THE DIALOGUE BETWEEN FORENSIC SCIENTISTS, STATISTICIANS AND LAWYERS ABOUT COMPLEX SCIENTIFIC ISSUES FOR COURT

EDITED BY: Sue Pope and Alex Biedermann

PUBLISHED IN: Frontiers in Genetics and Frontiers in Sociology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-049-0
DOI 10.3389/978-2-88966-049-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

THE DIALOGUE BETWEEN FORENSIC SCIENTISTS, STATISTICIANS AND LAWYERS ABOUT COMPLEX SCIENTIFIC ISSUES FOR COURT

Topic Editors:
Sue Pope, Principal Forensic Services, United Kingdom
Alex Biedermann, University of Lausanne, Switzerland

Citation: Pope, S., Biedermann, A., eds. (2020). The Dialogue Between Forensic Scientists, Statisticians and Lawyers about Complex Scientific Issues for Court. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-049-0

Table of Contents

frontiers in Genetics

EDITORIAL
published: 07 August 2020
doi: 10.3389/fgene.2020.00704

Editorial: The Dialogue Between Forensic Scientists, Statisticians and Lawyers About Complex Scientific Issues for Court

Sue Pope[1*†] *and Alex Biedermann*[2†]

[1] *Principal Forensic Services, Bromley, Kent, United Kingdom,* [2] *Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, Lausanne, Switzerland*

Keywords: criminal jurisprudence, expert evidence, DNA likelihood ratios, DNA evidence, principles of forensic interpretation

Editorial on the Research Topic

The Dialogue Between Forensic Scientists, Statisticians and Lawyers About Complex Scientific Issues for Court

OPEN ACCESS

Edited and reviewed by:
Andrew J. Mungall, Canada's Michael Smith Genome Sciences Centre, Canada

***Correspondence:**
Sue Pope
suepope
@principalforensicservices.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *12 February 2020*
Accepted: *10 June 2020*
Published: *07 August 2020*

Citation:
Pope S and Biedermann A (2020) Editorial: The Dialogue Between Forensic Scientists, Statisticians and Lawyers About Complex Scientific Issues for Court. Front. Genet. 11:704. doi: 10.3389/fgene.2020.00704

Courts across jurisdictions have seen a massive "scientification" of their evidential proceedings, fueled by permanent technological advances, in particular with the advent of modern DNA profiling analyses since the mid-1980s. Never before, in the history of forensic science, could analyses be extended to such small quantities of trace material, and never before have forensic experts had more powerful computational and data analytic devices at their disposal for handling the vast array of data that their analyses produce. At the same time, conceptual questions on how to assess the probative value of scientific findings have largely been settled: there is now a broad agreement that evaluating scientific evidence should adhere to the precepts of logic, balance, transparency, and robustness (e.g., Jackson, 2000; Association of Forensic Science Providers, 2009). But as much as there have been advances, modern scientific evidence has been, and is still, accompanied by challenges and contestation. What was once called the "DNA-wars" (Thompson, 1993) in the early 1990s, has developed during the last decade into refined discourses about selected aspects of scientific evidence, such as algorithmic transparency. While some of these debates are confined almost exclusively to scientific circles, they are also brought to the open by meticulous legal discussants, who care about the foundations of evidence and its ability to help discriminate between prosecution and defense views (e.g., Imwinkelried, 2017). What is more, paradoxically, much of the specialized discussion around these topics is confined to scientific journals whose deterring paywalls prevent vital information from being distributed among those practitioners—especially defense lawyers—for whom access to such information would be most beneficial. The purpose of this *Frontiers* Research Topic thus is twofold. On the one hand, the aim is to bring together a broad range of authors from various forensic science and legal disciplines (both academic and practice oriented) to elaborate on key topics that sit at the intersection between (forensic) science and the law. On the other hand, the purpose is to serve the scientific and legal community by providing this collection of contributions freely and fully accessible (open access, OA), a goal that is achieved through the *Frontiers* OA publishing model[1].

[1] This is the second *Frontiers* research topic on forensic science after "DNA, statistics and the law: a cross-disciplinary approach to forensic inference" (https://www.frontiersin.org/research-topics/1325).

This collection of papers focuses on so-called evaluative uses of evidence, in particular DNA evidence. That is, situations in which a potential source (i.e., reference material of known origin) for a given trace is available and the value of the results of the comparison between the trace and the reference needs to be assessed with respect to competing propositions regarding the source of the evidential material, or propositions regarding alleged activities (ENFSI, 2015; Gill et al., 2018). This is to be distinguished from so-called investigative uses of evidence, which are situations in which *no* potential source for recovered trace material is available. See, for example, Butler and Willis (2020) for a recent review on this topic, in particular investigative DNA genealogy as used, for example, in the "Golden State Killer" case. Developments in the latter field heavily rely upon large datasets generated by the expanding direct-to-customer genomic industry (e.g., Phillips, 2018).

Several papers in this collection address selected issues that affect the sound use of DNA profiling analyses in evaluative settings. Taylor et al. discuss matters that arise in connection with the use of modern computer software for biostatistical and the value of evidence computations, especially concerns raised by legal commentators. In turn, Roberts addresses general aspects of expert testimony, followed by a discussion of these aspects in the context of the use of low-template DNA profiling results by English and Northern Irish courts. Biedermann et al. and Biedermann and Hicks focus on recurrent misconceptions in the assessment of DNA profiling results, in particular the distinction between issues of source and alleged activities, and the importance of drawing this distinction carefully by acknowledging the circumstances of the case and the specific accounts provided by the prosecution and defense. The importance of these topics has recently been reiterated by guidelines published by the DNA Commission of the International Society for Forensic Genetics (Gill et al., 2018, 2020). Scientific evidence other than DNA is discussed in the legal commentaries by Caruso and Symes and Kotsoglou.

Aitken and Aitken et al. focus on statistical methodologies and concepts, in particular the likelihood ratio, which is now widely recognized as providing the most suitable framework for assessing the value of scientific evidence in a way that is logical, balanced, transparent, and robust. Both these articles address and rebut critiques (e.g., Lund and Iyer, 2017) that have recently been leveled against the likelihood ratio.

Finally, Taroni et al. discuss a case example that they consider demonstrates the gap that still exists between what academics consider sound evaluative procedures and what scientists in the field actually practice and convey to recipients of expert information. Burnier offers additional considerations regarding the same case.

AUTHOR CONTRIBUTIONS

Both authors have made equal contributions to the work and approved it for publication.

FUNDING

AB gratefully acknowledges the support of the Swiss National Science Foundation through Grant No. BSSGI0_155809.

REFERENCES

AFSP (Association of Forensic Science Providers) (2009). Standards for the formulation of evaluative forensic science expert opinion. *Sci. Justice* 49, 161–164. doi: 10.1016/j.scijus.2009.07.004

Butler, J., and Willis, S. (2020). Interpol review of forensic biology and forensic DNA typing 2016–2019. *Forensic Sci. Int. Synergy*. doi: 10.1016/j.fsisyn.2019.12.002

ENFSI (2015). *Guideline for Evaluative Reporting in Forensic Science*. Dublin. Available online at: http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

Gill, P., Hicks, T., Butler, J. M., Connolly, E., Gusmão, L., Kokshoorn, B., et al. (2018). DNA commission of the International society for forensic genetics: assessing the value of forensic biological evidence – Guidelines highlighting the importance of propositions Part I: evaluation of DNA profiling comparisons given (sub-) source propositions. *Forensic Sci. Int. Genet.* 36, 189–202. doi: 10.1016/j.fsigen.2018.07.003

Gill, P., Hicks, T., Butler, J. M., Connolly, E., Gusmão, L., Kokshoorn, B., et al. (2020). DNA commission of the International society for forensic genetics: assessing the value of forensic biological evidence – Guidelines highlighting the importance of propositions. Part II: Evaluation of biological traces considering activity level propositions. *Forensic Sci. Int. Genet.* 44:102186. doi: 10.1016/j.fsigen.2019.102186

Imwinkelried, E. J. (2017). Computer source code: a source of the growing controversy over the reliability of automated forensic techniques. *DePaul Law Rev.* 66, 97–132. Available online at: https://via.library.depaul.edu/law-review/vol66/iss1/6

Jackson, G. (2000). The scientist and the scales of justice. *Sci. Justice* 40, 81–85. doi: 10.1016/S1355-0306(00)71947-2

Lund, S., and Iyer, H. (2017). Likelihood ratio as weight of forensic evidence: a closer look. *J. Res. Natl. Inst. Stand. Technol.* 122:27. doi: 10.6028/jres.122.027

Phillips, C. (2018). The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Sci. Int. Genet.* 36, 186–188. doi: 10.1016/j.fsigen.2018.07.010

Thompson, W. C. (1993). Evaluating the admissibility of new genetic identification tests: lessons from the DNA war. *J. Criminal Law Criminol.* 84, 22–104.

Conflict of Interest: SP is a Director at the company Principal Forensic Services.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

OPINION
published: 08 February 2016
doi: 10.3389/fgene.2016.00008

The Importance of Critically Examining the Level of Propositions When Evaluating Forensic DNA Results

Alex Biedermann and Tacha Hicks*

Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, Lausanne, Switzerland

Keywords: level of propositions, evaluative reporting, European guidelines

INTRODUCTION

The proposal of a discussion about the use of software to help assign likelihood ratios for forensic DNA profiling results, and the use of their output in the legal process, is both timely and important (see also related contributions elsewhere in *Frontiers*, e.g., Biedermann et al., 2014). Ever since their introduction in forensic science, DNA profiling analyses have been accompanied with the results of calculations of various sorts. Their scope is well illustrated and documented in several reference monographs (e.g., Evett and Weir, 1998; Buckleton et al., 2005; Balding and Steele, 2015). This solid body of scholarly research and established practice has contributed to the widely held view among scientists and recipients of expert information that eliciting the probative strength of forensic DNA profiling results is *per se* a numerical task.

In this commentary, we intend—in a first part—to make the point that although calculations are, by virtue, an integral part of the quantification of probative strength, it is equally important at the outset to be clear about the question "Why are we doing a calculation?" (Buckleton et al., 2005, p. 151). We will argue that this is not a question that statistics can answer. Stated otherwise, we will contend that, as much it is important to be clear in any instance about what a particular computation exactly purports to do, it is essential to define the questions that are of interest in a particular case at hand. In a second part, we will emphasize on the extent to which, why and how recently issued guidelines (e.g., ENFSI, 2015) encourage such thinking about cases prior to conducting calculations, if any.

QUESTIONING DEFAULT CALCULATIONS

Experience demonstrates that many scientists working in operational laboratories decide on the use of particular computational procedures—often provided by ready-to-use software packages—based on the mere availability of those procedures at their workplace. This amounts to a convenience choice, but what is more is that proceeding in this way is considered the best one could do. This view may be reinforced if the software is based on Bayesian principles, because procedures that belong to this class of inferential methods are referred to as the most inferentially sound. But the sole fact that a procedure relies on Bayesian principles does not make it *per se* pertinent for the case at hand. As noted by Lindley (2004, p. 74), "[t]he main danger is that they [Bayesian methods; *added by the authors*] will be used automatically. (. . .) You must think about the real quantities involved, like temperature or blood pressure, and not about symbols that represent them. This distinction

OPEN ACCESS

Edited by:
Sue Pope,
Principal Forensic Services, UK

Reviewed by:
Antonio Amorim,
Institute of Molecular Pathology and Immunology of the University of Porto, Portugal

****Correspondence:***
Alex Biedermann
alex.biedermann@unil.ch

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *26 November 2015*
Accepted: *20 January 2016*
Published: *08 February 2016*

Citation:
Biedermann A and Hicks T (2016) The Importance of Critically Examining the Level of Propositions When Evaluating Forensic DNA Results. Front. Genet. 7:8. doi: 10.3389/fgene.2016.00008

between the thinking you and the unthinking, calculating personal computer is essential." This danger also exists in the context of interpreting and reporting forensic DNA results. Indeed, most of the commonly available computational procedures[1] lead to expressions of probative strength to help discriminate between so-called sub-source level propositions [e.g., "the person of interest (POI) is the source of the recovered DNA" vs. "an unknown person is the source of the recovered DNA"]. But, in many practical cases, the real question goes beyond this level, e.g., how the detected DNA got where it was found (Evett et al., 2002; Taroni et al., 2013), that is so-called activity level propositions. Cases of alleged rape where the competing versions only differ with respect to the activities that led to the trace illustrate this. This is of course not a critique of models being Bayesian in nature, but of the kind of questions to which some of these models are tailored.

Skeptics may invoke that none of the above problems are novel. But why then practice by and large remains unchanged? While some scientists openly acknowledge that expressions of probative strength of DNA considering sub-source level propositions may indeed be insufficient for the needs, some hold that it is for the Court to decide on that matter. We perfectly agree with this stance, of course, because whatever the level of the propositions, it is for the Court to decide on the probability of the propositions. Notwithstanding, scientists can add considerable value by assessing their results *given* activity level propositions.

Yet others contend that one can leave this debate until the Courtroom. However, this may raise issues from a quality management point of view, and render the situation very uncomfortable for the witness, because of the inevitable difficulty of the task. The challenge is real for a variety of case scenarios, in particular where only low quantities of DNA are detected and/or when POIs do not deny that the recovered DNA is theirs. We seriously doubt that members of the judiciary are able to properly appreciate the extent to which one can expect to obtain a low quantity of DNA, recovered at a certain position on the crime scene, the victim, or a POI, given one activity as compared to another activity. We would not recommend either doing this evaluation on the stand. This is because such assessments are very challenging even for experts, and require *scientific knowledge* about many factors, such as transfer, persistence, and the capacity of a given donor to shed detectable quantities of DNA[2]. Let us emphasize again that the question of whether the detected DNA is that of the POI may be entirely uncontested (and thus there would be no need for a likelihood ratio given sub-source propositions as there is no uncertainty about sub-source). What is really of interest is to assess the probability of observing such a result for a DNA trace, that is a trace found in a particular position, in a given quantity and leading to a profile of the observed quality given the alleged activities and given relevant information such as the time lapse between collection of trace material and the commission of the crime, environmental factors to which the trace was exposed (e.g., temperature, humidity) etc. Such assessments are highly case dependent, which calls for the generation of more research with experiments under controlled conditions, that can help build a community-wide knowledge base (Evett, 2015)[3]. To further emphasize the need for considering observations given activity level propositions, note again that the result which is to be assessed is *not only* the rarity of genetic features, but also extends to the very fact of finding, at a given position, a detectable quantity of DNA (Evett and Weir, 1998), which may be nil. Sub-source level propositions cannot deal with results that did not yield a DNA profile.

The mismatch between default evaluations given sub-source level propositions and the decision makers' interest in activity level propositions is a cause of concern because the strength of the observations in the former case can be radically different from that of the latter, so that inappropriate conclusions can result if the two are taken to be equivalent. We have seen this happen in cases where scientists report likelihood ratios in the order of $>10^{20}$ with propositions at sub-source level when in fact the real issue was one of activities and where the strength of the findings, given the conditioning information of the case at hand, was way more moderate[4].

CURRENT RECOMMENDATIONS

The above discussion is not intended to suggest that evaluation given (sub)-source level propositions is useless or detrimental in principle[5]. The point we seek to make is that it is crucial to assess the needs of the recipient of expert information prior to choosing a computational procedure. This seems like an obvious and moderate requirement, yet experience shows that often it is given little attention in practice. Recent works by forensic scientists from across Europe, published in the form of a guideline (ENFSI, 2015), seek both to strengthen awareness of this issue and help scientists and recipients of expert information proceed in a more sensible way. For example, in its Guidance Note 2 on propositions, the document specifies: "Source level propositions are adequate in cases where there is no risk that the court will misinterpret them in the context of the alleged activities in the case" (ENFSI, 2015, p. 12). To illustrate this idea, the following example is given: "A large fresh bloodstain is recovered at the point of entry at a burglary scene and delivered to the laboratory for DNA analysis. Combination of a presumptive test and appearance allows the scientist to safely assume that the stain is blood. A suspect says that he has never been in the premises. The set of propositions can be (1) the bloodstain came from the defendant and (2) the bloodstain came from another unknown individual" (ENFSI, 2015, p. 12). In this example, source level

[1]The scope of these procedures is large and includes topics such as complex modeling of products (e.g., stutters) of the PCR amplification of STRs (e.g., Bright et al., 2014; Gittelson et al., 2014) and the study of the sensitivity of expressions of probative value due to the use of particular statistical techniques (e.g., MCMC techniques, see for example Bright et al., 2015).

[2]On this topic, see for example http://www.telegraph.co.uk/news/science/9115916/The-case-against-DNA.html.

[3]For an example in other transfer traces, see Buckleton et al. (1989).

[4]Another issue, not pursued here, is whether likelihood ratios exceeding one over the earth population, and multiples of that, are reasonable. There is much argument to say they are not (e.g., Thompson et al., 2003; Hopwood et al., 2012).

[5]In fact, the strength of the DNA correspondence is so high that this will lead to situations where the source of the DNA will be admitted (leaving no uncertainty on the source of the DNA). This, then moves the issue to the activities.

propositions are *not* problematic because no expert knowledge is required regarding phenomena such as transfer and persistence, as well as background levels of DNA. Such factors do not impact, in this kind of circumstances, on the understanding of scientific findings relative to the alleged activities. In particular, it is not doubted that the bloodstain results from the act of breaking in. This example also illustrates that there is more to the collected trace than the DNA profile: there are aspects such as the freshness of the stain, the quantity of material and the position where the trace was found. In turn, it is clear that specialized knowledge regarding transfer, persistence and background *would matter* in the above scenario *if* DNA had been detected in low quantities, rather than from a rich bloodstain.

The above understanding has far reaching implications: the level of propositions depends on the factors and observations on which forensic scientists have expert knowledge. It is their duty to evaluate all their results so that the Court is not deprived of information that is necessary for a balanced view. For example, the ENFSI guideline explicitly advises against the changing of propositions from activity to (sub-) source level when relevant expert knowledge is not available: "In fact, the choice between (sub-) source and activity should not be influenced by the availability of data or expert knowledge but solely from the consideration of factors such as transfer, persistence, and background levels that could crucially affect the strength of the findings within the context of the case circumstances." (ENFSI, 2015, p. 13).

We acknowledge, from personal experience, that the implementation of the above perspective is challenging. It may be even more so in systems exposed to commercialisation where forensic providers that conduct DNA profiling analyses operate more and more separated from those entities that collected trace material at the crime scene (Jackson, 2013). Further obstacles may be operational constraints such as time and costs, because evaluation given activity level propositions does not rely on default computations, but generally requires a case-based approach. Regarding the latter point, some scientists deplore a lack of formulaic developments for evaluation given propositions at higher hierarchical levels. But this critique does fall short of the current state of developments. Formal likelihood ratio approaches exist (e.g., Evett, 1984; Evett and Weir, 1998), used also for other transfer materials (e.g., glass; Curran et al., 2000), and there are reports that demonstrate the relevance and practical feasibility (e.g., McKenna, 2013). Yet, other developments allow one to account for uncertainty about the relevance of the recovered material and the possibility that material was left for innocent reasons (e.g., Evett, 1993; Evett et al., 2002).

The role of statistics in evaluating DNA profiling evidence has always been important, but we now must realize that, increasingly often, the traditional perspective of sub-source level propositions, and the main focus on the rarity of the corresponding features (i.e., the so-called conditional genotype probability), may represent only a first step of the evaluative process. This does not make these evaluation approaches wrong, only less comprehensive. The fact is that the extrinsic characteristics of the trace material (i.e., low quantities of DNA) and the propositions of interest have changed, and it is important to realize that this represents the relevant starting point. This recognition of the needs cannot be answered by statistics, only the evaluative procedures that need to be built once the needs are properly elicited. The importance of statistics in this endeavor remains unaffected, and stands as noted by Lindley (2000, p. 38): "(. . .) the first task of a statistician is to develop a (probability) model to embrace the clients' interests' and uncertainties. It will include the data and any parameters that are judged necessary. Once accomplished, the mechanics of the calculus take over and the required inference is made."

AUTHOR CONTRIBUTIONS

Both authors have made substantial, direct, and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

AB gratefully acknowledges the support of the Swiss National Science Foundation through grant No. BSSGI0_155809 and the University of Lausanne.

REFERENCES

Balding, D. J., and Steele, C. D. (2015). *Weight-of-Evidence for Forensic DNA Profiles, 2nd Edn.* Chichester: John Wiley & Sons.

Biedermann, A., Vuille, J., and Taroni, F. (2014). DNA, statistics and the law: a cross-disciplinary approach to forensic inference. *Front. Genet.* 5:136. doi: 10.3389/fgene.2014.00136

Bright, J.-A., Buckleton, J. S., Taylor, D., Fernando, M. A., and Curran, J. M. (2014). Modeling forward stutter: toward increased objectivity in forensic DNA interpretation. *Electrophoresis* 35, 3152–3157. doi: 10.1002/elps.201400044

Bright, J.-A., Stevenson, K. E., Curran, J. M., and Buckleton, J. S. (2015). The variability in likelihood ratios due to different mechanisms. *Forensic Sci. Int. Genet.* 14, 187–190. doi: 10.1016/j.fsigen.2014.10.013

Buckleton, J. S., Pinchin, R., and Evett, I. W. (1989). *Computerised Assistance for Glass Evidence (CAGE): Experimental Knowledge Based System for Assisting in the Interpretation of Forensic Glass Examinations.* Technical Report, Forensic Science Service.

Buckleton, J. S., Triggs, C. M., and Walsh, S. J. (2005). *Forensic DNA Evidence Interpretation.* Boca Raton, FL: CRC Press.

Curran, J., Hicks, T. N., and Buckleton, J. S. (2000). *Forensic Interpretation of Glass Evidence.* Boca Raton, FL: CRC Press.

ENFSI (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science.* Available online at: https://www.unil.ch/esc/files/live/sites/esc/files/Fichiers%202015/ENFSI%20Guideline%20Evaluative%20Reporting

Evett, I. W. (1984). A quantitative theory for interpreting transfer evidence in criminal cases. *Appl. Stat. (Ber)* 33, 25–32. doi: 10.2307/2347659

Evett, I. W. (1993). Establishing the evidential value of a small quantity of material found at a crime scene. *J. Forensic Sci. Soc.* 33, 83–86. doi: 10.1016/S0015-7368(93)72985-0

Evett, I. W. (2015). The logical foundations of forensic science: towards reliable knowledge. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370, 1–10. doi: 10.1098/rstb.2014.0263

Evett, I. W., Gill, P. D., Jackson, G., Whitaker, J., and Champod, C. (2002). Interpreting small quantities of DNA: the hierarchy of propositions and the

use of Bayesian networks. *J. Forensic Sci.* 47, 520–530. doi: 10.1520/JFS15291J

Evett, I. W., and Weir, B. S. (1998). *Interpreting DNA Evidence*. Sunderland, MA: Sinauer Associates Inc.

Gittelson, S., Biedermann, A., Bozza, S., and Taroni, F. (2014). Decision analysis for the genotype designation in *low-template-DNA* profiles. *Forensic Sci. Int. Genet.* 9, 118–133. doi: 10.1016/j.fsigen.2013.11.005

Hopwood, A. J., Puch-Solis, R., Tucker, V. C., Curran, J. M., Skerrett, J., Pope, S., et al. (2012). Consideration of the probative value of single donor 15-plex STR profiles in UK populations and its presentation in UK courts. *Sci. Justice* 52, 185–190. doi: 10.1016/j.scijus.2012.05.005

Jackson, G. (2013). "The impact of commercialization on the evaluation of DNA evidence," in *DNA, Statistics and the Law: A Cross-Disciplinary Approach to Forensic Inference*, Vol. 4, eds A. Biedermann, J. Vuille, F. Taroni (Frontiers Media SA), 16–18. Available online at: http://www.frontiersin.org/books/DNA_statistics_and_the_law_a_cross-disciplinary_approach_to_forensic_inference/284

Lindley, D. V. (2000). The philosophy of statistics. *Statistician* 49, 293–337. doi: 10.1111/1467-9884.00238

Lindley, D. V. (2004). Bayesian thoughts. *Significance* 1, 73–75. doi: 10.1111/j.1740-9713.2004.027.x

McKenna, L. (2013). "Understanding DNA results within the case context: importance of the alternative proposition," in *DNA, Statistics and the Law: A Cross-Disciplinary Approach to Forensic Inference*, Vol. 4, eds A. Biedermann, J. Vuille, and F. Taroni (Frontiers Media SA), 16–18. Available online at: http://www.frontiersin.org/books/DNA_statistics_and_the_law_a_cross-disciplinary_approach_to_forensic_inference/284

Taroni, F., Biedermann, A., Vuille, J., and Morling, N. (2013). Whose DNA is this? How relevant a question? (a note for forensic scientists). *Forensic Sci. Int. Genet.* 7, 467–470. doi: 10.1016/j.fsigen.2013.03.012

Thompson, W. C., Taroni, F., and Aitken, C. G. G. (2003). How the probability of a false positive affects the value of DNA evidence. *J. Forensic Sci.* 48, 47–54. doi: 10.1520/JFS2001171

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 October 2016
doi: 10.3389/fgene.2016.00180

LTDNA Evidence on Trial

Paul Roberts [1, 2, 3]*

[1] School of Law, University of Nottingham, Nottingham, UK, [2] Collaborative Innovation Center of Judicial Civilization, Institute of Evidence Law and Forensic Science, China University of Political Science and Law, Beijing, China, [3] Faculty of Law, University of New South Wales, Sydney, NSW, Australia

Adopting the interpretative/hermeneutical method typical of much legal scholarship, this article considers two sets of issues pertaining to LTDNA profiles as evidence in criminal proceedings. The section titled Expert Evidence as Forensic Epistemic Warrant addresses some rather large questions about the epistemic status and probative value of expert testimony in general. It sketches a theoretical model of expert evidence, highlighting five essential criteria: (1) expert competence; (2) disciplinary domain; (3) methodological validity; (4) materiality; and (5) legal admissibility. This generic model of expert authority, highlighting law's fundamentally normative character, applies to all modern forms of criminal adjudication, across Europe and farther afield. The section titled LTDNA Evidence in UK Criminal Trials then examines English and Northern Irish courts' attempts to get to grips with LTDNA evidence in recent cases. Better appreciating the ways in which UK courts have addressed the challenges of LTDNA evidence may offer some insights into parallel developments in other legal systems. Appellate court rulings follow a predictable judicial logic, which might usefully be studied and reflected upon by any forensic scientist or statistician seeking to operate effectively in criminal proceedings. Whilst each legal jurisdiction has its own unique blend of jurisprudence, institutions, cultures and historical traditions, there is considerable scope for comparative analysis and cross-jurisdictional borrowing and instruction. In the spirit of promoting more nuanced and sophisticated international interdisciplinary dialogue, this article examines UK judicial approaches to LTDNA evidence and begins to elucidate their underlying institutional logic. Legal argument and broader policy debates are not confined to considerations of scientific validity, contamination risks and evidential integrity, or associated judgments of legal admissibility or exclusion. They also crucially concern the manner in which LTDNA profiling results are presented and explained to factfinders in criminal trials.

Keywords: expert evidence, criminal adjudication, LTDNA profiling evidence, comparative criminal procedure, Law-Science, interdisciplinarity

OPEN ACCESS

Edited by:
Sue Pope,
Principal Forensic Services, UK

Reviewed by:
Aurélie Mahalatchimy,
University of Sussex, UK
Bettina Bock Von Wülfingen,
Humboldt University of Berlin, Germany

***Correspondence:**
Paul Roberts
paul.roberts@nottingham.ac.uk

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *17 May 2016*
Accepted: *26 September 2016*
Published: *25 October 2016*

Citation:
Roberts P (2016) LTDNA Evidence on Trial. Front. Genet. 7:180. doi: 10.3389/fgene.2016.00180

INTRODUCTION: AN INTERDISCIPLINARY TOPIC OF CONVERSATION

As originally conceived, this Research Topic focused on "the interface between forensic scientists and statisticians when calculating likelihood ratios for low template and complex DNA results"[1]. The problem of calculating and interpreting likelihood ratios was thereby implicitly characterized as a bilateral conversation between statisticians and forensic scientists. Of course, on further reflection, there is self-evidently a third major conversation partner in this discussion, namely courts and legal professionals. We might add that criminal "courts" in jurisdictions such as England and Wales comprise a mixture of professional judges and lay fact-finders, though jurists and jurors alike are laypeople when it comes to DNA profiling science. Our Topic Editors observed that "[t]here is a danger for courts if [likelihood ratios] are produced by a black box where the reporting forensic scientist has little input into and less understanding" of statistical methods. The clear (and unsurprising) presupposition is that difficulties associated with Low Template DNA (LTDNA) evidence cannot simply be conceptualized as "technical" questions to be resolved between specialists according to mutually satisfactory methodological criteria. There remains the challenge of communicating the meaning and significance of technical fixes to lay audiences in criminal adjudication. This communicative dimension is a general feature of expert testimony, whether or not it concerns anything properly categorized as "science"[2].

This article considers two sets of issues pertaining to LTDNA profiles as evidence in criminal proceedings, which are pertinent to all modern legal systems in which this type of evidence is currently or might in future be adduced. The section titled Expert Evidence as Forensic Epistemic Warrant addresses some rather large questions about the epistemic status and probative value of expert testimony in general. The following section on LTDNA Evidence in UK Criminal Trials then examines English courts' attempts to get to grips with LTDNA evidence in recent cases. These efforts might or might not appear impressive to outsiders, but they do, generally speaking, follow a predictable judicial logic—a logic which might usefully be studied and reflected upon by any forensic scientist or statistician seeking to operate effectively in this system of justice. In the spirit of promoting more nuanced and sophisticated interdisciplinary dialogue this article examines judicial approaches to LTDNA evidence and begins to elucidate their underlying institutional logic.

The following discussion adopts the interpretative or hermeneutical method typical of much legal scholarship. It engages with primary and secondary institutional materials—prominently featuring reported criminal appeals in England and Wales—in an attempt both to understand judicial practice and to contribute to the best normative (re)interpretation of legal doctrine and institutions. It might fairly be conceptualized, in methodological terms, as an elucidation of the internal logics of legal argumentation and judicial reasoning, which are often opaque to non-lawyers, even to those such as forensic scientists who regularly participate in criminal investigations and are no strangers to courtrooms. In striking contrast to scientific knowledge, law is delimited by jurisdiction. The second half of this article discusses appellate decisions drawn from two specific, common law jurisdictions: Northern Ireland (applying Northern Ireland law), and England and Wales (applying English law). Whilst each legal jurisdiction has its own unique blend of jurisprudence, institutions, cultures, and historical traditions, there is considerable scope for comparative analysis and cross-jurisdictional borrowing and instruction, the more so in a shrinking world characterized by globalization and cosmopolitan legality. Better appreciating the ways in which UK courts have addressed the challenges of LTDNA evidence may offer some insights into parallel developments in other legal systems. Moreover, the practical challenges posed by forensic science and other expert witness testimony cut across conventional Comparative Law distinctions between "adversarial" and "inquisitorial" procedures or "common law" and "civilian" legal systems. In setting the scene for more detailed doctrinal analysis, the first part of this article presents a generic model of expert authority highlighting law's fundamentally normative mission which applies to all modern forms of criminal adjudication, across Europe and farther afield.

EXPERT EVIDENCE AS FORENSIC EPISTEMIC WARRANT

Criminal trials are practical exercises in reasoning under uncertainty. We want to know what happened; but material facts are contested (otherwise the accused would have pleaded guilty in common law systems). Relevant evidence rationally authorizes or "warrants" particular inferential conclusions. The more probative value evidence has, the more warrant it provides for the conclusion. In traditional common law thinking, the best evidence is the oral testimony of a percipient witness, given on oath, and tested through cross-examination. This category of evidence is regarded as providing the best epistemic warrant for the inferential conclusions supported by the witness's testimony. This doesn't mean to say, of course, that every witness in court is truthful, accurate and reliable. We know, for example, that there can be many kinds of difficulty with eyewitness testimony[3]. But it does not follow, as a general proposition, that we should therefore prefer the testimony of those who did *not* see the incident to testimony from witnesses who did. "Best" does not mean infallible. Evidence adduced in criminal trials is often contested

[1]This was the title of the Research Topic originally announced on the Frontiers website, and in my invitation to contribute to it, and was still the current version when I first submitted this paper (11 March 2016). Now see Sue Pope and Alex Biedermann, "Research Topic: The Dialogue Between Forensic Scientists, Statisticians and Lawyers about Complex Scientific Issues for Court," Frontiers, http://journal.frontiersin.org/researchtopic/4000/the-dialogue-between-forensic-scientists-statisticians-and-lawyers-about-complex-scientific-issues-f (accessed 17 May 2016).

[2]Generally, see Paul Roberts (ed.), *Expert Evidence and Scientific Proof in Criminal Trials* (Ashgate, 2014).

[3]See e.g., Richard A. Wise, Clifford S. Fishman and Martin A. Safer, "How to Analyze the Accuracy of Eyewitness Testimony in a Criminal Case" (2009) 42 *Connecticut Law Review* 435.

or contradictory, and the factfinder must make the best of it, resolving any enduring doubts in accordance with the applicable burden and standard of proof. In criminal litigation, most (not all) doubts are resolved in favor of the accused, in accordance with the presumption of innocence[4].

Expert evidence supplies inferential warrant through the argument from authority. The expert says to the court, *trust me, I'm an expert*. The authority paradigm underpinning the inferential logic of forensic expert testimony has five major components: (1) the expert is a genuine expert (competence); (2) in a field in which expertise can be obtained (domain); (3) and has correctly and conscientiously applied authentic domain-specific protocols to produce proffered evidence (methodological validity); (4) in relation to a legally relevant issue (materiality); (5) and in a form that is likely to provide legitimate epistemic warrant for legal adjudication (admissibility). The authority paradigm is generic. It applies to "sciences" as conventionally understood (and in idiomatic English, this generally means "the hard sciences" like physics and chemistry), but also to historical, social and psychological facts, and even to moral and theological reasoning. This 5-fold taxonomy, albeit a necessarily simplifying model, offers a powerful heuristic for teasing out theoretical complexities and practical challenges entailed by the familiar-sounding notion of forensic expertise[5]. For example, components (1) to (3)—competence, domain and methodological validity—interact in interesting ways. The authority paradigm obviously breaks down if the testimony is not proffered by a genuine expert; if the so-called expert is an "incompetent" witness in the common lawyer's sense. But sometimes, it is not so much the qualifications and experience of individual experts that are at stake, but the very possibility of domain expertise. The objection to "expert" witch-finders[6] or ghost hunters is more fundamental (and less *ad hominem*) than any criticism that individual exponents have not taken the appropriate training courses or gained enough job-related experience.

