validation to be conducted and the results reported.[5] We

---

[2] Federal Rules of Evidence as amended Apr. 17, 2000, eff. Dec. 1, 2000; Apr. 26, 2011, eff. Dec. 1, 2011.

[3] Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993); General Electric Co. v. Joiner, 522 U.S. 136 (1997); Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

[4] Criminal Practice Directions [2015] EWCA Crim 1567.

[5] Although "error margin" literally appears to specify precision (reliability), we doubt that such a narrow interpretation

think it perverse to title a paper "What is the error margin of your signature analysis?" but to recommend the use of a subjective-judgement-based method that has not undergone any empirical validation. The discussion of the use of Bayes' theorem in [1] §3 is factually correct, but it appears to us to have been used to obfuscate with respect to the court's request for meaningful empirical validation.

As stated in [22], we believe that "Much resistance to the adoption of the likelihood ratio framework is not to the idea of assessing the relative probabilities (or likelihoods) of the evidence under prosecution and defence hypotheses *per se*, but to what is perceived as unwarranted subjective assignment of those probabilities" ([22] p. 472). We are advocates of the use of the likelihood ratio framework, but not of likelihood ratio values based on uncalibrated and untested subjective judgement. Likelihood ratio value should preferably be based on relevant data, quantitative measurements, and statistical models. If they are based on subjective judgement, they should be empirically calibrated and procedures should be adopted to reduce the potential for cognitive bias. Irrespective of the basis of the method, its implementation should be empirically validated under conditions reflecting those of the case at hand.

Sincerely

Authors

Affiliations

Declarations of interest: none

Acknowledgments

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

s

was intended by the court, and it would not make sense to test precision without also testing accuracy (validity). Note that in *Daubert* footnote 9 "evidentiary reliability" was equated with "scientific validity" (Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579, 1993).

References

[1] R. Marquis, L. Cadola, W.D. Mazzella, T. Hicks (2017). What is the error margin of your signature analysis? *Forensic Science International*, 281, e1–e8. https://doi.org/10.1016/j.forsciint.2017.11.012

[2] Forensic Science Regulator (2014). Guidance on validation (FSR-G-201 Issue 1). Forensic Science Regulator, Birmingham, UK. https://www.gov.uk/government/publications/forensic-science-providers-validation

[3] G.S. Morrison, W.C. Thompson (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18, 326–434.

[4] G.S. Morrison (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, 54, 245–256. http://dx.doi.org/10.1016/j.scijus.2013.07.004

[5] G.S. Morrison, R.D. Stoel. (2014). Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: How far have we come? *Australian Journal of Forensic Sciences*, 46, 282–292. http://dx.doi.org/10.1080/00450618.2013.833648

[6] J.J. Koehler (2017). Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences. *Arizona State Law Journal*, 49(4), 1369–1416.

[7] J.J. Koehler (2018). How trial judges should think about forensic science evidence. *Judicature*, 102(1).

[8] G. Edmond, B. Found, K.A. Martire, K. Ballantyne, D. Hamer, R. Searston, M. Thompson, E. Cunliffe, R. Kemp, M. San Roque, J. Tangen, R. Dioso-Villa, A. Ligertwood, D. Hibbert, D. White, G. Ribeiro, G. Porter, A. Towler, A. Roberts (2016). Model forensic science. *Australian Journal of Forensic Sciences*, 48, 496–537. http://dx.doi.org/10.1080/00450618.2015.1128969

[9] B. Found (2015). Deciphering the human condition: The rise of cognitive forensics. *Australian Journal of Forensic Sciences*, 47, 386–401. http://dx.doi.org/10.1080/00450618.2014.965204G

[10] R.D. Stoel, C.E.H. Berger, W. Kerkhoff, E.J.A.T. Mattijssen, E.I. Dror (2015). Minimizing contextual bias in forensic casework. In Strom K.J., Hickman M.J. (Eds.), *Forensic Science and the Administration of Justice: Critical Issues and Directions*, Thousand Oaks, CA: Sage. pp. 67–86. http://dx.doi.org/10.4135/9781483368740.n5

[11] W.C. Thompson (2009). Painting the target around the matching profile: The Texas sharpshooter fallacy in forensic DNA interpretation. *Law, Probability and Risk*, 8, 257–276. https://doi.org/10.1093/lpr/mgp013

[12] J. Lindh, G.S. Morrison (2011). Forensic voice comparison by humans and machine: Forensic voice comparison on a small database of Swedish voice recordings. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China* (pp. 1254–1257).

[13] D. Ramos, J. Franco-Pedroso, J. González-Rodríguez (2011). Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST Human Aided Speaker Recognition 2010. In *Proceedings of the International Conference on Speech Signal Processing, Prague, Czech Republic* (pp. 5908–5911). http://dx.doi.org/10.1109/ICASSP.2011.5947706

[14] G.S. Morrison (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45, 173–197. http://dx.doi.org/10.1080/00450618.2012.733025

[15] A. O'Hagan, C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, T. Rakow (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470033312

[16] K.A. Martire, B. Growns, D.J. Navarro (in press). What do the experts know? Calibration, precision, and the

wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review*.

[17] President's Council of Advisors on Science and Technology (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/

[18] G.S. Morrison, D.H. Kaye, D.J. Balding, D. Taylor, P. Dawid, C.G.G. Aitken, S. Gittelson, G. Zadora, B. Robertson, S.M. Willis, S. Pope, M. Neil, K.A. Martire, A. Hepler, R.D. Gill, A. Jamieson, J. de Zoete, R.B. Ostrum, A. Caliebe (2017). A comment on the PCAST report: Skip the "match"/"non-match" stage. *Forensic Science International*, 272, e7–e9. http://dx.doi.org/10.1016/j.forsciint.2016.10.018

[19] G.S. Morrison (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51, 91–98. http://dx.doi.org/10.1016/j.scijus.2011.03.002

[20] D. Meuwly, D. Ramos, R. Haraksim (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, *Forensic Science International*, 276, 142–153. http://dx.doi.org/10.1016/j.forsciint.2016.03.048

[21] G.S. Morrison (2018). Admissibility of forensic voice comparison testimony in England and Wales. *Criminal Law Review*, (1), 20–33.

[22] G.S. Morrison (2017). What should a forensic practitioner's likelihood ratio be? II. *Science & Justice*, 57, 472–476. http://dx.doi.org/10.1016/j.scijus.2017.08.004