Methodological validity, component (3), embraces a set of important epistemic considerations arising even in relation to genuine experts in well-established disciplinary domains. The authority paradigm breaks down for different reasons when genuine experts succumb to personal or professional biases, fail to implement pertinent methodological protocols correctly, or purport to speak beyond the boundaries of their domain-specific expertise. It may be difficult in the general run of cases for courts to differentiate between genuinely well-credentialed experts, and plausible-sounding charlatans and shysters[7]. This practical challenge is so much greater, however, in relation to the types of failing encompassed by component (3), where genuine experts are over-reaching in one way or another[8]. Forms of expert testimony incorporating multiple specialist domains, including DNA profiling, pose such dilemmas acutely. Plainly, not every opinion or judgment expressed by an expert properly qualifies as *expert opinion*. Judges may be ill-equipped and trial procedure ill-suited to policing experts' disciplinary boundaries effectively.

Components (4) and (5) of the authority paradigm—materiality and admissibility—introduce further major complexities, in terms of managing the interface between expert knowledge and forensic objectives, concerns and values. A vital distinction is that, whereas components (1)–(3) are essentially epistemic matters, criminal adjudication is fundamentally *normative*. The overriding objective of criminal proceedings is doing justice;[9] and whilst epistemic considerations are vital ingredients in the mix—we want to convict the guilty, and only them, of the right offence(s)—epistemology is not the proof of the pudding. We only want to convict the guilty *in the right way* ("by due process of law"), not any which way—e.g., by vigilante lynch mob or Dirty Harry policing[10] in violation of the rule of law. Thus, all evidence, including expert witness testimony, must satisfy fundamental criteria of procedural fairness, transparency, exposure to adversarial testing, and compliance with other basic criteria of the right to a fair trial. Notwithstanding their divergent legal histories, idiosyncratic procedural traditions and distinctive institutional cultures, all 47 Council of Europe nations are bound by a common conception of the fundamental requirements of fair criminal trials under Article 6 of the ECHR[11] (which is entirely separate from parallel or overlapping EU legal frameworks applicable only to the 28 countries of the "smaller Europe"[12]).

[4]The leading case in English law is (still) *Woolmington* v *DPP* [1935] AC 462, HL. However, the concept of "presumption of innocence" is complex and open to a range of normative and epistemic interpretations. See further, Paul Roberts, "Loss of Innocence in Common Law Presumptions" (2014) 8 *Criminal Law and Philosophy* 317.

[5]Also see Susan Haack, *Evidence Matters: Science, Proof and Truth in the Law* (Cambridge UP, 2014); Harry Collins, *Are We All Scientific Experts Now?* (Polity, 2014); Harry Collins and Robert Evans, *Rethinking Expertise* (Chicago UP, 2007).

[6]For pertinent historical context, see Malcolm Gaskill, "Witchcraft and Evidence in Early Modern England" (2008) 198 *Past and Present* 33; Gregory Durston, *Witchcraft and Witch Trials: A History of English Witchcraft and its Legal Perspectives, 1542 to 1736* (Barry Rose, 2000).

[7]Although in no sense representative, the literature contains examples of truly egregious malpractice in pockets of forensics: see e.g., Paul C. Giannelli, "The Abuse of Scientific Evidence in Criminal Cases: The Need for Independent Crime Laboratories" (1997) 4 *Virginia Journal of Social Policy and the Law* 439.

[8]Cf. *Meadow* v *General Medical Council* [2007] QB 462, [2006] EWCA Civ 1390; Richard Nobles and David Schiff, "Misleading Statistics within Criminal Trials – The Sally Clark Case" (2005) 2(1) *Significance* 17.

[9]This is now an explicit normative requirement in English law: Criminal Procedure Rules 2015, r.1. For broader contextualizing discussion, see Paul Roberts, "Groundwork for a Jurisprudence of Criminal Procedure" in R. A. Duff and Stuart Green (eds.), *Philosophical Foundations of Criminal Law* (OUP, 2011).

[10]For idiomatic applications across the political spectrum, see e.g., Russell Dean Covey, "*Miranda* and the Media: Tracing the Cultural Evolution of a Constitutional Revolution" (2007) 10 *Chapman Law Review* 761; Joëlle Anne Moreno, "What Happens when Dirty Harry Becomes an (Expert) Witness for the Prosecution?" (2004) 79 *Tulane Law Review* 1; Richard Nobles and David Schiff, "Due Process and Dirty Harry Dilemmas: Criminal Appeals and the Human Rights Act" (2001) 64 *Modern Law Review* 911; Michael Stokes Paulsen, "Dirty Harry and the Real Constitution" (1997) 64 *University of Chicago Law Review* 1457; Carl B. Klockars, "The Dirty Harry Problem" (1980) 452 *Annals of the American Academy of Political and Social Science* 33.

[11]For general elucidation, see John D. Jackson and Sarah J. Summers, *The Internationalization of Criminal Evidence* (CUP, 2012); Paul Roberts and Jill Hunter (eds.), *Criminal Evidence and Human Rights* (Hart, 2012).

[12]Cf. Paul Roberts, "From Extradition to Surrender: EU Criminal Law and Comparative Legal Method" (2014) 53 *Howard Journal of Criminal Justice* 545; Jacqueline Hodgson, "EU Criminal Justice: The Challenge of Due Process Rights

It is at this point in the discussion that, in my experience, lawyers and scientists tend to see things differently; and misunderstandings easily arise. Science investigates empirical matters, and produces factual information about the world. It is epistemic to its core and overwhelmingly instrumental in outlook. The policy paradigm is "curing cancer." A new drug either works (in part), or it does not. It has particular side-effects (in some degree), or it does not. It can be manufactured by a particular process, or it cannot. Likewise, the DNA collected from the crime scene was either deposited by the accused, or by somebody else; the accused lacked capacity to form the required intention at the material time (e.g., because catatonic or sleepwalking), or he did not; and so on. These are all facts about the (empirical) world; they are either true or false; and they invoke or presuppose causal explanations. This is not to say or imply that "science" always provides unequivocal, certain answers to discrete, well-formulated questions. To the contrary, scientific investigation is inherently uncertain ("experimental"), and conclusions are typically framed in probabilistic terms—whether or not employing explicitly quantitative measurements of uncertainty in numbers or words. But the equivocation introduced by resorting to probability is *epistemic* not ontological: it relates to the status of our knowledge and beliefs about facts in the world, not to the facts themselves (setting aside complications arising from quantum physics and sub-atomic particles not pertinent to the present discussion). Judgments of justice are of an entirely different, normative, order. It is not merely doubtful or uncertain whether, say, it would be *just* or *fair* if Drug X cured cancer; or *just* or *fair* if the accused were the donor of crime scene DNA. Such questions are incoherent. They perpetrate a category error, confusing normative standards with empirical facts.

Criminal adjudication comprises a set of institutionalized practices for determining liability and censuring and punishing criminal wrongdoing[13]. This set of practices is normative through and through. It is not just that epistemic considerations are subject to normative side-constraints, as where we exclude relevant evidence procured by torture irrespective of its epistemic credentials[14]. Epistemic objectives are themselves normatively constituted, in the sense that the standard of adequate epistemic warrant is indexed to the institutionalized practices and objectives of criminal adjudication. So what we require is not "adequate warrant" (sufficient grounds) in the abstract, but *adequate warrant for the purposes of determining criminal liability and censuring and punishing criminal wrongdoers*. By reframing the issue in this way, it should become clearer why expert evidence cannot provide its own epistemic warrant for judicial purposes, no matter how highly the evidence scores on components (1)–(3) of the authority paradigm. Expert witnesses do not decide what evidence is "good enough" for the purposes of criminal adjudication. This is the role of the legally, indeed constitutionally, authorized fact-finder. Furthermore, issues of materiality and strategic application in individual cases are determined by *legal* standards and *judicial* decision-makers, not by the disciplinary standards embraced by particular sciences or expert witnesses. This insight goes to the heart of the truism that forensic science serves justice, not the other way around.

The logic of the authority paradigm and the priority of normative over epistemic considerations in criminal adjudication are general features of all modern legal systems. However, the ways in which resulting interfaces are organized, opportunities exploited, and tensions managed vary considerably from one legal system to the next, working with the grain of local procedural traditions, institutional practices and professional cultures. For example, in the common law world lay fact-finding is still regarded as significant (even though professional judges increasingly predominate), whilst lay input in criminal adjudication is diminished or even non-existent in most Continental juristic traditions. The roles, relationships, and distribution of powers between judges, prosecutors and defence lawyers also vary considerably across legal jurisdictions. In legal systems with stronger adversarial leanings, prosecutors and defence lawyers tend to play a more active role in shaping the course of the proceedings, whereas the "inquisitorial" judge is the dominant figure in other procedural models. Criminal procedure is dynamic and constantly evolving (we have seen major shifts toward a philosophy of activist judicial trial management in England and Wales in recent years, for example), and it is always perilous to over-generalize abstract formal models or to extrapolate too confidently from national traditions. It follows that approaches to expert evidence in general, or to particular types of scientific evidence such as LTDNA, which are utilized successfully in one jurisdiction cannot automatically be expected to operate with the same success, or at all, in a different procedural environment structured by alternative normative priorities. This observation holds *irrespective of the epistemic credentials of expert evidence*, encapsulated in authority paradigm components (1)–(3). Normative pluralism and jurisdictional diversity are inherent features of modern legality requiring detailed local knowledge and careful negotiation, not least on the part of expert witnesses operating in multiple jurisdictions. However, these elements of cultural relativity tend to provoke intuitive resistance form scientists accustomed to prioritizing universal (empirical) scientific truth over national ideology. In one sense, skepticism is justified: sacrificing science to ideology *in general*, and regardless of political variety, leads to Lysenkoism, authoritarianism, crop failure and mass starvation. Nonetheless, the subservience of science to normative criteria is an inherent, fully rationalized and legitimate *requirement* of penal justice.

Two kinds of recurrent problems with expert evidence call for practical solutions in all legal systems. The first is the problem of expert disagreement; the second, more fundamental problem concerns a recurrent dynamic between deference and education in reliance on expertise. What should a court do when expert witnesses disagree? An attractive first option would be to find that, on further investigation, there is no genuine

within a Framework of Mutual Recognition" (2011) 37 *North Carolina Journal of International Law and Commercial Regulation* 308.

[13]Generally, see Paul Roberts (ed.), *Theoretical Foundations of Criminal Trial Procedure* (Ashgate, 2014).

[14]*A* v *Secretary of State for the Home Department (No 2)* [2006] 2 AC 221, [2005] UKHL 71; *Gäfgen* v *Germany* (2011) 52 EHRR 1, ECtHR (GC). For discussion, see Paul Roberts, "Normative Evolution in Evidentiary Exclusion: Coercion, Deception and the Right to a Fair Trial" in Paul Roberts and Jill Hunter (eds.), *Criminal Evidence and Human Rights* (Hart, 2012).

disagreement to resolve, e.g., because one of the protagonists is not really an expert after all, or not an expert in the relevant domain, or because the experts have been fed different factual assumptions by their instructing lawyers, and once these discrepancies have been clarified the ostensible disagreement disappears. But this convenient resolution will not always be possible. In cases of genuine, well-informed, unshakeable disagreement between experts, various strategies are available to the court. One approach would be to accept the disagreement as a forensically significant fact in and of itself and resolve the issue in accordance with the burden and standard of proof (usually, but not invariably, giving the accused the benefit of the doubt in a criminal trial)[15]. A second strategy would be to side-step scientific disagreements by invoking individual experts' respective qualifications, experience and/or testimonial credibility as proxies for the reliability of their evidence, e.g., by adopting the working assumption that the professor or consultant is more likely to be correct than a laboratory technician or medical student. A third possibility is for the court to try to resolve the disagreement for itself. Strategies two and three exemplify the education/deference dynamic in factfinders' reliance on expert evidence. Strategy two entails deferring to the most authoritative expert, as judged by the factfinder (with or without the benefit of further judicial directions). The third strategy initially sounds the most attractive, because it comports with the factfinder's overarching responsibility for determining disputed questions of fact. The obvious problem is that, by definition, the factfinder lacks domain-specific expertise. Can the experts, through their courtroom testimony, effectively educate the factfinder to arrive at its own decision? Perhaps some element of "education" is possible, even in the constrained and most unpromising pedagogical environment of the criminal courtroom, but it seems quite implausible that factfinders in criminal trials could be equipped with sufficient knowledge and insight to resolve disputes between genuine experts with long years of study and extensive practical experience under their belts[16]. Worse, fact-finders' lack of domain-relevant expertise also undercuts strategy two, because how is it possible for jurors to assess the comparative merits of experts' disagreements when their own knowledge of the field is tenuous or non-existent? The worry is that, in the absence of rational criteria for making a determination, factfinders will fall back on irrational proxies for robust epistemic warrant, such as placing their faith in the expert with greater testimonial eloquence or the doctor with the most reassuring bedside manner.

Legal systems resolve these pervasive issues of expertise in their own distinctive ways. At the level of sweeping generalization, common lawyers tend to think that "civilians don't try"[17] and that inquisitorial judges too readily defer to authoritative court-appointed experts[18]. Civilians, for their part, tend to regard common law criminal procedure as irrational in its preferences for adversarial theatre, excessive technicality and lay over expert (including expert judicial) decision-making[19]. These longstanding debates implicate deep-rooted and enduring controversies, which it would not be profitable to dig into here; save to say that there is no reason for thinking that structures and cultures of criminal adjudication must conform to a single uniform pattern (so long as they adhere to fundamental standards of justice), any more than we should expect rigid, monotonous uniformity in national cuisine, manners or language. Recognition of legitimate scope for national cultural diversity, even in procedural fundamentals, is another major respect in which international, interdisciplinary conversations about law and justice differ markedly from international conversations about science and expertise.

LTDNA EVIDENCE IN UK CRIMINAL TRIALS

Recent attempts by criminal courts in England and Wales and Northern Ireland to get to grips with LTDNA profiling evidence must be interpreted in light of the conceptual, normative and juridical considerations summarized in the previous section. The five principal components of the authority paradigm and the diversity of national criminal procedures within a shared ECHR framework mandating fair trials, in particular, need to be borne in mind as the exposition unfolds. Just as the entirety of western philosophy has evolved in productive antagonism with skeptical doubt, legal recognition of LTDNA profiling was propelled by challenges to its methodology, epistemic status and evidential reliability. The evolution of English criminal jurisprudence on LTDNA profiling may find some resonances with parallel developments in other legal jurisdictions, and possibly inform

[15]This strategy was suggested in *R* v *Cannings* [2004] 1 WLR 2607, [178] (CA), where Judge LJ advised prosecutors and trial judges that "if the outcome of the trial depends exclusively or almost exclusively on a serious disagreement between distinguished and reputable experts, it will often be unwise, and therefore unsafe, to proceed." But cf. *R* v *Hookway and Noakes* [2011] EWCA Crim 1989.

[16]These issues are well-debated in the legal literature: see further, Ronald J. Allen, "Expertise and the *Daubert* Decision" (1994) 84 *Journal of Criminal Law and Criminology* 1157; Edward J. Imwinkelreid, "The Next Step in Conceptualizing the Presentation of Expert Evidence as Education: the Case for Didactic Trial Procedures" (1997) 1 *International Journal of Evidence and Proof* 128; Gary Edmond, "The Next Step or Moonwalking? Expert Evidence, the Public Understanding of Science and the Case Against Imwinkelreid's Didactic Trial Procedures" (1998) 2 *International Journal of Evidence and Proof* 13; Imwinkelreid, "Correspondence: Didactic Trial Procedures" (1998) 2 *International Journal of Evidence and Proof* 205.

[17]William Twining, "Civilians Don't Try: A Comment on Mirjan Damaska's "Rational and Irrational Proof Revisited"" (1997) 5 *Cardozo Journal of International and Comparative Law* 69.

[18]MN Howard QC, "The Neutral Expert: A Plausible Threat to Justice" [1991] *Criminal Law Review* 98. Such stereotypes are not without empirical foundation: cf. Chrisje Brants, "Wrongful Convictions and Inquisitorial Process: the Case of the Netherlands" (2012) 80 *University of Cincinnati Law Review* 1069, 1111 (observing that, "[j]udges may be inclined to give too much weight to expert testimony and forensic evidence (especially true of DNA).... [I]t is perhaps more problematic that judges will generally have at their disposal the evidence of only one expert.... [T]he routine absence of an expert for the defence means that the court is dependent upon its own, often amateur, evaluation of the evidence").

[19]See further, J. R. Spencer, "Court Experts and Expert Witnesses: Have We a Lesson to Learn from the French?" (1992) 45 *Current Legal Problems* 213; William T. Pizzi, *Trials Without Truth* (NYU Press, 1999); Gordon van Kessel, "Adversary Excesses in the American Criminal Trial" (1992) 67 *Notre Dame Law Review* 403; Mirjan R Damaška, *The Faces of Justice and State Authority: A Comparative Approach to the Legal Process* (Yale UP, 1986).

legal argument and policy debates in criminal trials and appeals elsewhere.

(a) Legal Recognition

Our story begins in 2007 with the judgment of the Northern Ireland Crown Court in *R* v *Hoey*[20], in which Weir J, sitting without a jury in a "Diplock" trial court[21], commented on the reliability of what was then generally referred to as Low Copy Number (LCN) DNA profiling evidence. LCN profiling evidence in this case, generated by the Forensic Science Service (FSS) laboratory in Birmingham[22], purported to link the accused to explosive devices used in a string of terrorist bombings across Northern Ireland, including "the infamous car bomb explosion that destroyed much of the shopping Centre of Omagh on the afternoon of Saturday, 15 August 1998 with... appalling consequence[s]... leaving permanent and widespread physical and psychological scars."[23] During the course of the trial, serious concerns were identified regarding the integrity of the evidential samples collected from crime scenes, which had not initially been taken, handled or stored with DNA profiling in mind. The Court found that "the arrangements within the police in 1998 and 1999 for the recording and storage of items were thoroughly disorganized"[24] and that "thoughtless and slapdash" exhibit handling and anti-contamination practices extended to the laboratories and staff of Forensic Service Northern Ireland (FSNI)[25]. Mr. Justice Weir considered it "extraordinary" that "knowing that these items had not been collected or preserved using methods designed to ensure the high degree of integrity needed not merely for DNA examination but for the more exacting requirements of LCN DNA, examinations were performed at Birmingham with a view to using them for evidential rather than solely intelligence gathering purposes." Yet analytical results from DNA profiling had then been "put forward and stoutly defended" at trial "as evidence that the Court might safely rely upon as tending to establish the guilt of the accused."[26] As a matter of legal logic, Weir J's decisive conclusion flowed almost ineluctably from the prosecution's failure to establish the integrity of its evidence:

> [O]ne police and SOCO witness after another and also Dr. Griffin [of FSNI] had candidly made clear that possible examination for DNA was not in their minds at all as they were collecting, storing, transmitting and dealing with these items in 1998. Why therefore would they then have had present to their minds and been complying with the exacting integrity requirements which reliable DNA examination and most especially that in its LCN form demands? All this [FSNI] must have known very well when it co-operated in searching for and collecting items for LCN examination in Birmingham and again later when the idea of using the results of those examinations as evidence in this trial must have been under discussion. By that stage the problems inherent in the need to prove integrity had plainly come to be appreciated by one or more police officers concerned in this investigation as was shown by the mendacious attempts to retrospectively alter... evidence so as to falsely make it appear that appropriate DNA protective precautions had been taken at that scene.... [H]aving carefully reviewed all the evidence on this issue, I am not in the least satisfied in relation to any one of the items upon which reliance is sought to be placed for the results of their LCN DNA examinations that the integrity of any of those items prior to its examination for that purpose has been established by the evidence. Accordingly I find that that DNA evidence... cannot satisfy me either beyond a reasonable doubt or to any other acceptable standard[27].

That is to say, in terms of the conceptual framework sketched in the previous section, the DNA profiling evidence lacked adequate epistemic warrant for grounding a criminal conviction, owing to the well-established fact that forensic samples were compromised—at least in the sense that their integrity could not be demonstrably assured.

Weir J's judgment might have stopped there, but instead briefly addressed the validity and merits of LCN profiling techniques themselves, since these had been extensively canvassed during the trial and conflicting expert views had been expressed. Weir J was "concerned at the wide variance in expert opinions, not only as between the Prosecution and Defence but also between the two experts called for the Prosecution,"[28] as well as by the "manner and content of the response" of the main FSS expert to defence criticisms. This witness appeared to Weir J to be "inappropriately combative as an expert witness and his unwillingness to debate constructively the various matters put to him was unhelpful in the extreme."[29] Notice that, in the absence of domain expertise, Weir J predictably falls back on general proxies for testimonial reliability, such as the (not unreasonable) working assumptions that a conscientious and objective scientist will display an open mind and be prepared to debate criticisms and objections in a fair-minded way. A second prosecution expert, by contrast, came over to the Court as "willing to carefully consider the propositions put to him" by defence counsel, such that "his evidence greatly helped to inform and bring some objectivity to the debate."[30] Weir J registered "concern about the present state of the validation of the science and methodology associated with LCN DNA and, in consequence, its reliability as an evidential tool" and expressed himself "not satisfied that the publishing of two journal articles describing a process invented by the authors can be regarded without more as having "validated" that process for the purpose of its being confidently used for evidential purposes."[31] These

[20] *R* v *Sean Hoey* [2007] NICC 49.

[21] See John Jackson and Sean Doran, *Judge Without Jury: Diplock Trials in the Adversary System* (OUP, 1995).

[22] The FSS was subsequently, and controversially, closed down to save public money: see House of Commons Science and Technology Committee, *Forensic Science*, Second Report of Session 2013–14, HC 610 (TSO, 2013).

[23] *R* v *Sean Hoey* [2007] NICC 49, [1].

[24] Ibid [51].

[25] "The position so far as [FSNI] is concerned is even more difficult to comprehend as everyone there must have been very well aware of the risks of improper labeling, storage and examination": Ibid [59].

[26] Ibid [60].

[27] Ibid [60], [61].

[28] Ibid [62].

[29] Ibid [63].

[30] Ibid.

[31] Ibid [64]. Notice that "confident use for evidential purposes" parallels, in my terminology, adequate epistemic warrant for (use in) criminal adjudication.

remarks were left hanging within the context of an unresolved broader discussion about enhancing procedural frameworks for regulating the admission and uses of scientific evidence in criminal trials. Weir J suggested that "the evidence given in this case by the FSS witnesses reinforces in the clearest way possible the need for urgent attention to this task."[32]

Being a decision at first instance, *R* v *Hoey* did not create a binding legal precedent (not even in Northern Ireland), and Weir J's remarks on LCN DNA profiling were strictly *obiter dicta*, i.e., not part of the formal legal holding in the case. It was nonetheless a widely reported judgment in a very high profile trial, which sparked much agitated discussion amongst forensic scientists and led the Association of Chief Police Officers (ACPO), in consultation with the Crown Prosecution Service (CPS), to recommend temporary suspension of LCN profiling techniques in criminal investigations and prosecutions pending further inquiry and review. The science of LTDNA analysis, defined as "[a]n ultra-sensitive technique that has the potential to yield a DNA profile from sub-optimal biological samples e.g., Low Copy Number DNA analysis,"[33] was subsequently examined by an expert panel established by the Forensic Science Regulator and chaired by Professor Brian Caddy. The Caddy Review concluded that "the science supporting the delivery of Low Template DNA (LTDNA) analysis is sound and that the three companies... providing this service to the Criminal Justice System have validated their processes in accord[ance] with accepted scientific principles."[34] But it was noted that "regardless of which signal enhancement method is selected, the problems of allele drop out due to stochastic effects in the presence of low quantities of template and that of increase[d] noise will occur in sub-optimal DNA samples,"[35] and the concerns expressed by Weir J in *Hoey* regarding the absence of reliable validation were characterized as "well-founded."[36] Furthermore:

> Interpretation of the results is complex for two reasons: the statistics are challenging and probably hard to comprehend by a non-specialist and the decision how and when to apply certain statistical methods has not yet reached a clear consensus... the challenges in terms of statistical interpretation of the data and in communicating them to a largely innumerate criminal justice system should not be under-estimated, nor should the importance of earning and maintaining public confidence in the system[37].

These observations appropriately acknowledge the broader institutional context and social expectations of evidential (epistemic) warrant for criminal verdicts. The Caddy Review recommended that "any LTDNA profile should always be reported to the jury with the caveats: that the nature of the original starting material is unknown; that the time at which the DNA was transferred cannot be inferred; and that the opportunity for secondary transfer is increased in comparison to standard DNA profiling."[38] Crucially, Caddy expressed the opinion that matches for LCN DNA profiling should be reported at the sub-source, genetic level only. Consequently, it would be "inappropriate to comment upon the cellular material from which the DNA arose or the activity by which the DNA was transferred."[39]

A follow-up report issued by the Forensic Science Regulator concurred that "the science underpinning the LTDNA analytical services, as provided to the CJS [criminal justice system], is sound and that... suppliers offering such services have properly validated their processes. There is no flaw inherent in the process which prevents its use within the CJS."[40] Although there remained "key areas where improvements can be made... probably most importantly, the interpretation of the evidence,"[41] the Regulator stressed that scope for improvement "does not mean that the approach should not be employed within the CJS":

> As long as the scientist reporting the results of LTDNA analysis complies with the duties and obligations placed on expert witnesses the CJS will appreciate the nature and value of the evidence provided[42].

This was the state of play, in technical and policy circles, when the Court of Appeal in England and Wales came to consider the status of LTDNA evidence in criminal trials in a clutch of criminal appeals in 2009 and 2010, beginning with *R* v *Reed*[43].

(b) Authoritative Rulings

The case popularly known as *Reed and Reed* actually comprised two conjoined appeals arising from separate trials, both of which involved challenges to LCN DNA profiling evidence. In *Reed* itself, genetic material was recovered from pieces of plastic which the prosecution contended had broken off from a knife handle employed as the murder weapon. In *Garmson*, the accused was identified as the rapist from small amounts of DNA deposited on the victim's lips, undergarments and tampon. Primed by Weir J's widely reported reservations in *Hoey* and the Caddy Review's findings, the Court of Appeal—led by Thomas LJ, who has since been promoted to Lord Chief Justice—embarked upon a thorough reconsideration of LCN and LTDNA profiling evidence, and directed the parties to assist the Court, within the framework of judicially managed pre-trial case preparation implemented by the Criminal Procedure Rules (CrimPR) since 2005 (and regularly updated). These exchanges produced the following fixed points of agreement: (1)

[32] Ibid.

[33] Brian Caddy, Graham R Taylor and Adrian M T Linacre, *A Review of Low Template DNA Analysis* (2008) [1.8], www.gov.uk/government/uploads/system/uploads/attachment_data/file/117556/Review_of_Low_Template_DNA_1.pdf

[34] Ibid, Executive Summary.

[35] Ibid [3.11].

[36] Ibid [3.15].

[37] Ibid [9.3], [8.1].

[38] Ibid [7.4].

[39] Ibid [7.5].

[40] Forensic Science Regulator, *Response to Professor Brian Caddy's Review of the Science of Low Template DNA Analysis* (FSR, 2008) [4.1.1].

[41] Ibid [4.1.2]. Also now see Peter Gill, June Guiness and Simon Iveson, *The Interpretation of DNA Evidence (Including Low-Template DNA)*, FSR-G-202 (Forensic Science Regulator, 2012); Roberto Puch-Solis, Paul Roberts, Susan Pope and Colin Aitken, *Assessing the Probative Value of DNA Evidence* (Royal Statistical Society, 2012), www.rss.org.uk/statsandlaw.

[42] Forensic Science Regulator, *Response to Professor Brian Caddy's Review of the Science of Low Template DNA Analysis*, [4.1.3].

[43] *R* v *Reed and Reed*; *R* v *Garmson* [2010] 1 Cr App R 23, [2009] EWCA Crim 2698.

the "standard kit" employed in DNA profiling using the SGM+ system was "designed optimally to produce a full profile on 1 ng which is the approximate equivalent of 160 human somatic cells which typically can be visualized in a tiny blood spot"[44]; (2) "[P]articularly where no identifiable body fluid is present, the amount of DNA present may be as low as the equivalent of that contained in one body cell. Where a sample is measured to be less than what is required to generate a profile using the standard SGM+ test, then Low Template DNA analysis is often undertaken;"[45] (3) "[T]he stochastic effects may be such that no reliable profile can be generated. The FSS had found that in a very high proportion of profiles obtained using the LCN process the profiles were not capable of robust and reliable interpretation because of stochastic variations."[46]

Drawing on the technical data and expert opinions canvassed before it and adduced in evidence, the Court of Appeal noted the importance of "the stochastic threshold" at which "the profile is unlikely to suffer from stochastic effects (such as allelic drop out...) which prevent proper interpretation of the alleles."[47] Below the stochastic threshold, it is, at the very least, debateable whether analytical results can support meaningful findings, owing to the "noise" generated by uncontrolled random effects. The question then becomes, how much DNA is required to meet this stochastic threshold? The Court of Appeal had heard differing expert views, but prevailing opinion ("in the absence of new scientific evidence") placed it within the range 100–200 picograms[48]. In the light of this (albeit, possibly temporary and unstable) scientific consensus, the Court of Appeal in *Reed* announced the following principles of admissibility:

> [A] challenge to the validity of the method of analysing Low Template DNA by the LCN process should no longer be permitted at trials where the quantity of DNA analysed is above the stochastic threshold of 100–200 picograms.... There may be cases where reliance is placed on a profile obtained where the quantity of DNA analysed is within the range of 100–200 picograms where there is disagreement on the stochastic threshold on the present state of the science. We would anticipate that such cases would be rare and that, in any event, the scientific disagreement will be resolved as the science of DNA profiling develops. If such a case arises, expert evidence must be given as to whether in the particular case, a reliable interpretation can be made. We would anticipate that such evidence would be given by persons who are expert in the science of DNA and supported by the latest research on the subject. We would not anticipate there being any attack on the good faith of those who sought to adduce such evidence[49].

Here we see the Court of Appeal (literally) laying down the law in relation to the admissibility of LTDNA profiling evidence. There is no general statutory test governing the admissibility of expert evidence in England and Wales. Admissibility is governed by common law principles,[50] which the courts are both entitled and duty-bound to develop[51]. Strictly speaking, the Court's remarks about evidence under the stochastic threshold are *obiter*, because the DNA evidence in both appeals in *Reed* was above the threshold, and the appeals were ultimately argued and determined on issues of transference and persistence of DNA traces, not on the validity of profiling techniques. However, this (technical) legal objection would predictably fail to gain judicial traction in subsequent cases, given the institutional status and authority of the *Reed* judgment. It would be perfectly evident to experienced lawyers that a senior court was deliberately articulating guidance to be followed in future criminal trials and appeals, with the firm expectation of compliance.

The admissibility principles propounded by the Court of Appeal in *Reed* are interesting at a number of levels. They are animated by the strong desirability of providing clear and reasonably determinate guidance to prosecutors, defence lawyers and trial judges in the conduct of criminal litigation. It would hardly be efficient to try to re-litigate the existence and calibration of stochastic thresholds in each and every criminal trial involving LTDNA profiling evidence, and it would—to say the least—be highly undesirable for individual courts to be improvising their own, quite possibly discrepant, thresholds, depending partly on which expert witnesses happened to testify in particular trials and how their evidence was received and assessed by individual trial judges in admissibility determinations (and hostage to further contingencies, including whether admissibility was, in fact, challenged in the instant case[52]). The problem is that there is no readily available institutional mechanism for establishing "legislative facts," such as the nature of stochastic thresholds for LTDNA profiling evidence, in English criminal proceedings. This is not the sort of thing that could be included in a Code of Criminal Procedure, even if we had one (which, if one has in mind the standard continental model, we don't). So the Court of Appeal is obliged to step into the void and take responsibility for standard-setting upon itself. However, this is slippery and even perilous territory. Can the law plausibly dictate standards of scientific validity, even for its own juridical purposes? The Court of Appeal is primarily concerned with adjudicating questions of *law*, not fact. There is some flexibility, inasmuch as the Court of Appeal makes classificatory choices as the arbiter of what qualifies, in law, as "questions of fact" and "questions of law." But scientific facts themselves, as opposed to the legitimacy of their

[44]Ibid [39].
[45]Ibid [45].
[46]Ibid [49].
[47]Ibid [74].
[48]Ibid.
[49]Ibid.

[50]The leading case remains *R* v *Turner* [1975] 1 QB 834 (CA), elucidating a generic test of "helpfulness"; though this must now be read through the quasi-legislative effect of *Criminal Practice Directions 2015* [2015] EWCA Crim 1567, CPD V Evidence 19A, propounding criteria of admissibility modeled on proposals by the Law Commission. See further, Ian Dennis, "Editorial: Tightening the Law on Expert Evidence" [2015] *Criminal Law Review* 1; Tony Ward, "Expert Evidence and the Law Commission: Implementation without Legislation?" [2013] *Criminal Law Review* 561.

[51]See Paul Roberts, "Expert Evidence and Criminal Trial Procedure" in Gerben Bruinsma and David Weisburd (eds.), *Encyclopedia of Criminology and Criminal Justice* (Springer, 2013) 1480–1494; Paul Roberts and Adrian Zuckerman, *Criminal Evidence* (OUP, 2/e 2010) ch 11.

[52]"[U]nless the admissibility is challenged, the judge will admit that evidence. That is the only pragmatic way in which it is possible to conduct trials": *R* v *Reed and Reed*; *R* v *Garmson* [2010] 1 Cr App R 23, [2009] EWCA Crim 2698, [113].

forensic uses, cannot be subjected to the normative dictates of law, on pain of reversion to ideology and Lysenkoism.

The judgment in *Reed* makes extensive reference to the Caddy Review[53] and the input of the Forensic Science Regulator, reflecting a notable emergent symbiosis. Systematic reviews of scientific issues by expert practitioners and independent regulators will almost certainly form sounder technical conclusions, and supply superior epistemic warrant for legal decision-making on scientific questions, than comparable efforts by appellate courts (even those staffed by relatively knowledgeable and scientifically literate judges), examining the facts of particular cases within the constrained institutional parameters of criminal litigation. Recognizing this, the Court of Appeal in *Reed* lent its judicial authority to conclusions arrived at extra-judicially which, in the absence of the Court's imprimatur, would have been relegated to the marginal jurisprudential status of supporting material for expert witness testimony. In this mutually obliging fashion, the Court of Appeal acquires credibility for its scientific conclusions whilst simultaneously conferring judicial kudos on the Regulator. Yet precisely because we are dealing with *scientific facts*, forensic closure cannot be permanent or complete. What happens if prevailing scientific opinion shifts?[54] This is hardly a remote possibility in rapidly developing fields such as DNA profiling technology (does anybody still remember "genetic fingerprinting," southern blotting and autoradiographs?[55]). Mindful of the risk of petrifying the law's approach to stochastic thresholds, the Court of Appeal inserted the rider "in the absence of new scientific evidence" into its admissibility principles. However, this is effectively an invitation for trial counsel to argue that they do, indeed, have new scientific evidence at their disposal; and it would not be entirely surprising if time spent researching the relevant academic journals reaped forensic rewards for enterprising trial lawyers. Conversely, if the opportunity for challenge in the light of new evidence had been closed down, later shifts in scientific understandings of stochastic thresholds would have resulted in appeals against conviction on the basis of "fresh evidence." It is bred into common lawyers that circumstances alter cases and that each set of facts presents its own unique, and eminently distinguishable, characteristics. The Court of Appeal consequently never says "never" in relation to the scope for challenging prevailing scientific wisdom; a caution entirely vindicated by the fact that some very longstanding practices of forensic science (not to mention transient enthusiasms[56]) have turned out to lack sound methodological foundations, and some (like the old "points" system for declaring fingerprint matches) have lately been abandoned[57].

The Court of Appeal's further "anticipation" (read: directive) that evidence pertaining to stochastic thresholds "would be given by persons who are expert in the science of DNA and supported by the latest research on the subject" might sound like no more than a reiteration of common sense legal orthodoxy, extrapolating from the competency and domain components of the standard authority paradigm for expertise outlined in the previous section. In fact, these loaded remarks were intended to signal the Court's impatience with defence testimony challenging the validity and inferential logic of DNA profiling evidence, based not on direct practical experience in profiling, but on generic principles of scientific validation, methodology and inferential logic. Of one defence expert witness, the Court of Appeal remarked:

> He gives evidence with a degree of gravitas and fluency that is impressive and is able to explain concepts clearly. However, his expertise on the interpretation of DNA profiles is limited, without any relevant first hand laboratory or research experience. He is not qualified to make a scene of crime investigation... Whilst it is impossible to understand how he had sufficient expertise to be able to give evidence in *R* v *Hoey*, let alone to assist in the attack made in that case on the LCN process, he has given evidence in so many Low Template DNA cases since then on the strength of the observations in *R* v *Hoey* that he has acquired a degree of experience from these cases, his discussion with others and his reading of papers. We retain clear reservations about the extent of his expertise in relation to DNA profiles...[58]

In relation to a second defence expert, the Court complained that "his experience is of a different jurisdiction where the scientist who gives evidence may have a narrower type of expertise and the scope of evidence an expert can give may not be the same as the scope in this jurisdiction.... [H]is experience was not based on the work of a forensic scientist in this jurisdiction who attends both the scene of the crime and supervises the laboratory work."[59] These attempts to prioritize hands-on forensic experience over academic research and theorizing are not entirely convincing. Given that expertise is domain-specific, careful attention needs

[53] For further critical discussion, see Christopher Lawless, *Forensic Science: A Sociological Introduction* (Routledge, 2016), 107–114 and further sources cited therein. As a legal scholar, I am patently unqualified to second-guess Caddy's evaluation of LTDNA profiling techniques, nor do I express any view on the matter. My analysis relates only to English courts' use of the Caddy Review as an authoritative source of information and public endorsement of LTDNA's scientific credentials.

[54] Cf. *R* v *Henderson; R* v *Butler; R* v *Oyediran* [2010] 2 Cr App R 24, [2010] EWCA Crim 1269, discussed in Paul Roberts, "Fue el Bebé Sacudido? Preuba, Pericia y Epistemologia Juridica en el Proceso Penal Inglés" in Carmen Vázquez (ed.), *Estandares de prueba y prueba cientifica* (Marcial Pons, 2013).

[55] On the evolution of DNA profiling techniques, see Michael Lynch, Simon A Cole, Ruth McNally and Kathleen Jordan, *Truth Machine: The Contentious History of DNA Fingerprinting* (Chicago UP, 2008).

[56] Cf. *R* v *McIlkenny* (1991) 93 Cr App R 287 (CA); *R* v *Maguire* (1992) 94 Cr App R 133 (CA); Sir John May, *Inquiry into the Circumstances Surrounding the Convictions Arising Out of the Bomb Attacks on Guildford and Woolwich in 1974*, HC Paper 556 (1990) and HC Paper 296 (1992); Mike Redmayne, "Expert Evidence and Scientific Disagreement" (1997) 30 *UC Davis Law Reviw* 1027; Clive Walker and Russell Stockdale, "Forensic Evidence and Terrorist Trials in the United Kingdom" (1995) 54 *Cambridge Law Journal* 69.

[57] *R* v *Smith (Peter)* [2011] 2 Cr App R 16, [2011] EWCA Crim 1296. See further, Simon A Cole and Andrew Roberts, "Certainty, Individualization and the Subjective Nature of Expert Fingerprint Evidence" [2012] *Criminal Law Review* 824; Jennifer L. Mnookin, "The Validity of Latent Fingerprint Identification: Confessions of a Fingerprinting Moderate" (2008) 7 *Law, Probability and Risk* 127; Robert Epstein, "Fingerprints Meet *Daubert*: The Myth of Fingerprint "Science" is Revealed" (2002) 75 *Southern California Law Review* 605.

[58] *R* v *Reed and Reed*; *R* v *Garmson* [2010] 1 Cr App R 23, [2009] EWCA Crim 2698, [107].

[59] Ibid [103].

to be given to the grounds, parameters and content of any particular expert's evidence. Courts should certainly be chary of receiving "expert" testimony about laboratory conduct and protocols from somebody who has never worked in a laboratory. By parity of reasoning, however, practitioners are not necessarily good authorities on questions of policy or theory. An example highly pertinent to the present discussion is that a geneticist could be highly accomplished and very experienced in DNA profiling techniques without necessarily having acquired a firm grasp of the statistical foundations and probabilistic methods employed in assessing likelihood ratios for complex mixtures or partial profiles (or even, for that matter, in generating random match probabilities for straightforward single profiles)[60]. Another background consideration possibly at work here is English courts' intuitive suspicion of expert testimony attempting to instruct fact-finders in relation to general considerations of logic and inferential reasoning. Such testimony is often viewed, not without justification, as potentially trenching on the jury's constitutional prerogatives in fact-finding[61]. But the practitioner/theoretician continuum is orthogonal to that concern; and expert competence must always be assessed relative to domain and materiality. The danger is that in ostentatiously rejecting one false proxy for testimonial reliability (a witness's gravitas and fluency), the Court of Appeal in *Reed* may have allowed itself to be gulled by another ("local practitioners always know best").

(c) After *Reed*

Lord Justice Thomas' prediction, or pious hope, that cases involving LTDNA profiles arguably under the stochastic threshold "would be rare" was soon put to the test. In *R* v *Broughton*[62] an animal rights activist was convicted of planting incendiary devices in buildings owned by two Oxford colleges. It was common ground that the attacks were a protest against animal experiments by university scientists. A central plank of the prosecution's (entirely circumstantial) case against Broughton was an LTDNA profile derived from match stalks which had formed part of the fuse mechanism of improvised incendiary devices (bottles filled with petrol) used in one attack. The amount of genetic material recovered from the crime scene was <100 picograms. This was insufficient to generate any usable results from standard profiling techniques. However, by running multiple enhanced LTDNA analyses and combining their results to produce a "cleaned up" profile, a forensic scientist was able to identify 20 alleles shared in common with the accused. This produced a random match probability, comparable to RMPs for standard profiling, of <1 in 1 billion.

One argument advanced on appeal was that *Reed* had already decided that profiles below the stochastic threshold range of 100–200 picograms are inadmissible in English law. The Court of Appeal in *Broughton* made short work of the faulty logic in this submission:

> The appellant's submission is... founded upon a misunderstanding of the decision in *Reed & Reed*. This court recognised that in the current state of technology there is a stochastic threshold between 100 and 200 picograms above which LTDNA techniques... can be used to obtain profiles capable of reliable interpretation. Specifically, the court observed that above this threshold a challenge to the validity of the method of analysing LTDNA by the LCN process should not be permitted in the absence of new scientific evidence. However, the court did not hold or make any observation to the effect that below the stochastic threshold DNA evidence is *not* admissible[63].

The defence argument that LTDNA profiles below the stochastic threshold are automatically inadmissible resurfaced in a second appeal heard later in the year, involving DNA mixtures from more than one donor, and it was once more emphatically rejected. The Court of Appeal reiterated that "what mattered was the quality of the minor profiles and not the quantity.... [P]rofiles obtained from <200 picograms can be reliable. It is reliability that is the issue, not the quantity, though plainly the quantity is relevant... to the consideration of stochastic effects."[64]

A more promising line of attack was to challenge the profiling evidence on its own scientific merits. The Court of Appeal recognized that the profiles adduced in *Broughton* "were derived from unquantified samples of DNA of <100 picograms and that this raised entirely legitimate grounds for scientific dispute which the appellant was right in testing before the judge."[65] Prevailing scientific understanding was summarized as follows:

> [T]here is now a considerable body of opinion from respected independent scientists and the Forensic Science Regulator that LTDNA techniques, including those used to generate the profiles relied upon by the Crown in this case, are well understood, have been properly validated and are accepted to be capable of generating reliable and valuable evidence. At these very low levels of DNA, the dangers presented by the possibility of stochastic effects, including allelic drop-out, drop-in and stutter are very real and must be fully appreciated, but they may often be addressed by repeating the process a number of times...[66]

Observe, again, the instrumental role of the Forensic Science Regulator in authenticating the underpinning science and validation processes. If—and for as long as—the Regulator is satisfied on these technical questions, the courts are likely to follow her[67] lead. Having noted the *potential* shortcomings of LTDNA profiles, however, the Court of Appeal was satisfied that the pertinent issues had been fully ventilated in the trial and that the evidence actually generated and adduced in the instant case had been properly explained and vindicated by competent experts:

[60] See the sources cited at n.41, above.

[61] This was the real issue in *R* v *Adams* [1996] 2 Cr App R 467 (CA); and *R* v *Adams (No 2)* [1998] 1 Cr App R 377 (CA).

[62] *R* v *Broughton* [2010] EWCA Crim 549.

[63] Ibid [31].

[64] *R* v *C* [2010] EWCA Crim 2578, [24], [27].

[65] *R* v *Broughton* [2010] EWCA Crim 549, [37].

[66] Ibid [34].

[67] The present incumbent is Dr Gillian Tully: "Appointment of New Forensic Science Regulator Announced," Home Office Press Release, 17 July 2014. Her predecessor, in post when *Broughton* was decided, was Andrew Rennison.

> [A]ll of the consensus alleles match those in the appellant's profile. In other words, the consensus profiles do not suggest the procedures suffered from drop-in or stutter such as to render the results inherently unreliable. Indeed, this is reflected in the statistics [sic] derived from the consensus profiles to which we have referred and about which there was no dispute. At their most powerful and when derived from all duplicated components, these give rise to the match probability of <1 in 1 billion. We believe that these were all matters properly admitted in evidence[68].

It is important to appreciate that criminal appeals in England and Wales are not re-trials, as they are in many continental European jurisdictions. The Court of Appeal in England and Wales performs an essentially reviewing function, and is primarily concerned with the legality and fairness of trial proceedings, not with the accuracy of their outcomes. Crucially, the Court of Appeal does not second-guess jury verdicts. Once the Court in *Broughton* had satisfied itself that the trial judge had adopted the correct approach to assessing the admissibility of LTDNA evidence, that question was settled for the purposes of this appeal. But the Court still had something it wanted to get off its chest, and the Judgment had a sting in the tail.

The Court signaled a general concern about defence tactics in challenging the credibility and "integrity" of experts presenting DNA profiling evidence:

> Whatever may be the position in other jurisdictions, it is the duty of an advocate and an expert in this jurisdiction not to embark upon an attack on the integrity of other experts unless there is an evidential basis for doing so. There was none in this case. The attack made on the integrity of LGC Forensics and Cellmark was without foundation and should never have been made... [T]here can well be a difference of opinion between experts on LTDNA, but there should be no question of the good faith of those involved in LTDNA being put in issue. This is a case where there is a proper disagreement between experts but the course taken by those giving evidence on behalf of the appellant went into matters for which there was no foundation. Not only was the attack on the good faith of the Crown's witness wholly deplorable and unwarranted, but it also was a great disservice to the appellant's case[69].

The Court is here saying that not only are such credibility attacks contrary to ethical standards of advocacy, and therefore liable to get counsel into hot water with their professional regulator[70], but also likely to back-fire by harming the defendant's prospects in the instant case. The threat is clear, but whether advocates will pay any attention to it, less so. In *Broughton*, specifically, "an attack was made... on the integrity of LGC Forensics; it was alleged that their commercial interests and influence over their case workers had tainted their professionalism and objectivity. LGC Forensics were underestimating the problems which were associated with LTDNA and promoting its viability for financial reasons."[71] Counsel presumably thought it legitimate to draw the jury's attention to possible conflicts of interest in the production of expert evidence, which is now an embedded structural feature of a marketplace dominated by commercial providers following the demise of the FSS[72]. The Court of Appeal's message is that unfocused and entirely unsubstantiated insinuations of commercial corruption will not be tolerated. The situation would presumably be different if there were material evidence that a particular expert's objectivity or impartiality might have been compromised by commercial incentives. In relation to *judicial* impartiality, the court must be manifestly, not merely actually, unbiased[73], so that justice is *seen* to be done. How much of this expectation carries over to expert witnesses utilizing techniques from which their employers derive a commercial advantage is a nicely balanced question.

Once evidence has been ruled admissible, attention shifts to its uses and probative value in the trial. An important dimension of evidentiary regulation, and one which has been assuming greater prominence in many common law jurisdictions including England and Wales over the last several decades, concerns judicial directions to the jury[74]. English law contains an expanding corpus of "forensic reasoning rules"[75] instructing factfinders how they must, may or should not utilize particular types and pieces of evidence, which inferences are rationally available and which are legally forbidden. A number of these rules or guidelines pertain to expert evidence in general[76], and to DNA evidence in particular[77]. This is where the case against *Broughton* unraveled on appeal.

[68] *R* v *Broughton* [2010] EWCA Crim 549, [37]. Note that RMPs are not themselves "statistics," but rather probabilistic extrapolations from allele frequency statistics sampled from reference populations.

[69] Ibid [38].

[70] The Bar Council and Bar Standards Board for barristers; the Law Society and Solicitors Regulation Authority for solicitor-advocates.

[71] *R* v *Broughton* [2010] EWCA Crim 549, [14].

[72] On structural features of "market forensics," see Christopher J Lawless, "Policing Markets: The Contested Shaping of Neo-Liberal Forensic Science" (2011) 51 *British Journal of Criminology* 671; Paul Roberts, "What Price a Free Market in Forensic Science Services? The Organization and Regulation of Science in the Criminal Process" (1996) 36 *British Journal of Criminology* 37.

[73] Also now a requirement of ECHR Article 6: "According to the Court's settled case law, the existence of impartiality for the purposes of Article 6(1) must be determined according to: (i) a subjective test, where regard must be had to the personal conviction and behavior of a particular judge—that is, whether the judge held any personal prejudice or bias in a given case; and (ii) an objective test, that is to say by ascertaining whether the tribunal itself and, among other aspects, its composition, offered sufficient guarantees to exclude any legitimate doubt in respect of its impartiality.... What is at stake is the confidence which the courts in a democratic society must inspire in the public": *Volkov* v *Ukraine* (2013) 57 EHRR 1, [104], [106]; *Borgers* v *Belgium* (1993) 15 EHRR 92.

[74] Also see Paul Roberts, Colin Aitken and Graham Jackson, "From Admissibility to Interpretation: New Guidance on Expert Evidence" (2015) 179 *Criminal Law and Justice Weekly* 538 (Part I) and 564 (Part II).

[75] Roberts and Zuckerman, *Criminal Evidence*, ch 15.

[76] See e.g., *R* v *Henderson* [2010] 2 Cr App R 24, [2010] EWCA Crim 1269, [215]–[220]; *R* v *Flynn and St John* [2008] 2 Cr App R 20, [2008] EWCA Crim 970; *R* v *Luttrell* [2004] 2 Cr App R 31, [2004] EWCA Crim 1344, [42], [43]: "The general principle... is that a "special warning" is necessary if experience, research or common sense has indicated that there is a difficulty with a certain type of evidence that requires giving the jury a warning of its dangers and the need for caution, tailored to meet the needs of the case. This will often be the case where jurors may be unaware of the difficulty, or may insufficiently understand it. The strength of the warning and its terms will depend on the nature of the evidence, its reliability or lack of it, and the potential problems it poses."

[77] Notably, *R* v *Doheny and Adams* [1997] 1 Cr App R 369 (CA). In relation to contested LTDNA profiling evidence, see *R* v *Thomas* [2011] EWCA Crim 1295.

The trial judge in *Broughton* was faced with the task of directing the jury in relation to a disagreement between the expert witnesses regarding the possibility that the DNA sample in question may have been a mixed profile. The expert witness called by the prosecution testified that she was satisfied, on the basis of her experience, that rogue profiling results obtained during the analytical process could be set aside as artefactual stochastic effects. Expert evidence adduced by the defence challenged this conclusion. It was argued that profiling results were consistent with the presence of an unidentified donor, and since the possibility of a mixed sample could not be ruled out, the match probabilities quoted by the prosecution's expert were invalid. The trial judge in his summing-up reminded the jury of this disagreement, which had been characterized as a legitimate difference of opinion between genuine experts. "In other words," he explained, "there is no, as it were, answer at the back of the book. There is no independent machine if people hold contrary views to tell you in these circumstances who is right and who is wrong. It is a question of expert evidence and scientific judgment..."[78] The judge added that, if the jury were not satisfied by the prosecution expert's expression of scientific judgment, then her "statistics"[79] could not be relied upon, and the jury could not substitute its own calculations "because you are not experts."[80] In that event, the jury would need to approach the matter cautiously, assessing the probative value of DNA evidence in the absence of any quantified RMP.

Readers of this scientific journal might well be thinking that this direction was incoherent, as a DNA "match" may be close to meaningless, or at least dangerously misleading, in the absence of a valid RMP. The Court of Appeal thought so, too, and concluded "with considerable regret"[81] that the appeal must be allowed and Broughton's conviction quashed on this, relatively narrow, ground:

> [T]he judge... fell into error in directing the jury that, in those circumstances, they could reach their own conclusions on the DNA evidence. It is fair to say that the judge urged the jury to exercise caution and be very careful in arriving at firm conclusions because they were not experts in statistics. However, we believe that only served to emphasise the void in which they were left. They had no guidance from the experts and no guidance from the court to enable them to conduct an evaluation of the evidence for themselves.... [T]he judge ought to have directed the jury that if [the prosecution's expert] was wrong in her conclusion that the DNA profiles were single rather than mixed, then on the only evidence before the court at the trial the DNA evidence must be disregarded. The judge having failed to do so, the jury may well have embarked upon a task of evaluation for which they were not equipped. This means their verdict cannot be regarded as safe[82].

Broughton underlines the point that admissibility is not the only important evidentiary issue raised by LTDNA profiles. The way in which profiling evidence is communicated to lay factfinders is also of fundamental importance if jury verdicts are to secure adequate epistemic warrant and broader normative legitimacy. In all other respects, the trial judge's "admirable summing up" in *Broughton* had "expertly addressed all the evidence and the complex issues in clear terms about which no complaint... could possibly be made."[83] A single slip was fatal. Summing up in relation to relatively novel and somewhat complex technologies like LTDNA profiling evidence is evidently a minefield for trial judges. Without a firm grasp of *both* the underlying science of profiling *and* the statistical foundations and probabilistic logic of valid RMPs, trial judges may inadvertently put a foot wrong, with potentially tragic consequences.

This case history might be interpreted, especially by readers more accustomed to inquisitorial procedures (scientists and civilian jurists alike), as a cautionary tale about the hazards of disaggregated tribunals in criminal adjudication and the perils of fastidiously microscopic appellate scrutiny of the wording of judicial directions to juries. These charges are not without substance; but the common lawyer has this riposte. In the absence of any parallel procedure in continental criminal trial proceedings, wherein lies the assurance that judges have any better understanding of the logical foundations of LTDNA evidence and can competently assess its probative value? Do reasoned judgments typically contain sufficiently detailed "motivations" to enable such assessments to be made, by an impartial observer or by the public at large? One could only begin to answer such questions through sustained research and on a jurisdiction-by-jurisdiction basis, but my own fragmentary and partly anecdotal acquaintance with judicial practice in continental Europe suggests that these are pertinent questions to add to our shared research agenda.

CONCLUSION

Forensic DNA profiling demands cooperative interdisciplinary expertise in forensic science, statistics and law. This article has reviewed UK courts' responses to LTDNA profiling, starting with initial skepticism in *R* v *Hoey*,[84] but—with the benefit of more considered official review and expert input—quickly producing authoritative statements endorsing admissibility. English courts proceeded in accordance with their tried-and-tested pragmatic method of *ad hoc* development of common law tests, approaching LTDNA profiling evidence in much the same way as DNA evidence itself was first addressed 30 years ago[85]. Some loose ends left dangling by the Court of Appeal in *Reed*[86] were tied up in *Broughton*, to produce the following doctrinal

[78] Quoted in *R* v *Broughton* [2010] EWCA Crim 549, [41].

[79] i.e., RMP calculations, (mis)characterized by the trial judge as "the statistical figure that has been given as a match probability": ibid [43].

[80] Ibid.

[81] Ibid [49].

[82] Ibid [48], [49].

[83] Ibid [49].

[84] *R* v *Sean Hoey* [2007] NICC 49.

[85] See e.g., *R* v *Gordon* [1995] 1 Cr App R 290 (CA). For historical discussion going back to the first trial in which DNA profiling evidence was adduced in 1987, see Paul Roberts, "Forensic Science and Criminal Justice" in Anthea Hucklesby and Azrini Wahidin (eds), *Criminal Justice* (OUP, 2/e 2013); Peter Alldridge, "Recognizing Novel Scientific Techniques: DNA as a Test Case" [1992] *Criminal Law Review* 687.

[86] *R* v *Reed and Reed*; *R* v *Garmson* [2010] 1 Cr App R 23, [2009] EWCA Crim 2698.

conclusion (if it is possible to create a legal precedent in relation to questions of fact, this is it):

> [T]he science of LTDNA is sufficiently well-established to pass the ordinary tests of reliability and relevance and it would be wrong wholly to deprive the justice system of the benefits to be gained from the new techniques and advances which it embodies, in cases where there is clear evidence... that the profiles are sufficiently reliable[87].

Reliability, moreover, is primarily a function of the *quality* of the profiling evidence in the instant case, as vouchsafed by experienced experts. There is no arbitrary stochastic threshold above which LTDNA evidence is admissible, and below which it is automatically excluded.

There is, however, much more to be gleaned from English jurisprudence on LTDNA profiling evidence than these "headline" *rationes decidendi* (formal legal holdings). Judgments rendered by common law courts are complex pieces of legal literature that must be interpreted against a backdrop of "thick" institutional practice and cultural meaning. Some of the factors in play, including the structural logic of the argument from authority sketched in the first part of this article and the priority of normative over epistemic considerations in criminal adjudication, are universal features of modern legal systems. Other factors reflect more local dynamics pertaining to the structural logic of criminal procedure, national legal traditions, and broader features of culture and society (including those features inflecting local apprehensions of adequate epistemic warrant for criminal verdicts). The second half of the article surveyed the principal arguments and judicial rationales that have been deployed in English criminal appeals concerned with LTDNA profiling evidence, pointing out their broader institutional context and resonances and explaining why some gained traction whilst others were rejected. The issues, we saw, are not confined to considerations of scientific validity, contamination risks and evidential integrity, and associated judgments of legal admissibility or exclusion. They also crucially concern the manner in which LTDNA profiling results are presented and explained to lay factfinders in criminal trials.

[87] *R* v *Broughton* [2010] EWCA Crim 549, [36].

If opinions differ concerning the adequacy of English courts' responses to LTDNA evidence, this may in part reflect divergent understandings of the deeper structural logic and values of criminal adjudication. These deeper structures are always engaged, and ought to be elucidated and consciously considered, whenever the admissibility and uses of expert evidence are placed under the policy microscope or raise novel legal issues for courts. Because policy questions are fundamentally normative (within the domain of political morality) rather than factually empirical or "scientific," legal jurisdictions must, in the final analysis, decide what is best for themselves, within the broad parameters of international legal consensus on fundamental rights and democratic values and in harmony with local juristic traditions and cultures. But just as surely as the fact that technical standards of DNA profiling or statistical science cannot dictate the terms of criminal justice, modern legal systems committed to post-Enlightenment conceptions of fact-finding and proof must necessarily rely on the best available scientific and other technical advice, communicated via competent, domain-specific expert evidence, to underpin the rationality (*qua* epistemic warrant) of criminal adjudication.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

I am grateful to two reviewers for constructive feedback and suggestions for more effective communication across disciplinary boundaries.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 12 December 2016
doi: 10.3389/fgene.2016.00215

Evaluation of Forensic DNA Traces When Propositions of Interest Relate to Activities: Analysis and Discussion of Recurrent Concerns

Alex Biedermann[1], Christophe Champod[1], Graham Jackson[2], Peter Gill[3,4], Duncan Taylor[5,6], John Butler[7], Niels Morling[8], Tacha Hicks[1], Joelle Vuille[9] and Franco Taroni[1]*

[1] Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, Lausanne, Switzerland, [2] School of Science, Engineering and Technology, Abertay University, Dundee, Scotland, [3] Norwegian Institute of Public Health, Oslo, Norway, [4] Department of Forensic Medicine, University of Oslo, Oslo, Norway, [5] Forensic Science South Australia, Adelaide, SA, Australia, [6] School of Biological Sciences, Flinders University, Adelaide, SA, Australia, [7] National Institute of Standards and Technology, Gaithersburg, MD, USA, [8] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, [9] Faculty of Law, University of Neuchâtel, Neuchâtel, Switzerland

OPEN ACCESS

Edited by:
Mogens Fenger, University Hospital of Copenhagen, Denmark

Reviewed by:
Wei-Min Chen, University of Virginia, USA
David Albert Lagnado, University College London, UK

***Correspondence:**
Alex Biedermann
alex.biedermann@unil.ch

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 August 2016*
Accepted: *23 November 2016*
Published: *12 December 2016*

Citation:
Biedermann A, Champod C, Jackson G, Gill P, Taylor D, Butler J, Morling N, Hicks T, Vuille J and Taroni F (2016) Evaluation of Forensic DNA Traces When Propositions of Interest Relate to Activities: Analysis and Discussion of Recurrent Concerns. Front. Genet. 7:215. doi: 10.3389/fgene.2016.00215

When forensic scientists evaluate and report on the probative strength of single DNA traces, they commonly rely on only one number, expressing the rarity of the DNA profile in the population of interest. This is so because the focus is on propositions regarding the source of the recovered trace material, such as "the person of interest is the source of the crime stain." In particular, when the alternative proposition is "an unknown person is the source of the crime stain," one is directed to think about the rarity of the profile. However, in the era of DNA profiling technology capable of producing results from small quantities of trace material (i.e., non-visible staining) that is subject to easy and ubiquitous modes of transfer, the issue of source is becoming less central, to the point that it is often not contested. There is now a shift from the question "whose DNA is this?" to the question "how did it get there?" As a consequence, recipients of expert information are now very much in need of assistance with the evaluation of the meaning and probative strength of DNA profiling results when the competing propositions of interest refer to different activities. This need is widely demonstrated in day-to-day forensic practice and is also voiced in specialized literature. Yet many forensic scientists remain reluctant to assess their results given propositions that relate to different activities. Some scientists consider evaluations beyond the issue of source as being overly speculative, because of the lack of relevant data and knowledge regarding phenomena and mechanisms of transfer, persistence and background of DNA. Similarly, encouragements to deal with these activity issues, expressed in a recently released European guideline on evaluative reporting (Willis et al., 2015), which highlights the need for rethinking current practice, are sometimes viewed skeptically or are not considered feasible. In this discussion paper, we select and discuss recurrent skeptical views brought to our attention, as well as some of

the alternative solutions that have been suggested. We will argue that the way forward is to address now, rather than later, the challenges associated with the evaluation of DNA results (from small quantities of trace material) in light of different activities to prevent them being misrepresented in court.

Keywords: interpretation, probative value, hierarchy of propositions, probability assignment

1. INTRODUCTION

1.1. Topic of the Discussion

This paper deals with perceived obstacles and potential solutions in the evaluation of the probative value of forensic biology results, such as DNA profiles[1], when the competing propositions of interest relate to activities rather than the source of the recovered trace material. So-called source level propositions deal with the origin of traces, for example, "The bloodstain on the broken window comes from Mr. A" vs. "The bloodstain comes from an unknown person[2]." In turn, examples of so-called activity level propositions, as they are understood here, are "Mr. A punched the victim" vs. "The person who punched the victim shook hands with Mr. A," or "Mr. A had sex with Ms. B" vs. "Mr. A and Ms. B attended the same party, and they had social interaction (i.e., shook hands) only[3]." At first sight, the evaluation of DNA results given (sub-) source level propositions is often more straightforward because it requires little more than a careful assessment of the rarity of the corresponding analytical features in the relevant population, and because well accepted models, data and software are available. This is different in the context of activities, as can be shown through formal analyses of expressions for probative strength (e.g., Evett, 1984; Evett et al., 2002). These formulaic developments show that it is necessary to extend the consideration to additional aspects, such as background presence of DNA and phenomena of transfer and persistence. Such additional factors are widely regarded as challenging and difficult to overcome (see Meakin and Jamieson, 2013 for a review). In essence, the concern perceived among practitioners is that the additional factors cannot be assessed appropriately (e.g., because of a lack of data). Therefore, the evaluation of DNA profiling results with respect to propositions regarding activities is considered not feasible or robust enough, and should be advised against. Clearly, following precepts and ethical considerations stipulated by codes of conduct (e.g., ENFSI Board, 2005; National Commission on Forensic Science, 2016), scientists are driven by their intention to inform recipients of expert information to the best of their knowledge so that no unwarranted conclusions will be reached. Laudable though this aim might be, there remains considerable diversity in opinions about the extent to which such results may be used, and how to report them.

Evaluation of scientific results with activity level propositions represents an important topic for current forensic science practice. Rather than dismissing the topic, we believe that it is necessary for the field as a whole to engage actively and submit the underlying issues to detailed analyses. The discussion presented in this paper aims at promoting and facilitating mutual understanding, which we hope will enable progress along new and feasible avenues. Not pursuing this topic bears the risk of leaving recipients of expert information without guidance. Reliance on recipients' own devices is prone to conclusions that are based on (sub-) source level propositions being wrongly carried over to conclusions about activity level propositions.

1.2. Objectives

The aim of this paper is twofold—firstly, to discuss recurrent concerns and reservations about, and sometimes fear of, evaluations of probative value with respect to propositions about activities and, secondly, to discuss alternative "solutions" that have been offered. Although we do not contest that challenges can arise in practice, we will argue that the central claims of the critiques cannot be sustained across the broad diversity of aspects of interpretation with activity level propositions. In particular, we will argue that some of the perceived drawbacks are sometimes the result of misunderstanding about the role of forensic scientists and forensic science in the legal process. Further, we will stress that it does not follow from the perceived deficits that evaluations given activity level propositions should be abandoned altogether, but areas need to be defined where additional research and support for practitioners is needed. The main motivation for this perspective is that it is by helping address activity level propositions that forensic science can offer more value to the criminal justice process, in terms of more focused and useful contributions. Moreover, this is a good way to assess all the scientific results[4] in any one case, ensuring that conclusions given by scientists do not run the risk of misleading at the evaluative stage[5]. The suggested framework provides a transparent way for experts, whether they be appointed by the court or hired by the prosecution or defense, to evaluate a case, where differences of opinion may be discussed and resolved. Courts need to provide a forum for such discussions to take place.

[1]DNA is chosen here because it is widely practiced. It goes without saying, however, that our discussion is equally applicable to other types of transfer trace materials, such as glass fragments, fibers or gunshot residues (GSR).

[2]Note that sub-source level propositions (Evett et al., 2002) are defined by replacing the bodily tissue "trace" (e.g., bloodstain) in the proposition by "DNA." For the remainder of this paper, we will focus on activity level propositions so that the distinction between source and sub-source levels is irrelevant. See also Cook et al. (1998) on the concept of hierarchy of propositions.

[3]It is worth indicating here that case circumstances are as important as propositions and that one will need indications such as alleged activities and timing.

[4]By all results we mean not only the DNA profile, but also aspects such as the quality and quantity of staining, and the position where it was found.

[5]Throughout this paper we will, at times, refer to the expression "evaluation in court" even though we intend our arguments to apply to the evaluative stage at large which, in some judicial systems, does not necessarily take place in court.

The paper is organized as follows. In Section 2 we present and discuss several recurrently expressed concerns. We broadly group these discussion points in three subsections, dealing with propositions (Section 2.1), data (Section 2.2) and aspects of reporting (Section 2.3). These three themes often arise in a hierarchy. Indeed, without propositions, it will be difficult for scientists to know how forensic science might help in a case. Next, with no or limited data, scientists may be reluctant to evaluate their findings. Finally, scientists may disagree about the form and content of scientific reporting, i.e., what exactly—if anything—should be reported. The issues and possible solutions explored in the three subsections are intimately linked and cannot truly be considered in isolation. Inevitably, there is some repetition between the subsections. Conclusions are presented in Section 3.

2. DISCUSSION OF SELECTED ISSUES

2.1. Propositions: "We Don't Know What the Exact Activities Are"[6]

It is often the case that scientists will be informed about the competing propositions regarding activities alleged by the parties only at trial, if at all (Risinger, 2013). It is generally understood that the propositions of interest are "(...) set by the specific case circumstances or as indicated by the mandating authority" (Willis et al., 2015, p. 6), but limited cooperation by the defense often represents an obstacle in practice[7]. If propositions are not available, or only one proposition is available, scientists should make every effort to obtain relevant information[8] regarding the position of each party involved in the process (see Willis et al., 2015 for complete guidance on how to deal with the absence of propositions), because without at least a pair of propositions, it is impossible to evaluate forensic observations in a balanced way.

It is a common misconception that the scientist who is evaluating the observations in light of competing posited activities needs to know every aspect of what has allegedly happened. For example, if it is alleged that the suspect grabbed the victim, aspects of DNA transfer will be important in considering activities. Only rarely will the scientist be provided with an exact recount of the position of grabbing, the force used to grab, the exact length of time the struggle lasted, and so on, to cover all aspects of the alleged encounter. However, there are several aspects that should be considered. Firstly, the manner in which the activities are, or have been, set up in controlled experiments to mimic the activity of interest are likely also to have similar aspects of uncertainty. Therefore, the uncertainty arising from many of the unmeasurable (and unknowable) aspects of the alleged activities will present themselves in the spread of the obtained data. Secondly, controlled experiments can be set up to study the impact that different factors have on the transfer of DNA during the activity in question. It may be that there is a large enough amount of variation in one aspect of the activity, such as the shedder status of the individual, that all others (such as the length or vigor of contact or when the person of interest last showered, etc.) have a negligible effect on the evaluations. In that case, this level of activity resolution is not required. Thirdly, if a number of aspects are found to have a considerable impact, then they can be included in the logical framework used to evaluate the findings. If the actual states of these important factors are not known (or not provided) by either party then they can be incorporated by considering all possible states within the evaluation, weighted by probabilities informed either by data from controlled experiments, supplemented by the analysts' knowledge, which should be available for disclosure and auditing (see also Section 2.2.2). If further information is provided later on, then the evaluation can be updated accordingly. Alternatively, sensitivity analyses can be used to determine how much of an effect any one of the unknown factors of the activities has on the value of the findings (Biedermann and Taroni, 2006). If the strength of the observations is particularly sensitive to some aspects then efforts should be made to find additional information about those aspects rather than every aspect of the activity. If scientists do not have such specialized scientific knowledge, the court will be even less likely to have such knowledge.

2.2. Data

2.2.1. "Because Each Case Has Its Own Features, the Use of Numerical Values from Experimental Studies Performed under Controlled (Laboratory) Conditions Cannot Be Used for Evaluation in Real-Life Cases"

This is a general claim (e.g., Jamieson, 2011) that conflicts with scientific practice. Throughout science, experiments are conducted in trials that reflect not all, but the essential, features of the problem at hand. Clearly, medical treatments administered to patients have not previously been "tested" on those particular individuals, but on other patients with the same disease. Similarly, the safety of consumer products (e.g., cars) is carefully assessed not by end-users but prior to marketing in a range of situations reflecting end-user profiles. Turning to forensic science, such as glass analysis, the phenomenon of transfer has been studied for a variety of factors, such as the mode of breaking (e.g., the number of blows), window dimension, etc. to build a model usable for assigning transfer probabilities in cases with features covered by this model (Curran et al., 1998, 2000). In the context of DNA, studies have been conducted to examine the rates of transfer, for example between shooters and guns (Polley et al., 2006), but also in more general situations (e.g., Phipps and Petricevic, 2007; Daly et al., 2012; Jones et al., 2016; Samie et al., 2016). So, when a scientist is faced with assigning a probability for finding trace material given the proposition of handling an object by a person of interest (e.g., the *activity* of discharging a firearm), we do see no harm in referring to studies that have focused on rates of transfer not exactly the same in the alleged circumstances of the case. Although some features of the individual case at hand may differ, nothing will

[6]The titles throughout Section 2 reflect our perceptions and summaries of recurrently encountered concerns. Specific references are cited within each subsection.

[7]In all criminal justice systems that we know of, it is the right of defendants to remain silent and not incriminate themselves. Also, it is often considered strategic for the defense to make a statement as to the evidence only after the results of the analyses are known.

[8]It is generally understood that scientists should concentrate on information that is relevant for the task at hand (i.e., so-called task-relevant information), keeping in mind challenges posed by human factors.

prevent the scientist from also judging that some additional case-tailored experiments should be conducted in order to extend their knowledge and understanding, but case backlogs and limited resources may render this difficult. Besides, if a scientist refuses to assign a probability of observing some finding under a given set of propositions, there is a risk that the fact-finder will nonetheless assign such probabilities according to their own unaided judgment which, as highly-publicized past cases suggest, will often play out to the detriment of the defendant.

Contrary to a widely held view (e.g., Budowle et al., 2012), the availability of "hard" (i.e., numerical) data is not a necessary requirement for probability assignment. That is, the absence of data does not mean that no probability can be assigned. How is this possible? To understand this, it is important to recall that scientists can also derive probability assignments from their understanding of the principles of the process at hand, formulated in terms of a model. For example, weather forecasters cannot "play" the next day over and over again to find the number of times there will be rain on exactly the next day. This makes no sense, essentially because there is only one tomorrow and its weather will be observable only once tomorrow arrives. And yet, scientists are able to formulate previsions about the state of the atmosphere based on related data (e.g., today's weather) and their understanding of the relevant science and technology. Similarly, forensic scientists can make probability statements for outcomes based on their general understanding of a phenomenon. In genetics, for instance, there is knowledge about population structure and the ways in which genetic traits are passed on between generations. Take, for example, a crime stain with a haplotype for which no other occurrence is found in a relevant database: clearly, the observed relative frequency (i.e., zero) in the database does not mean that we should retain a zero probability for observing the same haplotype in another person from the population of interest. Instead, a value for the haplotype population proportion can be obtained using reasonable assumptions (Brenner, 2010).

In reply to the above, it may be suggested that forensic genetics is not an insightful example because of the sophisticated mathematical models available in this area, so let us consider the case of DNA transfer phenomena. Here, many studies have found, for example, that the quantity of DNA recovered after touching (i.e., primary transfer) a surface with bare hands varies approximately between zero and more than 150 ng, depending on the experimental conditions (e.g., Daly et al., 2012). But can such knowledge help formulate probabilities of finding DNA under the assumption of secondary transfer? We contend that it can, by constructing an argument. It is known, for example, that secondary transfer is conditioned on the amount of DNA transferred initially. Hence, when a quantity above, say, 200 ng is found—something not typically expected when touching with bare hands (i.e., primary transfer)—it can be argued that this should be considered an even less probable event assuming secondary transfer. Thus, despite the fact that explicit data for a particular secondary transfer scenario may not be available, forensic scientists can still convey domain-relevant knowledge intelligibly in probabilistic terms. It is important, however, to ensure that the data are relevant to the analytical methods used in the case of interest, because detected quantities after secondary transfer will be sensitive to the method used to collect material and to detect DNA. It will also be necessary to ensure that probabilistic models, such as Bayesian networks (see also below), used to interpret findings, are informed by such relevant data and available for auditing.

There remain, however, justified questions as to where and how to obtain data. Currently, results from empirical research are mainly published in peer reviewed scientific journals, but there is no systematically organized body of research. The idea of developing a knowledge base (Evett, 2015), to be shared among scientists who all contribute to the system, would thus represent a major contribution to strengthen the data-supported evaluations.

In summary, probability assignment is feasible and justifiable even with limited data, but must be amenable to a critical analysis. Further, despite the fact that data are collected under conditions that do not exactly reflect all the features of the case at hand, this does not preclude, in principle, these data from being used at least to some extent[9]. Of course, this does not mean that any data are acceptable to support any claim, but—as noted above—data that the scientist regards pertinent. Moreover, expert assessment is not exclusively given by data alone; in fact, it never is because, while reference data have been collected in controlled studies, the probabilities we assign relate to one-off individual incidents. Instead, scientists use data to inform their judgment, by constructing an argument, explaining what data they have used, to what extent and why.

A topic related to the above viewpoint is the question of how to conduct assessments, i.e., reasoning in the face of uncertainty, whatever the data may be and the extent to which they are available. Often, one can note that scientists shy away from seeking support from conceptual devices that could help them structure their reasoning and, thus, avoid the impression of being overwhelmed by the inferential complexity of the evaluative task. It is thus worthy to mention one common method, known as Bayesian networks (BNs) (Evett et al., 2002; Biedermann and Taroni, 2012; Fenton and Neil, 2013; Taroni et al., 2014), for pulling together many aspects of information that need to be considered when activity level propositions are of interest. BNs are a graphical tool in which the problem can be constructed in a framework of logical inference[10]. The formulation of such a framework does not rely on having any data, it will in fact inform the analyst of what data is required in the evaluation of the findings. The formulation of a framework of inference should be the first step in any evaluation given activity level propositions. Sometimes, however, analysts claim that an insufficient amount of data exists, and they do so even before they know what data is

[9]On this point, see also Casey et al. (2016), who argue "(...) not evaluating DNA evidence in case work is potentially more dangerous and reckless than carrying out an evaluation based on limited datasets." We concede that this topic is delicate and it is important not to suggest that scientists allow themselves to suggest an answer as they please and hide behind statements such as "I am an expert."

[10]While there is substantial literature available on BNs for evaluating forensic DNA results (e.g., Biedermann and Taroni, 2012), BNs remain inaccessible to many biologists, mainly because of lack of training. An easily accessible repository with freely available examples that can be utilized with open-source software would be an asset to complement the idea of a knowledge base, as mentioned above.

actually required to help address the issues that are of interest in the case. Also, once constructed, and lack of data is found to be an issue for some aspects of the evaluation, there are still several avenues open to the scientists. These include the weighing of the various states that potential factors may take in terms of probabilities, informed by the scientist's documented knowledge and experience, and then conducting sensitivity analyses to determine how the evaluation changes as those probabilities vary over plausible ranges.

Reluctance to such introspective thinking, and expert probability elicitation in general, is surprising and odd to see among those scientists who would find no objection to be asked and to give opinions in court about the probabilities of various competing activities, regardless of whether relevant data exists, and regardless of whether they have undertaken some activity level consideration. In this situation, analysts are likely to express themselves non-numerically in the form of an answer such as "that is improbable," "I don't believe that is likely to have happened," "I think that sort of transfer is barely feasible," etc. So, analysts who would be willing to express themselves in such probabilistic terms about propositions, but refuse to provide probabilities for findings given propositions, exhibit an inherently contradictory position. We thus maintain that analysts who have considered their findings in a framework of logical inference, using experimentally derived data to assign probabilities and varying assignments for influencing factors will be in a much better position to usefully inform the court. This will include any limitations that characterize the data actually used (as well as detailing the information available to the scientist at the time of writing the report, see also Section 2.3.3).

2.2.2. "Expert Professional Experience Is Not Enough (Data) To Safely Assign Probabilities"

A recent exchange (Casey et al., 2016; Meakin and Jamieson, 2016) raised the latent issue of whether, and to what extent, expert experience forms an acceptable basis for assigning probabilities. As asserted in Champod (2014), and reiterated recently in Meakin and Jamieson (2016), the critical issue is disclosure of data and making it available early enough in the process in order to allow for a proper consideration by the defense. The deeper issue, however, appears to lie in the notions of expert experience and so-called "personal" probability assignments. The ENFSI Guideline, for example, mentions expert experience as one possible source for informing the process of probability assignment: "Such data can take, for example, the structured form of scientific publications, databases or internal reports or, in addition to or in the absence of the above, be part of the expert knowledge built upon experiments conducted under controlled conditions (including case-specific experiments), training and experience" (Willis et al., 2015, p. 19). This should not be read as meaning that a vague reference to personal experience is on a par with other, more structured data. This would amount to misconceiving the fact that there is actually a hierarchy in the data, with a clear preference given to scientific publications and otherwise widely accessible scrutinized data.

2.2.3. "Evaluations Given Activity Level Propositions Are Massively Vague and Hence Cannot Be Trusted"

This objection may be the result of the discomfort that can be experienced when faced with incomplete knowledge about factors that influence the assessment of the probative value of scientific observations. However, incomplete knowledge, and hence uncertainty, do not *per se* prevent the conduct of science and its operational use in legal proceedings. What is more, in all parts of legal and everyday practice, one needs—inevitably—to act *despite* knowledge being incomplete. It is the very task of science, thus, to quantify the extent of available knowledge so that it can be used in an informed way. The reply "it's possible"[11] when confronted with the event of transfer or contamination, as scientists still often do in criminal proceedings, is a vague statement and is not a quantification of knowledge as we understand it in the discussion here.

The view that partial knowledge can be used is challenged, for example, when outcomes are subject to variation and scientists refrain from addressing them, equating variation with "no knowledge" about the topic. Forensic examination of glass provides a telling illustration for this. Research has shown that the quantities of glass fragments transferred to the surfaces of the clothing of the breaker vary considerably even for experiments with the "same" controlled conditions (e.g., regarding mode of breaking, distance between the breaker and the window, etc.). But does this mean that we "know nothing"? Clearly, scientists *have* knowledge about the phenomenon of glass transfer[12] in the sense that they won't expect to encounter all possible numbers of fragments with the same probability. For example, depending on factors such as the distance between the breaker and the window, the mode of breaking, the time since window-breaking and seizing a suspect's clothing etc., scientists may consider it more probable to recover less than, say, five glass fragments, rather than more than five. It is the scientists' core task to elicit and convey such expressions of expert knowledge, because no one else in the proceedings is in a better position to do this. It may be a challenge for scientists to provide probabilities for recovering exactly 0, 1, 2, ... fragments (although simulation approaches exist e.g., Curran et al., 1998), but it is feasible also to choose a strategy going from the general to the particular, starting with probability assignments for apportionments of fragments such as "none," "few," "some," and "many." This helps break down the difficulty of probability assignment and make particular assignments more intersubjectively acceptable.

What is important is not the variation *per se* but how different the expected outcomes are given both propositions. Imagine that 2 min after a window is broken, a person is arrested and his sweater searched for glass. More than 80 fragments (sharing the same physical properties as the broken window) are recovered. Is this result more probable given that he broke the window or given that he had nothing to do with breaking incidents? In

[11] The same applies to "could have"; and there have been several notable judgments where courts have ruled against the unqualified use of such phrases.

[12] A further relevant factor is background presence. Regarding DNA, there is limited knowledge about naturally present DNA in the environment, which is especially important in cases where the defendant and victim cohabit (as in the Amanda Knox case, for example Gill, 2016).

some breaking experiments, it was observed that the number of fragments transferred was between 44 and 241 (with a mean of 127, see Hicks et al., 1996). So, there is variation, but when one looks at persons who have not broken windows and searches their garments, one finds that, in general, on sweaters, there are only between 0 and 2 fragments (sharing the same characteristics, see Coulson et al., 2001). Thus, clearly, finding 80 fragments is much more probable given one proposition than the other, *despite* the variation observed. Using the data from both surveys, one can assign a probability to the results given each proposition. In cases involving DNA, most studies focus on one activity only, but what is important is the comparison of the probability of the outcomes given both the alleged activity and at least one alternative. This comparison will then enable us to see if the variation that we have observed has an impact or not on our conclusions.

One further point that warrants a comment is "vagueness." We would strongly advise against the use of this term as a qualifier for a forensic evaluation *if* that evaluation has been conducted thoroughly. Let us recall again the undisputed starting point: there is variation in findings. What scientists do is to accommodate this *through probability*. What does this mean? It means that scientists will assign probabilities to the various outcomes depending on the extent to which they expect them to occur. For example, consider a case in which a victim has been punched several times to the head, resulting in profuse bleeding. Given the proposition that the suspect (arrested immediately after the assault) is the assailant, we can postulate one main, potential outcome: finding, on the suspect's fist, blood with a DNA profile corresponding to that of the victim. A fairly high probability (i.e., toward the upper end of the range between 0 and 1) may be assigned for this particular finding. By coherence, other findings such as no blood at all, or blood with a profile different from that of the victim will thus be assigned lower probabilities (i.e., toward the lower end of the range between 0 and 1). So, there is variation in the potential findings but there is a major finding that dominates our expectations. Consider now a different version of this case, in which the assault was less violent, did not result in the bleeding of any protagonist, but involved several victims. In such a case, it may be necessary to specify a broader range of potential findings, possibly including mixtures. In the event that the person of interest is the assailant, there may not be one outcome that stands out over all the others. Instead, one's probabilities for several of these potential outcomes may be quite similar. Thus, probability will be distributed over several outcomes, with no outcome receiving a probability assignment close 1. The assigned probabilities will *express less strong beliefs* (in the various outcomes) but—and this is the important point—*this does not mean that the assignment as such is vague*. Less strong beliefs simply reflect the fact that one is less affirmative. Every statement of probability expresses a particular state of uncertainty, that is a well defined opinion, but none of these expressions are deficient if they are derived properly to reflect the expert's current state of knowledge.

So, even if the scientist may report a neutral finding[13], due to limited expert knowledge (to enable the assignment of probabilities that are different in the numerator and the denominator), this is still an important evaluation to present to the fact-finder. If nothing else, it will inform the decision-maker that they need rely on non-DNA evidence to decide the case.

2.3. Reporting

2.3.1. "It Is Impossible to Know from the Quantity of DNA Obtained, and the Quality of the Profile, Whether the DNA Was Deposited by Direct Contact or Indirect Transfer."

The concern expressed in the section title is also sometimes seen as the problem of whether we can determine or, as noted by some discussants, "deduce" whether a given finding is the result of primary or secondary transfer. The misconception here is not to understand that the process is not deductive[14], but remains inductive. Hence one cannot "know for sure"—but one can offer guidance, in the form of probabilities for the results, to help fact-finders decide on the truth of the propositions of interest.

More generally, the claim that particular observations do not allow one to draw categorical conclusions about a particular activity is uncontested and also holds for many, if not all, types of forensic traces. Taking glass as an example, no proficient forensic scientist would conclude that finding a number x fragments is the result (or the probable result) of smashing a given window. Similarly, finding a number y particles of gunshot residue does not allow one to say that the person of interest discharged a firearm, to the exclusion of other propositions. The impossibility of such direct "jumps" from observations to conclusions in these examples does not derive, however, from the fact that the trace material is present in small quantities. The shortcoming in the reasoning also holds for the so-called macro-traces. To illustrate this point, imagine that large quantities of fresh blood are observed on the hands of a person of interest. Such a result does not entitle one to argue that the exclusive or probable cause is stabbing the victim. Depending on the case circumstances, trying to help the victim may also be a viable proposition.

As discussed, the scientists' task, when operating in evaluative mode, is not to "infer activities" but to provide expressions of probative strength to help the court discriminate between competing propositions regarding activities. This requires the scientist to assign probabilities for the DNA results as obtained in the case at hand given each of the propositions of interest. The fundamental question associated with probative value then is: "Under which of the competing propositions regarding activities do we consider the findings more probable?" It may be that scientists think that they have no reason to consider the observations more probable in one version of the events than another. But this will not be a defect of reporting given activity level propositions, nor of the framework of evaluation. It only means that, in the current state of knowledge, the findings do not have any discriminative capacity (in a technical sense, such results would have a likelihood ratio of 1). As discussed, this is a well-defined result and should be reported, so that people are not prosecuted on the basis of forensic results that are not probative at this stage. As much as it is useful for a recipient of expert

[13] In more technical language, this would correspond to a likelihood ratio of 1.

[14] On the notion of deductive logic, see also Jackson et al. (2013), for example.

information to hear when observations support one proposition rather than another, and (if possible) to what extent, it is useful for them to know when findings do *not* allow them to alter their beliefs in the propositions of interest.

In other terms, it is not a matter for the scientist to say whether a proposition, such as "Mr. A stabbed the victim (i.e., the DNA is from primary transfer)," is true, given the forensic observations, but the extent to which she expects to see these observations, given the proposition "Mr. A stabbed the victim." The scientist should be assessing the probability that DNA would be transferred, that it would persist and that a matching profile would be obtained, given the truth of this proposition. But there is one more dimension to the latter question. In order to be balanced, scientists must not only think about their results, given one activity, but also given at least one alternative activity (for example, that the suspect handled the knife innocently after the incident), and assess whether, and if so to what extent, the observations are more probable given one activity rather than another. It is therefore of paramount importance that scientists do not confuse the probability of primary transfer with the probability of observing the results if Mr. A stabbed the victim. We agree that the difference is very subtle and this is a reason why, in the propositions, one ought to describe activities and not use the terms "primary/secondary transfer[15]." This allows one to distinguish, on the one hand, what the court will assess (i.e., activities), and on the other hand, what the scientist will assess, that is the probability of observing the results given the activities. One of the terms used to assign the latter probability is commonly known in the literature as the "transfer probability" (Evett, 1984; Evett and Buckleton, 1989)[16]. We thus stress that a transfer probability focuses on the findings, given propositions, *not* the reverse.

2.3.2. "You Cannot Say That He Stabbed the Victim! The Only Thing That DNA Allows You to Say Is That He Had Recent Direct Contact"

This objection is often heard from recipients of expert information when the propositions of interest in the scientist's report are formulated closely to the specific actions that define the crime. For example, propositions such as "he handled the knife," "he punched the victim," "he fired the gun," may provoke such objections. It is often felt that less specific formulations such as "he is in contact with" are more appropriate. However, this objection stems from a misconception about the role of the scientist with respect to the propositions. As noted at the end of the previous Section, by writing down propositions in their report, scientists are only "setting the context" in which the findings will be assessed. That context is given to the scientists by the court and/or the parties and this context naturally relates to the alleged actions. Hence, scientists do not express any opinion directly on those propositions, notably regarding their truthfulness, adequacy or otherwise. The scientist's reporting only focuses on the weight to be assigned to the DNA findings in light of these propositions. Scientists should not suggest in any way that they are in a position to say, for example, that "he handled the knife," or that "he was in direct contact with the knife." If they do, they fall for the same fallacious thinking explained above. The only opinion they are allowed to express is in relation to the probability of the DNA findings if one or the other proposition is true. Specifically, when the scientist writes that "this amount of DNA is what we expect if Mr. A. stabbed the victim," the scientist is reporting only about the DNA results, and is not taking any stance on whether or not Mr. A. stabbed the victim. The latter is simply what is alleged by the parties in their own terms.

2.3.3. "Because Many Lawyers May Lack Awareness as to the Problem of Transfers, Analysts Should Flag the Issue in Their Reports Whenever the Analysis Process Suggests That Various Transfer Mechanisms May Explain the Findings"

Explaining the observations is a procedure that would be acceptable for the scientist to perform if they were at the investigative phase and not being asked to evaluate the forensic biological results in the context of the case, at court. To clarify this point, it is useful to recall the following two fundamentally different perspectives. In the investigative phase, observations are taken as a starting point. They are used to suggest what happened (i.e., activities). For example, one takes the finding of small quantities of DNA on the suspect's shoe as a starting point to suggest that the suspect was the person who kicked the victim. The other perspective takes propositions as a given (as it would be the case in court), to reason about the findings. One assumes that the suspect is the person who kicked the victim, and then one reasons about the kind of trace pattern one would expect to observe on the suspect's shoes. In evaluation, it is the latter perspective that is appropriate for forensic scientists. As noted by Margot, "[w]hether these results could be observed if one proposition for the event is true rather than another proposition is the central relevant matter on which the forensic scientist may comment" (Margot, 2011, p. 796). Note however that there may also be more than two propositions of interest.

At this juncture we would like to include a brief note on the distinction between explaining the observations[17] and evaluating them, as well as the difference between explanations and propositions (Evett et al., 2000a). We often hear that, after scientists do all the complex evaluations that activity level propositions may require, and provide their results on the stand, the defense are just going to suggest an explanation. For example, the defendant may argue that he spat on his hand as he was walking down the street and touched a bench on which the victim later sat, or some other explanation. It is worth stating that this is explaining the results, and that the defense[18] provides such explanations once the results are known. Therefore, such explanations are generally based on the results and may not be based on the relevant circumstantial information in the case. Such explanations do not count as acceptable,

[15] For more on this topic, we refer the reader to Hicks et al. (2015).

[16] Note that there are also phenomena of persistence and detection/recovery to be taken into account.

[17] While technically the word "observations" is to be preferred, we will be using the more colloquial word "results" to refer to the outcome of the scientist's analyses.

[18] The problem of *post-hoc* rationalizations is not restricted to the defense as explanations can also be brought up by the prosecution. See, for example, the bleach cleaning hypothesis in the Amanda Knox case (Gill, 2016).

formal propositions, because one cannot meaningfully assess the probability of the results given explanations that themselves are merged with the results (i.e., this would constitute circular thinking). Explanations are generated *post-hoc* in order to account for the results. They can be statements of the blindingly obvious, they can be speculative or they can be fanciful, having no logical connection with the circumstances of the case, even to the point of having no grounding in reality (see in particular Evett et al., 2000a and more recently Jackson et al., 2013 for examples and detailed discussion). In contrast, propositions are formal statements of competing allegations or suggestions that are dictated by the relevant circumstances of the case and not by the results themselves. So if changes are suddenly brought up at trial, scientists need to be careful not to give *ad-hoc* assessments where evaluation would require detailed checks with relevant literature and specialized knowledge. This is also why, for example, the ENFSI Guideline (Willis et al., 2015) emphasizes that scientists should mention in their report that their evaluation is based on their understanding of the relevant circumstances at the time of writing the report and that if any assumptions or information is incomplete or incorrect, they will have to re-evaluate their findings[19].

The above distinction between explanations and propositions is crucial and it is worth to summarize and relate it to standard notions from other inferential disciplines. Characteristically, explanations account for—or are made to "fit around"—the findings that have been made in a case. Explanations entail a deductive mode of reasoning as they seek to explain existing results, typically in causal terms. As such, forensic explanations are generated and considered after relevant observations are known. Often, the generating process for explanations results from abductive reasoning not limited to the forensic scientist, but may also extend to case investigators. As such explanations are theoretically open-ended, with no limit on their number though some of the explanations may be more or less fanciful (e.g., not testable in a logical sense), implausible or even speculative than others, to the point that no meaningful probability may be assigned to them. Unlike explanations, propositions are formal statements that can be clearly related to the case context and subjected to a proper inductive mode of reasoning.

We would suggest that consideration and proposal by a scientist of the various possible modes of transfer to account for DNA findings may be of use in the investigative phase of a case. But scientists should not *systematically* explore and comment on all conceivable mechanisms of transfer (so-called caveats) in their statements (but may do so within the lab-file, or in a "Technical Issues" section of the report Evett et al., 2000b). Moreover, when a case enters the evaluative phase, and particularly when in court, a scientist should resist offering a view on explanations for transfer but concentrate on evaluating probabilities for the results given formal propositions based on the circumstances of the case. Advancing explanations at the evaluative stage amounts to treating transfer dismissively, rather than considering its impact on probative value in a formal and explicit way.

2.3.4. "The Safest Course Is for an Analyst Simply to Report the Results of the DNA Test, Alert Both Counsel and the Jury to the Possibility of Transfer, and Leave the Jury or Factfinder to Assess Their Implications"

This argument is a similar to the previous one, about caveats, but here the burden of how to assess the implications of transfer is left to the factfinder. We are of the opinion that *leaving the factfinder to assess implications of transfer* threatens the appropriate conduct of the forensic findings: if scientists do not—or cannot—evaluate their results, then how could the factfinder do so? Hence we find this position problematic. Clearly, proceeding in this way is an easy course for scientists, because it reduces their task to technical reporting, but it could be very misleading for innocent defendants because findings will be left uninterpreted at the propositional level that really matters (i.e., activity level). Forensic scientists have (or should have) specialized knowledge on transfer and persistence, as shown by publications in forensic journals, and they therefore have the duty to report the value of their results at that level. If the knowledge is not sufficient, then scientists must tell the instructing magistrate or the court (or preferably even before, earlier in the process) that, as a consequence, their results do not help discriminate between the propositions at hand. We do not believe that—in the evaluative phase—scientists should provide a list of all theoretically possible modes of transfer of DNA (see also Section 2.3.3). If the scientist were to provide such a list, how does the court choose which is the most likely mode of transfer? This would leave the court in the difficult position of having to choose which of a potentially large number of possibilities (that are not necessarily exhaustive) is the most likely, without being able to rely on any specialized knowledge to do so. However, we do believe that it is the proper role of the scientist to talk generally about transfer and persistence of DNA (see also Section 2.3.3).

An intricacy related to the above is the use of the term "possible." As human beings, we refer to a lot of events as being "possible" (i.e., the probability of the event is not 0), but forensic scientists should be more informative than this: they should assess *how probable* their results are given the propositions at hand, just like they do when they assess the probability of observing a given DNA profile if it came from some unknown person. If a scientist were to be asked "what is the probability of obtaining a matching DNA profile if it came from Mr. A or if it came from someone else who happens to have, by coincidence, the same profile," which is an explanation, the scientist would have to answer that those two probabilities would be the same (i.e., approaching 1). Therein lies the problem for the scientist and the court, generating explanations leads to probabilities for the results of a value approaching 1. Provision of explanations is deeply rooted in general forensic science thinking and we regularly see reports in which the scientist writes "It is possible that this DNA comes from Mr. A. But, it is *also possible* that it comes from his brother or an unknown person." This sort of explanation-based answer is unsatisfactory because it leads the

[19]Interruptions in the proceedings can be granted by the court both in inquisitorial and adversarial proceedings, and will usually be granted if the question is of importance to the court.

scientist to opine directly on a proposition (see also Section 2.3.2 regarding the role of the forensic scientist).

2.3.5. "When Scientists Are Unable to Evaluate Their Observations Given Activity Level Propositions, Then They Should Retreat to Evaluations Given Source or Sub-source Propositions"

This claim rejoins Section 2.3.4, which refers to the claim that an evaluation given source level propositions is a "safe course" for forensic scientists, when they cannot help address activity level propositions. However, what a safe course of proceeding is for the scientist may again not be so for other participants in the process[20]. The problem is that the "safe course" for the scientist inevitably restricts the evaluation to the (sub-) source level. Consequently, the court is given no guidance about how to evaluate with respect to activity level propositions. If the court confuses the scientist's (sub-) source evaluation with the activity level evaluation, then there is a risk that this may lead to a miscarriage of justice[21].

For the above reason, recent recommendations by ENFSI specify that the choice of source level propositions is limited to well defined situations, that is "(...) cases where there is no risk that the court will misinterpret [the findings] (...) in the context of the alleged activities in the case" (Willis et al., 2015, p. 12)[22]. But for small quantities of trace material, this is rarely if ever the case, because such traces require expert knowledge "(...) to consider factors such as transfer mechanisms, persistence and background levels of the material which could have an impact on the understanding of scientific findings relative to the alleged activities" (Willis et al., 2015, p. 11). For all of these reasons, the ENFSI guideline concludes that "(...) the choice between (sub-) source and activity should not be influenced by the availability of data or expert knowledge but solely from the consideration of factors such as transfer, persistence and background levels that could crucially affect the strength of the findings within the context of the case circumstances" (Willis et al., 2015, p. 13). This includes a statement of limitations as to the data and the individual expert knowledge (see also Section 2.2.1).

An objection that may be raised against the position outlined above is its feasibility. That is, although activity level propositions may be recognized as the relevant propositional level, specialized knowledge necessary for evaluation given this level of propositions may be unavailable. A natural consequence of this starting point would be not to introduce the results at trial, in order to protect defendants against unwarranted interpretations in such cases. However, there appears to be no consensus among scientists on how to proceed in such situations. Some scientists maintain their intention to report findings given source level propositions *although* they are clearly unable to help with the issue of activities. As a consequence, we do not subscribe to this view *of retreating to evaluation given (sub)source level propositions*, and neither does the ENFSI Guideline, which requires scientists to clearly acknowledge that their evaluation falls short of the real needs. In Guidance note 2, the Guideline states that "(...) if the examiner chooses (...) to report the findings at source level (...), the examiner shall explicitly state that the rarity of the profile does not address the question of the relevance of the findings in relation to the alleged activity" (Willis et al., 2015, p. 14).

Proponents of the view according to which uninterpretable results should be mentioned at trial appear to misconceive the fact that different stages in the forensic process have different requirements (Jackson et al., 2006, 2013). It is beyond dispute that, at an investigative stage, scientists can help the process move on when they factually report about the observation that a defendant's traits are also observed in trace material (e.g., in the case of mixtures). This is useful information for selecting possible candidates on whom to focus further investigations. At trial, however, the requirements are different. At trial, the defendant has already been selected, and if DNA is to play any further role, it must be given a weight (Evett, 2015)—not against any propositions, but propositions at the relevant level.

2.3.6. "When Evaluating Forensic DNA Traces Given Activity Level Propositions, the Scientist Infringes on the Duties of the Court"

A perception encountered among legal practitioners, as noted earlier in Section 2.3.2, is that when evaluating DNA traces given postulated activities, scientists take on the role of the fact-finder. This observation is a cause of concern because it does not reflect the scientist's intention and laying bare this misconception is challenging. We think that there is merit in reiterating that it is not for the scientist to give an opinion on whether the transfer is primary or secondary (or the probability that the transfer is primary or secondary) because giving such an opinion would amount to giving an opinion on the propositions of interest, for example whether "Mr. A had sex with Ms. A" (transfer was primary) or "Mr. A and Ms. A attended the same party, but had no particular interaction" (transfer was secondary). Clearly, this is a question for the court.

The above distinction is very subtle, for all discussants, including scientists. It comes down, in one way or another, to the problem of the transposed conditional. Many authors have formally described the contribution of the scientist and the nature of expert opinion in the criminal justice system, with the one key aspect being that the scientists' role is to evaluate their results given the competing propositions regarding activities, and that it is for the court itself to assess the truth of the

[20] See for example cases such as Jama (Gill, 2014, p. 27) or the Ruelas Case (Murphy, 2015, p. 56) which illustrate that the sole consideration of sub-source issues does not accommodate and represent an evaluation of the joint probative value of the results from different swabs, the quantities of DNA found, nor the presence or absence of other trace materials. When considering activity level propositions, all observations should be assessed, which is what is needed.

[21] See http://www.scientificamerican.com/article/when-dna-implicates-the-innocent/ for recent case involving a stupefying situation in which a forensic DNA report found a correspondence between the DNA profile of a hospitalized person and DNA found on a murder scene. It appeared that the same paramedics treated the hospitalized person but also worked on the crime scene a few hours later (see also https://californiainnocenceproject.org/2013/06/how-an-innocent-mans-dna-was-found-at-a-crime-scene/).

[22] This may be the case, for example, when there is a large and fresh bloodstain at the point of entry (e.g., broken window) and it is not contested that the blood stain is relevant to the case (i.e., the trace is a direct consequence of committing the burglary).

propositions. Unless scientists are operating at the investigative stage, they should express probabilities only for their results given propositions, but not the reverse. An example of a relevant statement would be: "The probability of observing this quantity of DNA if Mr. A had sex with Ms. A as alleged by the prosecution is in the order of 0.6, whereas the probability of observing this quantity of DNA if Mr. A had social interactions as alleged by the defense is in the order of 0.01. This means that it is about 60 times more probable to observe this DNA result if the prosecution's case is true rather than if the case of the defense is true." However, there is the risk that the receiver of this information will interpret the low probability for the findings, given the alternative proposition, as meaning there is a low probability for the proposition, given the findings—a reasoning error that is known as "transposing the conditional," and which was not intended by the scientist. This is why some reporting agencies explicitly mention, in their written reports, examples of sentences of what their conclusions do *not* mean.

3. CONCLUSIONS

From the discussion presented throughout Section 2, three main points emerge:

First and foremost, forensic interpretation, as conducted by the scientist, focuses on the observations, not on the propositions. Stated otherwise, the question for the scientist is "What is the strength of these findings with regards to the propositions of interest?" The scientist should *not* attempt to answer the question "How probable are the propositions *given* the findings?" Scientists do not evaluate and provide an assessment of the probative strength of scientific findings when they express opinions on propositions. Hence—for the scientist—evaluation given activity level propositions does *not* mean to opine, probabilistically, on competing activities that may have "caused" the findings. Evaluating forensic results means to provide information that helps the recipient of expert information discriminate between propositions, whatever their belief in those propositions is prior to hearing the scientific findings.

Second, reporting on the probability of the observations given competing versions of the case, regarding activities, does not exclusively depend on numerical data, but is also informed by expert knowledge and experience, for which scientists can provide appropriate documentation and demonstrate how it shapes their opinion. What is more, the scientist invokes information that is available for disclosure and auditing. An important corollary of this is that even though task specific data may be unavailable or scarce, it does not mean that no probability can be assigned. In particular, this is not to suggest that any opinion, or mere guesswork, is a valid substitute for thorough scientific assessment. It highlights the need for the elicitation of expert probabilities and knowledge through formal methods and techniques, known also in other areas of specialization, such as risk and safety assessments (e.g., O'Hagan et al., 2006; Aven and Reniers, 2013). There is merit in further developing these approaches for forensic science applications, as well as strengthening the body of structured knowledge (i.e., relevant data on phenomena such as transfer and persistence) for various types of forensic traces (Evett, 2015). This rejoins the idea of developing a knowledge base system that would include experiments and exemplar probabilistic models for evaluation (e.g., BNs). This is widely considered a critical step that the field needs to take now.

Third, variability in the observations (e.g., with respect to quality and quantity of transferred material) observed in experiments under controlled conditions, is both natural and expected. It does not mean that such data cannot be used for evaluation in actual cases, nor does it mean that no conclusion may be drawn. This view is also supported by professionally organized forensic caseworkers (Casey et al., 2016). Variability is an inevitable feature of scientific experiments, observations and measurements, and produces uncertainty. The scientific approach to such uncertainty is to capture it by probability and to take it into account in the scientist's evaluation (e.g., it will be ensured that the data used come from experiments that relate directly to the analytical methods used in the case of interest). Therefore, variation *per se* is not a primary matter of concern; what does matter for the scientist is to see whether the probability of the outcomes given different propositions varies. That is, for the results to be useful, the outcomes need to be more probable given one version of the case (i.e., proposition) than given an alternative version of the case. It is on this latter issue that the scientists need to focus their attention.

The above observations diffuse the call for so-called "unpredictable" forensic DNA traces, in particular low quantities, to be withheld from being used in the process. This is so because the perceived drawbacks, although inspired by known difficulties, do not properly acknowledge additional levels of scientific observations (e.g., extrinsic features such as the quality and the quantity of recovered material, and the position in which it was found) that may be available and that characterize a comprehensive evaluation of forensic results. This perspective goes beyond the mere assessment of the rarity of the analytical features (i.e., the genetic profile). Indeed, for decades, forensic scientists and recipients of expert information have found comfort in seeing forensic DNA analyses provide "constant" and "stable" results in the sense that the DNA profile observed for a sample from a person of interest will, broadly speaking, be observed to be the same for a stain left by that person - as long as quality and quantity of the staining are appropriate, and the chain of custody is impeccable. To a large extent, this has led to technical efforts and investments being spent on ensuring that analyses will reveal the same profile for materials that come from the same source. This is, undoubtedly, an important preliminary requirement for use in forensic science. Unfortunately, however, this perspective was accompanied by the idea that all that is necessary to assess the strength of the findings is an assignment for the probability of observing the profile of interest for an unknown person. This focus on analytical accuracy and rarity of features conflicts with the intricacy of additional dimensions

that DNA profiling entails, such as the very fact of finding DNA at a particular place on a receptor surface (i.e., extrinsic aspects). Stated otherwise, what we have come to see now are conventional interpretation schemes conditioned mainly on source (or sub-source, e.g., in the UK) level propositions being applied to questions, issues and challenges for which these schemes have not been designed, and this has the potential to create stupefying situations in which reports on forensic DNA results are at odds with the case as a whole[23]. Worse still, evaluation given sub-source and source propositions alone can lead to an over-valuing of the scientific evidence, risking miscarriages of justice (Gill, 2014; Jackson, 2014).

The above calls for a readjustment of perspective. To ensure that forensic DNA results are meaningfully used in the legal process, scientists must work on improving their knowledge and understanding about additional factors that characterize not only intrinsic features (e.g., DNA profile) but also extrinsic features (e.g., location where DNA was found). This call is not new (e.g., Evett and Weir, 1998; Taroni et al., 2013; Champod, 2014), but we see that the field is rather reluctant and awareness increases only slowly. At the same time, reports accumulate on real cases (e.g., Gill, 2016) where DNA turned out to be a source of conflict essentially because the key issues for the court related to activities whereas the scientist evaluated the findings in light of questions of source. Thus, evaluation given activity level propositions corresponds to a real need and we foresee that both prosecution and defense counsels will intensify their probing of forensic science regarding this propositional level, not least because recently issued guidelines (i.e., Willis et al., 2015) on evaluation and reporting explicitly set this forth as the standard of interpretation. Achieving this standard is a delicate and challenging endeavor because it operates at the frontiers of current knowledge. However, by gathering, sharing and organizing specialized knowledge in a structured and systematic way (i.e., a shared knowledge base), the forensic community as a whole has the potential to work toward (i) increasing the number of cases in which findings can be assessed given activity level propositions, and (ii) rendering activity level evaluations more trustworthy in those cases where such evaluations are feasible.

In this paper, a discussion format has intentionally been chosen. The aim was to concentrate and restate replies to recurrent objections to emphasize on the need to pursue this topic from a broad perspective, associating both forensic scientists and lawyers. In view of all the arguments presented, our view is that evaluation given activity level propositions represents a main point of the agenda of future research. Besides the justified calls for more structured expert knowledge, we also recognize the need to report on more practical case examples that demonstrate the feasibility of this perspective in a way that practitioners can understand. Such reports on practical examples exceed the space available in this communication, but is the object of ongoing collaborative work between the authors.

[23] See also the case mentioned in footnote 21.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

Research reported in this paper was presented at the 2nd International Symposium on Sino Swiss Evidence Science (2nd ISSSES), University of Lausanne, School of Criminal Justice, September 5th to 9th 2016. The 2nd ISSSES was supported by a Grant from the Swiss National Science Foundation (SNSF Grant no. IZ32Z0_168366; awarded to CC). The work of AB was supported by the SNSF through Grant No. BSSGI0_155809 and the University of Lausanne. The work of JV was supported by the SNSF under the Ambizione Grant No. PZ00P1_154955.

REFERENCES

Aven, T., and Reniers, G. (2013). How to define and interpret a probability in a risk and safety setting. *Safety Sci.* 51, 223–231. doi: 10.1016/j.ssci.2012.06.005

Biedermann, A., and Taroni, F. (2006). Bayesian networks and probabilistic reasoning about scientific evidence when there is a lack of data. *Forensic Sci. Int.* 157, 163–167. doi: 10.1016/j.forsciint.2005.09.008

Biedermann, A., and Taroni, F. (2012). Bayesian networks for evaluating forensic DNA profiling evidence: a review and guide to literature. *Forensic Sci. Int. Genet.* 6, 147–157. doi: 10.1016/j.fsigen.2011.06.009

Brenner, C. H. (2010). Fundamental problem of forensic mathematics – the evidential value of a rare haplotype. *Forensic Sci. Int. Genet.* 4, 281–291. doi: 10.1016/j.fsigen.2009.10.013

Budowle, B., Ge, J., Chakraborty, R., and Gill-King, H. (2012). Response to: use of prior odds for missing persons identifications - author's reply. *Invest. Genet.* 3:3. doi: 10.1186/2041-2223-3-3

Casey, D. G., Clayson, N., Jones, S., Lewis, J., Boyce, M., Fraser, I., et al. (2016). A response to Meakin and Jamieson DNA transfer: review and implications for casework. *Forensic Sci. Int. Genet.* 21, 117–118. doi: 10.1016/j.fsigen.2015.12.013

Champod, C. (2014). "DNA transfer: informed judgment or mere guesswork?," in *DNA, Statistics and the Law: A Cross-Disciplinary Approach to Forensic Inference*, Vol. 4, eds A. Biedermann, J. Vuille, and F. Taroni (Lausanne: Frontiers Media S.A.), 22–24.

Cook, R., Evett, I. W., Jackson, G., Jones, P. J., and Lambert, J. A. (1998). A hierarchy of propositions: deciding which level to address in casework. *Sci. Justice* 38, 231–239. doi: 10.1016/S1355-0306(98)72117-3

Coulson, S. A., Buckleton, J. S., Gummer, A. B., and Triggs, C. M. (2001). Glass on clothing and shoes of members of the general population and people suspected of breaking crimes. *Sci. Justice* 41, 39–48. doi: 10.1016/S1355-0306(01)71847-3

Curran, J. M., Hicks, T. N., and Buckleton, J. S. (2000). *Forensic Interpretation of Glass Evidence*. Boca Raton, FL: CRC Press.

Curran, J. M., Triggs, C. M., Buckleton, J. S., Walsh, K., and Hicks, T. (1998). Assessing transfer probabilities in a Bayesian interpretation of forensic glass evidence. *Sci. Justice* 38, 15–21. doi: 10.1016/S1355-0306(98)72068-4

Daly, D. J., Murphy, C., and McDermott, S. D. (2012). The transfer of touch DNA from hands to glass, fabric and wood. *Forensic Sci. Int. Genet.* 6, 41–46. doi: 10.1016/j.fsigen.2010.12.016

ENFSI Board (2005). *Code of Conduct, BRD-GEN-003.* ENFSI Board.

Evett, I. (2015). The logical foundations of forensic science: towards reliable knowledge. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370, 1–10. doi: 10.1098/rstb.2014.0263

Evett, I. W. (1984). A quantitative theory for interpreting transfer evidence in criminal cases. *Appl. Stat.* 33, 25–32. doi: 10.2307/2347659

Evett, I. W., and Buckleton, J. S. (1989). Some aspects of the Bayesian approach to evidence evaluation. *J. Forensic Sci. Soc.* 29, 317–324. doi: 10.1016/S0015-7368(89)73271-0

Evett, I., Gill, P., Jackson, G., Whitaker, J., and Champod, C. (2002). Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks. *J. Forensic Sci.* 47, 520–530. doi: 10.1520/JFS15291J

Evett, I. W., Jackson, G., and Lambert, J. A. (2000a). More on the hierarchy of propositions: exploring the distinction between explanations and propositions. *Sci. Justice* 40, 3–10. doi: 10.1016/S1355-0306(00)71926-5

Evett, I. W., Jackson, G., Lambert, J. A., and McCrossan, S. (2000b). The impact of the principles of evidence interpretation and the structure and content of statements. *Sci. Justice* 40, 233–239. doi: 10.1016/S1355-0306(00)71993-9

Evett, I. W., and Weir, B. S. (1998). *Interpreting DNA Evidence*. Sunderland: Sinauer Associates Inc.

Fenton, N., and Neil, M. (2013). *Risk Assessment and Decision Analysis with Bayesian Networks*. Boca Raton, FL: CRC Press.

Gill, P. (2014). *Misleading DNA Evidence: Reasons for Miscarriages of Justice*. London: Academic Press.

Gill, P. (2016). Analysis and implications of the miscarriages of justice of Amanda Knox and Raffaele Sollecito. *Forensic Sci. Int. Genet.* 23, 9–18. doi: 10.1016/j.fsigen.2016.02.015

Hicks, T., Vanina, R., and Margot, P. (1996). Transfer and persistence of glass fragments on garments. *Sci. Justice* 36, 101–107. doi: 10.1016/S1355-0306(96)72574-1

Hicks, T., Biedermann, A., de Koeijer, J. A., Taroni, F., Champod, C., and Evett, I. W. (2015). The importance of distinguishing information from evidence/observations when formulating propositions. *Sci. Justice* 55, 520–525. doi: 10.1016/j.scijus.2015.06.008

Jackson, G. (2014). "The impact of commercialization on the evaluation of DNA evidence," in *DNA, Statistics and the Law: A Cross-Disciplinary Approach to Forensic Inference*, Vol. 4, eds A. Biedermann, J. Vuille, and F. Taroni (Lausanne: Frontiers Media S.A.), 16–18.

Jackson, G., Jones, S., Booth, G., Champod, C., and Evett, I. W. (2006). The nature of forensic science opinion - a possible framework to guide thinking and practice in investigations and in court proceedings. *Sci. Justice* 46, 33–44. doi: 10.1016/S1355-0306(06)71565-9

Jackson, G., Aitken, C. G. G., and Roberts, P. (2013). *Case Assessment and Interpretation of Expert Evidence (Practitioner Guide No. 4), Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society's Working Group on Statistics and the Law*. Available online at: www.rss.org.uk/Images/PDF/influencing-change/rss-case-assessment-interpretation-expert-evidence.pdf

Jamieson, A. (2011). LCN DNA analysis and opinion on transfer: R v Reed and Reed. *Int. J. Evid. Proof* 15, 161–169. doi: 10.1350/ijep.2011.15.2.375

Jones, S., Scott, K., Lewis, J., Davidson, G., Allard, J. E., Lowrie, C., et al. (2016). DNA transfer through nonintimate social contact. *Sci. Justice* 56, 90–95. doi: 10.1016/j.scijus.2015.10.004

Margot, P. (2011). Commentary on 'The need for a research culture in the forensic sciences'. *Univ. Calif. Law Rev.* 58, 795–801. Available online at: http://www.uclalawreview.org/pdf/58-3-6.pdf

Meakin, G. E., and Jamieson, A. (2013). DNA transfer: review and implications for casework. *Forensic Sci. Int. Genet.* 7, 434–443. doi: 10.1016/j.fsigen.2013.03.013

Meakin, G. E., and Jamieson, A. (2016). A response to a response to Meakin and Jamieson DNA transfer: review and implications for casework. *Forensic Sci. Int. Genet.* 22, e5–e6. doi: 10.1016/j.fsigen.2016.02.010

Murphy, E. (2015). *Inside the Cell: The Dark Side of Forensic DNA*. New York, NY: Nation Books.

National Commission on Forensic Science (2016). *Directive Recommendation: National Code of Professional Responsibility for Forensic Science and Forensic Medicine Service Providers* (version 22/03/16). Available online at: https://www.justice.gov/ncfs/file/839711/download

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J. et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*, *Statistics in Practice*. Chichester: John Wiley & Sons.

Phipps, M., and Petricevic, S. (2007). The tendency of individuals to transfer DNA to handled items. *Forensic Sci. Int.* 168, 162–168. doi: 10.1016/j.forsciint.2006.07.010

Polley, D., Mickiewicz, P., Vaughn, M., Miller, T., Warburton, R., Komonski, D., et al. (2006). An investigation of DNA recovery from firearms and cartridge cases. *Can. Soc. Forensic Sci. J.* 4, 217–228. doi: 10.1080/00085030.2006.10757145

Risinger, D. M. (2013). Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'). *Law Probabil. Risk* 12, 63–73. doi: 10.1093/lpr/mgs011

Samie, L., Hicks, T., Castella, V., and Taroni, F. (2016). Stabbing simulations and DNA transfer. *Forensic Sci. Int. Genet.* 22, 73–80. doi: 10.1016/j.fsigen.2016.02.001

Taroni, F., Biedermann, A., Vuille, J., and Morling, N. (2013). Whose DNA is this? How relevant a question? (a note for forensic scientists). *Forensic Sci. Int. Genet.* 7, 467–470. doi: 10.1016/j.fsigen.2013.03.012

Taroni, F., Biedermann, A., Bozza, S., Garbolino, G., and Aitken, C. G. G. (2014). *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*, *Statistics in Practice, 2nd Edn.* Chichester: John Wiley & Sons. doi: 10.1002/9781118914762

Willis, S. M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., et al. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science, Strengthening the Evaluation of Forensic Results Across Europe (STEOFRAE)*. Dublin.

Disclaimer: Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the National Institute of Standards and Technology. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

GENERAL COMMENTARY
published: 15 March 2017
doi: 10.3389/fgene.2017.00033

Commentary: A "Source" of Error: Computer Code, Criminal Defendants, and the Constitution

Duncan A. Taylor[1,2]*, *Jo-Anne Bright*[3] *and John Buckleton*[3]

[1] *Forensic Science South Australia, Adelaide, SA, Australia,* [2] *School of Biological Sciences, Flinders University, Adelaide, SA, Australia,* [3] *Institute of Environmental Science and Research, Auckland, New Zealand*

Keywords: source code, STRmix, DNA profile interpretation, closed source, court challenge

A commentary on

A "Source" of Error: Computer Code, Criminal Defendants, and the Constitution
by Chessman, C. (2017). Calif. Law Rev. 105, 101–193.

Chessman (2017) warns of the current trend to admit into court unchallenged the results of complex computerized calculations. He provides a number of examples and arguments claimed to demonstrate the need for open source software to remove the "black box" element. We agree with parts of this sentiment, and the topic of this special issue, that there is a danger with those using and receiving information from black box systems.

Some care however is needed with simple diagnoses and prescriptions such as these.

Modern probabilistic genotyping software are replacing methods previously applied manually. We have great confidence in the forensic community with regard to both integrity and dedication. The previously applied processes are usually a composite of standard operating procedure and human judgment. The difference between these and probabilistic software is largely that the processes in the software are encoded.

Many disciplines are sufficiently broad that practitioners need to rely, in part, on the work of others. This is not new (for a discussion on this point see Taylor, 2016). The risk to which Chessman refers arises when the individual using the system has so little understanding that they do not know how to use the system, or when it has not worked[1]. Chessman provides some helpful suggestions for how breaking down black box barriers can be addressed on an individual and systemic scale. As developers of expert system STRmix™[2] (Taylor et al., 2013), we wish to address some of the alarmist points in Chessman (and echoed by others[3]) that gives the impression that producers of expert systems are all either incompetent or corrupt.

We first wish to correct a couple of points in (Chessman, 2017). Regarding the "erroneous assumption" referenced by footnotes 49–51: This miscode, and indeed any miscode found that has been identified in STRmix™ development or use, was identified by examination of the program's output and not the source code. It would be nearly impossible to identify subtle errors in code by viewing the code. The identification has always been a result of comparison of the results produced by a program to some known control[4]. The results of these comparisons then trigger the examination of a specific section of the code in order to discover the source of the discrepancy.

[1] Note that this is not an issue with just computer programs, recent history has numerous examples within forensic biology showing that a misunderstanding of the way a system works at a fundamental level can cause issues even when the calculations themselves are relatively simple and able to be done by hand (Budowle and Bieber, 2015).
[2] An expert system that analyses STR DNA profile data.
[3] For example see EPIC (https://epic.org/state-policy/foia/dna-software/).
[4] Commonly a "by-hand" recreation of the expected value(s).

OPEN ACCESS

Edited by:
Sue Pope,
Principal Forensic Services, UK

Reviewed by:
Peter Ronald Gunn,
University of Technology Sydney, Australia

***Correspondence:**
Duncan A. Taylor
Duncan.Taylor@sa.gov.au

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *03 November 2016*
Accepted: *28 February 2017*
Published: *15 March 2017*

Citation:
Taylor DA, Bright J-A and Buckleton J (2017) Commentary: A "Source" of Error: Computer Code, Criminal Defendants, and the Constitution. Front. Genet. 8:33. doi: 10.3389/fgene.2017.00033

Even as developers, during the developmental validation of new versions of STRmix™, we utilize the extended outputs of the software to validate, and do not validate by examination of code. A further reference (footnote 98) makes the same incorrect assumption that it was code review that lead to the discovery of a programming error. Our experience has been that even more crucial than a review of source code, is the ability to have access to outputs that demonstrate each step of a calculation. We should also note that our ongoing evaluation and testing of the software is a marker of continuous validation and refinement, rather than just fixing "errors" and "blunders."

The second point we wish to make is that the type and magnitude of miscodes are important to consider. The majority of programming errors will lead to instances of a program "crashing" or failing to produce an answer. These types of errors are arguably inconsequential as they will not lead to any erroneous results being produced. More serious are miscodes where no errors are identified or displayed by the software. These can be split into those that will be clearly identifiable[5] and those that are more subtle and may go initially unnoticed. Even in this latter category, the question should be asked "What effect does this error have?" If the magnitude of the difference in the result caused by the miscode is small compared with the natural variability in the results being produced[6] then arguably the consequences are minimal. We are by no means suggesting that that these types of errors are acceptable, they should be rectified as soon as found. We simply suggest that they tend to be used for scaremongering in a manner disproportionate to their impact. Case in point is the oft quoted article (David Murray, 2015), which contains the never quoted sentence "*The DNA likelihood ratios in both the new and original statements appear to be the same.*"

We agree with the suggestion of Chessman that source code should be available for scrutiny. STRmix™ abides by one of the mechanisms that Chessman suggests, namely the ability for code to be disclosed under confidentiality agreements[7]. We note that running of STRmix™ is just the final step in a long journey of computerized activities that ultimately lead to an answer. A true challenge of all steps in the process would require the examination of the source code underlying the Java programming language in which STRmix™ is written, the Windows™ operating system on which it is run, the software used to process the raw electrophoretic data, the software used to collect the raw electrophoretic data from the electrophoresis instrument, the code used to run the electrophoresis instrument, the PCR thermocycler, the quantification instrument and a myriad of no doubt thousands of blocks of code that sit within the numerous Peripheral Interface Controllers that control hardware components.

With the advent of complex computerized evaluation of evidence, there is a shift from a time where an expert can testify to all aspects of the evaluation, to one where, at some level, the workings of an expert system are accepted without absolute understanding. This may initially seem frightening, but an examination of the bigger picture suggests otherwise. It would be difficult to argue that the use of computerized breathalyzers is a backwards step from the reliability of the Field Sobriety Test. Similarly, virtually all senior advisory bodies relating to DNA profile evaluation recognize the clear benefits of the probabilistic interpretation systems (which by nature of their complexity require computerized implementation) over the preceding manual or binary interpretation methods (Coble et al., 2015; SWGDAM, 2015). In our efforts to ensure that software is not the "source" of errors, it is important to recognize that even with the noted occurrences of these errors, the current computerized solutions, when used by trained experts, represent a vast improvement to the quality and reliability of evidence presented in court.

AUTHOR CONTRIBUTIONS

All authors contributed to the discussions and writing of the manuscript. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the author's organizations.

FUNDING

Funding to write this manuscript was provided by the author's institutions only in the sense of allowing work time to be used to develop the document.

[5]Such as value of a probability greater than one, or a negative amount of some substance.

[6]This may either be in the raw results due to inherent variability in the laboratory process or it may be variability in the statistical result due to an evaluation method that utilizes random number generation (Bright et al., 2015).

[7]The code of STRmix™ has been viewed under such conditions in the past.

REFERENCES

Bright, J.-A., Stevenson, K. E., Curran, J. M., and Buckleton, J. S. (2015). The variability in likelihood ratios due to different mechanisms. *Forensic Sci. Int. Genet.* 14, 187–190. doi: 10.1016/j.fsigen.2014.10.013

Budowle, B., and Bieber, F. R. (2015). *Final Report on Review of Mixture Interpretation in Selected Casework of the DNA Section of the Forensic Science Laboratory Division, Department of Forensic Sciences, District of Columbia*. Available online at: http://dfs.dc.gov/page/usao-report-april-2015.

Chessman, C. (2017). A "source" of error: computer code, criminal defendants, and the constitution. *Calif. Law Rev.* 105, 101–193.

Coble, M. D., Buckleton, J., Butler, J. M., Egeland, T., Fimmers, R., Gill, P., et al. (2015). DNA Commission of the International Society for Forensic Genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications. *Forensic Sci. Int. Genet.* 25, 191–197. doi: 10.1016/j.fsigen.2016.09.002

David Murray (2015). *Queensland Authorities Confirm 'Miscode' Affects DNA Evidence in Criminal Cases [Online]*. Available online at: http://www.news.com.au/national/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92b [Accessed].

Scientific Working Group on DNA Analysis Methods (SWGDAM). (2015). *Guidelines for the Validation of*

Probabilistic Genotyping Systems [Online]. Available online at: http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf [Accessed 3 October 2016].

Taylor, D. (2016). Is technology the death of expertise? *Forensic Sci. Int. Genet.* 24, e1–e3. doi: 10.1016/j.fsigen.2016.06.006

Taylor, D., Bright, J.-A., and Buckleton, J. (2013). The interpretation of single source and mixed DNA profiles. *Forensic Sci. Int. Genet.* 7, 516–528. doi: 10.1016/j.fsigen.2013.05.011

Conflict of Interest Statement: Authors are technical developers of commercial software STRmix^{TM} but do not benefit financially from STRmix^{TM}.

frontiers in Sociology

GENERAL COMMENTARY
published: 18 May 2017
doi: 10.3389/fsoc.2017.00006

General Commentary: Federal Labour Court [2009] – 8 AZR 1012/08

*Kyriakos N. Kotsoglou**

Law Department, Liverpool Hope University, Liverpool, UK

Keywords: evidence and proof, law of evidence, labour law, forensic sciences, reference class problem

A commentary on

Federal labour court [2009] – 8 AZR 1012/08

For Silke K., it must have been one of these days that define someone's story beyond her life. After years of suffering gender discrimination and mobbing at work, the Berlin-Brandenburg's Labour Court (LAG) had just issued a game-changing decision in her favor. The question before the court essentially boiled down to whether the claimant had discharged her burden of proof simply by employing formal statistical methods. LAG's answer was positive. It awarded damages that included the difference (€1.468) between Silke K.'s monthly salary and a director's one, damages for sex discrimination (€28.214,66), and damages for violating rights of personality (€20.000) (Second Instance Land Labour Court, 2008).

The decision did not only implement the new policy enshrined in the General Equal Treatment Act 2006 (Allgemeines Gleichbehandlungsgesetz)—whose purpose was to put an end to discrimination against vulnerable social groups of workers and vertical segregation in labor market creating steep asymmetries between the percentage of women in higher positions and the total labor force (ILO Director-General, 2011)—but also most importantly constituted a paradigm shift regarding the legal protection of female employees. More specifically, whereas women were making up at that time the majority (69%) of the defendant's workforce (Second Instance Land Labour Court, 2008, para 23), not a single one of them was in the board of directors (Second Instance Land Labour Court, 2008, para 14), exemplifying thus in a clear way that women are "widely underrepresented" in decision-making positions in the private sector of Germany (EU Document of the Directorate General for Internal Policies, 2015). But what exactly propelled this long expected success and engineered the closing of the persistent gender pay gap?[1] Silke K.'s team of lawyers had hired a mathematician to calculate the probability that it is not random that the board of directors includes no women. The probability, the statistical analysis (Monte–Carlo simulation) showed, was between 98.7% and 100% (Second Instance Land Labour Court, 2008, para 34). Therefore, it was formal statistical calculations that gave thrust to the gender equality machinery, since the LAG explicitly equated this statistical result with the probability of discrimination against the claimant (Second Instance Land Labour Court, 2008). By employing a rigorous framework to draw inferences from data, courts broke the "glass ceiling", i.e., the "unseen, yet unbreakable barrier that keeps minorities and women from rising to the upper rungs of the corporate ladder, regardless of their qualifications or achievements" (Federal Glass Ceiling Commission, 1995).

Alas all good things come to an end. The Federal Labour Court (BAG) quashed the decision as it held that statistics are not conclusive for the individual case (Federal Labour Court [2009]). Although there was no explicit mention of the reference class problem in the decision, the Federal Court raised once again questions of sufficiency of proof by making clear that proof of unlawful behavior hinges

[1] According to the EU report (EU Document of the Directorate General for Internal Policies, 2015, p. 15): "In general, Germany is ranked by the European Gender Equality Index (GEI) lower than the EU average; its performance in achieving gender equality is 'mediocre'."

OPEN ACCESS

Edited by:
Alex Biedermann, University of Lausanne, Switzerland

Reviewed by:
Marcello Di Bello, Lehman College, USA

***Correspondence:**
Kyriakos N. Kotsoglou
kotsogk@hope.ac.uk

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Sociology

Received: *02 March 2017*
Accepted: *11 April 2017*
Published: *18 May 2017*

Citation:
Kotsoglou KN (2017) General Commentary: Federal Labour Court [2009] – 8 AZR 1012/08. Front. Sociol. 2:6. doi: 10.3389/fsoc.2017.00006

on "statistical data being conclusive for the employer in question", to wit, on *specific evidence* (Federal Labour Court [2009], para 68): proof of membership to a reference group is inconclusive.

This brings us to our main issue, the meaning of "specific evidence" and the renowned reference class problem (Colyvan et al., 2001). Can statistical information—accurate as they may be—motivate action? The question is at its kernel whether an epistemic inference from a relevant population—which serves as a basis for calculating and assigning probabilities—to an individual is valid—given that we only have information about the reference class. Since we deal with the problem of factual generalization and individualization, we rather unwillingly have to raise fundamental questions about the nature of our reasoning processes, in both law and elsewhere. Unsurprisingly, these issues have spawned an extensive debate Allen and Roberts (2007). For very good reasons, since legal adjudication aspires to be rational. However, there is no consensus on what lessons to draw. The debate between the opposing parties has stalled. It would not be exaggerating to say that we have reached the point "where one would like just to emit an inarticulate sound" (Wittgenstein, 1958, § 261).

This short commentary suggests that we should not be so pessimistic. From Aristotle who observes that "it is evidently equally foolish to […] demand from a rhetorician scientific proofs" (Aristotle, The Nicomachean Ethics, Book I, Ch. 3) to modern forensic scientists who are at pains to stress that the idea "of a frequency being attached to an outcome for a single event is ridiculous" (Lucy, 2006, 5), scholars have continuously rejected (bogus) aspirations of generality when it comes to (judicial) decisions. True, statistics enable us to validate knowledge-claims about the world; but at the same time, we resort to quantitative evidence in order to gain an understanding of a population *in its entirety*. This simple expression—"in its entirety"—destroys the riddle. Courts and decision makers do not formulate general rules. They give answers to questions such as "Is the defendant guilty?" to which we do not have scientific answers, not because they are intractable, profound mysteries, but simply because decision-making is not a scientific process yielding a *generally* valid solution. Of course, statistics should inform the evidential basis of decisions and help settle arguments. However, judges do have discretion when they apply the law, so that we have to willy-nilly reject the idea(l) of a mechanical jurisprudence. Extending statistical ideas and methods to procedural and forensic contexts can broadly be classified as scientism.

The Federal Labour Court made a move in the right direction. It criticized the transgression of the bound of "specific evidence" and reaffirmed the individualistic character of legal adjudication by authoritatively cutting the Gordian knot (reference class problem). The academic community has to deliver *ex post* facto the theoretical framework that (dis-)solves this problem.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

Allen, R., and Roberts, P. (eds). (2007). The special issue on the reference class problem. *Int. J. Evidence Proof* 11, 243–317.

Aristotle. *The Nicomachean Ethics, Book I: The Good for Man*, Chap. 3.

Colyvan, M., Regan, H. M., and Ferson, S. (2001). Is it a crime to belong to a reference class? *J. Political Philos.* 9, 168–181; discussion 171. doi:10.1111/1467-9760.00123

EU Document of the Directorate General for Internal Policies. (2015). *The Policy on Gender Equality in Germany*. 8. Available at: http://www.europarl.europa.eu/studies

Federal Glass Ceiling Commission. (1995). *Solid Investments: Making Full Use of the Nation's Human Capital*. Washington, DC: US Department of Labor, 13–15.

Federal Labour Court [2009] – 8 AZR 705/08.

ILO Director-General. (2011). *Equality at Work: The Continuing Challenge*. Report I(B). Geneva: International Labour Conference 100th Session.

Lucy, D. (2006). *Introduction to Statistics for Forensic Scientists*. New York: Wiley.

Second Instance Land Labour Court. Berlin-Brandenburg [26.11.2008] – 15 Sa 517/08; NZA 2009, 43.

Wittgenstein, L. (1958). *Philosophical Investigations*. Oxford: Basil Blackwell, § 261.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Sociology

GENERAL COMMENTARY
published: 20 December 2017
doi: 10.3389/fsoc.2017.00020

Commentary: Van Beelen [2016] SASCFC 71

David R. A. Caruso and Brigid Symes*

Litigation Law Unit, The University of Adelaide, Adelaide, SA, Australia

Keywords: second appeal, scientific advance, fresh evidence, compelling evidence, substantial

A commentary on

Van Beelen [2016] SASCFC 71

INTRODUCTION

On 15 July 1971, Deborah Leach was found dead on the beach. Dr. Manock, the State forensic pathologist, used stomach content analysis to determine the time of Leach's death. Manock concluded that death was between 4:00 p.m. and 4:30 p.m. Only Mr. Van Beelen had been witnessed on the beach during this time. Van Beelen was convicted of Leach's murder in 1973.[1] His conviction was upheld on appeal.[2]

In 2013, South Australian appellate law was amended to permit second and subsequent appeals if there was "fresh" and "compelling" evidence.[3] This reversed the common law position that allowed for only a single, perfected appeal.[4] Van Beelen's case was subsequently appealed in 2016.[5] The defense argued that scientific development concerning stomach content analysis constituted fresh and compelling evidence. The Appeal Court rejected that argument by majority.

STOMACH CONTENT ANALYSIS AT THE 1973 TRIAL

Manock relied on Leach's schoolmates for the content and timing of her last meal: a pasty, a glass of milk, and a slice of pie between 12:15 p.m. and 12:30 p.m. Manock found that Leach's stomach was three-quarters empty at death. He concluded that death could not have occurred before 3:30 p.m. or after 4:30 p.m. Dr. Pocock testified, for the defense, that it would "indeed be a rash, irresponsible man, who would dare pronounce the exact time of death in the witness-box; or for that matter, be ready to estimate the time even to within an hour on either side of the actual time."[6]

SCIENTIFIC DEVELOPMENTS IN THE 2016 APPEAL

In 2016, Professor Horowitz testified that stomach content analysis was now considered to be an unreliable way to accurately measure the time of death. This was due to the enormous variations in digestion between individuals. Horowitz testified that gastric emptying rates cannot (now) be estimated to within an hour.[7] He said that even in the 1970s, it was completely unreliable science to provide an estimate of time of death once gastric emptying had commenced.[8]

OPEN ACCESS

Edited by:
Alex Biedermann,
University of Lausanne, Switzerland

Reviewed by:
Paul Roberts,
University of Nottingham,
United Kingdom

****Correspondence:***
David R. A. Caruso
david.caruso@adelaide.edu.au

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Sociology

Received: *20 October 2017*
Accepted: *28 November 2017*
Published: *20 December 2017*

Citation:
Caruso DRA and Symes B (2017) Commentary: Van Beelen [2016] SASCFC 71. Front. Sociol. 2:20. doi: 10.3389/fsoc.2017.00020

[1] See R v Van Beelen (1973) 4 SASR 353.
[2] R v Van Beelen (No 3) (1973) 7 SASR 125.
[3] Statutes Amendment (Appeals) Act 2013 (SA) (No 9 of 2013) section 7.
[4] R v Edwards (No 2) (1931) SASR 376; Grierson v The King (1938) 60 CLR 431.
[5] R v Van Beelen (2016) 125 SASR 253.
[6] R v Van Beelen (2016) SASCFC 71 (129) Vanstone and Kelly JJ.
[7] R v Van Beelen (2016) SASCFC 71 (50) Kourakis CJ.
[8] R v Van Beelen (2016) SASCFC 71 (31) Kourakis CJ.

COURT'S REASONING

Kourakis CJ, Vanstone and Kelly JJ comprised the secondary appeal court. They agreed that the evidence of Horowitz was "fresh" within the meaning of the legislation. The point of difference was that the Majority (Vanstone and Kelly JJ) did not consider the evidence "compelling"; the Chief Justice did.

"Fresh" and "compelling" are defined. Evidence relating to an offense is "fresh" if: (i) *it was not adduced at the trial of the offense*; and (ii) *it could not, even with the exercise of reasonable diligence, have been adduced at the trial*. That same evidence is "compelling" if: (i) *it is reliable*; *and* (ii) *it is substantial*; *and* (iii) *it is highly probative in the context of the issues in dispute at the trial of the offense*.[9]

The Majority held that the evidence was not "compelling" because Horowitz's evidence was not "substantial."[10] This was due to the doubt already placed on Manock's evidence by Pocock in the original trial.[11] The Majority did not consider the prosecution case turned on the expert evidence regarding time of death. The Majority gave particular attention to the civilian evidence at the 1973 trial that set the parameters for Leach's death as occurring between 4:00 p.m. and 4:40 p.m., or at the latest 4:50 p.m. The fresh evidence was only capable of showing that time of death could have been a mere 10 or 20 min later. The Majority considered that, as a result, the evidence of Horowitz was not substantial and, therefore, not compelling.[12]

OPINION

Common law courts vested with secondary appeal powers are obliged to review scientific advance within the confines of its possible meaning and interpretation in an historic trial. Appellate rights must be constrained by the trial issues to realize the utility in litigation ending. That is a narrow lens for what may be wide-ranging gains in scientific knowledge. The purpose of secondary appellate legislation should be to permit scientific advancement to expose errors of fact-finding at trial. The conclusion of the Majority requires the relevant evidence to be substantial in light of the trial as a whole. The provisions of section 353A, however, do not require the evidence to be considered in light of the trial holistically. The fresh evidence itself must be substantial. Kourakis CJ identified why the evidence was substantial: its reputable source and basis in current science.

Compelling evidence under the legislative regime is "highly probative in the context of the issues" at the trial. Scientific advance is almost ubiquitously probative with respect to evolution of human knowledge, but the courts are concerned with a more restricted notion of probity. The "probative value" of expert evidence to the law concerns the effect that evidence would have on the rational assessment of issues before the court. In the case of secondary appeals, the probative value lies in the extent to which the expert evidence would compel re-assessment of the original trial issues. This symbiotic relationship between the probative value of expert evidence and the disputed issues reveals why the reasoning of Kourakis CJ is to be preferred.

Probative value is to be assessed in the context of the issues, not the evidence, in dispute. The Majority rejected Horowitz's evidence as compelling because there was other lay evidence, which diminished the import of stomach contents emptying regarding time of death. The legislation does not invite the appellate court to examine the issues in dispute having regard to the evidence in the trial court. The requirement that evidence justifying a secondary appeal must be "fresh" guards against evidence being re-litigated or agitated on secondary appeal. If evidence adduced on secondary appeal relates to an issue in the trial below which the context of the trial reveals to be an important issue in the dispute at trial, then, the definition of "compelling" indicates that the subject evidence should be regarded as highly probative *in the context of the issues in dispute*. The time of death was a key element in the prosecution case and, without its certainty, a possibility arose that someone other than Van Beelen committed the offense.[13] The Chief Justice found the relationship of Horowitz's evidence to the time of death was highly probative as the timing of death was an issue central to the context of the dispute at the 1973 trial.

In a trial where time of death is in issue and the context of the trial places emphasis on the timing of death, fresh evidence concerning errors in that timing should satisfy the criterion of "compelling." Secondary appellate legislation should be, unless a contrary parliamentary intention can be clearly shown, read with a view to assess whether the fresh evidence is compelling given the *issues* in the court below, not the *evidence* in the court below.

Van Beelen appealed the split decision to the High Court of Australia and the matter was heard in June 2017[14]; final judgment is pending.

AUTHOR CONTRIBUTIONS

This commentary was jointly written with a 80/20 division of work in reviewing the subject case.

[9] Criminal Law Consolidation Act 1935 (SA), section 353A (6).
[10] R v Van Beelen (2016) SASCFC 71 (159) Vanstone and Kelly JJ.
[11] R v Van Beelen (2016) SASCFC 71 (162) Vanstone and Kelly JJ.
[12] R v Van Beelen (2016) SASCFC 71 (164) Vanstone and Kelly JJ.
[13] R v Van Beelen (2016) SASCFC 71 (72) Kourakis CJ.
[14] R v Van Beelen (2017) HCATrans 19.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 16 April 2018
doi: 10.3389/fgene.2018.00126

Bayesian Hierarchical Random Effects Models in Forensic Science

*Colin G. G. Aitken**

School of Mathematics and Maxwell Institute, The University of Edinburgh, Edinburgh, United Kingdom

Statistical modeling of the evaluation of evidence with the use of the likelihood ratio has a long history. It dates from the Dreyfus case at the end of the nineteenth century through the work at Bletchley Park in the Second World War to the present day. The development received a significant boost in 1977 with a seminal work by Dennis Lindley which introduced a Bayesian hierarchical random effects model for the evaluation of evidence with an example of refractive index measurements on fragments of glass. Many models have been developed since then. The methods have now been sufficiently well-developed and have become so widespread that it is timely to try and provide a software package to assist in their implementation. With that in mind, a project (SAILR: Software for the Analysis and Implementation of Likelihood Ratios) was funded by the European Network of Forensic Science Institutes through their Monopoly programme to develop a software package for use by forensic scientists world-wide that would assist in the statistical analysis and implementation of the approach based on likelihood ratios. It is the purpose of this document to provide a short review of a small part of this history. The review also provides a background, or landscape, for the development of some of the models within the SAILR package and references to SAILR as made as appropriate.

Keywords: Bayes' Theorem, evidence evaluation, forensic science, hierarchical models, likelihood ratios, random effects, SAILR, statistics

OPEN ACCESS

Edited by:
Sue Pope,
Principal Forensic Services,
United Kingdom

Reviewed by:
Robert Brian O'Hara,
Norwegian University of Science and Technology, Norway
Ricardo De Matos Simoes,
Dana-Farber Cancer Institute,
United States

***Correspondence:**
Colin G. G. Aitken
cgga@ed.ac.uk

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *27 December 2017*
Accepted: *29 March 2018*
Published: *16 April 2018*

Citation:
Aitken CGG (2018) Bayesian Hierarchical Random Effects Models in Forensic Science. Front. Genet. 9:126. doi: 10.3389/fgene.2018.00126

1. INTRODUCTION

Statistical analyses for the evaluation of evidence have a considerable history. It is the purpose of this document to provide a short review of a small part of this history. It brings together ideas from the last forty years for statistical models when the evidence is in the form of measurements and thus of continuous data. The data are also hierarchical with two levels. The first level is that of source, the origin of the data. The second level is of items within a source. The models used to represent the variability in the data are random effects models. The models are chosen from analyses of samples of sources from some relevant population. Finally, the analysis is Bayesian in nature with prior distributions for the parameters of the within-source distributions. The nature of the prior distributions is informed from training data based on the samples from the relevant population.

The remainder of the document is structured as follows. Section 2 provides a general introduction to the likelihood ratio as a measure of the value of evidence. Section 3 provides a framework for models for comparison and discrimination. Section 4 discusses the assessment of model performance. An Appendix gives formulae for some of the more commonly used models.

2. THE VALUE OF EVIDENCE

Part of the role of a forensic scientist is to interpret evidence found at a crime scene in order to aid fact-finders in a criminal case (e.g., the judge or jury) in their decision making. The forensic scientist may be asked to comment on the value of the evidence in the context of various competing statements about the evidence, each of which may be true or false.Generally, a forensic scientist must consider two competing statements relating to the evidence, one put forward by the prosecution in a criminal case, and one put forward by the defense (Cook et al., 1998b). These statements are known as *propositions*[1]. They generally come in pairs that are mutually exclusive, though not necessarily exhaustive. For a debate about the requirement, or otherwise, for the propositions to be exhaustive (see Biedermann et al., 2014; Fenton et al., 2014a,b).

One member of the pair is associated with the prosecution and conventionally denoted H_p. The other member of the pair is associated with the defense and conventionally denoted H_d. The evidence to be evaluated is denoted E[2]. The value of evidence is taken to be the relative values of the probability of the evidence if a proposition put forward by the prosecution is true and the probability of the evidence if a proposition put forward by the defense is true. However, evidence is not evaluated in isolation. There is always other information to be taken into account, including, for example, personal knowledge of the fact-finder. Denote this information by I. The value of the evidence, denoted V say, can then be written formulaically as

$$\mathrm{V} = \frac{\Pr(E \mid H_p, I)}{\Pr(E \mid H_d, I)},$$

where Pr denotes *Probability*. This ratio is known as the *likelihood ratio*.

The likelihood ratio is the method used by SAILR to evaluate evidence. SAILR (Software for the Analysis and Implementation of Likelihood Ratios) is a user-friendly Graphical Interface (GUI) to calculate numerical likelihood ratios in forensic statistics and its development under the direction of the Netherlands Forensic Institute (NFI) was funded by the European Network of Forensic Science Institutes through their Monopoly programme. The likelihood ratio is a generally accepted measure for the value of evidence in much forensic case-work.

This representation of the value of evidence has a very good intuitive interpretation. Consider the odds form of Bayes' Theorem in the forensic context of the evaluation of evidence. The odds form of Bayes' Theorem then enables the prior odds(i.e., prior to the presentation of E) in favor of the prosecution proposition H_p relative to the defense proposition H_d to be updated to posterior odds given E, the evidence under consideration. This is done by multiplying the prior odds by the likelihood ratio. The odds form of Bayes' Theorem may then be written as

$$\frac{\Pr(H_p \mid E, I)}{\Pr(H_d \mid E, I)} = \frac{\Pr(E \mid H_p, I)}{\Pr(E \mid H_d, I)} \times \frac{\Pr(H_p \mid I)}{\Pr(H_d \mid I)}. \quad (1)$$

The likelihood ratio (LR) is the ratio

$$\frac{\Pr(H_p \mid E, I)/\Pr(H_d \mid E, I)}{\Pr(H_p \mid I)/\Pr(H_d \mid I)} \quad (2)$$

of posterior odds to prior odds. It is the factor which converts the prior odds in favor of the prosecution proposition to the posterior odds in favor of the prosecution proposition. The representation in Equation (1) also emphasizes the dependence of the prior odds on other information I. Values of the $LR > 1$ are supportive of H_p, the proposition put forward by the prosecution. Values of the $LR < 1$ are supportive of H_d, the proposition put forward by the defense. The word "odds" should be used advisedly. If H_p and H_d are not exhaustive then the component probabilities $\Pr(H_p \mid E, I)$ and $\Pr(H_d \mid E, I)$ cannot be derived from this ratio. All that can be said is that the posterior ratio is different from the prior ratio by a factor V.

An advantage of this formulation of evidence evaluation is the ease with which the effect of the addition of new evidence can be determined. The posterior odds for one piece of evidence, E_1 say, can be the prior odds for a second piece of evidence, E_2 say. Then Equation (1) may be rewritten as

$$\frac{\Pr(H_p \mid E_1, E_2, I)}{\Pr(H_d \mid E_1, E_2, I)} = \frac{\Pr(E_2 \mid H_p, E_1, I)}{\Pr(E_2 \mid H_d, E_1, I)} \times \frac{\Pr(H_p \mid E_1, I)}{\Pr(H_d \mid E_1, I)}, \quad (3)$$

where the conditioning of the evaluation of E_2 on E_1 is made explicit.

An illustration of the effect of evidence with a value V of 1,000 on the odds in favor of H_p relative to H_d is given in **Table 1**.

The following quote is very pertinent.

> 'That approach does not ask the jurors to produce any number, let alone one that can qualify as a probability. It merely shows them

TABLE 1 | Effect on prior odds in favor of H_p relative to H_d of evidence E with value V of 1,000.

Prior odds $Pr(H_p)/Pr(H_d)$	V	Posterior odds $Pr(H_p \mid E)/Pr(H_d \mid E)$
1/10,000	1,000	1/10
1/100	1,000	10
1 (evens)	1,000	1,000
100	1,000	100,000

Reference to background information I is omitted.

[1]Other writers use the term *hypothesis* (see section 2.7). The term *proposition* will he used except when there is an explicit need for the term *hypothesis*; see, for example, section 3.1

[2]In ENFSI guidelines ENFSI (2015) "findings" are distinguished from "evidence." "Findings are the result of observations, measurements and classification that are made on items of interest." "[E]vidence refers to outcomes of forensic examinations (findings) that, at a later point, may be used by legal decision-makers in a court of law to reach a reasoned belief about a proposition." However, the word "evidence" will be used in this document to refer to both situations for ease of nomenclature.

> how a "true" prior probability would be altered, if one were in fact available. It thus supplies the jurors with as precise and accurate an illustration of the probative force of the quantitative data as the mathematical theory of probability can provide. Such a chart, it can be maintained, should have pedagogical value for the juror who evaluates the entire package of evidence solely by intuitive methods, and who does not himself attempt to assign a probability to the "soft" evidence.' Kaye (1979).

The "it" in this context is a chart depicting, in numerical terms, how much the prior odds in favor of a proposition is enhanced by the evidence being evaluated. This is a graphical equivalent of **Table 1**. The mathematical tool for devising such a chart is Bayes' Theorem. These remarks of Kaye's refer to characteristics of the general method for the evaluation of evidence that is the likelihood ratio. They do not refer to a particular case. For example, it is not possible to comment on the accuracy of a likelihood ratio estimation in a particular case because the true value of the likelihood ratio is not known nor can it be known. It is, however, possible to refer to the accuracy of a method and performance assessment in general is discussed in section 4.

The use of a likelihood ratio for the evaluation of evidence is not a new idea. In the Dreyfus case (Champod et al., 1999), it was argued that

> . . . since it is absolutely impossible for us [the experts] to know the *a priori* probability, we cannot say: this coincidence proves that the ratio of the forgery's probability to the inverse probability is a real value. We can only say: following the observation of this coincidence, this ratio becomes X times greater than before the observation (Darboux et al., 1908).

The "ratio" in this quotation is the odds in favor of one proposition over another, The X refers to the likelihood ratio. The posterior odds in favor of the proposition is then X times the prior odds.

The ideas were also used in the work of I.J. Good and A.M. Turing as code-breakers at Bletchley Park during World War II (Good, 1979).

2.1. Background Information

The likelihood ratio updates the prior odds, those before consideration of evidence E, to posterior odds, which take E into account. The posterior odds are the odds with which, ultimately, the fact-finder is concerned. If the likelihood ratio multiplied by the prior odds is larger than one, then the probability of H_p given the evidence is larger than that of H_d given the evidence. As these propositions may not be exhaustive their explicit values, rather than their relative value, may not be known. It is the responsibility of the fact-finder to determine a value for the prior odds. The prior odds can then be combined with the likelihood ratio to obtain posterior odds. A forensic scientist is concerned only with the value of the evidence as expressed by the likelihood ratio so cannot usually comment on the value of the posterior odds. The likelihood ratio is considered as the strength of support of the evidence for one of the two propositions H_p or H_d.

The application of this form to a specific case is crucially dependent on the background information I. However, the background information available to each person is different. In part, this is because each person is different. In part it is because of professional differences. The information that a forensic scientist should use for their determination of the likelihood ratio is different from that which a fact-finder, such as judge or jury member, should use for their determination of the odds in favor of the prosecution proposition. There are differences in the background information available to these participants in the judicial process but these differences have no effect on the posterior odds in favor of the prosecution proposition.

Let $I = I_a \cup I_b$ where I_a is background information available to the forensic scientist and I_b is background information available to the fact-finder. There will be information available to both, the intersection $I_a \cap I_b$ is not empty. It can then be shown (Aitken and Nordgaard, 2017) that the posterior odds may be written in the form

$$\frac{\Pr(H_p \mid E, I)}{\Pr(H_d \mid E, I)} = \frac{\Pr(E \mid H_p, I_b)}{\Pr(E \mid H_d, I_b)} \times \frac{\Pr(H_p \mid I_a)}{\Pr(H_d \mid I_a)}.$$

The fact-finder and the forensic scientist have to treat the common information ($I_a \cap I_b$) with appropriate discretion.

2.2. Uniqueness of the Likelihood Ratio

The role of the likelihood ratio as the factor that updates the prior odds to the posterior odds has a very intuitive interpretation. There is also a mathematical derivation that shows it, or a function of it such as the logarithm, is the only way to update evidence. It was shown many years ago by I.J.Good in two brief notes in the *Journal of Statistical Computation and Simulation* (Good, 1989a,b) repeated in Good (1991) and in Aitken and Taroni (2004) that, with some very reasonable assumptions, the assessment of uncertainty inherent in the evaluation of evidence leads inevitably to the likelihood ratio as the only way in which this can be done.

Consider evidence E which it is desired to evaluate in the context of two mutually exclusive propositions H_p and H_d. Denote the value of the evidence by V. As always, the value will depend on background information I but this will not be stated explicitly. There are other assumptions implicit in this approach, namely that there is a probability that can be associated with evidence and one that is dependent on propositions and only on propositions (and background information). Another assumption is that V is a function only of the probability of E, given H_p to be true, and of the probability of E, given H_d to be true.

Let $x = \Pr(E \mid H_p)$ and $y = \Pr(E \mid H_d)$ where I is omitted for ease of notation. The assumption that V is a function only of these probabilities can be represented mathematically as

$$V = f(x, y)$$

for some function f.

Now, consider another piece of evidence T which is irrelevant to E, to H_p and to H_d. Irrelevance is taken in the probabilistic context to be equivalent to independence so that T may be taken to be independent of E, of H_p and of H_d. It is then permissible for

$\Pr(T)$ to be given notation which does not refer to any of E, H_p or H_d. Thus, let $\Pr(T)$ be denoted by θ. Then

$$\begin{aligned}\Pr(E, T \mid H_p) &= \Pr(E \mid H_p)\Pr(T \mid H_p) && \text{by the independence of } E \text{ and } T\\ &= \Pr(E \mid H_p)\Pr(T) && \text{by the independence of } T \text{and } H_p\\ &= x\theta.\end{aligned}$$

Similarly,

$$\Pr(E, T \mid H_d) = y\,\theta.$$

The value of (E, T) is $f(\theta x, \theta y)$ by the definition of f. However, evidence T is irrelevant and has no effect on the value of evidence E. Thus, the value of the combined evidence (E, T), $f(\theta x, \theta y)$, is equal to the value V of E, $f(x, y)$, and

$$V = f(x, y) = f(\theta x, \theta y)$$

for all θ in the interval [0,1] of possible values of $\Pr(T)$.

The only class of functions of (x, y) for which this can be said to be the case is the class which are functions of x/y or

$$\Pr(E \mid H_p)/\Pr(E \mid H_d)$$

which is the likelihood ratio. Hence the value V of evidence has to be a function of the likelihood ratio. It has been argued (Lund and Iyer, 2017) that the forensic community view the likelihood ratio as only one possible tool for communication with decision makers. The argument of Good shows that it is the only logically admissible form of evaluation.

2.3. Weight of Evidence

An interesting note of terminology can be mentioned here. It is common in some legal circles to talk of the *weight of evidence*. The concept of weight of evidence is an old idea. The term *weight of evidence* should be used for the logarithm of the likelihood ratio. The terminology was first used by Peirce (1878). The likelihood ratio is the *value* of the evidence and its logarithm is the *weight* of the evidence. The logarithm of the likelihood ratio has the pleasingly intuitive operation of additivity when converting the logarithm of the prior odds in favor of a proposition to the logarithm of the posterior odds in favor of the proposition.

$$\log\left\{\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)}\right\} = \log\left\{\frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}\right\} + \log\left\{\frac{\Pr(H_p)}{\Pr(H_d)}\right\}, \quad (4)$$

with I omitted. When considering the scales of justice it is the logarithm of the probabilities of the evidence given each of the two competing propositions that should be put in the scales, not the probabilities.

2.4. Terminology for Evidence

The evidence under consideration in this document and within the SAILR project is evidence that could have been transferred either from the crime scene to the criminal or from the criminal to the crime scene. Evidence that could have been so transferred is in the form of traces. Thus it has two names *transfer* or *trace* evidence. The evidential material discussed here is in the form of individual items. Thus, there may be a finite number of items, such as tablets or sachets of drugs or fragments of glass. Alternatively, the evidence may be a single measurement such as that of a DNA profile.

Consider the situation in which a crime has been committed, there is a crime scene and the investigation has reached the stage where a suspect has been identified. Trace evidence, denoted E, of a particular type has been found at the crime scene and on the suspect and its value is of interest. The evidence E may be partitioned into two parts, that found at the crime scene and that found in association with the suspect. In practice, the terminology takes a different form which depends on whether the source of the evidence is known or not known. A distinction is also drawn between evidential material and the evidence for evaluation. Evidence for evaluation is the observations made on the material. Only evidence which is in the form of measurements and thus represented by continuous data is considered here. Other factors such as the locations in which the material was found and the quantity of the material are not considered. Evidence of a discrete nature such as binary data as in the presence or absence of striation marks is also not considered.

Evidence whose source is known is called *control* evidence E_c. Evidence whose source is not known is called *recovered* evidence E_r. Measurements on E_c are conventionally denoted $\mathbf{x}$ where $\mathbf{x} = (x_1, \ldots, x_m)$ are m sets of measurements and where $x_i, i = 1, \ldots, m$ may be univariate or multivariate. Measurements on E_r are conventionally denoted $\mathbf{y}$ where $\mathbf{y} = (y_1, \ldots, y_n)$ are n sets of measurements and where $y_j, j = 1, \ldots, n$ may be univariate or multivariate [3].

For an evaluative comparison of $\mathbf{x}$ and $\mathbf{y}$, background data $\mathbf{z}$ are needed. These background data should be a representative sample of all possible sources from the population of interest, known as the *relevant* population. Ideally, the sample should be a random sample but this is rarely possible for practical reasons. The sample is often what might be called a *convenience* sample. If the convenience sample can be demonstrated to be composed of sources chosen in a manner independent of the case under investigation then the inference based on the comparison of $\mathbf{x}$ and $\mathbf{y}$ informed on $\mathbf{z}$ should be valid. Computation of the likelihood ratio requires data files from $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$.

One example of evidence in the form of multivariate data relates to glass elemental content. Such data are often subjected to a logarithmic transformation after taking the ratios of a particular elemental content to the oxygen content, for example, $\log_{10}(NaO) = \log_{10}(Na/O)$. These measurements can be for each of m fragments of control evidence and for each of n fragments of recovered evidence (Zadora et al., 2014). This evidence can be multivariate as there can be several ratios measured for each fragment, e.g., $\log_{10}(NaO), \log_{10}(MgO)$ and $\log_{10}(AlO)$. The control evidence is the measurements from a number m of fragments of glass from a broken window at a crime scene; the source of the fragments is known to be the window, items within source are the fragments. The recovered

[3]The use of $\mathbf{x}$ and $\mathbf{y}$ here is not to be confused with the use of $x = \Pr(E \mid H_p)$ and $y = Pr(E \mid H_d)$ in section 2.2.

evidence is the measurements from a number n of fragments of glass found in association with a suspect, for example on clothing identified as theirs. The source of the fragments of glass from the suspect is unknown. It may or may not have come from the window at the crime scene. A second example could be the measurements of color chromaticity coordinates on fibers and the evidence is bivariate (Martyna et al., 2013). There are three color chromaticity coordinates. The sum of their values is fixed so given the values of any two, the third is known. Control evidence is the measurements of color chromaticity coordinates from a number m of fibers from an article of clothing belonging to a suspect; the source is the article, the items are the fibers. Recovered evidence is the measurements of color chromaticity coordinates from a number n of fibers found at a crime scene. Thus control evidence may be found at a crime scene or in association with a suspect. Similarly, recovered evidence may be found at a crime scene or in association with a suspect.

Often the number m of control items can be chosen by the investigator. The number n of recovered items may be determined by what is available and the investigator has little choice in the selection of this number. If the number of recovered items is large, in some sense, and perhaps so large as for it to be impractical to count or analyse them, then the investigator may decide to select n items where n is less than the number available. Procedures for the choice of n and the manner of selection of the items are not discussed in this document or SAILR other than to note that the evidence selected should be representative of the total evidence available as far as is possible. Further information is available in Aitken and Taroni (2004) and references therein.

The likelihood ratio V for the comparison of $\{\mathbf{x}, \mathbf{y}\}$ where E is replaced by $\{\mathbf{x}, \mathbf{y}\}$ is then

$$\mathrm{V} = \frac{\Pr(\mathbf{x}, \mathbf{y} \mid H_p)}{\Pr(\mathbf{x}, \mathbf{y} \mid H_d)}, \tag{5}$$

where again the conditioning on I, the background information, has been omitted for clarity of notation.

Often, the propositions being considered are H_p that the control and recovered evidence are from the same source and H_d that the control and recovered evidence are from different sources. In such a circumstance, $\mathbf{x}$ and $\mathbf{y}$ may be assumed independent if H_d is true as they come from different sources. Then Equation (5) may be written as

$$\mathrm{V} = \frac{\Pr(\mathbf{x}, \mathbf{y} \mid H_p)}{\Pr(\mathbf{x} \mid H_d) \Pr(\mathbf{y} \mid H_d)}. \tag{6}$$

If $\mathbf{x}$ and $\mathbf{y}$ are continuous data, as is the case when the evidence is in the form of measurements rather than counts, the probabilities in the numerator and denominator are replaced by probability density functions, denoted say $f(\mathbf{x}, \mathbf{y})$ for the joint density and $f(\mathbf{x})$ and $f(\mathbf{y})$ for the marginal distributions. The continuous analog of Equation (6) can then be written as

$$\mathrm{V} = \frac{f(\mathbf{x}, \mathbf{y} \mid H_p)}{f(\mathbf{x} \mid H_d) f(\mathbf{y} \mid H_d)}. \tag{7}$$

In most cases, the full specification of the probability density function is unknown. The form of the distribution may be known or a reasonable assumption of its form may be made. For example, it may be known or can be assumed that the appropriate distribution is a Normal distribution. This assumption may be based on the unimodal, symmetric nature of the distribution. If the distribution has a positive skew then a transformation to normality with a logarithmic transformation of the data may be possible before consideration of the likelihood ratio. However, the parameters may neither be known nor able to be assumed known.

The numerator of Equation (7) may be written as $f(\mathbf{x}, \mathbf{y} \mid H_p) = f(\mathbf{y} \mid \mathbf{x}) \mid H_p) f(\mathbf{x} \mid H_p)$. Since the distribution of $\mathbf{x}$ is independent of whether H_p or H_d is true, $f(\mathbf{x} \mid H_p) = f(\mathbf{x} \mid H_d)$ and Equation (7) may be written as

$$f(\mathbf{y} \mid \mathbf{x}, H_p) / f(\mathbf{y} \mid H_d).$$

See Equation (18) in Appendix for an example.

2.5. Training Data

When parameters are not known, information about their possible values may be obtained from data independent of the crime but thought to be relevant for consideration of the variability in the measurements of the data comprising the evidence. These data are the *training data* or *background data* and are conventionally denoted $\mathbf{z}$. These data are considered to be a sample from a population, known as a *relevant* population. There is considerable continuing debate as to how to choose a population that is relevant for a particular crime and, once chosen, how a sample may be chosen from it to be a representative sample of the population. See, for example, *R. v. T* [2010] EWCA 2439, where the debate related to the choice of populations of shoes relevant for the consideration of evidence of shoeprints. Often the sample is a convenience sample; see section 2.4.

An alternative procedure would be to sample anew each time from a population deemed relevant to the case under investigation. A relatively early example of this is the investigation of a murder in Biggar, a town near Edinburgh, in 1967. A bite mark found on the breast of a young girl who had been murdered had certain characteristic marks, indicative of the conformation of the teeth of the person who had bitten her. A 17-year-old boy was found with this conformation and he became a suspect. Examination of 90 other boys of the suspect's age showed that the particular conformation was not at all common. The 90 other boys could be considered as a sample from a relevant population. Further details are available in Harvey et al. (1968). However, in most individual investigations it is not practical to obtain such a bespoke relevant population.

2.6. Hierarchy of Evidence

Often, with measurements, the training data can be thought of as a set of sources of items. Measurements are made of one or more characteristics of the items. For example, consider again the composition of the elemental ratio of various elements of

glass to oxygen for glass fragments from a set of windows. The items are glass fragments. A source would be a window. The training set is a set of windows. The set of windows is a sample from some population of windows, deemed relevant for crimes involving windows. The measurements are said to be *hierarchical* with two levels. One level is the fragment of glass within a window. Variation amongst measurements of fragments within a window is known as *within-group* or *within-source variation*. The second level is the window. Variation amongst measurements between windows is known as between-group or between-source variation. Measurements are taken from an item (fragments of glass) within a source (window). Notationally, the training data $\mathbf{z}$ has two indices, one for each level and may be represented as $\mathbf{z} = \{z_{k\ell}; k = 1, \ldots, g,\ \ell = 1, \ldots, h\}$ where g is the number of sources in the training set and h is the number of measurements within sources. The number of measurements within sources need not necessarily be constant though it is computationally convenient if this can be arranged during the compilation of the training set. Occasionally there may be further levels, for example measurement error.

2.7. Propositions

As well as evidence (E) and background information I, evidence evaluation depends on propositions H_p and H_d. There are different types of propositions, also known as *levels*. Both propositions (H_p and H_d) in any particular situation for the evaluation of evidence are at the same level. There are four different levels of propositions, known respectively as *offense* level, *activity* level, *source* level and *sub-source* level (Cook et al., 1998a; Evett et al., 2000).

The levels, with examples, are described as follows.

- *Offense level*: the propositions may be that the defendant is guilty of an offense (truly guilty, not just declared guilty) and that the defendant is innocent (truly innocent, not just declared not guilty).
- *Activity level*: the propositions concern an activity by the defendant which may or may not be a criminal act. An example of a pair of activity level propositions could be that the defendant hit the victim and that the defendant did not hit the victim.
- *Source level*: the propositions concern the source of evidential material. There is no consideration of the activity that may have led to the material being where it was found. An example of a pair of source level propositions could be that blood found at the scene of a crime came from the defendant and that the blood found at the scene of the crime came from some other source, unrelated to the defendant. Note that this example is one in which the two propositions are not exhaustive; relatives of the defendant are not included. SAILR can only be used for likelihood ratio computation on source level.
- *Sub-source level*: the propositions concern material for which it is not possible to identify a source. An example of a pair of sub-source level propositions could be that DNA found at a crime scene came from the defendant and that DNA found at the crime scene came from some other source, unrelated to the defendant. The quantity of material found is insufficient to identify its source, e.g., whether it came from blood or semen.

3. FRAMEWORK FOR MODELS

The likelihood ratio may be used in the context of forensic science in two different ways, that of comparison and that of discrimination. For comparison, two pieces of evidence found in different places are compared to see if they had a common source. For discrimination, one piece of evidence is compared with several sets of training or background data from different sources to see from which source the evidence may have come.

Most of the models described here are so-called *feature-based models*. These are models developed from the measurements (features) on the evidential material. Other models described are so-called *score-based models*. There may be occasions with multivariate data when a feature-based model is not tractable, e.g., multidimensional binary data where the number of possible models is unmanageable. On such occasions, the distance, denoted $d(\mathbf{x}, \mathbf{y})$, between control ($\mathbf{x}$) and recovered ($\mathbf{y}$) data can be used instead.

3.1. Comparison for Feature-Based Models

3.1.1. The Likelihood Ratio Approach for Continuous Univariate Evidential Data With Normal Distributions for the Means and Known Variances

A common problem occurs in forensic science when the prosecution and defense propositions concern whether two objects are from the same source or from different sources. For example, if a glass fragment is found on a suspect and there is a broken window at the crime scene, one proposition might be that the glass fragment found on the suspect came from the window at the crime scene, and the other proposition might be that the glass fragment came from some other window. The evidence is given by a set of measurements from the glass fragment found on the suspect (the recovered sample) and a set of measurements from one or more glass fragments from the crime scene (the control sample). The problem is one of comparison.

The structure of these models reflects the hierarchical nature of the underlying data (measurements and variation within a source and then variation between sources). Using a distribution for the means θ_1 and θ_2 in this way accounts for variance within source (σ^2) and variance between sources (τ^2).

The problem for the fact-finder is to determine which of the two propositions (H_p or H_d) is more likely, given all of the evidence in the case. Denote the other evidence and background information by I as before. The fact-finder can consider which proposition is more likely by considering the relative size of the two probabilities $\Pr(H_p \mid \bar{x}, \bar{y}, I)$ and $\Pr(H_d \mid \bar{x}, \bar{y}, I)$ (technically, in cases where the statistical assumptions include knowledge of the variances σ^2 and τ^2 and of a Normal distribution for the measurements, the means of the control and recovered samples are sufficient statistics so can be used in place of the measurements $\mathbf{x}$ and $\mathbf{y}$). Let $f(\bar{x}, \bar{y} \mid H_p, I)$ be the joint probability density function of $\bar{x}$

and $\bar{y}$, given proposition H_p and I and let $f(\bar{x}, \bar{y} \mid H_d, I)$ be the joint probability density function of $\bar{x}$ and $\bar{y}$ given proposition H_d and I. In this context Equation (1) may be represented as

$$\frac{P(H_p \mid \bar{x}, \bar{y}, I)}{P(H_d \mid \bar{x}, \bar{y}, I)} = \frac{f(\bar{x}, \bar{y} \mid H_p, I)}{f(\bar{x}, \bar{y} \mid H_d, I)} \times \frac{P(H_p \mid I)}{P(H_d \mid I)}, \tag{8}$$

where E is replaced by $(\bar{x}, \bar{y})$. For examples where the within-source variance is not known, the sample variances of $\mathbf{x}$ and $\mathbf{y}$ will also be included in the representation.

Denote the common mean of the measurements under the prosecution proposition by $\theta_1 = \theta_2 = \theta$. The likelihood ratio V is given by Equation (7). This may be rewritten as

$$\mathrm{V} = \frac{\int f(\bar{x} \mid \theta) f(\bar{y} \mid \theta) f(\theta) \mathrm{d}\theta}{\int f(\bar{x} \mid \theta_1) f(\theta_1) \mathrm{d}\theta_1 \int f(\bar{y} \mid \theta_2) f(\theta_2) \mathrm{d}\theta_2}, \tag{9}$$

where the dependence on I has been suppressed for ease of notation. The analytical form of this likelihood ratio, given the independence and Normality assumptions detailed above, is given by Lindley (1977). The density functions $f(\bar{x} \mid \theta)$ and $f(\bar{y} \mid \theta)$ are taken to be density functions of a Normal distribution. Note that when the prosecution proposition is chosen the random variables $\bar{X}$ and $\bar{Y}$, of which $\bar{x}$ and $\bar{y}$ are realizations, are conditionally independent, conditional on θ. They are independent if it is known they are from the same source. The distributions associated with these density functions are termed the within-source distributions, because they account for the within-source variability. The distribution associated with the density function $f(\theta)$ is termed the between-source distribution because it accounts for between-source variability, and it is a prior distribution for θ. The use of a between-source distribution allows the rarity of the data $\mathbf{x}$ and $\mathbf{y}$ to be taken into account when assessing the strength of the evidence; see Equation (13) in the Appendix for an example. Information to assist with the estimation of the prior distribution is contained in the training set. If the control and recovered samples have similar means, and the mean is unusual, then the strength of evidence supporting the proposition that the samples are from the same source should be stronger than if the mean is relatively common.

A solution to this problem of the comparison of sources in the case where the measurements are univariate and are assumed to be independent and Normally distributed was developed by Lindley (1977). Some details are given in the Appendix; see Equations (12) and (13) in the Appendix. Denote the m measurements on the control sample by $\mathbf{x} = (x_1, \ldots, x_m)$ and the n measurements on the recovered sample by $\mathbf{y} = (y_1, \ldots y_n)$. The corresponding means of each of these samples are denoted $\bar{x}$ and $\bar{y}$. The two propositions to be considered are at the source level and are:

- H_p: the control and recovered sample are from the same source.
- H_d: the control and recovered sample are from different sources.

Lindley's solution assumes that the means $\bar{x}$ and $\bar{y}$ of the control and recovered samples are sample means of data, whose corresponding random variables have Normal distributions with means θ_1 (control) and θ_2 (recovered), respectively, and variances σ^2/m (control) and σ^2/n (recovered). The variance σ^2 is a within-group (e.g., within window) variance. The means θ_1 and θ_2 are the means of the groups associated with $\mathbf{x}$ and $\mathbf{y}$ in the terminology of hierarchical data. Variability between groups has also to be considered. This is done with consideration of the variation in the group means. The two means θ_1 and θ_2 are also assumed to be realizations of a random variable which is Normally distributed, this time with mean μ and variance τ^2. At present the variances σ^2 and τ^2 are assumed known. Also, the within-group variance σ^2 is assumed constant within groups. An expression for the likelihood ratio if the between-group distribution is not Normal but is represented with a general distribution $p(\cdot)$, with second derivative $p''(\cdot)$ is given by Equation (14) in Appendix.

An extension using kernel density estimation has been derived to allow for a general non-Normal between-group distribution Equation (15) in Appendix. Checks of the distributional assumptions and estimation of hyperparameters are made using a training set of groups which are assumed to be a random sample of groups (sources) from some relevant population. Later work (e.g., Bozza et al., 2008 with an extension to multivariate data, Equation 24 in Appendix) relaxes the assumption that σ^2 and τ^2 are known.

The likelihood ratio can be used to assess evidence in a criminal trial and hence is a solution to the comparison of sources problem; Lindley (1977).

This approach for evidence evaluation based on the likelihood ratio is different from an approach based on hypothesis testing. The likelihood ratio approach has many advantages; a discussion of these can be seen in Aitken and Stoney (1991) and Aitken and Taroni (2004). One such advantage is that the likelihood ratio has no dependence on an arbitrary cut off point (e.g., 5% significance). Another advantage is that the use of a likelihood ratio reduces the risk that a transposition of the conditional probabilities (also known as the prosecutor's fallacy) occurs, a transposition which confuses the probability of finding the evidence on an innocent person with the probability of the innocence of a person on whom the evidence has been found. In addition, the likelihood ratio provides a method of comparing the likelihood of the evidence under the propositions of both the prosecution and the defense. This guards against potentially misleading situations when the likelihood under only one of these propositions is considered. Finally, an approach based on the likelihood ratio ensures equality of treatment of both propositions. In a procedure based on hypothesis testing, a null hypothesis is assumed true unless sufficient evidence is found to reject it at a pre-specified significance level. Often, the null hypothesis is that of a common source, $\theta_1 = \theta_2$ in Lindley's example. This is the prosecution proposition. Thus the burden of proof is placed on the defense to put forward sufficient evidence to enable rejection of the prosecution proposition, contrary to the dictum of "proof beyond reasonable doubt." The prosecution need prove nothing.

3.1.2. The Likelihood Ratio Approach for Other Forms of Continuous Evidential Data, Including Multivariate Data

Later work on evidence evaluation has extended the work done in Lindley (1977) to cover other data types, allowing for different forms of the within and between source distributions (Aitken and Lucy, 2004; Aitken et al., 2006, 2007a). In Bozza et al. (2008) and Alberink et al. (2013), extensions are given so that the between-source distribution in Equation (9) becomes a function of both the mean and the variance. This allows for variation in the variance of samples from different sources. All of these extensions assume that the m measurements $\mathbf{x}$ are independent and that the n measurements $\mathbf{y}$ are independent. Methods for autocorrelated data types, such as measurements associated with drug traces on banknotes are described in Wilson et al. (2014, 2015).

For multivariate measurements which are independent and which have a multivariate Normal distribution the analytical form is derived in Aitken and Lucy (2004). The likelihood ratio is given for two forms of the distribution of the mean between sources. The first form assumes multivariate Normality, and the second form uses nonparametric kernel density estimation. The within-source variance is assumed constant over all sources.

When there are several variables graphical models may be used to reduce the number of parameters needing to be estimated. The kernel density approach given in Aitken and Lucy (2004) can then be used to calculate likelihood ratios for the subsets of variables as indicated by the graphical models. The graphical model considers partial correlations amongst the variables and partitions these variables into overlapping subsets known as *cliques*. The overall distribution may then be represented as a function of the distributions over the cliques. These clique distributions have very few variables each (e.g., one, two or three; and the overall likelihood ratio is then a product of likelihood ratios which are based on one-, two- or three-dimensional data Aitken et al., 2007). Such a process for the reduction of dimension is necessary to avoid the curse of dimensionality whereby very large data sets are needed for the estimation of parameters in a multi-dimensional parameter set.

In Aitken et al. (2006) the multivariate methods used in Aitken and Lucy (2004) assuming Normality are extended further to allow for another level of variance (e.g., measurement error) to be taken into account, giving a three-level model. A model assuming an exponential distribution for between-sources in a three-level model is assumed in Aitken et al. (2007a) and the analytical form of the likelihood ratio is derived. Variation between the means of samples from different sources, variation between the means of different samples taken from the same source and variation within repeated measurements on the same sample are taken into account.

Relaxation of the assumption that samples from different sources will have the same variance means that an analytical solution is not available. Measurements are assumed multivariate and independently Normally distributed as before but the between-source (prior) distribution is taken to be the product of a multivariate Normal distribution (for the mean of the between-source distribution) and an inverse Wishart distribution (for the covariance of the between-source distribution). In this way, variation of covariances, as well as means, between different sources is taken into account. An analytical form of the likelihood ratio is not available so Markov chain Monte Carlo (MCMC) methods are used to estimate it (Bozza et al., 2008) (Equation 24 in the Appendix).

A similar approach to Bozza et al. (2008) for the evaluation of the likelihood ratio for the comparison of sources problem is used by Alberink et al. (2013) in that variation in the variance parameter between sources is modeled as well as variation in the mean parameter, although in Alberink et al. (2013) the data are univariate. As with all of the other approaches discussed, the within-source distribution is Normal, and the data are assumed independent. There are two main extensions seen in Alberink et al. (2013). The first is that three different distributions are used for the between-source distribution. One is the univariate equivalent of the between-source distribution used in Bozza et al. (2008) (a semi-conjugate prior), one is a non-informative prior, proportional to the inverse of the variance, and one is the conjugate prior distribution seen on p. 74 of Gelman et al. (2004). This conjugate prior distribution gives a between-source distribution for the parameter (μ, σ^2), denoting group mean and variance, of

$$\mu \sim N(\mu_0, \sigma^2/\kappa_0)$$
$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

where μ_0, κ_0, ν_0 and σ_0^2 are hyperparameters to be estimated and the notation Inv-χ^2 corresponds to a scaled inverse chi-squared distribution. The difference between this and the univariate equivalent of the between-source distribution used in Bozza et al. (2008) is that the variance of the parameter μ is proportional to σ^2. An analytical form of the likelihood ratio for the two cases when the between-source distribution is given by the non-informative prior and when the between-source distribution is given by the conjugate prior (Alberink et al., 2013) who also show that no analytic solution exists if a semi-conjugate prior is used. (See Equations (16,17) in the Appendix.)

As in Bozza et al. (2008), Alberink et al. (2013) use MCMC methods to evaluate the likelihood ratio when the between-source distribution is given by the semi-conjugate prior, although there are differences in the implementation, leading to the second main extension. Alberink et al. (2013) use prior distributions on the hyperparameters of the between-source distribution and then combine these prior distributions with training data to obtain a posterior distribution for the hyperparameters, conditional on the training data. All of the other methods discussed estimate the parameters of the between-source distribution directly from the training data using summary statistics. The methods used in Alberink et al. (2013) allow for a Bayesian approach for the estimation of the between-source distribution. One disadvantage of this approach is that the method for estimating the likelihood ratio used in Bozza et al. (2008) is no longer feasible because, instead of having a known analytic form for the between-source density function, draws from the between-source distribution are obtained using MCMC methods. Monte Carlo integration is used by Alberink et al. (2013) to estimate the likelihood ratio.

All of the literature discussed in sections 3.1.1 and 3.1.2 evaluates likelihood ratios for continuous evidential data. There are some common assumptions. All assume that measurements are independent and that the within-source distribution is Normal (univariate or multivariate). Constant variation between sources of the within-source distribution is assumed by Lindley (1977), Aitken and Lucy (2004), Aitken et al. (2007) and Aitken et al. (2006). This assumption is relaxed by Bozza et al. (2008) and Alberink et al. (2013), allowing the variance to vary between sources. A Bayesian approach is used by Alberink et al. (2013) to obtain the parameters of the between-source distribution.

Methods for the evaluation of continuous, autocorrelated data are described in Wilson et al. (2014) and Wilson et al. (2015). The data used for illustration are the quantities of drugs on banknotes where quantities on adjacent notes cannot be considered independent. Some work has also been done on the evaluation of evidence for discrete data, particularly in the field of DNA profiling (Buckleton et al., 2005) and more recently on data relating to clicks in speech (Aitken and Gold, 2013) and the presence or absence (binary data) of striation marks for screwdrivers (Aitken and Huang, 2015).

3.2. Discrimination

Forensic scientists are not only interested in comparisons of two pieces of evidence, such as control and recovered evidence, under different propositions, that of same source vs. that of different source, without attention being paid to the identity of the source. There is also interest in the source of one piece of evidence. The support of the evidence for a proposition of source is of interest. The problem concerns the determination of whether a sample of data is more likely to be from one population (source) or another. Of course, such a determination is the concern of the fact-finder. The scientist is concerned with the probability of the measurements on the evidential material if the material came from one source or if it came from another. If there are more than two possible sources, then prior probabilities, that is, probabilities for each source under consideration before the material is examined, are needed in order to obtain a likelihood ratio. In this problem there is only one set of evidential data compared with the two sets (control and recovered) in the comparison problem. The aim is to assist the decision-maker as to the population of origin of the evidential data. This is a problem of *discrimination*, as distinct from a problem of *comparison*.

An example of the use of likelihood ratios in a problem of this sort can be seen in Zadora et al. (2010) which looks at the discrimination of glass samples and in Wilson et al. (2014, 2015) which considers discrimination between banknotes assocated with a person associated with criminal activity and banknotes associated with a person not associated with criminal activity. As with the problem of comparison of sources, the likelihood ratio alone cannot determine whether a set of data is more likely from one population or another; it must be considered in conjunction with the prior odds. The derivation of the likelihood ratio for such discrimination problems is discussed in Taroni et al. (2010) (Chapter 8). The likelihood ratio for a set of evidence consisting of n measurements, $\mathbf{z} = (z_1, \ldots, z_n)$, under two propositions, H_p and H_d, is considered.[4] The two propositions are given by

- H_p : data $\mathbf{z}$ are from population 1, and
- H_d : data $\mathbf{z}$ are from population 2.

The likelihood ratio V for the discrimination problem, where I is the background information as usual, is given in Taroni et al. (2010) by

$$\mathrm{V} = \frac{f(\mathbf{z} \mid H_p, I)}{f(\mathbf{z} \mid H_d, I)}. \tag{10}$$

This expression can be compared with Equation (7) and the comparison problem. In the comparison context, the joint density function of control and recovered data is considered. In the discrimination problem, two (or more) possible sources (populations) are identified.

Assume as for the comparison problem that the data are hierarchical and that there are two possible sources. The probability density function of groups of data from source i is parameterized by $\theta_i,\ i = 1, 2$ (possibly multivariate). If the value of θ_i (for $i \in \{1, 2\}$) varies between different groups in population i then by conditioning on θ_1 in the numerator and θ_2 in the denominator, the likelihood ratio V can be written

$$\mathrm{V} = \frac{\int f(\mathbf{z} \mid \theta_1) f(\theta_1) \mathrm{d}\theta_1}{\int f(\mathbf{z} \mid \theta_2) f(\theta_2) \mathrm{d}\theta_2}. \tag{11}$$

The probability density function $f(\theta_i)$ models the variability of the parameter θ_i between groups in population i, and is termed the between-group density function (the associated distribution function will be termed the between-group distribution function). This is analogous to the between-source distribution used to model variability between sources in the comparison of sources problem. Similarly, the density function $f(\mathbf{z} \mid \theta_i)$ is termed the within-group density function (with the associated distribution function termed the within-group distribution function).

Using this formulation for the likelihood ratio, the methods discussed previously for the evaluation of the likelihood ratio for the comparison of sources problem can be adapted to evaluate the value of evidence for discrimination problems. The limitations and assumptions of these methods still apply.

In the context of discrimination, training data are a random sample of groups from each or both of the sources. Variation is between groups within each of the sources. There is an abuse of terminology here. In the comparison problem with the proposition of common source, the control and recovered evidence are deemed to be from the same source but without specification of the source. The source is a member of a population of sources. In the discrimination problem, support for a particular source is assessed. The distinction between comparison and discrimination problems is emphasized in Zadora et al. (2014) where the two problems are discussed in

[4]Note the change of use of notation. In this section, $\mathbf{z}$ refers to evidential data and not to training data.

different chapters (and note that discrimination is there noted as classification).

3.3. Score-Based Models

Return now to consideration of the problem of comparison of sources with a p-dimensional control measurement $\mathbf{x} = (x_1, \ldots, x_p)$ and a p-dimensional recovered measurement $\mathbf{y} = (y_1, \ldots, y_p)$. For those occasions when a feature-based model is not tractable (e.g., multidimensional binary data), the distance $d(\mathbf{x}, \mathbf{y})$, known as a *score* can be used instead. The value of the evidence is then

$$\mathrm{V} = \frac{f(d(\mathbf{x}, \mathbf{y}) \mid H_p, I)}{f(d(\mathbf{x}, \mathbf{y}) \mid H_d, I)}.$$

Rarity is not considered. Inference may then continue as before but using the score, which is univariate, as the statistic of interest. Score-based approaches estimate the probability distribution function of a calculated score. Score-based approaches have been used for handwriting (Hepler et al., 2012) and speech recognition (Brümmer and Du Preez, 2006; Gonzalez-Rodriguez et al., 2006; Morrison, 2011). Score-based methods do not require the distributional assumptions (such as within-source Normality) needed to fit the models described above but do still require a function to be chosen to model the probability distribution function of the score.

There are various distance measures that may be used. Three examples are

- Euclidean: $d = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$;
- Manhattan: $d = \sum_{i=1}^{p} \mid x_i - y_i \mid$;
- Pearson correlation distance: $100(1 - r)/2$ with

$$r = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \sum_{i=1}^{p}(y_i - \bar{y})^2}}.$$

Other examples are available in SAILR. For multiple control and recovered data $\mathbf{x}_i, i = 1, \ldots, m$ and $\mathbf{y}_i, i = 1, \ldots, n$, respectively, pairwise score measurements or means can be used.

For the calculation of score-based likelihood ratios, distributions of scores of same-source comparisons and of different-source comparisons are required. Determination of the same-source distribution can be made by comparing every measurement in a training set $\mathbf{z}$ with every other measurement within its own source except with itself for which the distance is zero. For the different-source distribution, every measurement is compared with all measurements from other sources. These results may then be used to estimate the distributions of same-source and between-source comparisons. The distributions can be represented initially by histograms. They may then be smoothed with a kernel density estimation or an appropriate parametric distribution. The current choice of parametric distribution in SAILR is a Gamma distribution or a Weibull distribution. The chosen distribution functions, one for same-source comparisons and one for different-source comparisons, then can be used to determine the density calculation of the evidence score for both distributions and hence calculate a likelihood ratio.

3.4. Comparison of Feature-Based and Score-Based Models

Models for discrimination and for comparison that use the original data are feature-based models. The models discussed in sections 3.1 and 3.2 are all feature-based. Feature-based multivariate Normal models compare the probability of observing the evidence given that the evidential samples (control and recovered) measured, and compared, come from the same source or come from different sources. In contrast, the score-based model compares the probability of observing the pairwise similarity between two samples (control and recovered) given that they come from the same source with the probability of the pairwise similarity given that the samples come from different sources. A comparison of the performances of score-based and frequency-based likelihood ratios for forensic MDMA comparisons is given in Bolck et al. (2015).

The benefits and shortcomings of both methods are given by Bolck et al. (2015) as:

- Feature-based benefits:
  - Original data dimensionality preserved; no information loss.
  - Rarity and similarity of the features relate directly to the magnitude of the likelihood ratio.
- Feature-based shortcomings:
  - Covariance estimation is difficult when limited data are available relative to the dimensionality of the variables.
  - The feature-based method is often less robust than the score-based model when there are limited population samples.
- Score-based benefits:
  - Covariance estimation between sources is possible with few samples available.
  - The method is robust and able to be generalized to new samples.
- Score-based shortcomings:
  - There is a loss of information because of a reduction of dimensionality.
  - The value of the likelihood ratio is based on the similarity of pairwise scores rather than the similarity and rarity of features.

3.5. Summary of Feature-Based Models

References for details of a selection of feature-based two-level models with within-group measurements independent and Normally distributed are listed here. Equation numbers are given for models for which further details are given in the Appendix.

- Univariate:
  - Within-group Normal,
    Between-group Normal for between-group mean (assume within-group variance known)
    (Lindley, 1977, see Equations 12, 13 in the Appendix).

  - Within-group Normal,
    Between-group Taylor expansion for between-group mean (assume within-group variance known)
    (Lindley, 1977, see Equation 14 in the Appendix).
  - Within-group Normal,
    Between-group kernel for between-group mean (assume within-group variance known),
    (Aitken and Taroni, 2004, see Equation 14 in the Appendix).
  - Within-group Normal,
    Between-group distribution:
    (a) Normal distribution - semi-conjugate prior,
    (b) Non-informative prior, proportional to the inverse of the variance,
    (c) Conjugate prior - Normal, scaled inverse chi-squared
    (Alberink et al., 2013, see Equations 15, 16 in the Appendix).

- Bivariate:
  Numerator (predictive distribution) (Bernardo and Smith, 1994),
  Denominator (kernel),
  (Evett et al., 1987, see Equation 18 in the Appendix).
- Multivariate, within-group measurements independent and Normally distributed
  - Within-group Normal,
    Between-group kernel for distribution of group means,
    Within-group variance assumed common and estimated from training data, (Aitken and Lucy, 2004, 4.1, see Equation 19 in the Appendix).
  - Within-group Normal
    Between-group Normal for distribution of group means,
    Inverse Wishart for the covariance of within-source distribution, (Bozza et al., 2008, see Equation 24 in the Appendix).
  - With graphical models:
    See section 3.1.1; Aitken et al. (2007).
  - In the presence of zeros, that is when no measurement of a specific characteristic has been made on certain members of the control data set, the recovered data set or the training data set: both Normal and kernel between-group distributions considered. Estimation of covariance matrices by imputation and by available cases (Zadora et al., 2010).
  - In addition, when within-group measurements are autocorrelated and Normally distributed (see Wilson et al., 2014, 2015).

4. MODEL PERFORMANCE

Model performance for the comparison problem is assessed with a training set and associated data **z** as discussed in section 2.6. If possible, another set, known as a *validation* set could be used. The training set and validation set should both comprise several sources of data from a relevant population. Within each source, measurements are taken on each of several items. The source of each member of the two sets is known. Models and parameters can be fitted using the training set. The performance can be assessed using the validation set. Thus when a method for comparison or discrimination is tested using members of the data set it is known if the correct answer is given. In the absence of a validation set, the performance can be assessed through a second use of the training set (e.g., with a leaving-one-out method). Validation enables the provision of measures of performance based on calculated likelihood ratios.

For a comparison of two members of the validation (or training) set a likelihood ratio is calculated. There are two conclusions that may be drawn by the fact-finder: they are from the same source or they are not from the same source. If the likelihood ratio is greater than 1, then this is support for the proposition of a common source for the two members of the validation (training) set being compared. If they are truly from the same source then this is counted as a correct result. Similarly, if its value is less than 1, then this is support for the proposition of different sources for the two members of the validation (training) set being compared. If they are truly from different sources then this is counted as a correct result. However, if the two members have a value for the likelihood ratio of greater than 1 when they are from different sources, this is an incorrect result and the result is known as a *false positive*. Similarly, if the two members have a value for the likelihood ratio of less than 1 when they are from the same source, this is an incorrect result and the result is known as a *false negative*.

For discrimination with two groups, say A and B, the member of the data set may be classified by the fact-finder as belonging to group A or to group B. False positives and false negatives can be defined in a manner analogous to that of the comparison procedure. A likelihood ratio is calculated. If its value is greater than 1, then this is support for the proposition that the member of the training set belongs to group A, say. If the member is truly from group A then this is counted as a correct result. Similarly, if its value is less than 1, then this is support for the proposition that the member is from group B. If it is truly from group B, then this is counted as a correct result. However, if the member has a value for the likelihood ratio of greater than 1 when it is from group B, this is an incorrect result and the result is a false positive, say. Similarly, if the member has a value for the likelihood ratio of less than 1 when it is from group A, this is an incorrect result and the result is a false negative.

For both comparison and discrimination problems, the strength of the support is measured by the value of the likelihood ratio. As noted in section 2.3 if the logarithm is taken this is known as the weight of evidence. Given the existence of a validation (training) set it is possible to measure the performance of a method for comparison or discrimination as the correct answer is known. It is not possible to assess the result in an individual case; the correct answer in an individual case is not known.

The likelihood ratio, or a function of it such as the logarithm, has been shown by Good, 1989a,b (section 2.2) to provide the best (only) value of the evidence. Attempts to express the uncertainty associated with this assessment (e.g. with a confidence interval) are attempts to put a probability on a probability and should not be done (Taroni et al., 2016). This view is not universally agreed, see discussion issues of *Law, Probability and Risk* (2016, volume 15, issue 1) and *Science and Justice* (2017, volume 56).

Note also the quote from Kaye (1979) in section 2: "It thus supplies the jurors with as precise and accurate an illustration of the probative force of the quantitative data as the mathematical theory of probability can provide". It is not necessary to provide an interval estimate.

There are several measures of performance.

- *The percentage of false positives and of false negatives amongst all the comparisons or discriminations tested.* Often, in a criminal case, one of the propositions is associated with the prosecution, hence the notation H_p, and other is associated with the defense, with the notation H_d. In such a circumstance, the burden of proof lies with the prosecution. It is a more serious error to support the prosecution proposition wrongly than to support the defense proposition wrongly. Let support for the prosecution proposition be known as a positive result. Thus, when considering the performance of a test, it is better to choose a test in which there is a low false positive rate and a high false negative rate rather than one in which there is a high false positive rate and low false negative rate. Ideally, zero false positive and zero false negative results are best but such an ideal is rarely achieved.
- *A Tippett plot.* See Evett and Buckleton (1996) and Tippett et al. (1968). This is a graphical measure of rates of misleading evidence for comparisons. It is the complement of empirical cumulative distribution functions for same-source and different-source comparisons. The plots come in pairs, one for same-source comparisons and one for different-source comparisons. The $log(LR)$ is plotted on the x-axis and, for a particular value x_0 of the $log(LR)$, the y-axis is the relative frequency of the number of comparisons greater than x_0. For same-source comparisons, it is to be hoped that all $log(LR)$ values are greater than 0. Thus for $x < 0$, it is hoped the corresponding value on the y-axis will be 1 (or 100%). Similarly, for different-source comparisons, it is to be hoped that all $log(LR)$ values are less than 0. Thus for $x > 0$, it is hoped the corresponding value on the y-axis will be 0 (or 0%).

  The vertical distance from the intersection of the same-source plot with the line $log(LR) = 0$ and the line $y = 1$(100%) is the rate of misleading evidence for same-source comparisons, the proportion of same-source comparisons that have a value of $log(LR) < 0$ ($LR = 1$). The vertical distance from the intersection of the different-source plot with the line $log(LR) = 0$ and the line $y = 0$(0%) is the rate of misleading evidence for different-source comparisons, the proportion of different-source comparisons that have a value of $log(LR) > 0$ ($LR = 1$).
- *Detection error trade-off (DET) curves.* See Meuwly et al. (2017). A detection error trade-off (DET) plot is a 2-dimensional graphical representation in which the proportion of false positives is plotted as a function of the proportion of false negatives. The closer the curves to the coordinate origin, the better are the discriminating capabilities of the method. The intersection of a DET curve with the main diagonal of the DET plot marks the Equal Error Rate (EER) which is the point when the proportions of false positives and false negatives are equal.
- *Empirical cross-entropy.* See Meuwly et al. (2017), Ramos et al. (2013) and Ramos and Gonzalez-Rodriguez (2013).
  The performance of probabilistic assessments has been addressed by *strictly proper scoring rules* (SPSR). Consider two propositions about a parameter θ, one that $\theta = \theta_p$ and one that $\theta = \theta_d$, with $\Pr(\theta = \theta_p) = 1 - \Pr(\theta = \theta_d)$ For evidence evaluation, the *logarithmic* SPSR is used and defined as

$$C(\Pr(\theta_p \mid I), \theta) = -\log_2(\Pr(\theta_p \mid I)) \text{ if } \theta = \theta_p,$$

$$= -\log_2(1 - \Pr(\theta_d \mid I)) \text{ if } \theta = \theta_d,$$

  The measure of accuracy for evidence evaluation based on the SPSR is a weighted average value of the logarithmic scoring rule, and is known as the *empirical cross-entropy* (*ECE*):

$$\begin{aligned} ECE = & -\frac{\Pr(\theta_p \mid I)}{N_p} \sum_{\theta_{(i)}=\theta_p} \log_2 \Pr(\theta_p \mid E_i, I) \\ & -\frac{\Pr(\theta_d \mid I)}{N_d} \sum_{\theta_{(j)}=\theta_d} \log_2 \Pr(\theta_d \mid E_j, I) \\ = & \frac{\Pr(\theta_p \mid I)}{N_p} \sum_{\theta_{(i)}=\theta_p} \log_2 \left(1 + \frac{1}{LR_i \times O(\theta_p)}\right) \\ & +\frac{\Pr(\theta_d \mid I)}{N_d} \sum_{\theta_{(j)}=\theta_d} \log_2 \left(1 + LR_j \times O(\theta_p)\right), \end{aligned}$$

  where $LR_i(LR_j)$ is the likelihood ratio for the i-th (j-th)E_i (E_j) piece of evidence where $\theta = \theta_i(\theta_j)$, respectively, and $O(\theta_p)$ denotes the prior odds $\Pr(H_p)/Pr(H_d)$. For the discrimination problem with two sources, the parameters θ_p and θ_d represent the parameters of the two sources. For the comparison problem θ_p represents same-source comparisons and θ_d represents different-source comparisons in the validation dataset.

  This measure tends to indicate better performance when the likelihood ratio leads to the correct decision. The numerical value will be lower as the performance increases. The ECE can be represented as an ECE-plot, showing its value for a certain range of priors.

4.1. Conclusion

The development of methods for the evaluation of evidence for frequency-based continuous two-level models is described from the hierarchical model for univariate continuous data developed by Lindley (1977) to multivariate models with unknown means and covariances (Bozza et al., 2008). This development is of interest in its own right as a compilation of some thirty years of development. However, it also provides a background to the development of the SAILR package, a package which extends these ideas to include score-based models.

Formulae for many of these are given in the Appendix and may also be found in many books on the subject (e.g., Aitken and Taroni, 2004; Zadora et al., 2014).

There is much more that can be reviewed. References for some of the omissions of this paper are given here. It is hoped they are useful. There have been few papers on models for discrete data; see Aitken and Gold (2013) for an example. Score-based models have received a lot of attention recently and are included in SAILR; see Bolck et al. (2015) for examples. Graphical models provide an approach for a reduction in the dimensionality of multivariate problems; see Zadora et al. (2014) for examples.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work was supported by the European Network of Forensic Science Institutes 2013 Monopoly programme grant T6 for Software for the Analysis and Implementation of Likelihood Ratios (SAiLR), the Leverhulme Trust, grant number EM2016-027, and the Swiss National Science Foundation, grant number BSSGI0_155809.

ACKNOWLEDGMENTS

The author acknowledges very helpful contributions from Annabel Bolck and all other members of the SAILR group including Leon Aronson, David Lucy, Jonas Malmborg, Petter Mostad, Tereza Neocleous, Anders Nordgaard, Jane Palmborg, Amy Wilson and Grzegorz Zadora. An early version of this document was written as an internal landscape document for the SAILR project. Further information about the project is available from Dr. Jeannette Leegwater at the Netherlands Forensic Institute (jleegwater@nfi.minvenj.nl). Details of the software are available on-line from https://downloads.holmes.nl/sailr/sailr. Operation of SAILR requires at least Java 8 to be installed. Java 8 can be downloaded from http://www.oracle.com/technetwork/pt/java/javase/downloads/jre8-downloads-2133155.html.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00126/full#supplementary-material

REFERENCES

Aitken, C. G. G., and Huang, C. (2015). Evidence evaluation for hierarchical, longitudinal, binary data using a distance measure. *Stat. Appl. Ital. J. Appl. Stat.* 27, 213–223. Available online at: http://sa-ijas.stat.unipd.it/sites/sa-ijas.stat.unipd.it/files/Aitken%20and%20Huang.pdf

Aitken, C. G. G., and Nordgaard, A. (2017). Letter to the editor – the roles of participants' differing background information in the evaluation of evidence. *J. Forensic Sci.* 63, 648–649. doi: 10.1111/1556-4029.13712

Aitken, C. G. G., Shen, Q., Jensen, R., and Hayes, B. (2007a). The evaluation of evidence for exponentially distributed data. *Comput. Stat. Data Anal.* 51, 5682–5693. doi: 10.1016/j.csda.2007.05.026

Aitken, C. G. G., and Stoney, D. A. (1991). *The Use of Statistics in Forensic Science*. Chichester: Ellis Horwood Limited.

Aitken, C. G. G., and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists, 2 Ed* . Chichester: Wiley.

Aitken, C. G. G., Zadora, G., and Lucy, D. (2007). A two-level model for evidence evaluation. *J. Forensic Sci.* 52, 412–419. doi: 10.1111/j.1556-4029.2006.00358.x

Aitken, C. G. G., and Gold, E. (2013). Evidence evaluation for discrete data. *Forensic Sci. Int.* 230, 147–155. doi: 10.1016/j.forsciint.2013.02.042

Aitken, C. G. G., and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. R. Stat. Soc. C Appl. Stat.* 53, 109–122. doi: 10.1046/j.0035-9254.2003.05271.x

Aitken, C. G. G., Lucy, D., Zadora, G., and Curran, J. M. (2006). Evaluation of transfer evidence for three-level multivariate data with the use of graphical models. *Computat. Stat. Data Anal.* 50, 2571–2588. doi: 10.1016/j.csda.2005.04.005

Alberink, I., Bolck, A., and Menges, S. (2013). Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data. *J. Appl. Stat.* 40, 2579–2600. doi: 10.1080/02664763.2013.822056

Bernardo, J., and Smith, A. (1994). *Bayesian Theory*. Chichester: John Wiley and Sons.

Biedermann, A., Hicks, T., Taroni, D., Champod, C., and Aitken, C. G. G. (2014). On the use of the likelihood ratio for forensic evaluation: response to Fenton et al. [2014a]. *Sci. Just.* 54, 316–318. doi: 10.1016/j.scijus.2014.04.001

Bolck, A., Ni, H., and Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law Probab. Risk* 14, 243–266. doi: 10.1093/lpr/mgv009

Bozza, S., Taroni, F., Marquis, R., and Schmittbühl, M. (2008). Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *J. R. Stat. Soc. C Appl. Stat.* 57, 329–341. doi: 10.1111/j.1467-9876.2007.00616.x

Brümmer, N., and Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Comput. Speech Lang.* 20, 230–275. doi: 10.1016/j.csl.2005.08.001

Buckleton, J., Triggs, C., and Walsh, S. (2005). *Forensic DNA Evidence Interpretation*. Boca Raton, FL: CRC Press.

Champod, C., Taroni, F., and Margot, P. (1999). The Dreyfus case - an early debate on experts' conclusions (an early and controversial case on questioned document examination). *Int. J. Forensic Doc. Exam.* 5, 446–459.

Cook, R., Evett, I.W., Jackson, G., Jones, P. J., and Lambert, J. A. (1998a). A hierarchy of propositions: deciding which level to address in casework. *Sci. Justice* 38, 231–239. doi: 10.1016/S1355-0306(98)72117-3

Cook, R., Evett, I., Jackson, G., Jones, P. J., and Lambert, J. A. (1998b). A model for case assessment and interpretation. *Sci. Just.* 38, 151–156. doi: 10.1016/S1355-0306(98)72099-4

Darboux, J., Appell, P., and Poincaré, J. (1908). "Examen critique des divers systèmes ou études graphiques auxquels a donné lieu le bordereau," in *L'affaire DREFUS - la révision du procès de Rennes - enquête de la chambre criminelle de la Cour de Cassation* (Paris: Ligue française des droits de l'homme et du citoyen), 499–600.

ENFSI (2015). *Guideline for Evaluative Reporting in Forensic Science*. Available online at: http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

Evett, I., and Buckleton, J. (1996). "Statistical analysis of STR data," in *Advances in Forensic Haemogenetics 6*, eds A. Carracedo, B. Brinkmann, and W. Bär (Berlin: Springer Verlag), 79–86.

Evett, I., Cage, P., and Aitken, C. G. G. (1987). Evaluation of the likelihood ratio for fibre transfer evidence in criminal cases. *Appl. Stat.* 36, 174–180. doi: 10.2307/2347549

Evett, I., Jackson, G., and Lambert, J. A. (2000). More on the hierarchy of propositions: exploring the distinction between explanations and propositions. *Sci. Just.* 40, 3–10. doi: 10.1016/S1355-0306(00)71926-5

Fenton, N., Berger, D., Lagnado, D., Neil, M., and Hsu, A. (2014a). When 'neutral' evidence still has probative value (with implications from the Barry George case). *Sci. Just.* 54, 274–287. doi: 10.1016/j.scijus.2013.07.002

Fenton, N., Lagnado, D., Berger, D., Neil, M., and Hsu, A. (2014b). Response to 'On the use of the likelihood ratio for forensic evaluation: response to Fenton et al.' *Sci. Just.* 54, 319–320. doi: 10.1016/j.scijus.2014.05.005

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis, 2 Edn*. London: Chapman and Hall.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., and Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Comput. Speech Lang.* 20, 331–355. doi: 10.1016/j.csl.2005.08.005

Good, I. (1979). Studies in the history of probability and statistics. XXXVIII A. M. Turing's statistical work in World War II. *Biometrika* 66, 393–396. doi: 10.1093/biomet/66.2.393

Good, I. J. (1989a). C312: Yet another argument for the explication of weight of evidence. *J. Stat. Comput. Simul.* 31, 58–59. doi: 10.1080/00949658908811115

Good, I. J. (1989b). C319: Weight of evidence and a compelling metaprinciple. *J. Stat. Comput. Simul.* 31, 121–123. doi: 10.1080/00949658908811131

Good, I. J. (1991). "Weight of evidence and the Bayesian likelihood ratio," in *The Use of Statistics in Forensic Science*, eds C. G. G. Aitken and D. A. Stoney (Chichester: Ellis Horwood), 85–106.

Harvey, W., Butler, O., Furness, J., and Laird, R. (1968). The Biggar murder: dental, medical, police and legal aspects. *J. Forensic Sci. Soc.* 8, 157–219. doi: 10.1016/S0015-7368(68)70474-6

Hepler, A. B., Saunders, C. P., Davis, L. J., and Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Sci. Int.* 219, 129–140. doi: 10.1016/j.forsciint.2011.12.009

Kaye, D. (1979). The laws of probability and the law of the land. *Univ. Chic. Law Rev.* 47, 34–56. doi: 10.2307/1599414

Lindley, D. V. (1977). A problem in forensic science. *Biometrika* 64, 207–213. doi: 10.1093/biomet/64.2.207

Lund, S. P., and Iyer, H. (2017). Likelihood ratio as weight of forensic evidence: a closer look. *J. Res. Natl. Inst. Stand. Technol.* 122:27. doi: 10.6028/jres.122.027

Martyna, A., Lucy, D., Zadora, G., Trzcinska, B., Ramos, D., and Parczewski, A. (2013). The evidential value of microspectrophotometry measurements made for pen inks. *Anal. Methods* 5, 6788–6795. doi: 10.1039/c3ay41622d

Meuwly, D., Ramos, D., and Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Sci. Int.* 276, 142–153. doi: 10.1016/j.forsciint.2016.03.048

Morrison, G. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic phonetic data multivariate kernel density (mkvd) versus Gaussian mixture model-universal background model (gmm-ubm). *Speech Commun.* 53, 91–98. doi: 10.1016/j.specom.2010.09.005

Peirce, C. (1878). "The probability of induction," in *The World of Mathematics, 1956*, Vol. 2, ed J. Newman (New York, NY: Simon Schuster), 1341–1354.

Ramos, D., and Gonzalez-Rodriguez, J. (2013). Reliable support: measuring calibration of likelihood ratios. *Forensic Sci. Int.* 230, 156–169. doi: 10.1016/j.forsciint.2013.04.014

Ramos, D., Gonzalez-Rodriguez, J., Zadora, G., and Aitken, C. G. G. (2013). Information-theoretical assessment of the performance of likelihood ratio computation methods. *J. Forensic Sci.* 58, 1503–1518. doi: 10.1111/1556-4029.12233

Taroni, F., Bozza, S., Biedermann, A., and Aitken, C. G. G. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law Probab. Risk* 15, 1–16. doi: 10.1093/lpr/mgv008

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., and Aitken, C. G. G. (2010). *Data Analysis in Forensic Science: a Bayesian Decision Perspective*. Chichester: Wiley.

Tippett, C., Emerson, V., Fereday, M., Lawton, F., and Lampert, S. (1968). The evidential value of the comparison of paint flakes from sources other than vehicles. *J. Forensic Sci. Soc.* 8, 61–65. doi: 10.1016/S0015-7368(68)70442-4

Wilson, A., Aitken, C. G. G., Sleeman, R., and Carter, J. (2014). The evaluation of evidence relating to traces of cocaine on banknotes. *Forensic Sci. Int.* 236, 67–76. doi: 10.1016/j.forsciint.2013.11.011

Wilson, A., Aitken, C. G. G., Sleeman, R., and Carter, J. (2015). The evaluation of evidence for autocorrelated data in relation to traces of cocaine on banknotes. *Appl. Stat.* 64, 275–298. doi: 10.1111/rssc.12073

Zadora, G., Martyna, A., Ramos, D., and Aitken, C. G. G. (2014). *Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. Chichester: John Wiley and Sons Ltd.

Zadora, G., Neocleous, T., and Aitken, C. G. G. (2010). A two-level model for evidence evaluation in the presence of zeros. *J. Forensic Sci.* 55, 371–384. doi: 10.1111/j.1556-4029.2009.01316.x

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

GENERAL COMMENTARY
published: 22 June 2018
doi: 10.3389/fgene.2018.00224

Commentary: Likelihood Ratio as Weight of Forensic Evidence: A Closer Look

Colin Aitken[1]*, *Anders Nordgaard*[2], *Franco Taroni*[3] and *Alex Biedermann*[3]

[1] School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom, [2] National Forensic Centre (Sweden), Linköping, Sweden, [3] School of Criminal Justice, Université de Lausanne, Lausanne, Switzerland

Keywords: likelihood ratio, value of evidence, forensic science, logarithm, forensic reporting

A commentary on

Likelihood Ratio as Weight of Forensic Evidence: A Closer Look
by Lund, S. P., and Iyer, H. (2017). J. Res. Natl. Inst. Stand. Technol. 122:27. doi: 10.6028/jres.122.027

A recent article (Lund and Iyer, 2017) provides, in the words of its title, a closer look at the likelihood ratio as the weight of forensic evidence. This note comments critically on two aspects of the article.

The first aspect concerns two related statements. In the abstract the statement is made that "[W]e find the likelihood ratio paradigm to be unsupported by arguments of Bayesian decision theory, which applies only to personal decision making and not to the transport of information from an expert to a separate decision maker." The idea presented in this statement of lack of support for the likelihood ratio as a means of transport of information is repeated in the conclusion where it is stated that ". . . we hope the forensic science community comes to view the *LR* as one *possible*, not normative or necessarily optimum, tool for communicating to DMs (decision makers)" (Lund and Iyer's emphasis). Despite this opinion of these authors, it was shown many years ago by I.J.Good in two brief notes in the *Journal of Statistical Computation and Simulation* (Good, 1989a,b) repeated in Good (1991) and in Aitken and Taroni (2004) that, with some very reasonable assumptions, the assessment of uncertainty inherent in the evaluation of evidence leads inevitably to the likelihood ratio as the only way in which this can be done.

In order to show that the likelihood ratio is the only way to evaluate evidence, it is necessary to introduce some mathematical notation. This is a device to ease the presentation of the argument. The argument could be made verbally but would be lengthy and more difficult to follow. Consider evidence E which it is desired to evaluate in the context of two mutually exclusive propositions H_p and H_d. Denote the value of the evidence by V. Of course, this statement makes the implicit assumption that evidence has a value that can be measured. The value will depend on background information I. Four and only four factors have been introduced, E, H_p, H_d and I. Thus, V is a function of these four factors, $V = f(E, H_p, H_d, I)$. There is uncertainty about E, so it should be analyzed probabilistically. Use of the argument of conditional probability leads to $f(E \mid H_p, H_d, I)f(H_p, H_d, I)$, rather than forms such as $f(H_p \mid H_d, E, I)$ or variants of it. The expression $f(H_p, H_d, I)$ does not involve the evidence, which reduces considerations further to $f(E \mid H_p, H_d, I)$. Propositions H_p and H_d are mutually exclusive so if E is to be a function of both H_p and H_d then $f(E \mid H_p, H_d, I)$ is a combination of two functions, one that involves H_p and not H_d and one that involves H_d and not H_p. Value may thus be expressed as a function of the probabilities of E given H_p (and I) and of E given H_d (and I). Again, this makes implicit assumptions, namely that there is a probability that can be associated with evidence and that is dependent on a proposition and background information. For ease of notation explicit mention of I will be omitted from notation in what follows.

OPEN ACCESS

Edited by:
Mariza De Andrade, Mayo Clinic, United States

Reviewed by:
Daniel Ramos, Universidad Autonoma de Madrid, Spain

***Correspondence:**
Colin Aitken
cgga@ed.ac.uk

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 November 2017*
Accepted: *06 June 2018*
Published: *22 June 2018*

Citation:
Aitken C, Nordgaard A, Taroni F and Biedermann A (2018) Commentary: Likelihood Ratio as Weight of Forensic Evidence: A Closer Look. Front. Genet. 9:224. doi: 10.3389/fgene.2018.00224

Let $x = \Pr(E \mid H_p)$ and $y = \Pr(E \mid H_d)$. The assumption that V is a function only of these probabilities can be represented mathematically as

$$V = f(x, y)$$

for some function f.

Now, consider another piece of evidence T which is irrelevant to E, to H_p and to H_d. Irrelevance is taken in the probabilistic context to be equivalent to independence so that T may be taken to be independent of E, of H_p and of H_d. It is then permissible for $\Pr(T)$ to be given notation which does not refer to any of E, H_p or H_d. Thus, let $\Pr(T)$ be denoted by θ. Then

$$\begin{aligned}&\Pr(E, T \mid H_p)\\ &\quad = \Pr(E \mid H_p)\Pr(T \mid H_p) \quad \text{by the independence of } E \text{ and } T\\ &\quad = \Pr(E \mid H_p)\Pr(T) \quad \text{by the independence of } T \text{and } H_p\\ &\quad = x\theta.\end{aligned}$$

Similarly,

$$\Pr(E, T \mid H_d) = y\,\theta.$$

The value of (E, T) is $f(\theta x, \theta y)$ by the definition of f. However, evidence T is irrelevant and has no effect on the value of evidence E. Thus, the value of the combined evidence (E, T), $f(\theta x, \theta y)$, is equal to the value V of E, $f(x, y)$, and

$$V = f(x, y) = f(\theta x, \theta y)$$

for all θ in the interval [0,1] of possible values of $\Pr(T)$.

The only class of functions of (x, y) for which this can be said to be the case is the class which are functions of x/y or

$$\Pr(E \mid H_p)/\Pr(E \mid H_d)$$

which is the likelihood ratio. Hence the value V of evidence has to be a function of the likelihood ratio. Lund and Iyer wish the forensic community to view the likelihood ratio as one possible tool for communication with decision makers. We hope that we have shown here through the argument of Good that it is the only logically admissible form of evaluation. Incidentally, note that no recourse has been made to arguments of Bayesian decision theory. The support of these arguments for the likelihood ratio paradigm, as suggested in the abstract, is not necessary.

The second aspect is minor and concerns a definition. The concept of weight of evidence is an old idea. The term *weight of evidence* for the logarithm of the likelihood ratio was given by Charles Sanders Peirce (Peirce, 1878). It is not the likelihood ratio that should be referred to as the weight of evidence as is done in the title of the article. It is better to refer to the likelihood ratio as the *value* of the evidence and its logarithm as the weight of the evidence. The logarithm of the likelihood ratio has the pleasingly intuitive operation of additivity when converting the logarithm of the prior odds in favor of a proposition to the logarithm of the posterior odds in favor of the proposition.

$$\log\left\{\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)}\right\} = \log\left\{\frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}\right\} + \log\left\{\frac{\Pr(H_p)}{\Pr(H_d)}\right\}. \quad (1)$$

When considering the scales of justice it is the logarithm of the probabilities of the evidence given each of the two competing propositions that should be put in the scales, not the probabilities. Equation (1) can be rewritten as

$$\begin{aligned}&\log\{\Pr(H_p \mid E)\} - \log\{\Pr(H_d \mid E)\} =\\ &\quad \log\{\Pr(E \mid H_p)\} - \log\{\Pr(E \mid H_d)\} + \log\{\Pr(H_p)\} - \log\{\Pr(H_d)\}\\ &= [\log\{\Pr(E \mid H_p)\} + \log\{\Pr(H_p)\}] - [\log\{\Pr(E \mid H_d)\} + \log\{\Pr(H_d)\}]\end{aligned}$$

Expressions to the left of the negative sign in the last line are associated with one pan in the scales, expressions to the right with the other pan. Thus $\log(\Pr(E \mid H_p))$ is added to the prior log probability for H_p in one scale and $\log(\Pr(E \mid H_d))$ is added to the prior log probability for H_d in the other scale. The difference in the sums of the two pairs of log probabilities is a more intuitive characteristic of the evidence to which the term *weight* may be applied than the ratio of the probabilities of the evidence given the respective propositions.

AUTHOR CONTRIBUTIONS

CA drafted this commentary which results from equal and direct intellectual contributions of all listed authors.

FUNDING

The authors gratefully acknowledge the support of Leverhulme Trust through the Emeritus Award EM-2016-027 (CA), the Swiss National Science Foundation through grant No. BSSGI0_155809 and the University of Lausanne (AB).

REFERENCES

Aitken, C. G. G., and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd Edn*. Chichester: John Wiley and Sons Ltd.

Good, I. J. (1989a). C312: yet another argument for the explication of weight of evidence. *J. Stat. Comput. Simul.* 31, 58–59. doi: 10.1080/00949658908811115

Good, I. J. (1989b). C319: weight of evidence and a compelling metaprinciple. *J. Stat. Comput. Simul.* 31, 121–123. doi: 10.1080/00949658908811131

Good, I. J. (1991). "Weight of evidence and the Bayesian likelihood ratio" in *The Use of Statistics in Forensic Science*, eds C. G. G. Aitken and D. A. Stoney (Chichester: Ellis Horwood), 85–106.

Lund, S. P., and Iyer, H. (2017). Likelihood ratio as weight of forensic evidence: a closer look. *J. Res. Natl. Inst. Stand. Technol.* 122:27. doi: 10.6028/jres.122.027

Peirce, C. S. (1878). "The probability of induction," *in The World of Mathematics, 1956,* Vol. 2, ed J. R. Newman (New York, NY: Simon Schuster), 1341–1354.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Sociology

OPINION
published: 20 August 2018
doi: 10.3389/fsoc.2018.00025

Statistical Adhockeries Are No Criteria for Legal Decisions—The Case of the Expert Medical Report on the Assessment of Urine Specimens Collected Among Athletes Having Participated to the Vancouver and Sochi Winter Olympic Games

Franco Taroni[1]*, *Alex Biedermann*[1], *Joëlle Vuille*[1] and *Silvia Bozza*[2,1]

[1] School of Criminal Justice, University of Lausanne, Lausanne, Switzerland, [2] Department of Economics, University Ca' Foscari of Venice, Venice, Italy

Keywords: legal decision making, doping in sports, statistics, forensic interpretation, Olympic Games

OPEN ACCESS

Edited by:
Athanasios Alexiou, Novel Global Community Educational Foundation (NGCEF), Hebersham, Australia

Reviewed by:
Donald Arthur Berry, University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Franco Taroni
Franco.Taroni@unil.ch

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Sociology

Received: *25 May 2018*
Accepted: *30 July 2018*
Published: *20 August 2018*

Citation:
Taroni F, Biedermann A, Vuille J and Bozza S (2018) Statistical Adhockeries Are No Criteria for Legal Decisions—The Case of the Expert Medical Report on the Assessment of Urine Specimens Collected Among Athletes Having Participated to the Vancouver and Sochi Winter Olympic Games. Front. Sociol. 3:25. doi: 10.3389/fsoc.2018.00025

Scientific literature and practice, notably expert reports, commonly involve misinterpretations of standard statistics, such as the *p*-value, or the calculation of so-called "3-standard deviation intervals," elements upon which decisions in medicine, physics, or legal matters are based[1]. Such instances of expert reporting reflect a misreading of the way in which scientists should assist the judiciary in assessing results coming from analytical laboratories. A recent example of such a practice are the conclusions of the international report on urine specimens collected among athletes participating in the Vancouver and Sochi Winter Olympic Games (report dated 5th October, 2017)[2,3].

Uncertainty is a complication that accompanies participants of the justice system who face inference and decision-making as core aspects of their activities. Inference relates to the use of incomplete information, as given by scientific findings, in order to reason about propositions of interest, such as whether the quantity of a given substance in some bodily fluid is larger than a legal threshold. In turn, decision-makers, notably judges, are required to make practical decisions, such as declaring whether or not an athlete has used a performance-enhancing substance. Inference and decisions of this kind abound in the legal field. Toxicology laboratories, across jurisdictional systems, are regularly asked to quantify the amount of target substances (e.g., alcohol, illegal drugs, biological markers) detected in, for example, blood samples taken from persons of interest (e.g., Karkazis and Jordan-Young, 2015).

Inference and decision require logical assistance because unaided human reasoning is liable to bias and misinterpretation. These represent causes of concern because fallacious reasoning and erroneous conclusions in legal proceedings risk endangering the fairness of the proceedings and can lead to miscarriages of justice. Statistical approaches are often used to support expert conclusions

[1] Hence the use of "adhockeries" in the title. The term adhockery was introduced by I. J. Good, and used also by de Finetti (e.g., de Finetti, 1993a,b), to denote the use of improvised measures rather than a robust and logic methodology.

[2] Report available at https://stillmed.olympic.org/media/Document%20Library/OlympicOrg/IOC/Who-We-Are/Commissions/Disciplinary-Commission/IOC-DC-Schmid/Appendix-VIII-CHUV-Report-Prof-Burnier-06-10-2017.pdf. Hereafter, the Report.

[3] For a previous example of the use of statistics in a case of alleged doping see, e.g., the Andrus Veerpalu case (Fischer and Berry, 2014)

but inferential misunderstandings regrettably plague disciplines such as forensic science and medicine when scientists report on statistical analyses conducted as part of their casework or research.

The case on which we intend to comment can be briefly summarized as follows. The International Olympic Committee requested statistical analyses on results of urine examinations performed on samples coming from the XXII Olympic Winter Games in Sochi, with the aim of identifying athletes who had used prohibited substances. Specifically, the question was "[t]o determine [...] if the values are within the reference values obtained from the control population at the XXII Olympic Winter Games and in agreement with data published." (Report at p. 3). A potentially doped athlete, called in the Report "true outlier," was defined as a person having a given bio-chemical parameter value—for example, urinary sodium concentration—greater than a reference mean plus three standard deviations. The reference mean and standard deviation were calculated using data from reference athletes, considered *not* doped, of the Vancouver Olympic Winter Games.

As an illustration, consider a measured target substance (e.g., urinary sodium concentration), where the experimental unit is a urine sample from a person under investigation. The mean values reported for the sodium concentration in urines in the reference male (female) population of athletes are 95.4 (67.39) mmol/l. The reported standard deviation values are 49.37 and 40.88 mmol/l, respectively (see Report at p. 6). According to the "three standard deviation rule," athletes with values greater than 243.51 (or 190, for women) are considered to be outliers. Thus, the measurement would be said to meet the requirement for establishing the presence of an unrealistic level of a target substance in urine *if* the measurement for the investigated sample were larger than the upper value of the bound in a reference population. It was noted in the report that "[w]ith this approach, we identify 13 samples (of 5 men and 8 women) which are definitively out of the range." (Report at p. 2).

As a preliminary, it is worth noting that such an approach for the treatment and reporting of experimental results does not address the inferential and decisional issues at stake. Instead, it is merely descriptive. This does not mean, though, that it is intrinsically wrong: scientists widely rely on effective descriptive methods of exploratory data analysis to illustrate, for example, how population data are distributed, where given sample data are located and how they spread. However, such a description does not fully address the questions of interest for the decision-maker, which are: *How can we use data (or a summary of them) on the Vancouver Olympic athletes to infer something about the value of the urinary sodium concentration in the reference population?* and: *How can we conclude that a new measurement from a given athlete is in fact an outlier (or an anomalous value) with reference to this population?* These are intrinsically inferential questions, not descriptive ones, and remain unresolved with the approach taken in the report, as we explain below.

It is commonly understood, and unquestioned, that measurements on urinary samples taken from individuals of a given population will show some variation. Stated otherwise, the results will, in some sense, distribute. Basic statistics such as the mean and the standard deviation of a quantity of interest (e.g., the sodium concentration) used to describe the reference samples from the Vancouver Olympic athletes represent indeed succinct and informative summaries. The mean provides a measure of location and the standard deviation provides a measure of dispersion (spread) of the available measurements. In this context, the "three standard deviation rule" may have some appeal. Provided that data distribute symmetrically around the mean, at least approximately, then values within one standard deviation of the mean account for about 68% of the observations, while two standard deviations account for about 95% and three standard deviations account for about 99.7% of the values. The 68-95-99.7 rule is a shorthand used to remember the approximate percentage of values that lie within a band around the mean with a width of two, four and six standard deviations, respectively. It is a rule to describe the available data (i.e., measurements from Vancouver athletes), but not to infer something about a new value coming from a new athlete, as emphasized also in Berry (2008).

However, does this rule allow one to conclude that values outside this range are necessarily "outliers"—lying at an abnormal distance from other values? Obviously, *any* set of observations contains extremes: the minimum and the maximum value are extremes. Notwithstanding, it is understandable to express concerns in situations, such as the case discussed here, where the highlighted extremes are not only the largest (or, in other cases, the smallest) observation, but are actually "*extremely* extreme": they are apparently inconsistent with the reference observations and therefore candidates for being considered "outliers." It is no accident to term these values "candidates." Several reasons can, in isolation or combined, account for extreme observations: first, natural variation, beyond the currently known bands, but also laboratory measurement or recording errors, or even intentional tampering, such as the addition of a target substance (here salt). Since these are potential accounts for the observations, it is—by definition—a matter of a *personal* judgment on the part of the scientist to decide *when* a given observation appears to be inconsistent with the remainder set of data. One way to avoid a rigid and intrinsically arbitrary threshold is to consider at least one explicit alternative account for the findings: the scientist could then provide a statistical measure called a "likelihood ratio" that represents an expression of how the measurements, whatever their value, extreme or otherwise, are capable of discriminating amongst competing propositions of interest. When no discernible alternative hypothesis can be specified (as in the case of interest) several ways of categorizing suspicious observations are available (see, e.g., Barnett and Lewis, 1994).

It is common to distinguish between frequentist (or classical) and Bayesian approaches. Statistical data analyses in the forensic and medical contexts commonly rely on a so-called "frequentist" perspective, associated with the idea that statistical conclusions could be entirely objective, with known error rates. Consider, for instance, the problem of hypothesis testing, where attempts at drawing conclusions about competing propositions often rely on a comparison between the significance level of the test and the observed significance level, i.e., p-value. A large majority of

papers published nowadays still propose statistical treatments based on this quantity. Controversial discussion was initiated by an editorial of Basic and Applied Social Psychology (Trafimow and Marks, 2015), expressing the intention to ban from publication in their journal any paper containing procedures advocating *p*-values. This announcement has echoed widely, from general weekly science journals (e.g., Nuzzo, 2014; Leek and Peng, 2015) to specialist groups such as the International Society for Bayesian Analysis (Schmidt et al., 2015). The main concern expressed in these reactions is not the correctness or usefulness of frequentist statistical procedures, but rather the misinterpretations surrounding the use of such procedures and their consequences. There is a need to emphasize what exactly the various approaches allow scientists to draw as a conclusion, and what they do not allow them to say.

One of the major misunderstandings found in the reporting on significance testing through a *p*-value consists in interpreting this value as the probability that the null hypothesis (e.g., as previously stated a difference between populations mean values) is true. This fallacious conclusion is also known as the fallacy of the transposed conditional. The temptation to believe that, if an observation is rare under a given hypothesis it can be regarded as evidence against that hypothesis, must be resisted[4]. Bayesian approaches avoid these intricacies by relying on the fundamental tenet of capturing, using probability, all uncertainties characterizing a problem. According to these approaches, discordancy can be assessed by means of a predictive probability to observe a value greater than the particular (suspicious) observation given the rest of the reference sample, which allows one to restrict attention to manifestly extreme (unlikely) observations (Geisser, 1998).

Despite struggles over philosophical stances regarding statistical inference and decision-making, the restriction of attention to the sole question of outliers still falls short of the fundamental problem that the case in question poses. Among the ultimately disputed questions is the issue of whether there is sufficient evidence to conclude that a given urine value is an outlier. The answer to this question cannot rely on scientific findings only, because it requires the assessment of all available information, scientific, and other, in a given case. What is more, it cannot be reduced to a descriptive (statistical) account of scientific findings, but extends to inference and decision-making, and associated decision criteria. The latter are not given by *ad-hoc* statistical thresholds, but are intimately related to the decision-maker's preferences and policy values, which are even further beyond the scientist's area of competence.

[4]"Researchers often rely on the seeming objectivity of the $p<0.05$ criterion without realizing that theory behind the p-value is invalidated when analysis is contingent on data." (Gelman and Hennig, 2017).

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

AB acknowledges the support provided by the Swiss National Science Foundation through Grant No. BSSGI0_155809 and the University of Michigan Law School, Ann Arbor, MI (Michigan Grotius Research Fellowship).

REFERENCES

Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data, ed*. 3. Chichester: Wiley.

Berry, D. (2008). The science of doping. *Nature* 454, 692–693. doi: 10.1038/454692a

de Finetti, B. (1993a). "The role of probability in the different attitudes of scientific thinking," in *Probabilità e induzione – Induction and probability*, eds P. Monari and D. Cocchi (Bologna: Editrice Clueb), 491–511.

de Finetti, B. (1993b). "Bayesian statistical inference" in *Probabilità e Induzione – Induction and Probability*, eds P. Monari and D. Cocchi (Bologna: Editrice Clueb), 513–524.

Fischer, K., and Berry, D. A. (2014). Statisticians introduce science to international doping agency: the Andrus Veerpalu case. *Chance* 27, 10–16. doi: 10.1080/09332480.2014.965625

Geisser, S. (1998). "Some uses of order statistics in Bayesian analysis," in *Handbook of Statistics 17 - Order Statistics*, eds N. Balakrishnan and C.R. Rao (Amsterdam: Elsevier), 379–399.

Gelman, A., and Hennig, C. (2017). Beyond subjective and objective in statistics. *J. R. Stat. Soc. A* 180, 967–1033. doi: 10.1111/rssa.12276

Karkazis, K., and Jordan-Young, R. (2015). Debating a testosterone 'sex gap'. *Science* 348, 858–860. doi: 10.1126/science.aab1057

Leek, J. T., and Peng, R. D. (2015). Statistics: *P*-values are just the tip of the iceberg. *Nature* 520, 612. doi: 10.1038/520612a

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature* 506, 150–152. doi: 10.1038/506150a

Schmidt, A., Berger, J., David, P., Kadane, J., O'Hagan, T., and Pericchi, L. (2015). Banning null hypothesis significance testing. *ISBA Bull*. 22, 5–9. Available Online at: https://bayesian.org/wp-content/uploads/2016/09/1503.pdf

Trafimow, D., and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psychol*. 37, 1–2. doi: 10.1080/01973533.2015.1012991

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Sociology

GENERAL COMMENTARY
published: 20 December 2019
doi: 10.3389/fsoc.2019.00085

Commentary: Statistical Adhockeries Are No Criteria for Legal Decisions—The Case of the Expert Medical Report on the Assessment of Urine Specimens Collected Among Athletes Having Participated to the Vancouver and Sochi Winter Olympic Games

*Michel Burnier**

Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland

Keywords: expertise, statistical approach, plausibility, Bayesian (subjective) probability, coherence

Edited by:
Sue Pope,
Principal Forensic Services,
United Kingdom

Reviewed by:
Perikles Simon,
Johannes Gutenberg University
Mainz, Germany

***Correspondence:**
Michel Burnier
Michel.Burnier@chuv.ch

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Sociology

Received: *22 October 2018*
Accepted: *05 December 2019*
Published: *20 December 2019*

Citation:
Burnier M (2019) Commentary: Statistical Adhockeries Are No Criteria for Legal Decisions—The Case of the Expert Medical Report on the Assessment of Urine Specimens Collected Among Athletes Having Participated to the Vancouver and Sochi Winter Olympic Games. Front. Sociol. 4:85. doi: 10.3389/fsoc.2019.00085

A Commentary on

Statistical Adhockeries Are No Criteria for Legal Decisions—The Case of the Expert Medical Report on the Assessment of Urine Specimens Collected Among Athletes Having Participated to the Vancouver and Sochi Winter Olympic Games
by Taroni, F., Biedermann, A., Vuille, J., and Bozza, S. (2018). Front. Sociol. 3:25. doi: 10.3389/fsoc.2018.00025

As cited by Gelman and Hennig, "*Decisions in statistical data analysis are often justified, criticized or avoided by using concepts of objectivity and subjectivity*" (Gelman and Hennig, 2017).

The paper by Taroni et al. (2018), that I have read with interest and some surprise, does not escape this principle. Indeed, it heavily and namely criticizes the statistical approach that was used in a recent expertise that I performed on behalf of the Medical and Scientific Department of the International Olympic Committee (Taroni et al., 2018). The authors indirectly suggest that the analysis could provide "*erroneous conclusions in legal proceedings risk endangering the fairness of the proceedings and can lead to miscarriages of justice*" (Taroni et al., 2018). I understand the worry of the authors who belong to a School of Criminal Justice, to provide as much as possible reliable and unbiased expert conclusions to assist judiciary in their decision-making processes, and I do have the same preoccupation. I do not want to debate on whether a purely statistical approach is more or less appropriate than another statistical method using a Bayesian approach and the calculation of a probability to make an odd observation. Indeed, it is well possible that a Bayesian approach could be superior and more useful for the judges although this remains to be demonstrated in the particular case. The main reason why I would like to react on the content of Taroni's publication is because it seems obvious that the authors have not read the expertise completely and have not clearly understood its purpose and its analysis. Thus, they have not taken into account what Gelan and Hennig call "*the context dependence*" (Gelman and Hennig, 2017).

The first important issue is the question asked to the expert. In this case, the first demand was "to determine reference values for various urinary analytes (sodium, potassium, chloride, calcium, creatinine, and urinary density) coming from samples taken from top athletes tested at the time of Vancouver XXI Winter Olympic Games." This goal could be achieved only with a statistical approach taking into account the distribution of the values of athletes having participated in the Vancouver Games. The second objective was to examine the distribution and statistics of each sample collected from the XXII Olympic Winter Games, which occurred in Sochi and to evaluate them in the light of the reference values obtained in Vancouver. As Taroni et al. correctly pointed out, the populations were different, the former containing athletes of all countries, including Russia, and the latter only samples from Russia and this might have explained some differences due to country-specific diets. This is reason why two analyses were done, one within each population, and one between populations. With this approach, some values were clearly outside the distribution of both the Vancouver and the Sochi populations of athletes and could be considered as "outliers" or extremes of extremes as Taroni et al. name them.

Of note, our objective was not to determine who was doped or not, identifying the presence of a prohibited substance. The baseline hypothesis was that some samples had been manipulated and urine perhaps reconstituted with an excess of salt to match the initial urinary density that was the only parameter available. Therefore, the expert focused on samples with very high sodium and chloride concentrations, which could fit with the hypothesis. Samples eventually manipulated but with a normal sodium chloride concentration would of course escape from this strategy.

Now, if no Bayesian analysis was performed in this expertise to assess the probabilities of extremes to be real outliers, other aspects of plausibility were considered in my analysis using an approach fitting with the abductive approach discussed recently by Simon and Dettweiler (2019). One of them is the coherence between several measured analytes. Indeed, humans are not eating sodium chloride but a diet containing salt but also potassium and calcium. Consequently, humans on a very high salt diet also ingest more calcium and potassium. Interestingly, in the outliers of the Sochi group of athletes, there was a clear gap between urinary sodium and chloride excretions and the excretion of potassium, this latter being in the normal range and comparable to the athletes tested in Vancouver. To a certain degree, the same was true for calcium. Thus, there appears to be incoherence between the urinary content of analytes in subjects recognized as outliers based on urinary sodium concentrations. In addition, one must also take into account the physiological plausibility when examining samples. In some of the athletes, the measured urinary sodium concentrations were so high that they were incompatible with human physiology and were therefore more than suspect. At last, the level of plausibility became extremely high when one noted that several outliers were not isolated athletes but fellow-members of the same competition team.

Thus, Taroni et al. had the impression that our conclusions were based only on a statistical analysis but this is clearly wrong. Interestingly, in their publication Taroni et al. do not propose any alternative for this kind of analysis and even suggest in their publication that results might be the same using a Bayesian approach. Today, the data are available and the authors are welcome to confront their approach with the one used in my expertise. But, as long as no comparison has been performed, the results of my expertise must be considered as correct and reliable and judges can use these data to integrate them in the overall set of evidence, to make their opinion and finally take their decisions.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Gelman, A., and Hennig, C. (2017). Beyond subjective and objective in statistics. *J. R. Stat. Soc. A* 180 (Part 4), 967–1033. doi: 10.1111/rssa.12276

Simon, P., and Dettweiler, U. (2019). Current anti-doping crisis: the limits of medical evidence employing inductive statistical inference. *Sports Med.* 49, 497–500. doi: 10.1007/s40279-019-01074-0

Taroni, F., Biedermann, A., Vuille, J., and Bozza, S. (2018). Statistical adhockeries are no criteria for legal decisions—The case of the expert medical report on the assessment of urine specimens collected among athletes having participated to the vancouver and sochi winter olympic games. *Front. Sociol.* 3:25. doi: 10.3389/fsoc.2018.00025

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership