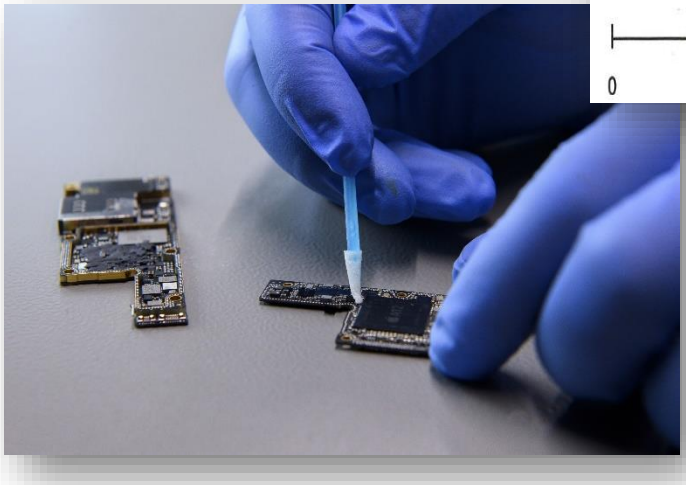
Polisen

LUND UNIVERSITY | FACULTY OF LAW

11th International Conference on Forensic Inference and Statistics (ICFIS)

Lund, Sweden, June 12-15, 2023

Hosts: National Forensic Centre (NFC) and Lund University, Faculty of Law



Welcome!

Dear Attendees,

We welcome you to Lund and the 11th International Conference on Forensic Inference and Statistics, ICFIS2023.

The conference is hosted by the Swedish National Forensic Centre - NFC in collaboration with the Faculty of Law at Lund University. The conference venues are located in the Faculty of Law main building and lecture halls. Lund University was founded in 1666 and the Faculty of Law is one of the four original faculties, situated in the medieval heart of Lund.

ICFIS2023 brings scientists, legal scholars and practitioners together to discuss and improve the scientific foundations of evidence evaluation and forensic statistics. The ICFIS conference series is a platform for networking and building future collaborations in this field, comprising all kinds of forensic evidence and related legal rules and issues.

The conference program shows a great variety of topics spanning from the philosophy of legal evidence to the use of machine learning methods for automatic identification. The ongoing global development on statistical modelling of fingerprint evidence is specially addressed as is the growing interest in score-based approaches. A round-table discussion of healthcare professionals being accused of committing murders puts lights on a serious problem, where statisticians play a most important role to quash incorrect reasoning about statistics. Probabilistic genotyping is another current field that is the topic of a workshop during the conference. And there is much more.

We hope you will enjoy your days in Lund and wish you a fruitful conference.

Anders Nordgaard
Forensic specialist, NFC

Christian Dahlman
Professor of Law, Lund University

General information

Conference venue ★
"Tryckeriet" ("The Print Shop") at the Faculty of Law, Lund University (www.law.lu.se), see Map. The building once housed a printers' business and is situated across the street from the main faculty building "Juridicum", at walking distance both from the railway station and Grand Hotel (about 10 min). There will be signs directing you to the entrance.

Visiting address:
Lilla Gråbrödersgatan 3 D, Lund
Phone:
+46 46 222 10 00 (exchange)

Map

Registration
Outside Pufendorf Hall ("Pufendorfsalen") on the first floor of "Tryckeriet":

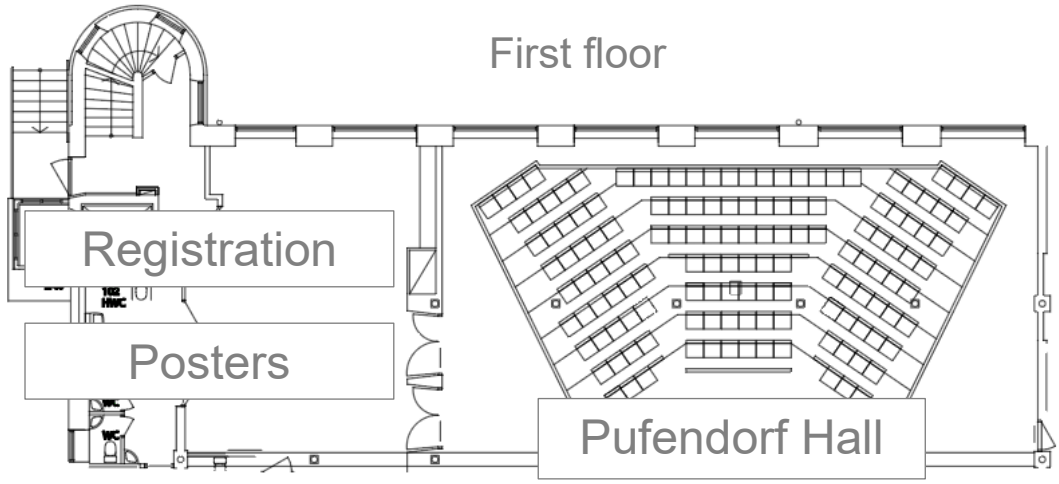
Monday, June 12	08:00 – 17:30
Tuesday, June 13	08:00 – 17:30
Wednesday, June 14	08:00 – 13:00

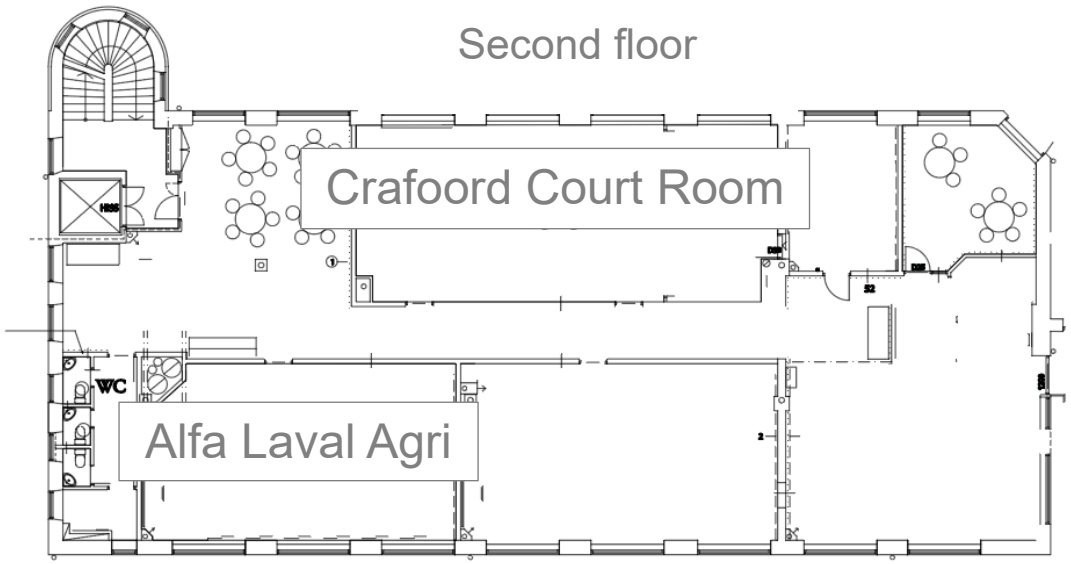
The registration desk will be happy to answer any questions. You are also welcome to contact a conference host or a member of the local organising committee, who will wear name tags with a red bar.

Name tag
Please wear your name tag on at all times when accessing conference services or sessions.

Lecture halls

- Pufendorf Hall ("Pufendorfsalen"); "Tryckeriet", first floor
- Alfa Laval Agri; "Tryckeriet", second floor
- The Crafoord Court Room ("Crafoordska Rättegångssalen"); "Tryckeriet", second floor





Poster session
The posters will be displayed outside Pufendorf Hall (“Pufendorfsalen”).

Restrooms
Restrooms are located on the first, second and third floor of “Tryckeriet”

WiFi
Choose **LU-guest**. Access is free but you need to register your e-mail address.

Lunch
Lunch will be served in "Gallerian", on the fourth floor of the main faculty building (i.e. opposite "Tryckeriet"). Entrance is organised since the doors are locked.

Coffee and tea
Coffee and tea will be served on the second floor of "Tryckeriet" i.e. at the same floor as Alfa Lava Agri and The Crafoord Court Room.

Welcome reception
The Welcome reception is held in "Gallerian", see above.

City tour and Conference dinner on June 14
City tour
Assembly outside "Tryckeriet" at 18:30.
Conference dinner
Venue: "AF-borgen", Sandgatan 2, Lund, see Map
Drinks will be served from 19:30.
Dinner starts at 20:00.

Health
For emergencies, call 112
Non-emergencies, call 1177

Taxi
Call +46 46 12 12 12 or +46 46 970 00

Payment
In principle, all shops, restaurants and taxis in Sweden accept MasterCard or Visa. Signs on the doors of shops and restaurants show whether other modes of payment are accepted.
Smaller shops and restaurants may not accept cash.

ATM
The ATM:s have blue signs ("Bankomat"). The closest ATM to the conference venue is at Stora Södergatan 2.

Language
English works perfectly everywhere in Sweden.

Dining in the city centre

There are plenty of good places to eat and drink in Lund. Italian (e.g. *GASTROnome, Gattostretto*) and Asian restaurants (e.g. *Ihsiri, Thuy, DubbelDubbel*) are popular. The restaurants in and around the food market ("Saluhallen") at Mårtenstorget (e.g. *Malmstens Fisk & Kök*) are also worth a visit, but close early (8 p.m.) on weekdays.
In the pubs (e.g. *John Bull*) you will find many types of beer, burgers, fish & chips etc. There are also several centrally located Sushi places (e.g. *Aiko Sushi*). In addition, most coffee shops serve salads, savory pies, sandwiches etc.
Klostergatans vin och delikatess is a good choice if you are looking for a wine bar/bistro.
Mat & Destillat and the restaurant(s) at *Grand Hotel*, serve dishes with fresh shrimp, fish, roe, local meat and local vegetables. We recommended to book a table in advance.
P.S. If you want to try pickled herring or gravlax, look for it on the Grand Hotel breakfast buffet.

The App **Tripadvisor** will show you the ratings, price range and location of most restaurants and pubs.

About Lund

The city of Lund was founded around AD 990. The southernmost part of Sweden (Scania or Skåne in Swedish) was then part of Denmark. From 1103 Lund was the seat of a Catholic archdiocese and Lund cathedral has been a central feature in the city since 1145. Denmark ceded Scania (including Lund) to Sweden in 1658 in The Roskilde Treaty, but its´ status as part of Sweden was not formalized until 1720.

Lund University, established in 1666, is one of Scandinavia's oldest and largest universities. The Faculty of Law was founded in 1668 and its first Dean was Samuel Pufendorf, originally from Heidelberg.

Today, Lund has around 120 000 inhabitants of which 40 000 are students. The Öresund Region, which comprises Copenhagen, Malmö, Lund and several other Swedish and Danish cities within commuting distance, is home to 4.1 million people.

Lund Cathedral. Photo by Kristina Strand Larsson

Stephen E. Fienberg
CSAFE Young Investigator Award

Who was Stephen Fienberg?

Stephen E. Fienberg was born in Toronto, Canada, on November 27, 1942, and died in his home in Pittsburgh, Pennsylvania, on December 14, 2017. Steve received a PhD in Statistics from Harvard University, and held faculty positions in the University of Chicago, the University of Minnesota, and finally in Carnegie Mellon University, where he spent most of his career.

Steve was a pioneer in the development of Bayesian theory and methods, and in the principled application of statistics in a wide range of disciplines, including the law and public policy making. He was a tireless proponent of the use of Bayesian methods worldwide, and was among the first to implement them in forensic science and criminal justice problems. In particular, Steve was a founding co-Director of CSAFE, and its intellectual leader until his untimely death. Much of the research that is conducted by CSAFE researchers today was suggested by Steve's far-reaching ideas and grand vision for the Center, so his influence persists. In addition to being a ground-breaking statistician, Steve was a dedicated mentor for graduate students and young researchers in the United States and elsewhere, and helped many as they were launching their own careers. Thus, it is really fitting to have an award named after him that recognizes research excellence in forensic statistics among new investigators.

Who qualifies to receive the award?

Young investigators will be considered for the Stephen E. Fienberg CSAFE Young Investigator Travel Award. This award is supported by the Center for Statistics and Applications in Forensic Evidence (CSAFE) and is intended to (partially) cover travel expenses for young investigators who attend the 2023 ICFIS conference and present a poster or presentation. CSAFE will support up to $1500 in travel expenses for the award winner. A committee will review relevant posters and presentations at the conference and the winner will be awarded a plaque at the end of the conference.

Young investigators are students at any level or scientists within 5 years of their terminal degree.

Program

Monday June 12

08:00 – 17:30	Registration (outside Pufendorf Hall)
09:30 – 10:30	*Overture session* (Pufendorf Hall) Welcome addresses by Dean Eva Ryrstedt, Faculty of Law, Lund University and Emil Hjalmarsson, National Forensic Centre - NFC Talk by a Guest of Honour
10:30 – 11:30	*Plenary lecture 1* (Pufendorf Hall) Charles Berger, Netherlands Forensic Institute and University of Leiden: **What's our standard going to be?** Chair: Christian Dahlman Charles Berger is principal scientist at the Netherlands Forensic Institute (NFI), and professor of Criminalistics at Leiden University. He specializes in subjects such as evidence interpretation and inference. At the NFI he is active in a number of areas such as education, R&D strategy, and his own research. He also supports the NFI experts, advises the direction, and guards the scientific quality. Before he developed into a generalist, he was trained as a forensic document examiner at the NFI. Charles has a background in the exact sciences: he has a PhD in applied physics from the University of Twente. After completing his PhD thesis he gathered 6 years of experience abroad with projects at the universities of California and Bordeaux and the École Normale Supérieure in Paris. At the moment he is also involved in promoting logically correct reasoning, introducing more objective methods, and reducing the risks of cognitive bias. For such improvements it is essential to explain them as often and as well as possible to all the stakeholders in the justice system. It's an exciting challenge at the interfaces of the worlds of science, police, and law.
11:30 – 12:30	Lunch (Gallerian)
12:30 – 13:30	*Plenary lecture 2* (Pufendorf Hall) Jeannette Leegwater, Netherlands Forensic Institute: **LR Systems for Fingermark Comparison in The Netherlands: a Unique Playing Field** Chair: Jonas Malmborg Jeannette Leegwater studied Mathematics and Forensic Science at the University of Amsterdam. During her work in the team Latent Fingerprint Research at the Netherlands Forensic Institute (NFI), she and her colleagues developed an LR system for fingerprints, based on an AFIS algorithm. Currently, Jeannette is a member of team Forensic Big Data Analysis (FBDA) at the NFI. Here, she uses data science and artificial intelligence for classic forensic domains and investigates whether new technologies and models can be applied in forensic science.

13:45 – 15:15	Parallell sessions: *Invited session 1* (Pufendorf Hall) Statistical Models for Fingerprint Analysis: Thinking Broadly about the Future (for detailed program, see Sessions) *Contributed session 1* (Alfa Laval Agri) Legal and Evidence Theory 1 (for detailed program, see Sessions)
15:15 – 15:45	Coffee/Tea
15:45 – 17:15	Parallell sessions: *Special call, Session 1* (Pufendorf Hall) Applications and Foundational Methods in the Presentation and Interpretation of Evidence with Score-Based Approaches (for detailed program, see Sessions) *Contributed session 2* (Alfa Laval Agri) Legal and Evidence theory 2 (for detailed program, see Sessions)
18:00 – 20:30	**Welcome reception** (Gallerian)

Tuesday June 13

08:00 – 17:30	Registration (outside Pufendorf Hall)
09:00 – 16:45	*Workshop 1* (Pufendorf Hall) **North Sea Group workshop on Theory of Legal Evidence** (for detailed program, see Workshops)
08:30 – 10:00	Parallell sessions: *Special call, Session 2* (Alfa Laval Agri) Applications and Foundational Methods in the Presentation and Interpretation of Evidence with Score-Based Approaches (for detailed program, see Sessions) *Contributed session 3* (Crafoord Court Room) DNA Evidence 1 (for detailed program, see Sessions)
10:00 – 10:30	Coffee/Tea
10:30 – 12:00	Parallell sessions: *Special call, Session 3* (Alfa Laval Agri) Applications and Foundational Methods in the Presentation and Interpretation of Evidence with Score-Based Approaches (for detailed program, see Sessions) *Contributed session 4* (Crafoord Court Room) Statistical modelling (for detailed program, see Sessions)
12:00 – 13:30	Lunch (Gallerian)
13:30 – 15:15	Parallell sessions: *Contributed session 5* (Alfa Laval Agri) Impression evidence (for detailed program, see Sessions) *Invited session 2* (Crafoord Court Room) The SAILR software (for detailed program, see Sessions)
15:15 – 15:45	Coffee/Tea

15:45 – 17:15	Parallell sessions: *Contributed session 6* (Alfa Laval Agri) Forensic value of evidence 1 (for detailed program, see Sessions) *Contributed session 7* (Crafoord Court Room) Forensic value of evidence 2 (for detailed program, see Sessions)
18:00 – 19:30	**Poster session** (outside Pufendorf Hall)

Wednesday June 14

08:00 – 13.00	Registration (outside Pufendorf Hall)
08:30 – 10:00	*Invited session 3* (Pufendorf Hall) Round-table discussion on "Healthcare serial killer or coincidence?" (for detailed program, see Sessions)
10:00 – 10:30	Coffee/Tea
10:30 – 11:30	*Plenary lecture 3* (Pufendorf hall) Paul Roberts, University of Nottingham **A DNA Profile is Never Sufficient Proof of Guilt in a Criminal Trial: What's the Problem?** Chair*:* Marjan Sjerps Paul Roberts is Professor of Criminal Jurisprudence, University of Nottingham School of Law, UK, and Adjunct Professor of Law, China University of Political Science and Law (CUPL), Beijing. His research spans criminal procedure and evidence, forensic science, criminal justice and legal theory, including comparative, international and philosophical perspectives. Roberts has been an advisor to the English and Scottish Law Commissions, the Crown Prosecution Service, and the UK Forensic Science Regulator. Recent publications include Roberts and Zuckerman's Criminal Evidence (OUP, 3/e 2022); Roberts and Stockdale (eds), Forensic Science Evidence and Expert Witness Testimony (EE, 2018); and Hunter, Roberts, Young and Dixon (ed), The Integrity of Criminal Process (Hart, 2016).
11:45 – 12:45	Parallell sessions: *Contributed session 8* (Pufendorf Hall) DNA evidence 2 (for detailed program, see Sessions) *Contributed session 9* (Alfa Laval Agri) Activity level (for detailed program, see Sessions)
12:45 – 13:45	Lunch (Gallerian)
13:45 – 17:15	*Workshop 2* (Alfa Laval Agri) **Opportunities with Models for Probabilistic Genotyping** (for detailed program, see Workshops)
13:45 – 15:15	Parallell sessions: *Contributed session 10* (Pufendorf hall) Forensic value of evidence 3 (for detailed program, see Sessions) *Contributed session 11* (Crafoord Court Room) Forensic chemistry (for detailed program, see Sessions)

15:15 – 15:45	Coffee/Tea
15:45 – 17:45	<u>Parallell sessions:</u> *Contributed session 12* (Pufendorf Hall) Forensic value of evidence 4 (for detailed program, see Sessions) *Contributed session 13* (Crafoord Court Room) Evaluative reporting (for detailed program, see Sessions)
18:30 – 22:30	**City tour** and **Conference dinner** (AF Borgen)

AF Borgen, Photo by Lund University

Thursday June 15

08:30 – 10:30

Parallell sessions:

Contributed session 14 (Pufendorf Hall)
Health and misuse of statistics (for detailed program, see Sessions)

Contributed session 15 (Alfa Laval Agri)
DNA evidence 3 (for detailed program, see Sessions)

10:30 – 11:00

Coffee/Tea

11:00 – 12:00

Plenary lecture 4 (Pufendorf Hall)
Nathalie Hicks-Champod, Forensic Genetics Unit, University Center of Legal Medicine & Formation Continue, School of Criminal Justice, University of Lausanne:
Understanding the Value of Evidence: Challenges met in Education.
Chair: Birgitta Rådström

Nathalie Hicks-Champod earned both her M.Sc. and Ph.D. in Forensic Science from the University of Lausanne. During the early stages of her career, she focused on assessing forensic results given on activity level propositions, specializing in the analysis of glass during her Ph.D. research. She gained valuable experience during her three-year tenure at the former Forensic Science Service, working in the R&D department of the Physical Science division. Following her postdoctoral work in the field of DNA at the University of Lausanne, Tacha has been dedicated to delivering online interpretation courses tailored for forensic caseworkers since 2010.
Currently, she also serves as the interpretation leader at the Forensic Genetic Unit of the University Center of Legal Medicine in Lausanne - Geneva, actively participating in casework. Her extensive publications focus on interpretation issues, and she has contributed her expertise as a member of various commissions on the interpretation of forensic evidence.

12:15 – 13:15

Plenary lecture 5 (Pufendorf Hall)
William C. Thompson, University of California, Irvine:
Forensic Scientists' Decision Thresholds and the Accuracy of Verdicts
Chair: Julia Mortera

William C. Thompson is professor emeritus at the University of California, Irvine, where he has held academic appointment in criminology, psychological science and law. He has a PhD in psychology (Stanford) as well as a JD (UC Berkeley). His career has largely been devoted to improving forensic science through scholarly publications, test case litigation, and work with advisory and standard-setting bodies. He currently co-chairs a committee appointed by the Attorney General of Maryland to design audit procedures for the medical examination system of Maryland. He helped draft the recently-issued report of the Royal Statistical Society on the investigation of alleged serial misconduct by medical professionals.
He has also served as Special Master for the United States District Court for the District of Minnesota, advising the court on the scientific status of methods for probabilistic genotyping. He chaired a panel of the American Association for the Advancement of Science (AAAS) that produced a major report on latent fingerprint examination. And he served for five years on the Forensic Science Standards Board

of the Organization of Scientific Area Committees for Forensic Science (OSAC), which is the major government-supported body responsible for development of forensic science standards. He also chaired the OSAC Human Factors Task Group. At UCI he continues to do research on cognitive and contextual factors that affect the production and communication of forensic science evidence, with support from the National Institute for Standards and Technology (NIST) through the Center for Statistical Applications in Forensic Evidence (CSAFE).

13:15 – 13:45 *Closing ceremony* (Pufendorf Hall)

13:45 – 14:45 Lunch (Gallerian)

Sessions

Monday June 12

Invited session 1
13:45 – 15:15 (Pufendorf Hall)

Statistical Models for Fingerprint Analysis: Thinking Broadly about the Future

Organizer and chair: Simon A Cole, University of California, Irvine

Sponsored by the Center for Statistics and Applications in Forensic Evidence (CSAFE)

Over the past several years, increasing attention has been drawn to the need for statistical models in forensic science. It has been argued that statistical models are necessary to assist forensic scientists by enabling them to evaluate and report the significance of their findings in a logical and scientifically defensible manner. However, the actual development and use of such models has been slow and subject to controversy. Forensic DNA profiling is perhaps the area in which such models have been most fully developed, but the ascendence of probabilistic genotyping has been characterized by vigorous controversy and debate.

Aside from DNA profiling, perhaps the most obvious candidate discipline for a statistical model is friction ridge ("fingerprint") analysis. Fingerprint analysis remains widely used and highly trusted. Conceptual work on statistical models for fingerprint analysis has been done (e.g.,Neumann et al., 2012), and at least two working models are available (e.g., Swofford et al., 2018).

The focus of this session is to envision what a future with a fingerprint statistical model will look like and to anticipate the scientific, legal, sociological, and ethical issues that such a future may entail.

Contributors:
Alex Biedermann, University of Lausanne:
The importance of the distinction between reliability and reliance in the use of algorithmic output in forensic science

Caroline Gibb, University of Twente:
The game changer: fingerprint statistical modelling

Geoffrey Stewart Morrison, Aston University:
What a future forensic-data-science model for fingermark-fingerprint comparison might look like

Contributed session 1

Legal and Evidence theory 1

Chair: Julia Mortera

13:45 Henrike Neumann, Ria Langbehn Jensen, Pernille Skovbo Carøe, Birgit Feldtmann:
Overseeing the Handling of Technical Evidence by Police and Prosecution in Denmark

14:15 Michał Sikorski:
Is Forensic Science in Crisis*?*

14:45 Joseph B Kadane, Anders Nordgaard:
Using Bayes Factors to Limit Forensic Testimony to Forensics: Composite Hypotheses

Special call, session 1

Applications and Foundational Methods in the Presentation and Interpretation of Evidence with Score-Based Approaches

Chair: Danica Ommen

15:45 JoAnn Buscaglia, Christopher P Saunders, Janean Hanka, Danica Ommen:
Introduction to the Interpretation of Evidence with Score-based Approaches Illustrated using Trace Element Concentrations in Aluminum

16:30 Peter Vergeer:
Comparing a Trace to a Known Reference: the Applicability of Score-based and Common-source Likelihood Ratios

Contributed session 2

Legal and Evidence theory 2

Chair: Christian Dahlman

15:45 Marcello Di Bello, Rafal Urbaniak:
Higher-order Legal Probabilism

16:30 Frans Alkemade:
How should we further encourage the still hesitant, but growing acceptance of Bayesian reasoning in the Dutch (and other) courts?

Tuesday June 13

Special call, session 2

Applications and Foundational Methods in the Presentation and Interpretation of Evidence with Score-Based Approaches

Chair: JoAnn Buscaglia

08:30 Danica Ommen, Christopher P Saunders, JoAnn Buscaglia:
Development and Evaluation of a Contrastive Learning Framework with Applications to Micromorphometry of Aluminum Powder used in Explosives

09:15 Tuomas Korpinsalo, Kaisa Jalava, Martin Söderström, Olli Laine:
Producing likelihood ratios based on expert knowledge for oil spill identification comparisons through elicitation

Contributed session 3

DNA evidence 1

Chair: Therese Graversen

08:30 Petter Mostad, Andreas Tillmar, Daniel Kling:
Improved relationship inference when using low-coverage sequencing data

09:15 Peter Green, Julia Mortera:
Inference about complex relationships using peak height data from DNA mixtures

Special call, Session 3

Applications and Foundational Methods in the Presentation and Interpretation of Evidence with Score-Based Approaches

Chair: Christopher P Saunders

10:30 Joseph Zemmels, Susan Vanderplas, Heike Hofman:
Automatic Cartridge Evidence Scoring

11:00 Wauter Bosma:
Addressing typicality in score-based LR systems

11:30 Geoffrey Stewart Morrison:
Similarity-score-based likelihood ratios do not take account of typicality

Contributed session 4

Statistical modelling

Chair: Roberto Puch-Solis

10:30 Joyce Klu, Roberto Puch-Solis, Niamh Nic Daeid:
Measurement Uncertainty Estimation and Calculator Software (MUCalc)

11:00 Camille van Dijk, Anoek van Someren, Richard Visser, Marjan Sjerps:
Evidential value of duct tape comparison using loopbreaking patterns

11:30 Andrew Simpson, Semhar Michael:
Finite Mixture Modeling for Hierarchically Structured Data with Application to Keystroke Dynamics

Contributed session 5

Impression evidence

Chair: Birgitta Rådström

13:30 Danyela Kellett, D Lagnado, S Nakhaeizadeh, R Morgan:
Use of Bayesian Networks as a model for the Evaluation of Footwear Evidence

14:00 Muxin Hua, Susan VanderPlas
Automatic Acquisition and Identification of Footwear Class Characteristics

14:30 Ruoyun Hui, Anjali Mazumder, Lisa J Hall, Graham Jackson, Ian W Evett:
The evolution of fingerprint evaluative opinions in the UK: embracing themodern Paradigm

Invited session 2

13:30 – 15:15 (Crafoord Court Room)

The SAILR software

SAILR is a software package that calculates likelihood ratios (LRs) for source level propositions that addresses both whether a trace comes from a known source (evidential evaluation) and also whether two traces come from an unknown source (classification).

An example of evidential evaluation propositions is:
H1: The glass fragment from the suspect's garment came from the broken window
H2: The glass fragment from the suspect's garment came from another source

An example of classification propositions is:
H1: The glass fragments from the crime scene came from Class A
H2: The glass fragments from the crime scene came from Class B

SAILR incorporates two types of statistical models: score- and feature-based models. Within the score-based models, there are several scores that can be selected. Once the scores have been calculated, there are several one-dimensional continuous probability distributions that can also be selected.

SAILR also incorporates multidimensional Gaussian models for the calculation of likelihood ratios from several available features. Typically, a multivariate Gaussian distribution is assumed for the within-source distribution, while the between-source distribution can be estimated using either a multivariate Gaussian distribution or multivariate kernel density estimation.

In the session we will present the models in SAILR as well as a couple of casework examples of multivariate feature-based models. The first example will discuss elemental composition (as percentage data) of glass samples. Secondly, we will also look at the transformation of chromatographic data (as ratios of peak heights) into likelihood ratios using SAILR.

Contributors:
Roberto Puch-Solis, University of Dundee
Tereza Neocleous, University of Glasgow
Jonas Malmborg, Swedish National Forensic Centre

Contributed session 6

Forensic value of evidence 1

Chair: Christopher P Saunders

15:45 Alex Biedermann, Timothy T Lau:
Decisionalising the problem of reliance on expert and machine evidence

16:15 Gerhard Wevers:
Do verbal probability equivalents have any meaning?

16:45 Semhar Michael, Andrew Simpson, Dylan Borchert, Christopher P Saunders, Larry Tang:
Detection and Characterization of subpopulations and the study of algorithmic bias in Forensic Identification of Source Problems

Contributed session 7

Forensic value of evidence 2

Chair: Grzegorz Zadora

15:45 Geoffrey Stewart Morrison:
A bi-Gaussian method for calibration of likelihood ratios

16:15 Paolo Akira Kunii:
Considering the possibility of disguise in forensic handwriting examination with a Bayesian network

Posters

Eduardo Avila, Márcio Dorn, Carlos Eduardo Ibaldo Gonçalves, Alessandro Kahmann Clarice Sampaio Alho.
BR-EASE: Brazilian eye and skin estimator. A forensic webtool for DNA phenotyping

Eduardo Avila, Alessandro Kahmann, Cássio Augusto Bettim, Alexsandro de Vasconcellos, Clarice Sampaio Alho, Márcio Dorn:
MC1R and age heteroclassification of face phenotypes in the Rio Grande do Sul population

Eduardo Avila, Alessandro Kahmann, Clarice Sampaio Alho, Márcio Dorn:
Pareto Front solutions as a tool for feature selection in SNP-based forensic DNA phenotyping panels design

Janean Hanka, Christopher Saunders, JoAnn Buscaglia, Danica Ommen
SLR Properties for the Specific Source Problem

Catherine Holland:
A Bayesian Hierarchical Model for Compositional Data with Structural Zeros, Applied to Forensic Glass Data for Classification and Evidence Evaluation

Ruoyun Hui, Gail Robertson, Amy Wilson, Ivan Birch, Graham Jackson, Colin Aitken
Evaluating the probative value of forensic gait evidence

Peter Vergeer, Dylan Borchert, Christopher P. Saunders:
Probabilistic foundations for the use of the logistic regression Bayes factor in forensic source identification

Grzegorz Zadora, Rafał Borusiewicz,, Agnieszka Martyna:
Verification of (un)common source of capsaicinoid profiles of oleoresin capsicum sprays

Wednesday June 14

Invited session 3
08:30 – 10:00 (Pufendorf Hall)

Round-table discussion on "Healthcare serial killer or coincidence?"

Organiser and chair: Julia Mortera, Roma Tre University and University of Bristol

Suspicions about medical murder often arise due to a surprising or unexpected series of events, such as an unusual number of deaths among patients under the care of a particular professional. There have been several high-profile cases where healthcare professionals are accused of murdering patients where statistical evidence has been misused.

There are major concerns about the analysis and interpretation of the evidence in these types of investigations and whether it can be guaranteed that the data have been compiled in an objective and unbiased manner.

When interpreting such data, investigators need to consider:

- Could the deaths have occurred for reasons other than murder?
- If murder was the cause, is the person under suspicion responsible?

Attention is rarely given to ensuring that unconscious bias has not influenced the selection of these cases. Experts informing an investigation should be kept "blind" to all aspects of the case. Blinding is a key tool in minimising prejudicial subjective effects such as unconscious bias.

There is a need for better collaboration between the legal and statistical communities to prevent miscarriages of justice happening in the future.

Contributors:
Peter Green, University of Bristol
Jane Hutton, University of Warwick
William C. Thompson, University of California, Irvine
Lena Wahlberg, Lund University

Contributed session 8

DNA evidence 2

Chair: Ronny Hedell

11:45 Rolf J F Ypma, P Maaskant, S van Soest, M Sjerps, M van den Berge:
Bringing an LR system from concept to practice - mRNA analysis

12:15 Charles E H Berger, Maarten Kruijver, Tacha Hicks, Christophe Champod, Duncan Taylor, John Buckleton:
Some of the lower LRs are misleading! On the German recommendations for (not) reporting LRs from FCMs.

Contributed session 9

Activity level

Chair: Anders Nordgaard

11:45 Ruoyun Hui, Anjali Mazumder, Lisa J. Hall, Graham Jackson, Ian Evett:
Evaluating activity level propositions using fingerprint evidence: an outline of challenges

12:15 Marouschka Vink, Marjan Sjerps:
A Collection of Idioms for Modeling Activity Level Evaluations in Forensic Science

Contributed session 10

Forensic value of evidence 3

Chair: Anders Nordgaard

13:45 Karen Kafadar, Sydney Campbell, Jordan Rodu:
Estimating error rates in binary decisions with inconclusive outcomes

14:30 Dan Spitzner:
Recent methodological advances in Bayes factors for use in forensic analysis and reporting

Contributed session 11

Forensic chemistry

Chair: Jonas Malmborg

13:45 Grzegorz Zadora, Alicja Menżyk, Agnieszka Martyna, Alessandro Damin, Marco Vincenti:
Linking theory to forensic practice: A likelihood ratio-based approach for bloodstains dating

14:30 Pablo Ramirez-Hereza, Juan Maroñas, Daniel Ramos, Jose Almirall:
Hierarchical Bayesian Models to Improve Likelihood Ratio Calibration in Forensic Glass Comparison

Contributed session 12

Forensic value of evidence 4

Chair: Peter Vergeer

15:45 Amanda Luby, Erwin Mattijssen:
Understanding Factors in Forensic Decision-Making using Item Response Theory

16:30 Michael Puthawala:
Likelihood Estimation and Uncertainty Quantification

17:15 Roberto Puch-Solis, Muthu Rama Krishnan Mookiah, Niamh Nic Daeid:
Automated Segmentation of Breech Face and Firing Pin Images of Cartridge Cases and Firearm Identification using Deep Learning Methods

Contributed session 13

Evaluative reporting

Chair: Birgitta Rådström

15:45 Aurélie Barret:
Evaluative reporting: Why is it so difficult to implement? Observations from Belgium and other countries

16:30 Ullrika Sahlin:
Practical certainty and approximate probability - solutions for expressing uncertainty in scientific advice from food safety

Thursday June 15

Contributed session 14

Health and misuse of statistics

Chair: Colin Aitken

08:30 Dimitra Eleftheriou, Tereza Neocleous, Thomas Piper, Mario Thevis:
Doping control analysis in athletes' steroid profile: a multivariate Bayesian learning approach

09:15 Jane L Hutton:
Estimating life expectancy after injury or disease

10:00 Marjan Sjerps, Leen van der Ham, Peter Vergeer, Ivo Alberink, Patricia de Bruin:
Incident series: can this be just coincidence?

Contributed session 15

DNA evidence 3

Chair: Petter Mostad

08:30 Therese Graversen:
Advanced statistical reasoning about mixed DNA profiles

09:15 Klaas Slooten:
Differentiating between monozygotic twins on the basis of a mixed offender-victim DNA profile

Workshops

North Sea Group Workshop on Theory of Legal Evidence

Tuesday June 13, 09:00 – 16:45 (Pufendorf Hall)

Organizer and chair: Christian Dahlman, Faculty of Law, Lund University

Workshop on theoretical challenges at the intersection of law and forensic science. Example of questions to be discussed: What methodology should legal fact-finders use for assessing the combined effect of forensic results and other kinds of evidence? How can the legal standard of proof be understood in probabilistic terms?

09:00 – 10:00	Marjan Sjerps, University of Amsterdam and Dyon Doensen, Forensic Advisor at Court Limburg What should Legal Practitioners know about the Bayesian Framework for Evidence Evaluation?
10:00	Coffee/Tea
10:30 – 11:30	Alicia Carriquiry, Iowa State University, Director of the Center for Statistics and Applications in Forensic Evidence On the Design and Analysis of Black-Box Studies
12:00	Lunch (Gallerian)
13:00 – 14:00	Franco Taroni and Silvia Bozza, University of Lausanne The Primacy of the Bayes Factor in the Evaluation of Evidence
14:00 – 15:00	William C. Thompson, University of California at Irvine Dealing with Issues of Conditional Relevancy When Algorithmic Evidence is Presented to Lay Fact-Finders
15:15	Coffee/Tea
15:45 – 16:45	Anne Ruth Mackor, University of Groningen Preventing Miscarriages of Justice

Opportunities with Models for Probabilistic Genotyping

Wednesday June 14, 13:45 – 17:15 (Alfa Laval Agri)

Organizer and chair: Therese Graversen, IT University of Copenhagen, DK

Recent statistical advances have allowed more complex DNA profiles to be used as evidence in court. While probabilistic genotyping comes with major challenges, adopting a statistical approach also creates the opportunity for presenting a much more detailed story about the DNA evidence. Many of the questions naturally posed about the evidence, are equally naturally addressed as part of a statistical analysis.

My DNAmixtures software (https://dnamixtures.r-forge.r-project.org/) has been used for statistical evaluation of mixed DNA profiles in criminal cases both in Denmark and in the United Kingdom. I will demonstrate how I have used statistical reasoning in witness statements to give a detailed picture of my interpretation of the DNA evidence and to argue the reliability of the stated LR.

I will use DNAmixtures for the demonstration and hands-on tutorials, but the statistical reasoning applies broadly to other models for probabilistic genotyping.

As a participant you should ideally **bring your own laptop** to facilitate practical exercises.

Knowledge of statistics and the R/RStudio software is an advantage, but not required.

Abstracts[1]

A, B, C, D

How should we further encourage the still hesitant, but growing acceptance of Bayesian reasoning in the Dutch (and other) courts?

Frans **Alkemade,** AFR Alkemade Forensic Reasoning

Despite the benefits and growing recognition of Bayesian reasoning in domains such as science and technology, courts worldwide have struggled to embrace this approach. Applying Bayesian reasoning in evidence evaluation would require dealing with degrees of probability and uncertainty, involving a significant shift in thinking for some judges and prosecutors. While Bayesian reasoning and analyses have often been used to evaluate criminal cases, this was typically done after the final verdict had already been reached, and almost always in a more academic setting.
However, in the Netherlands there appears to be a growing acceptance of Bayesian reasoning in the practice of criminal law. Since 2014, integral Bayesian analyses have played a role – in my opinion a crucial role – in the outcomes of several criminal cases, including some high-profile ones.
To the best of my knowledge, it is fairly unique that courts (or prosecutors, or defence lawyers) have asked for integral Bayesian analyses. Such an analysis not only includes forensic, but also tactical evidence, and not only likelihood ratios but also priors. The fact that this eventually yields posteriors, raises important questions about the role of the analyst versus the role of the judge. Possibly as a result of the visibility of these integral analyses, an increasing number of judges in the Netherlands are now not only becoming aware of Bayesian thinking and its value, but are also willing to learn how to apply Bayesian principles to their own evidence evaluation. As a practitioner who has conducted a number of these integral analyses, I would like to share my experiences in solving some of the problems that I have encountered in situations where Bayesian reasoning had to be explained to (and accepted by) lawyers and judges.
To address the issue of Bayesian networks being perceived as too much of a black box by many legal professionals, I have adopted a more linear approach to presenting Bayesian analyses. This method allows for an intuitive understanding of concepts such as priors, likelihood ratios, and conditional dependences. Starting with a set of mutually exclusive and also – as much as reasonably possible – collectively exhaustive hypotheses, the available forensic, tactical, and circumstantial evidence can be added step by step, taking possible dependencies into account. This approach, while mathematically equivalent to a Bayesian network, offers full transparency, making it accessible even to laypeople. This allows courts to really 'judge' the analysis: They can decide for themselves to what extent, and possibly on which parts, they agree or disagree with the analysis. The judges can also, if desired, provide their own detailed probability estimates as inputs. I believe that integral Bayesian analyses are both viable and necessary for improving evidence evaluation in criminal cases. However, not everyone shares this viewpoint, and there have been fierce criticisms, both from within the judiciary and from the academic world. Some of these criticisms may stem from misunderstandings, but others are more fundamental in nature. I will conclude by briefly touching upon these fundamental disagreements, inviting further discussion on the matter.

[1] The abstracts are sorted alphabetically based (with one exception) on the last name of the first (presenting) author.

BR-EASE: Brazilian Eye and Skin Estimator. a forensic webtool for DNA phenotyping

Eduardo **Avila**[1,2], Márcio Dorn* [1,3], Carlos Eduardo Ibaldo Gonçalves [1], Alessandro Kahmann [1,4], Clarice Sampaio Alho [1,5].

1 INCT FORENSE National Council for Scientific and Technological Development, Brazil.

2 Setor Técnico-Científico, Superintendência Regional da Polícia Federal do Rio Grande do Sul, Porto Alegre, Brazil.

3 Structural Bioinformatics and Computational Biology Laboratory, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

4 Institute of Mathematics and Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

5 INCT BioOncoPed National Council for Scientific and Technological Development, Brazil

* Correspondence: marcio.dorn@inf.ufrgs.br

The evaluation of DNA evidence for human identification is not restricted to STR profile confront of questioned samples. Especially when no reference sample is available, DNA can be a source to the inference of Externally Visible Characteristics - EVC of a suspect or victim, thus Forensic DNA Phenotyping - FDP systems has been efficiently validated to predict skin, eye, and hair color. Molecular markers panel as well as the interpretation, evaluation and application of statistical models were successfully applied in homogeneous populations when people have extreme (light vs dark) phenotypes. However, these systems reached lower predictive values when intermediate phenotypes, or populations with major admixture features, are targeted. Eyes and Skin colors are highly heritable genetic traits and are the most obvious and distinguishable externally visible characteristics to be used in human identification. We investigated 68 pigment production-related SNPs and their efficiency to predict eyes and skin color, for forensic purposes, in 465 admixed individuals with different phenotypes from Southern Brazil, using statistical and bioinformatic tools. In this work, we introduce a simple system for eyes and skin phenotype prediction through DNA: BR-EASE, Brazilian Eye and Skin Estimator, available at ‘http://sbcb.inf.ufrgs.br/brease’. Eye color predictor uses 10 SNPs to segregate three phenotype classes (blue, intermediate, and brown) with an overall accuracy of 87%, and sensitivities of 98% for brown and 94% for blue, intermediate color has lower sensitivity of 31%. Skin color predictor uses 25 SNPs to segregate three phenotypic classes (light, intermediate and dark) with a general accuracy of 97.5%, and sensitivities of 100% for light, 98% for dark and 90% for intermediate color. BR-EASE is a simple interface and easy to use tool that can be used by genetics and anthropological researchers, in the classroom as an exercise of multiple gene inheritance, and especially by law enforcement agencies to assist in their investigations.

Key words: BR-EASE, Human Pigmentation, Eyes and Skin color, Forensic DNA Phenotyping.

Pareto Front solutions as a tool for feature selection in SNP-based forensic DNA phenotyping panels design

Eduardo **Avila**[1,2], Alessandro Kahmann[2,3], Clarice Sampaio Alho[1,2], Márcio Dorn[2,4]
[1]Agronomy College, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
[2]National Institute of Science and Technology - Forensic Science, Porto Alegre, RS, Brazil
[3]Interdisciplinary Departament, Federal University of Rio Grande do Sul, Tramandaí, RS, Brazil
[4]Institute of Informatics, Federal University of Rio Grande do Sul

Designing forensic-oriented genetic panels based on SNP variants presents a challenge to researchers: how to achieve maximum classification performance while employing the minimal number of loci possible? This issue can be addressed as a multi-objective optimization problem. As such, the Pareto front provides a tool to evaluate results aiming to maximize conflicting objectives simultaneously and can assist the choice of genetic variants included in a panel. The present work evaluated 77 DNA polymorphisms located in genes associated with pigmentation traits (eyes, skin, and hair) in 650 admixed individuals from Southern Brazil. Several machine-learning and statistical methods were used to generate pigmentation trait prediction panels and their respective classification efficiency. The Pareto front was then used to select the best-performing panels, considering the trade-off between classification effectiveness and the smallest number of SNPs included. Pareto dominant solutions seem to consist of appropriate and efficient prediction models and can be adopted as a framework for variant selection in panel design. Criteria of selection must, however, include a minimal classification threshold, since models presenting low classification efficiency are also present in the Pareto front. Such solutions are usually associated with a small number of genetic variants included in the panel.

Keywords: Forensic DNA Phenotyping; Forensic DNA panel design; Pareto Front; Multi-objective optimization.

MC1R and age heteroclassification of face phenotypes in the Rio Grande do Sul population

Eduardo **Avila**[2,3], Alessandro Kahmann[3,4], Cássio Augusto Bettim[1], Alexsandro de Vasconcellos[2,3], Clarice Sampaio Alho[3] and Márcio Dorn[1,3,5]

[1] Center of Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
[2] Technical Scientific Section, Federal Police Department in Rio Grande do Sul State, Porto Alegre, RS, Brazil
[3] National Institute of Science and Technology - Forensic Science, Porto Alegre, RS, Brazil
[4] Interdisciplinary Departament, Federal University of Rio Grande do Sul, Tramandaí, RS, Brazil
[5] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Forensic DNA Phenotyping (FDP) consists in the use of methodologies for predicting externally visible characteristics (EVCs) from the genetic material of biological samples found in crime scenes and has proven to be a promising tool in aiding human identification in police activities. Currently, methods based on multiplex assays and statistical models of prediction of EVCs related to hair, skin and iris pigmentation using panels of SNP and INDEL biomarkers have already been developed and validated by the forensic scientific community. As well as traces of pigmentation, an individual's perceived age (PA) can also be considered an EVC and its estimation in unknown individuals can be useful for the progress of investigations. Liu et al. (2016) were pioneers in evidencing that, in addition to lifestyle and environmental factors, the presence of SNP and Indel variants in the MC1R gene - which encodes a transmembrane receptor responsible for regulating melanin production - seems to contribute to an individual's PA. The group highlighted the association between these MC1R gene polymorphisms and the PA in the European population, where carriers of risk haplotypes appeared to be up to 2 years older in comparison to their chronological age (CA). Understanding that genotype-phenotype relationships cannot be extrapolated between different population groups, this study aimed to test this hypothesis and verify the applicability of this variant panel in the Rio Grande do Sul admixed population. Based on genomic data from a sample of 261 volunteers representative of gaucho population and using a multiple linear regression (MLR) model, our group was able to verify a significant association between 9 intronic variants in locus adjacent to MC1R (e.g.: AFG3L1P, TUBB3, FANCA, etc.) and facial age appearance, whose PA was defined after age heteroclassification of frontal face images through 11 assessors. Different from that observed in European populations, our results show that the presence of effect alleles (R) of the selected variants in our sample influenced both younger and older face phenotypes. The influence of each variant on PA is expressed as β values. For example, homozygotes for AFG3L1P rs112220510 appear to be 3.16±1.34 years younger than non-carriers of the variant, while homozygotes for AFG3L1P rs201156703 appear to be 7.54±3.59 years older than non-carriers.

Keywords: Forensic DNA Phenotyping; External Visible Characteristics; Perceived Age; MC1R; Single Nucleotide Polymorphisms.

Evaluative reporting: Why is it so difficult to implement? Observations from Belgium and other countries

Aurélie **Barret**, Forensic Advisor, National Institute of Forensic Science, Belgium

Like many forensics institutes, our multidisciplinary national forensic institute (NICC) performs a systematic evaluation of DNA results. Furthermore, some of our other labs already use the global Bayesian approach. After our institute's involvement in the defining of the ENFSI guideline for evaluative reporting in forensic science, we wanted to build on this momentum by developing a global approach within our institute, sharing our experience with evaluative reporting and by working on better communication with the different actors of the judicial system.

Where are we ten years after the guideline was implemented? The experts who already perform evaluative reporting do so within the framework of the ENFSI guideline and therefore also help to spread the guideline. More of the NICC experts are being trained in evaluative reporting. However we are still faced with the difficulty of moving towards a global approach for evaluative reporting considering our diversity. We were able to spread our knowledge to our national colleagues and partners. But lots of judicial actors and courts deal with keys questions at an activity level unfortunately without the forensic experts so without a necessary context-led evaluation of forensic results.

To evaluate the current situation, we reached out to our ENFSI partners and other forensic institutes. Through a national networking project (BELSPO project 2022) we were able to discuss their experiences with evaluative reporting and these discussions led us to believe that the forensic environment in Belgium is ready for a context-led approach in judicial cases, built on a close collaboration between forensics experts, magistrates and police officers. However, evaluative reporting is a complex thing to explain to investigating judges and even more so to a jury in court. A dedicated training, as performed by the Netherlands forensic institute (NFI), can assist with this but the plus minus ten cases per year that are subject to evaluative reporting are a poor argument for the necessary investments. To increase the number of cases concerned and therefore better motivate investments, a routine application of the evaluative approach to sexual assaults cases, as seen at the Forensic Institute in Ireland (FSI), could be an interesting direction for the NICC. Ultimately, a mentoring role of the practicing institutes and of the experienced experts could be explored through the ENFSI network.

Keywords: evaluative reporting, Bayesian approach, management, communication, education

Some of the lower LRs are misleading! On the German recommendations for (not) reporting LRs from FCMs

Charles E.H. **Berger** [1,2], Maarten Kruijver [3], Tacha Hicks [4,5], Christophe Champod [6], Duncan Taylor [7,8], and John Buckleton [3,9]

1. Netherlands Forensic Institute, The Hague, the Netherlands
2. Institute of Criminal Law and Criminology, Leiden University, Leiden, the Netherlands
3. Institute of Environmental Science and Research Limited, Auckland, New Zeeland
4. Forensic Genetics Unit, University Center of Legal Medicine and University Hospital and University of Lausanne, Lausanne, Switzerland
5. Fondation pour la Formation Continue Universitaire Lausannoise (UNIL-EPFL) and School of Criminal Justice, Lausanne, Switzerland
6. Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, Lausanne, Switzerland
7. Forensic Science SA, Adelaide, Australia
8. School of Biological Sciences, Flinders University, Adelaide, Australia
9. Department of Statistics, University of Auckland, Auckland New Zeeland

The German project group “Biostatistical DNA Calculations” and the German Stain Commission recently gave recommendations for the reporting (or not) of the LRs output by Fully Continuous Models (FCMs) that compare DNA profiles. In previous work, they observed a lack of reproducibility in the LRs produced. This led the working group and commission to five recommendations. In this presentation, we will show that the basis of the recommendations is not scientifically valid. We discuss the underlying common misconception about what an LR is and how LRs and the systems that generate them can and should be evaluated. We will also discuss some problems associated with the recommendations themselves. While we commend the Germans for their use of fully continuous models, implementation of the recommendations will lead to a lot of relevant evidence never reaching the courts.

Keywords: Forensic DNA, likelihood ratios, calibration, system performance, reporting recommendations, Germany

References:

1. Berger, C.E.H.; Kruijver, M; Hicks, T.; Champod, C.; Taylor, D.; Buckleton, J.S.B. Reaction to the recommendations by Hahn et al. regarding the interpretation and reporting of LRs from Fully Continuous Models. Manuscript in preparation.
2. Ramos, D.; Meuwly, D.; Haraksim, R.; Berger, C.E.H. Chapter 7: Validation of forensic automatic likelihood ratio methods. In: Banks, D.; Kafadar, K.; Kaye, D.; Tackett, M. (eds) Handbook of Forensic Statistics, 2020, Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

What's our standard going to be?

Charles E.H. **Berger**, Netherlands Forensic Institute and Institute of Criminal Law and Criminology, Leiden University

How can we raise the standard in forensic science? What do we want our standard to be?
This presentation discusses the upcoming ISO 21043 Forensic Sciences standard. It is from the perspective of one of the drafters, and it is not about quality management. There will be special emphasis on the ISO 21043-4 Interpretation standard.
The latest meeting of the technical committee was from May 29 until June 2, so you will get the news hot off the press.

Decisionalising the problem of reliance on expert and machine evidence

Alex **Biedermann**, University of Lausanne, School of Criminal Justice, Lausanne, Switzerland and Timothy T. Lau, Federal Judicial Center, Washington DC, US

This paper analyzes and discusses the problem of reliance on expert and machine evidence, including AI output, from a decision-analytic point of view. Machine evidence is understood here broadly as the result of computational approaches, with or without a human-in-the-loop, applied to analysing and assessing the probative value of forensic traces such as fingermarks.
We treat reliance as a personal decision for the factfinder; specifically, we define it as a function of the congruence between expert output in a particular case and ground truth, combined with the decision-maker's preferences among accurate and inaccurate decision outcomes. The originality of this analysis lies in its divergence from mainstream approaches based on standard, aggregate performance metrics for expert and AI systems such as aggregate accuracy rates as the defining criteria for reliance. Using fingermark analysis as an example, we illustrate that our decision-theoretic criterion for the reliance on expert and machine output has a twofold advantage. On the one hand, it focuses on what is really at stake in reliance on such output and, on the other hand, it has the ability to assist the decision-maker with the fundamentally personal problem of deciding to rely. In essence, our account represents a model- and coherence-based analysis of the practical questions and justificatory burden encountered by anyone required to deal with computational output in forensic science contexts. Our account provides a normative decision structure, i.e., a reference point against which intuitive viewpoints regarding reliance can be compared, which complements standard and essentially data-centered assessment criteria. We argue that these considerations, though being a primarily theoretical contribution, are fundamental to discourses about how to use algorithmic output in areas such as fingerprint analysis.

Keywords: Machine evidence, AI output, normative decision structures, decision theory, fingerprints

The importance of the distinction between reliability and reliance in the use of algorithmic output in forensic science

Alex **Biedermann**, University of Lausanne, School of Criminal Justice, Lausanne, Switzerland

In my contribution to this round table discussion on the role of statistics-based methods in forensic practice, I will argue that the current and foreseeable future emphasis on method *reliability*, while important, is insufficient in itself to address the salient question of how a decision maker ought to proceed when receiving expert evidence. I shall call this *the question of reliance* and propose it to be understood as a personal decision for recipients of expert information. I will outline why analysing and understanding reliance as a case-specific decision holds importance for clarifying what is fundamentally at stake for fact finders when they encounter computational results.

Addressing typicality in score-based LR systems

Wauter **Bosma**, Netherlands Forensic Institute

Score-based likelihood ratio systems have considerable advantages over feature-based systems. They can compare traces by measuring their similarity and are less demanding than feature based systems with respect to data requirements. A prominent *disadvantage* is that they lose typicality information in the process. Intuitively, when comparing a person's height, it is more meaningful if both the suspect and the perpetrator measured 203 and 205 cm than if they measured 180 and 182, even though the difference is only 2 cm in both cases, because many people are approximately 180 while a height of 203 cm is a rare find. Although an LR system using similarity only may be valuable, it is suboptimal because information on typicality is lost. We are not aware of an adequate answer to this problem. In an attempt to address it, we experimented with transforming features into their percentile rank before measuring similarity. In the person height example, this means we define (dis)similarity not as the difference in length directly, but as the percentage of people who are in between the measured lengths. Although this may be intuitive and easy to explain, we are not aware of earlier experiments to this extent. We measured performance on simulated and forensic data sets and found promising results.

Keywords: score-based likelihod ratio, percentile rank, typicality

Introduction to the Interpretation of Evidence with Score-based approaches Illustrated using Trace Element Concentrations in Aluminum

Dr. JoAnn **Buscaglia**, FBI Laboratory, Research and Support Unit, Dr. Christopher P. Saunders, South Dakota State University; Janean Hanka, South Dakota State University; Dr. Danica Ommen, Iowa State University.

There are three main interpretation methods that examiners follow when evaluating evidence in the context of an identification of source problem. The first method is largely frequentist, often referred to as the Two-Stage or Classical approach and uses coincidence probabilities to characterize the evidential value. The second method uses a likelihood ratio/Bayes Factor focused on logical and coherent decisions. The final class of methods broadly corresponds to Royall's likelihood approach, which is closely related to modern machine learning techniques. However, in many settings, it is difficult to construct these probability distributions with respect to the natural feature space in which the evidence is observed; therefore, it is becoming more common to rely on a scoring function that represents the results of the comparison of two sets of the evidence as either a univariate similarity or dissimilarity score. Then, the induced relevant probability distributions are used for the probabilistic interpretation of evidence. Additionally, the score-based methods have an attractive advantage in that they allow the forensic scientist to determine how to discriminate (or characterize the similarity) between sources and the statistician to focus on the probabilistic statements induced by the interaction of the forensic propositions and a score function.
We will present our desired properties for score-based methods used in each of these interpretation methods using illustrations from the analysis of trace element concentrations in aluminum used in improvised explosive devices.

Keywords: coherency; exclusionary difference; trace evidence; identification of source

On the Design and Analysis of Black-Box Studies

Alicia **Carriquiry**, Distinguished Professor and President's Chair, Department of Statistics, and Director of the Center for Statistics and Applications in Forensic Evidence, Iowa State University

Evidence from a crime scene including fingerprints and firearm markings on bullets is evaluated by examiners by comparing their image to one or more reference images. Typically, the examination is purely visual and results in a categorical conclusion such as “the print was made by the suspect’s finger”. How valid are these conclusions and the methods that lead to them? For pattern comparison disciplines, *black box* studies are considered the “gold standard” for assessing validity. In this type of study, participants are presented with a series of test kits and are asked to reach a conclusion as they would in real case work. Black box studies have been conducted in multiple forensic disciplines in the last few years, and published results suggest that examiners hardly ever make an error. Or not?

We argue that most of the studies violate basic experimental design rules and lack statistical justification. Furthermore, estimation of error rates computed using only complete cases results in unrealistically low values, that are at odds with moderate to low repeatability and reproducibility, estimated using the same data.

We propose some minimal statistical criteria for black box studies and describe some of the data that need to be available to plan and implement such studies. Depending on the data that are published, we formulate different hierarchical models to jointly estimate the missing data, and the set of plausible average error rates.

Collaborators in this work include Profs. Kori Khan and Heike Hofmann from Iowa State University and Susan VanderPlas from University of Nebraska-Lincoln.

Higher-order Legal Probabilism

Marcello **Di Bello**†, Rafal Urbaniak*
†Arizona State University - mdibello@asu.edu *University of Gdansk and Basis.ai,

The problem

It is standard to assess the strength of evidence using likelihood ratios, at least for quantitative evidence such as genetic matches. For example, the likelihood ratio of interest can be P(match evidence / source hypothesis)/P(match evidence / alternative to source hypothesis). Simplifying a bit, this likelihood ratio can be approximated by 1/f where f is the expected frequency (or proportion) of the matching genetic profile in a reference population. The 1/f ratio captures some of the uncertainty associated with the match in relation to the source hypothesis, but also leaves out crucial information. After all, the expected frequency may have been arrived at in different ways, say, via a larger or smaller sample. Should an additional measure of uncertainty accompany the likelihood ratio itself?

The state of play

A debate is ongoing in the forensic science literature on whether likelihood ratios should also be accompanied by a measure of precision, confidence or an interval estimation. (See, e.g., the 2016 special issue of Science and Justice, "Special issue on measuring and reporting the precision of forensic likelihood ratios", edited by G.S. Morrison.) Some scholars note that the likelihood ratio is not a parameter to be estimated, and thus all the uncertainty – including sampling uncertainty – should be encapsulated into a single number. Others believe that sampling uncertainty should be modelled separately, for example, via an interval estimation. Interestingly, this debate among forensic scientists parallels a philosophical debate on how probabilities can model a rational agent's evidence-based beliefs. One approach, known in the philosophical literature as precise probabilism, posits that an agent's credalstate is modeled by a single, precise probability measure. Another approach, known as imprecise probabilism, replaces precise probabilities by sets of probability measures. The philosophical literature contains arguments for and against each of these views.

Contribution

We favor a third approach, what we call higher-order probabilism, which relies ondistributions over probabilities. We show that there are good theoretical reasons to abandon both precise and imprecise probabilism and endorse higher-order probabilism. This claim has applications to the debate in forensic science. We argue that a second-order uncertainty should be taken into account when we assess, in probabilistic terms, the strength of match evidence. Instead of single-number likelihood ratios or intervals, we propose that higher-order likelihood ratios be used. This approach is consistent with Bayesianism in epistemology and does not require treating likelihood ratios as parameters to be estimated. In addition, we show that higher-order probabilism can be scaled to model complex bodies of evidence. Standard, single-number likelihood ratios associated with different pieces can be combined together to model complex structures of evidence by means of Bayesian networks. The same is possible for higher-order likelihood ratios. To this end, we sketch a formalism for constructing what we call higher-order Bayesian networks. We illustrate this higher-order approach by revisiting familiar cases in the literature such as Sally Clark and Charles Shonubi.

Keyworks: likelihood ratio; strength of evidence; Bayesian network; higher-order probability

Evidential value of duct tape comparison using loopbreaking patterns

Camille **van Dijk**[a,b], Anoek van Someren[a], Richard Visser[a], Marjan Sjerps[a,c]
[a]Netherlands Forensic Institute, [b]Utrecht University, The Netherlands, [c]University of Amsterdam, The Netherlands

A new method for the evaluation of duct tape ends is proposed. This method is based on the breaks of the loops in the warp yarns, when duct tape with a scrim of chain-stitched warp yarns and weft-insertion is torn. After tearing, the loop at the end of each warp yarn can be in one of four states: open, closed, complex or missing. Additionally, the horizontal position of each warp yarn can be expressed in terms of weft yarns. The evidential strength of these loopbreaking patterns is expressed in terms of likelihood ratios. We construct a likelihood ratio system to determine these likelihood ratios. This consists of three dynamic Bayesian networks, which are based on the main assumption that the loopbreaking patterns are a stochastic process which comply to the Markov property. Two of the networks are used to determine the probabilities for each loopbreaking pattern to occur, given that the two pieces were not connected. The third dynamic Bayesian network is used to determine the probability to find these two loopbreaking patterns given that the two ends used to be connected. The LR-system is trained and tested using a small dataset. Based on these initial results, it is found that the loopbreaking patterns contain very strong evidence. More data is needed to further develop this “proof of concept”.

Key words: LR-system, Dynamic Bayesian network, Markov property, Duct tape, Physical fit

E, F, G, H

Doping control analysis in athletes' steroid profile: a multivariate Bayesian learning approach

Dimitra **Eleftheriou** [1,2] d.eleftheriou@lacdr.leidenuniv.nl, Dr Tereza Neocleous [2] Tereza.Neocleous@glasgow.ac.uk, Dr Thomas Piper [3] t.piper@biochem.dshs-koeln.de and Dr Mario Thevis [3,4] thevis@dshs-koeln.de

1 Leiden Academic Centre for Drug Research (LACDR), Faculty of Science, Leiden University, Netherlands.
2 School of Mathematics and Statistics, University of Glasgow, United Kingdom.
3 Center for Preventive Doping Research/Institute of Biochemistry, German Sport University Cologne, Germany.
4 European Monitoring Center for Emerging Doping Agents (EuMoCEDA) Cologne/Bonn, Germany.

New methodology is developed for anomaly detection in repeated measurements of biomarkers sampled from the same individual. In anti-doping research, detecting abnormal biomarker values within the athletes' steroid profiles is very challenging if there is information on the negative cases while data on the positive cases is either limited or not available. An algorithm based on the one-class classification (OCC) concept is used to address this problem, hence, only negative cases (normal data) are used to train the developed classifier. Using a Bayesian hierarchical learning approach, the goal is to produce robust adaptive decision boundaries around normal concentrations as new recordings become available, and differentiate them from the abnormal ones. Probability density functions are specified as prior information on the parameters and the hyperparameters in a two stage Bayesian mixed effects model. This approach is a generalisation of the current single-biomarker test in forensic toxicology, which is powerful to monitor and assess repeated measurements from multiple biomarkers. We applied the developed model to detect anomalous values in an athlete's steroid profile consisting of repeatedly measured biomarkers and their concentrationratios from urine samples of professional athletes. Bayesian inference was based on Markov chain Monte Carlo sampling methods. A model validation procedure was used to compare the model results to the test data. Improved values of evaluation metrics suggest that the proposed approach may provide an effective tool in doping detection.

Keywords: Anomaly detection, Doping, Multivariate Bayesian mixed effects model, One-class classification, Steroids.

The game changer: fingerprint statistical modelling

Caroline **Gibb**, University of Twente

Several factors contribute to the challenges faced in successfully implementing rule changes, particularly in the field of forensic fingerprint analysis and interpretation. Rule changes are difficult to implement due to the complexity of systems and processes involved, resistance to change from individuals and stakeholders, legal and regulatory hurdles, resource constraints, the risk of unintended consequences, and the need for effective communication and education.

To advance current reporting practices in fingerprint analysis, it is essential to dissect and evaluate the examiner, the process, and the system. This involves carefully considering different approaches, such as frequency-based, AFIS score-based, and subjective probabilistic methods. Ideally, a well-calibrated system should yield consistent results, regardless of the chosen approach. However, the presence of different perspectives, strong opinions, and a somewhat competitive approach towards determining the superior model hinder an objective playing field. Overcoming resistance to change and achieving consensus among players in the criminal justice system becomes crucial in navigating the complexities of implementing updated rules and practices.

Let us consider a rule change in a game of sport. Rule changes are implemented to enhance the flow of the game and ensure player safety. Umpires are entrusted with enforcing these rule changes, while coaches play a crucial role in informing and training the players accordingly. The players themselves must adapt to the new rules, or they risk penalties and potential defeat. Although comparing the criminal justice system to a sport may not be a perfect analogy, it highlights the importance of coaching. In this context, the question arises: Who assumes the role of the umpire? Or perhaps more importantly, how do we get everyone on the same team?

Advanced statistical reasoning about mixed DNA profiles

Therese **Graversen**, IT University of Copenhagen, DK

Recent statistical advances have allowed more complex mixed profiles to be used as evidence in court. I will demonstrate how I have used statistical reasoning in witness statements to argue the reliability of the stated LR, in particular through various methods for checking the applicability of the probabilistic genotyping model. My DNAmixtures software has been used for statistical evaluation of mixed DNA profiles in criminal cases both in Denmark and in the United Kingdom. The software is open source and extremely efficient – a normal laptop enables you to evaluate mixtures of as many as 5 people. I will discuss witness statements based on calculations performed by DNAmixtures, but the statistical reasoning applies also to any other model for probabilistic genotyping.

Keywords: mixed DNA profiles, probabilistic genotyping, statistical methodology

Inference about complex relationships using peak height data from DNA mixtures

Peter J. **Green***, University of Bristol and Julia Mortera., Universit`a Roma Tre.

In both criminal cases and civil cases there is an increasing demand for the analysis of DNA mixtures that involve relationships, whether these relationships are the prime focus of interest, or an ancillary factor that needs to be taken into account. The goal might be, for example, to identify the contributors to a DNA mixture where the donors may be related, or to infer the relationship between individuals based on a mixture. We introduce a novel approach to modelling and computation for DNA mixtures involving contributors with arbitrarily complex relationships. It builds on an extension of Jacquard's condensed coefficients of identity, to specify and compute with joint relationships, not only pairwise ones, including the possibility of inbreeding. In this model, a Bayesian network (BN) is used as a modelling framework and computational device. The methodology developed is applied to two casework examples involving a missing person. In the first case we compute the likelihood ratio that two unknown contributors to a mixture are related compared to unrelated, testing relationships such as parent-child, sibs, first cousins, etc. The second case concerns a DNA sample taken from a murder weapon which appeared to be a mixture from the victim and possibly a close relative of the victim. In addition, we use simulation studies to demonstrate the ability of the methodology to recover complex relationship information from synthetic data with known 'true' family structure. The methods used to analyse the examples are implemented in the new KinMix R package, that extends the DNAmixtures package to allow for modelling DNA mixtures with related contributors. Rreeware R versions of both packages will be released imminently. KinMix inherits from DNAmixtures the capacity to deal with mixtures with many contributors, in a time- and space-efficient way.

Key words: Bayesian networks, coefficients of identity, DNA mixtures, inbreeding, kinship.

SLR Properties for the Specific Source Problem

Janean **Hanka** (South Dakota State University), Dr. Chris Saunders (South Dakota State University) Dr. JoAnn Buscaglia (FBI Laboratory, Research and Support Unit) and Dr. Danica Ommen (Iowa State University)

In the identification of source problem, a likelihood ratio (LR) is used to quantify the value of evidence under two competing models for how the evidence has arisen. When the feature space of this evidence is very complex, a score-based likelihood ratio (SLR) can be used as a surrogate for the value of evidence. Using a SLR results in the use of simpler underlying densities due to the score function mapping the complex evidence to a univariate score; however, it is expected that some information is lost when using a score. Hence, the SLR can perform slightly differently than the LR. In this poster, we discuss four reasonable properties that should be expected of a SLR when used for the specific source identification problem: first, that the SLR can be constructed when the background population consists of one alternative source. Second, when the background population consists of a single alternative source, and we invert the role of the specific source and the alternative source, the full SLR is also inverted. Third, when the alternative source population is composed of multiple sources, the inverse of the omnibus SLR can be written in terms of the average of the inverse of the simple SLR, where the simple SLR is the SLR of the specific source vs one alternative source. Finally, that the SLR does not provide stronger support for either model than a LR. These properties will be formally written and demonstrated on trace element concentrations in aluminum foil sources.

Keywords: Score-based likelihood ratio; trace evidence; explosives

Understanding the value of evidence: challenges met in education.

Nathalie **Hicks-Champod**, Forensic Genetics Unit, University Center of Legal Medicine & Formation Continue, School of Criminal Justice, University of Lausanne

Although there are only three principles of interpretation, and the breadth of probabilistic knowledge required to assess forensic findings in a logical way is reasonable, assessing and communicating forensic findings in casework is challenging. Why is it so? And how can education and training remedy the situation? In this presentation, we will share our experience in teaching interpretation principles over the last 10 years and highlight the most common pitfalls.

One principle of forensic interpretation that proves difficult to implement is the formulation of propositions. While it may appear straightforward, our experience has shown otherwise. We will delve into how we teach the important aspects of formulating propositions meaningfully, as well as those that are not as crucial.

It is important to emphasize that on this journey, experienced colleagues and mentors are essential. There will be storms, and at times, one may find themselves in uncharted waters, needing rescue in order to explain, for example, why scientists cannot provide their opinion on the source of DNA or whether they can determine if the transfer was direct or indirect. Avoiding fallacies, clarifying our role, and rephrasing conclusions that have been transposed are challenging tasks. It not only requires a shared culture among forensic scientists but also the ability to engage constructively with the end-users of our forensic reports.

Another challenge encountered during practical implementation, once scientists have acquired the necessary knowledge and data, is securing the required support, such as a conducive work environment, appropriate software, and colleagues engaged in interpretation. We will propose several avenues that have proven effective in facilitating the use of a logical framework for interpreting forensic evidence in casework.

A Bayesian Hierarchical Model for Compositional Data with Structural Zeros, Applied to Forensic Glass Data for Classification and Evidence Evaluation

Catherine **Holland**, University of Glasgow
c.holland.1@research.gla.ac.uk

Compositional data are those which are expressed as parts of some whole, such as the percentages of each element present in the chemical composition of glass. Analysing compositional data using standard statistical techniques poses several challenges, including a sum constraint of 100% for the different parts of the whole, and complex correlation structures between those parts. The analysis of compositional data is further complicated by the presence of zeros. This study focuses on a forensic elemental glass database that contains a significant number of structural zero values, which are considered to be true zeros as opposed to below the detection limit.

Previous research on this database examined splitting the data by the presence and absence of compositional elements to account for the structural zeros present. However, this approach is limited by the need for manual separation of the elements, which requires prior knowledge of the data. In this study, hierarchical clustering and k-means clustering were applied to automate the splitting of the data by the presence and absence of compositional elements. A multivariate Bayesian hierarchical model was then fitted to each subset of the data. The model was used for classifying glass fragments into use types and for evidence evaluation under two competing propositions about the source of two sets of glass fragments.
To reduce computation time, the R package NIMBLE was used for a flexible Markov Chain Monte Carlo (MCMC) implementation. The performance of this method was assessed in terms of classification accuracy using a five-fold cross-validation approach. Additionally, false positive and false negative error rates were obtained in the context of evidence evaluation.

The results of applying a clustering approach prior to implementing a Bayesian model are comparable to those of prior research. The overall correct classification rate of the two clustering methods ranges from 70.9% to 75.6%, while the false negative and false positive error rates for evidence evaluation range from 5% to 7.5% and 1.5% to 1.7%, respectively. Using NIMBLE reduced the computation time of the model by approximately 9 hours, and the addition of a clustering approach to split the data based on the presence and absence of the elements has made the model more accessible to users by adding this automated step.

Keywords: Compositional data; Bayesian hierarchical model; forensic glass; classification;evidence evaluation.

Automatic Acquisition and Identification of Footwear Class Characteristics

Muxin **Hua** and Susan VanderPlas, University of Nebraska Lincoln

One common problem when establishing the utility of footwear impression evidence is that it is difficult to characterize the comparison population. Footwear intelligence programs are difficult to implement, in part because there are relatively few ways to easily gather and process data from the local population. In this presentation, we discuss a scanner developed to passively acquire footwear data from the general population and a computer model which can identify features from the captured images to automatically label the shoes with appropriate class characteristics. The scanner has been designed to be embedded in sidewalks or hallways, and it takes color photos of the shoe outsole and side as someone walks across the surface. This captures more data than similar systems which are no longer manufactured, such as the EverOS scanner, and is designed to operate both indoors and outdoors across a range of weather conditions. In order to process the data gathered by the scanner, we employ transfer learning across multiple stages, using a Faster Region-based Convolutional Neural Network (Faster R-CNN) to find and identify features in images. This model has the capability to extract and identify features of an image the same way human brains can, differentiating shoes from the background, locating regions of interest, and labeling the regions of interest with appropriate features. The photos taken by the scanner have sufficient resolution to identify class characteristics, so we have designed the FRCNN model to identify tread shapes using labels such as circles, rectangles, bowties, and lines. Each prediction comes with a probabilistic score; when combined, the score, location, and prediction form a feature set which can be used to classify a shoe. Combining pretrained neural networks and newly gathered and labeled training data, this method bridges the gap between unfriendly numerical features, descriptors used by other algorithmic approaches, and features used by examiners in practice.

This presentation will describe progress made in automatic identification of relevant footwear features - brand, shoe size, and tread pattern elements, as well as complications that arise when combining machine learning algorithms with human-friendly features. Leveraging both clean training data and ”messy” data gathered from the local community using newly developed footwear surveillance devices, the author will present developments in footwear forensics which will enable examiners to testify as to the frequency of class characteristics in the local population in the very near future.

Keywords: Computer vision, Forensics, Machine learning, Classification

Evaluating activity level propositions using fingerprint evidence: an outline of challenges

Ruoyun **Hui**[a], Anjali Mazumder[a], Lisa J. Hall[b], Graham Jackson[c], Ian Evett[d]
a The Alan Turing Institute, British Library, London, UK
b Metropolitan Police: MO4 – Forensic Service, London, UK
c Advance Forensic Science, St. Andrews, UK; Abertay University, Dundee, UK
d Principal Forensic Services Ltd, Bromley, UK

Fingerprint examiners are sometimes asked to address propositions relating to the activities that may have led to a particular fingermark, or set of fingermarks, being deposited at a crime scene. This is indeed an area where they have a lot to contribute from their expertise, and practitioners routinely go through such reasoning when prioritising casework too. Frameworks around activity level questions in other types of evidence (e.g. DNA), however, do not readily translate into fingerprint evidence. It is not clear how practitioners should approach these questions to conform with regulatory requirements and reliably evaluate the weight of evidence.

In this study we outline some main challenges that complicate evaluative reporting of activity level propositions using fingerprint evidence. Considering the large variety of activity level questions around fingerprint evidence, we start with classifying them into three common types: TPPR (transfer, persistence, prevalence, and recovery), handling objects, and navigating through the environment. Examples are drawn from requests received in case work. The complexity of reasoning generally increases in this order, as more entities and case-specific information become relevant. We then discuss the challenges around framing (specifying the propositions and evidence under consideration), knowledge (obtaining empirical data or expert judgement), and inference (reasoning through complex scenario). Different types of activity level questions tend to bring out different types of challenges, although exceptions always exist.

Finally, we discuss some tools and methods that can help to address these challenges, including conducting experiments, assigning subjective probabilities, and applying probabilistic graphical models. Their use is demonstrated using case examples. The more tools fingerprint practitioners have at their command, the better they can flexibly adapt and combine them according to the case circumstances.

Keywords: Fingerprint evidence, activity level propositions, evaluative reporting, probabilistic reasoning

Evaluating the probative value of forensic gait evidence

Ruoyun **Hui**[a], Gail Robertson[b], Amy Wilso [a,b], Ivan Birch[c], Graham Jackson[d], Colin Aitken[b]
a The Alan Turing Institute, British Library, 96 Euston Road, London, UK
b School of Mathematics, University of Edinburgh, Edinburgh, UK
c FGA Services, UK
d Advance Forensic Science, St. Andrews. UK; Abertay University, Dundee, UK

In recent years substantial progress has been made to validate the tools and procedures in forensic gait analysis. However, there has been limited research on statistical methods to evaluate the probative value of gait evidence. To do so requires a better understanding of the variation of features of gait between individuals in the population and within the same individuals. We addressed this question using a previously described population dataset and newly collected datasets with repeated observations of the same individuals on separate occasions. Within the same individual, we observed little variation when participants walked under controlled conditions and were filmed from clear viewing angles, compared to a pilot study where participants were merely asked to film themselves walking without a clear protocol. We subsequently developed a likelihood ratio (LR) model through recoding ordinal gait features as binary variables and performing principal component analysis on them to transform the original data into independent numeric variables, which were then fed into a two-level model assuming the same within-individual variability. This model produced misleading LRs in less than 10% of the comparisons using the first four principal components; but we note that the risk increases when within-individual variability is mis-specified, especially in same-source comparisons when it is wrongly assumed to be low.
To account for factors that either affect the features of gait or the observation of gait, we built a Bayesian network that encodes expert knowledge about their effects on observed gait features, using footwear as an example. Once the gait expert provides case-specific information on footage quality, observed gait feature and an assessment of whether the subject is wearing their usual footwear, the model produces an LR for that feature.
There is still a lot of scope to further develop both models, especially on combining them to accommodate multiple correlated features of gait as well as the modifying factors. But it is clear that supervision by gait experts is indispensable in the use of statistical tools.

Keywords: gait analysis, likelihood ratio, gait variability, Bayesian network

The evolution of fingerprint evaluative opinions in the UK: embracing the modern paradigm

Ruoyun **Hui**[a], Anjali Mazumder[a], Lisa J. Hall[b], Graham Jackson[c], Ian W. Evett[d]
[a] The Alan Turing Institute, London, UK
[b] Metropolitan Police: MO4 – Forensic Service, London, UK
[c] Advance Forensic Science, Abertay University, Dundee, UK
[d] Principal Forensic Services Ltd, Bromley, UK

The categorical opinion of absolute certainty with regard to fingerprints source level comparisons is living on borrowed time. Not only has it been under external pressure for the lack of a logical basis but an increasing number of practitioners are also dissatisfied about not being able to communicate their opinions in a more nuanced way that would allow more useful evidence to be adduced.

In anticipation of this, we have considered how a new approach to formulating opinions may develop under control from within the profession. This approach is based on the principles of the Case Assessment and Interpretation framework (CAI), as already adopted or supported by the Association of Forensic Science providers (AFSP), European Network of Forensic Science Institutes (ENFSI) and the UK Forensic Science Regulator. Central to the approach is the notion of weight of evidence (WoE), which is formally defined as the logarithm of the likelihood ratio, but can be expressed qualitatively following a verbal scale. The exa-miner evaluates observations in the light of two competing propositions and considers whether their observations and evaluations support the prosecution proposition (Hp)or the alternative defence proposition (Hd).

The initiative for this paradigm change should emanate from consensus within the fingerprint profession. We recognise that the implementation will involve multiple stages of consultation, training, and evaluation, alongside the development of a national collection of ground-truth cases. In parallel, statistical models for quantitative assessment of likelihood ratios will continue to be developed. This proposed change will also facilitate the incorporation of such models in the future, through familiarising practitioners with the CAI framework and quantitative reasoning, as well as establishing the ground-truth collection to monitor the calibration, reliability and consistency of both statistical models and expert opinions. The primacy of expert judgement will always be recognised. The interaction of personal judgement with robust mathematical modelling should be seen as a powerful synergy.

Keywords: Fingerprint evidence, evaluative reporting, case assessment and interpretation, weight of evidence, calibration

Estimating life expectancy after injury or disease

Jane L **Hutton**, Department of Statistics, University of Warwick, Coventry, UK
J.L.Hutton@warwick.ac.uk

People who are injured in a traffic, industrial or medical accident, or contract a disease or cancer through negligence of an employer, doctor or local authority, can bring a civil case for compensation. Approaches to establishing causation differ between medical and legal professionals. In Britain, some judgements have been sceptical of epidemiological evidence, others have relied on finding research which shows relative risks greater than 2: “doubling the risk”. Statisticians and epidemiologists typically consider a wider range of issues, such as the source and quality of data and analysis Hutton (2018). After responsibility for injury or negligence is decided, the value of compensation will often depend on how long the person is expected to live. The two main challenges are finding reliable data or estimates of mortality risks (Hutton, 2018; Pharoah and Hutton, 2006), and deciding on a sensible method of estimation. In the UK, for common injuries, there are life tables for use in Personal Injury and Fatal Accident Cases Government Actuary’s Department (2020). For some particular injuries, such as spinal cord or brain injuries, an expert witness might be instructed to provide individual report. Such a report might also take into account other factors, such as alcohol consumption or medical history of diabetes or depression. I will compare the approaches used by actuaries, statisticians and medical doctors to estimating an individual person’s life expectancy. The underlying assumptions of independence of different factors, or additive or multiplicative modifiers of mortality rates differ substantially. Some medical doctors simply select a number. Others refer to research articles. Estimates are based on substracting years from, or taking a percentage of, the population life expectancy. Another method is the “Rating of Substandard Lives”. This involves selecting an excess mortality rate for each factor, adding the rates and referring to a set of life tables.
Statisticians base their estimate either directly on condition specific cohorts, or modify population estimates of annual mortality rates. Relative risks for obesity, say, are extracted from research publication. Such risks might be modified to reflect population rates of obesity. Mortality rates are modified by excess or relative risks, with and without loglinear decline to unity at old age. Another approach uses constant proportional life expectancy Strauss et al. (2005). There are open questions about criteria for discriminating between approaches, and professional judgement in the choice of risk factor estimates.

Keywords: Compensation, injury, life expectancy, methods of estimation.

References:

Government Actuary’s Department. Actuarial Tables With explanatory notes for use in Personal Injury and Fatal Accident Cases. The Stationery Office, London, 8 edition, 2020. The Ogden Tables.
JL Hutton. Expert evidence: civil law, epidemiology and data quality. Law, Probability And Risk, 17: 101–110, 2018. doi: doi.org/10.1093/lpr/mgy004.
P O D Pharoah and J L Hutton. Life expectancy in severe cerebral palsy. Arch. Dis. Ch., 91:254–258, 2006. doi:0.1136/adc.2005.075002.
DJ Strauss, PJ Vachon, and RM Shavelle. Estimation of future mortality rates and life expectancy in chronic medical conditions. J, Ins Med, 37:20–34, 2005.

I, J, K, L

Using Bayes Factors to limit forensic testimony to forensics: composite hypotheses

Joseph B. **Kadane**, Carnegie Mellon University and Anders Nordgaard, Swedish National Forensic Centre (NFC)

Limiting forensic testimony to forensics is mandated by many western legal systems. Using Bayes Factors under that restriction leads to constraints if the hypotheses used in the Bayes Factor are composite. In the light of the requirement of “guilt beyond a reasonable doubt,” attention is drawn to a lower bound on the Bayes Factor formed as the ratio of the least probability of the evidence in among prosecution scenarios divided by the greatest probability of the evidence among defense scenarios. DNA, toolmarks and shoe prints are discussed in this light.

Estimating error rates in binary decisions with inconclusive outcomes

Karen **Kafadar**, Sydney Campbell, Jordan Rodu, Department of Statistics, University of Virginia, Charlottesville VA, US

Binary decision-making occurs in many areas of science and policy; e.g., medicine (tumor present or absent), forensics (ID or exclusion), finance (good or bad credit risk), agriculture (healthy or diseased plant). Lab or field studies may be conducted to assess the error rates in such binary decision-making processes; e.g., proficiency tests or "error-rate studies" of the latent print examination process. In such tests, a true outcome is known (e.g., latent print and file print did or did not come from the same source), but study outcomes allow three responses (e.g., "same," "different," "inconclusive"). Many articles in forensic science report results of such studies by completely ignoring "inconclusive" responses, which can artificially increase the apparent accuracy rate. In this talk, we will discuss ways of estimating error rates in such studies that more fairly account for "inconclusive" decisions and enable fair comparisons of results across studies.

Key words: statistical hypothesis tests, direct standardization, confidence intervals, blind/double-blind experiments

Use of Bayesian Networks as a Model for the Evaluation of Footwear Evidence

Kellett D[1,2], Lagnado D[3], Nakhaeizadeh S[1,4] and Morgan R[1,4]
[1]Department of Security and Crime Science, University College London, London, UK
[2]Scientific Support Department, Lancashire Constabulary, Hutton, UK
[3]Department of Experimental Psychology, University College London, London, UK
[4]UCL Centre for the Forensic Science, University College London, UK

In October 2010, the Court of Appeal (Criminal Division) in England overturned the murder conviction in the case of R v T, in part because the Court considered that the process by which the strength of the footwear evidence was adduced and presented in the trial "lacked transparency" and was "inherently unreliable". In the decade since this ruling, the forensic footwear community in England have for the most part continued to use reference data and a likelihood ratio approach to inform the subjective opinion of the relative strength of the evidence. At the same time there has been commitment to demonstrate transparency and communicate that such methods are part of a holistic approach using the examiner's experience, judgement and knowledge as well as the data to provide a subjective opinion.

Footwear evidence can be a key tool in the investigation of crimes. The interpretation and evaluation of footwear evidence is a subjective process based on the experience and knowledge of the examiner, supported by the use of representative databases to provide an indication of the significance of any observed correspondence. In this presentation, we suggest the use of Bayesian Networks to model footwear evidence and provide a better tool for evaluating and understanding the probative value of footwear evidence and communicating this to the courts.

This presentation will introduce an initial model that has been produced based on the circumstances of the R v T case. This model has been populated with information from different databases, to explore the effect on the posterior probabilities. Whilst there is clearly value in modelling scientific evidence to understand better the effect on outcomes, it is important to build and view these models from the perspective of the forensic science profession. This would then make it possible to address real-life challenges and use the outcomes to communicate the findings successfully to the courts. We hope, therefore, that the model can be adapted for more generic modelling of footwear cases, to include different variables and databases. Ideally, it can then be used for other forms of evidence and insights from other forensic science fields. Ultimately, we anticipate that this model could become the first building block in an automated tool to provide a systematic and repeatable method to enable forensic scientists to consider the probative value of their findings and demonstrate a more transparent and objective means for evaluation and interpretation.

Keywords: Footwear, Interpretation, Bayesian Network

Measurement Uncertainty Estimation and Calculator Software (MUCalc)

Joyce **Klu**, Roberto Puch-Solis and Niamh Nic Daeid; Leverhulme Research Centre for Forensic Science, University of Dundee, UK

It is becoming a necessity for forensic toxicologists to assess and report the measurement uncertainty (MU) associated with their methods for drugs analysis [1]. For accreditation bodies such as ISO/IEC 17025 [1], it's now a fundamental requirement of quality management systems that forensic toxicology laboratories need to adhere to, not just to identify the sources of uncertainty but also to estimate the size of their contribution.

In this study, a method is developed in accordance ISO/IEC 17025 [2] requirements, for the quantitative analysis of delta-9-tetrahydrocannabinol (THC) in blood using solid phase extraction and LC/MS/MS [3]. A bottom-up approach is used to evaluate and estimate the MU of the analytical method. The underpinning method in deriving the overall uncertainty is used to develop a freely accessible software; MUCalc (Measurement Uncertainty Calculator) [4].

MUCalc is an interactive software for estimating the measurement uncertainty associated with a laboratory's analytical measurement. MUCalc provides a detailed methodical analysis with a transparent step by step calculation of how each uncertainty component is estimated in a user-friendly approach.

At the present, MUCalc estimates the measurement uncertainty associated with the component's, homogeneity, method precision, calibration curve, calibration standard and sample preparation. These are then put together to estimate the combined and the expanded uncertainty. These calculations are in accordance with standards set out by accreditation organisations including ISO/IEC 17025[2].

MUCalc is unique in the sense that it is a white-box application. It displays on screen all formulas together with step-by-step guide calculations in an easy to follow approach. This makes it easy for users to understand and cross examine every result generated by MUCalc. It also provides reference links to published articles to assist users make informed parameter choices. Such a transparent display of the workings of the calculation is essential for the purposes of disclosure within the legal domain.

The current version is suitable for toxicological samples (e.g. blood, urine), and seized drug samples (e.g. tablets, powder). We are currently testing MUCalc.

The study and the developed software provide a more detailed methodical analysis of how each uncertainty component can be calculated, providing full data sets for each computation to make it easy for practitioners needing to calculate MU to follow.

Keywords: Measurement uncertainty; Toxicology; MUCalc software

References:

[1] de Souza Eller, S. C. W., de Oliveira, F., and Yonamine, M. (2016). Measurement uncertainty for the determination of amphetamines in urine by liquid-phase microextraction and gas chromatography-mass spectrometry. Forensic science international, 265:81{88. Elsevier.

[2] ISO/IEC 17025:2017 (2017). General requirements for the competence of testing and calibration laboratories.

[3] Klu, J. K., Officer, J. A., Park, A., Mudie, R., & Nic Daeid, N. (2021). Measurement uncertainty in quantifying delta-9-tetrahydrocannabinol (THC) in blood using SPE and LC/MS/MS. Forensic Science International, 322, [110744]. https://doi.org/10.1016/j.forsciint.2021.110744

[4] Klu, J., Nic Daeid, N. and Mudie, R. (2022). Measurement Uncertainty Calculator (MUCalc) Software, University of Dundee.

[3] EURACHEM /CITAC Guide CG 4 (2012). Quantifying uncertainty in analytical measurement. Third Edition. International Organization for Standardization (ISO).

[4] JCGM 100:2008 (2008). Evaluation of measurement data: Guide to the expression of uncertainty in measurement. GUM 1995 with minor corrections.

Producing likelihood ratios based on expert knowledge for oil spill identification comparisons through elicitation

Tuomas **Korpinsalo**, Kaisa Jalava, Martin Söderström and Olli Laine; National Bureau of Investigation, Forensic Laboratory, Finland.

Investigation of oil spills in the environment typically involves comparing a chemical sample from a suspected source to a sample recovered from the spill with the intention of using chemical analysis to help answer whether the two samples contain the same oil. While standard methodology exists for the chemical analysis of samples and the interpretation of the chemical attributes in terms of indicating similarity between the samples, there is no obvious way to assess the strength of these findings in the form of a likelihood ratio (LR). A fully data-driven way of obtaining LRs would require assessing the distributions of high dimensional chemical data under assumptions of both same and different oils. However, producing such data in large enough quantities is difficult due to technical limitations of the standard analysis method.

Against this background, the ÖVERI project was conducted at the National Bureau of Investigation Forensic Laboratory in Finland in 2022 to develop a systematic method for assigning LRs in oil comparisons. To solve the issues mentioned above, an approach involving expert elicitation was adopted. Firstly, the data was reduced into a small number of scores describing similarity between two samples according to expert opinion. Secondly, variables representing these scores and their dependencies as well as relevant background factors were arranged into a Bayesian network. In the third and final phase, the Bayesian network was mathematically optimized to provide similar likelihood ratios as expected by the experts for given cases. The experts were involved in every stage of the process to ensure that the final model corresponds to the experts' knowledge regarding oil comparisons. The produced model allows experts to assign LRs on a likelihood ratio scale to cases based on assessments of similarity from chemical data in a consistent and transparent way. Additionally, the network codifies the expert reasoning into a form which allows easy scrutiny and further improvements on the model.

In this presentation, the produced Bayesian network is introduced as well as the elicitation process that was used to obtain it. Furthermore, the mathematical optimization of the model is described. The presentation covers the practical challenges and implications of the method and possible future extensions.

Keywords: likelihood ratio, evidence evaluation, elicitation, oil spills, subjective probability

Considering the possibility of disguise in forensic handwriting examination with a Bayesian network.

Paulo Akira **Kunii**, Brazilian Federal Police

Despite a growing body of research on the use of statistics and datasets in forensic handwriting examination (FHE), most practitioners still rely on two subjective processes to compare handwritings and evaluate the findings: perception and cognition. This should not prevent forensic handwriting examiners from providing opinions that help investigators and triers of fact—a view that is supported by ENFSI in its "Best Practice Manual for the Forensic Handwriting Examination", and by the FSR in its "Codes of Practice and Conduct — Development of Evaluative Opinions", for instance. The adoption of the likelihood approach to evidence interpretation is an important step to guarantee that the interpretation phase of FHE is balanced, logic, robust and transparent. An important feature of this approach is the possibility to formally consider different sources of uncertainty, which allows the examiner to combine them in a coherent manner. A particular challenge in FHE is the possibility that a questioned writing is the product of a disguise. This issue has been addressed by Marquis, Hicks, and Mazzella in the article 'How to Account for the Possibility of Disguise When Assessing Signature Comparisons', in which they propose three different approaches. In this work, I propose a fourth approach, one that combines ideas from the aforementioned paper and others, and is the result of a development of the likelihood ratio that includes uncertainties regarding the 'mode of writing' (usual or unusual) and the occurrence of disguise. The final expression of the likelihood ratio is then used to implement and validate a Bayesian network that represents the inferential problem. This model allows the evaluation of the evidence in different scenarios, helping one to address questions like "who wrote the questioned writing?" and "given that it was written by Mr. S, is the questioned writing the product of disguise?".

Keywords: forensic handwriting examination, disguise, likelihood ratio, Bayesian network.

LR systems for fingermark comparison in The Netherlands: a unique playing field

Jeannette **Leegwater**, Netherlands Forensic Institute

In forensic science, there has been a shift from claims about the posterior odds or probabilities of the hypotheses, to the Likelihood Ratio (LR) framework [Saks, Koehler (2005)]. Another development in the forensic field is the emergence of data science (artificial intelligence and machine learning) and statistical models. These techniques can be employed to construct LR systems.

The field of fingerprint analysis in The Netherlands has turned out to be a unique playing field encompassing all these aspects. In The Netherlands, the National Police tends to report conclusions that are formulated slightly towards an LR, yet still strongly resemble categorical statements. On the other hand, at the Netherlands Forensic Institute (NFI) a verbal likelihood ratio is reported. Moreover, at the NFI the experts' conclusion is supported by a feature-based model [De Jongh *et al.* (2019)]. A score-based LR system based on an AFIS-algorithm was developed [Leegwater *et al.* (2017)], however it is not used in the report. Although there is another model available [Swofford *et al.* (2018)], this is also not being used.

This presentation will examine the situation of the reporting of fingerprint evidence in The Netherlands. The discussion will focus on LR systems for fingermarks and explore the underlying reasons for the resistance against these systems. Insights from two other forensic domains, namely face and voice comparison, will be considered to gain further understanding. Finally, the presentation will contemplate on how the developers of such LR systems and fingerprint experts can collaborate to advance fingerprint analysis to the next level.

References:

1. Saks, M.J., & Koehler, J.J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892-895.
2. De Jongh, A., Lubach, A. R., Lie Kwie, S. L., & Alberink, I. (2019). Measuring the rarity of fingerprint patterns in the Dutch population using an extended classification set. *Journal of forensic sciences*, *64*(1), 108-119.
3. Leegwater, A.J., Meuwly, D., Sjerps, M., Vergeer, P., & Alberink, I. (2017). Performance study of a score-based likelihood ratio system for forensic fingermark comparison. *Journal of forensic sciences*, *62*(3), 626-640.
4. Swofford, H.J., Koertner, A.J., Zemp, F., Ausdemore, M., Liu, A., & Salyards, M.J. (2018). A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic science international*, *287*, 113-126.

Understanding Factors in Forensic Decision-Making using Item Response Theory

Amanda **Luby**, Swarthmore College, aluby1@swarthmore.edu and Erwin Mattijssen, Netherlands Forensic Institute, e.mattijssen@nfi.nl

"Black Box" studies have emerged as the gold standard in feature comparison disciplines to estimate error rates, with recent studies having been performed in latent prints, handwriting, and firearm comparisons. These studies provide error rate estimates aggregated over all examiners and comparison tasks, but the variability in error rates is also important for understanding the overall reliability within the discipline, since errors are not equally likely across every examiner and comparison. This variability was observed in a recent study of firearms examiners (Mattiijssen et al, 2020) even though examiners as a whole performed quite well on a relatively difficult comparison set. In this study, examiners were also asked to report the degree of similarity and degree of support for each comparison, and these ratings varied across examiners even when they agreed on the source conclusion. Correlation between the human decision-makers and an automated measure was generally mild, suggesting a high degree of variability in reporting styles. In a domain that relies on precise and consistent reporting, it is essential that this variability is understood, as small differences in early decision-making could have downstream impact. Item Response Theory (IRT), a class of statistical methods used prominently in educational testing, is one approach that accounts for differences in proficiency among participants as well as varying difficulty among comparisons. Using results from the 2020 firearms study discussed above, we demonstrate recent advances in IRT-based approaches for analyzing examiner behavior, including simple Rasch models, more elaborate decision tree models, and extensions that incorporate degree of support as a second response variable. Our results show that while there are differences in examiner proficiency, the largest variability in examiner decisions occurs in the tendency to make inconclusive decisions. We also find major differences in examiners' reporting styles, even after accounting for their proficiency and item difficulty; and examiners' reporting styles are generally unrelated to their proficiency and tendency to make inconclusive decisions. These results underscore the importance of better understanding the cognitive mechanisms involved with feature comparison and subsequent reporting tasks.

Keywords: firearms, Bayesian statistics, item response theory, psychometrics.

M, N, O, P

Detection and Characterization of subpopulations and the study of algorithmic bias in Forensic Identification of Source Problems

Semhar **Michael**[1], Andrew Simpson[1], Dylan Borchert[1], Christopher Saunders[1], and Larry Tang[2]
(1) Mathematics and Statistics, South Dakota State University, USA
(2) Statistics and Data Science, University of Central Florida, USA

The forensic source identification problem involves providing the summary of the forensic evidence to a decision-maker via the value of that evidence.This can be done via the forensic likelihood ratio which in turn requires modeling of a relevant background population. Some of the commonly used methods involve the assumption of normality. However, there might exist a latent variable representing an underlying subpopulation structure.
In this work, we will focus on identifying and characterizing subpopulations in the relevant population when there are hierarchically structured data. This will be done through semi-supervised finite mixture models that are adjusted for the hierarchical sampling procedure. In addition, we will study systematic algorithmic biases that can occur as measured by rates of misleading evidence for each of the subpopulations when the subpopulation structure is not accounted for. We will illustrate this based on a simulation study using synthetic data and a classical glass datasets.

Keywords: subpopulation modelling, hierarchically structured data, likelihood ratio, the value of forensic evidence, semi-supervised mixture models

A bi-Gaussian method for calibration of likelihood ratios

Geoffrey Stewart **Morrison**, Forensic Data Science Laboratory, Aston University Forensic Evaluation Ltd

The following is a perfectly-calibrated forensic-evaluation system: A system that outputs natural-log likelihood ratios, ln(LR), for which the distribution of ln(LR) in response to different-source inputs and the distribution of ln(LR) in response to same-source inputs are both Gaussian, the two distributions have the same variance, σ2, and the means of the different-source and same-source distributions are −σ2/2 and +σ2/2 respectively (hereinafter, we refer to this as a "perfectly-calibrated bi-Gaussian system"). Given a such a system, for any LR value, the probability density of the same-source distribution evaluated at the corresponding ln(LR) value divided by the probability density of the different-source distribution evaluated at the corresponding ln(LR) value will equal that LR value.

A perfectly-calibrated bi-Gaussian system for which σ2 is larger will have better performance than a perfectly-calibrated bi-Gaussian system for which σ2 is smaller. Performance can be measured using the log-likelihood-ratio cost (Cllr). Different empirical sets of system outputs could map to the same Cllr value (a many-to-one mapping), but there is a bidirectional one-to-one mapping between the σ2 value of a perfectly-calibrated bi-Gaussian system and its Cllr value.

The proposed bi-Gaussian calibration method consists of the following steps:

1. Using a forensic-evaluation system, calculate uncalibrated LRs for a set of input pairs for which the different-source or same-source status of each pair is known.
2. Calibrate the output of Step 1 using a traditional monotonic calibration method, e.g., logistic regression.
3. Calculate the Cllr value for the output of Step 21.
4. Determine the σ2 value for the perfectly-calibrated bi-Gaussian system with the same Cllr value as calculated in Step 3.
5. Ignoring same-source and different-source labels, determine the mapping function from the empirical cumulative distribution of the output of Step 1 to the cumulative distribution of a two-Gaussian mixture corresponding to the perfectly-calibrated bi-Gaussian system with the σ2 value determined in Step 4.
6. Using the mapping function from Step 5, map the output of Step 1 to calibrated LRs.

The uncalibrated LR value corresponding to an input pair for which the same-source versus different-source status is in question can be inserted at Step 5, and the calibrated LR value obtained from Step 6.

The presentation will explore the behaviour of the bi-Gaussian method compared to other calibration methods, logistic regression and pool-adjacent violators (PAV), and will highlight some potential advantages and limitations.

Keywords: bi-Gaussian; calibration; likelihood ratio; logistic regression; pool-adjacent violators

Similarity-score-based likelihood ratios do not take account of typicality

Geoffrey Stewart **Morrison**, Forensic Data Science Laboratory, Aston University Forensic Evaluation Ltd

There is confusion in the literature between "scores" which are uncalibrated likelihood ratios, which take account of both the similarity between the items of questioned and known source and their typicality with respect to a sample of the relevant population, and "scores" which only take account of similarity. The latter we will refer to as "similarity scores". Uncalibrated likelihood ratios can be converted to calibrated likelihood ratios (this has been common practice in forensic voice comparison since at least as early as 2007). Similarity scores, however, cannot be converted to likelihood ratios that meaningfully address source-level forensic propositions because they do not take account of typicality with respect to a sample of the relevant population.

Using the same data, the presentation will demonstrate the application of a common-source likelihood-ratio model and a similarity-score-based likelihood-ratio model. It will demonstrate that, for equally similar pairs of items, the common-source model results in a lower likelihood-ratio value for a pair of items that is typical with respect to the relevant-population distribution, and results in a higher likelihood-ratio value for a pair of items that is atypical with respect to the relevant-population distribution. In contrast, the presentation will demonstrate that a similarity-score-based model results in the same likelihood-ratio value for both pairs of items. The similarity-score-based model overvalues the typical pair of items and undervalues the atypical pair of items.

Keywords: likelihood ratio; score; typicality

What a future forensic-data-science model for fingermark-fingerprint comparison might look like

Geoffrey Stewart **Morrison**, Forensic Data Science Laboratory, Aston University Forensic Evaluation Ltd

Morrison (2022, https://doi.org/10.1016/j.fsisyn.2022.100270) described a paradigm shift in evaluation of forensic evidence in which methods based on human perception and subjective judgement are replace by methods based on relevant data, quantitative measurements, and statistical models / machine-learning techniques. The new paradigm methods are transparent and reproducible, intrinsically resistant to cognitive bias, use the logically correct framework for evaluation of evidence, and are calibrated and validated under casework conditions. The presenter will discuss what a future forensic-data-science model for fingermark-fingerprint comparison might look like, how it might be used by practitioners, and some of the challenges for getting there.

Improved relationship inference when using low-coverage sequencing data

Petter **Mostad**, Chalmers University, Sweden; Andreas Tillmar, National Board of Forensic Medicine, Sweden; and Daniel Kling, National Board of Forensic Medicine, Sweden.

Pedigree inference, for example determining whether two persons are second cousins or unrelated, can be done by comparing their genotypes at a selection of genetic markers. We consider this problem when test data for one or more of the persons is from low-coverage next generation sequencing (lcNGS), and also consider issues related to low-quality samples. Currently available computational methods either ignore genetic linkage or do not take advantage of the probabilistic nature of lcNGS data, relying instead on first estimating the genotype.

We provide a method and software bridging the above gap. Simulations indicate how our results are considerably more accurate than some previously available alternatives. Our method, resembling the Lander-Green algorithm, uses a group of symmetries to speed up calculations. This group may be of further interest in other calculations involving linked loci.

Keywords: lcNGS; pedigree inference; linkage

Overseeing the Handling of Technical Evidence by Police and Prosecution in Denmark

Henrike **Neumann**[1], Ria Langbehn Jensen[1], Pernille Skovbo Carøe[1] and Birgit Feldtmann[2]
[1]Danish Independent Evidence Oversight Board, Aarhus, Denmark
[2]Aalborg University, Aalborg, Denmark

In Denmark, legal actors can draw upon a large variety of evidence throughout the course of criminal proceedings. There are relatively few regulations regarding which types of evidence can be put forward and how they should be presented in court. Both legal actors and the general public in Denmark tend to have high confidence in the accuracy and appropriate application of evidence in criminal cases. However, in recent years incidents have occurred regarding DNA evidence and telecommunications data, which have affected the aforementioned confidence. Due to these incidents, the Danish Independent Evidence Oversight Board was established in 2022. The background for establishing the Board was a wish to strengthen the public's confidence in evidence being handled appropriately by police and prosecution. The Board's task is to objectively oversee how police and prosecution in Denmark handle technical evidence. The Board defines technical evidence as evidence that cannot be included in criminal proceedings in its original form and therefore first has to undergo some sort of technical analysis. This includes for instance DNA evidence or fingerprints, which are subjected to forensic analysis, the results of which in turn can be applied in criminal proceedings. The Board can inspect specific types of evidence on its own initiative in so-called thematic oversights. Evidence types are selected based on a discretionary assessment of associated risk factors and significance for legal certainty. Moreover, police and prosecution are obliged to notify the Board whenever they encounter (possible) errors in their handling of technical evidence, which may be of systematic nature and have implications for legal certainty. Following such notifications, the Board will examine whether police and prosecution initiate appropriate measures to counter the discovered issues. In its first thematic oversight, the Board examines how police and prosecution handle DNA evidence. This selection was mainly due to the application of DNA evidence involving a complex probability calculation as well as being associated with well-known risk factors such as DNA transfer and fallacies potentially committed when interpreting and presenting evidence. A thematic oversight concludes with an overall report that contains an assessment of the current practice of police and prosecution as well as potential recommendations for future practice. In order to make these assessments and recommendations, the Board collects not only data on current practice in the handling of DNA evidence by police and prosecution but also information on international best practice.

Keywords: Evidence Oversight Board, oversight, technical evidence, DNA evidence

Reference: Feldtmann, B., Jensen, R.L., & Neumann, H. (in press). DNA als Beweismittel im Dänischen Strafverfahren: Wie sichert man Qualität von Beweisen in der Strafjustiz? [DNA as Evidence in Criminal Proceedings in Denmark: How to ensure the Quality of Evidence in the Criminal Justice System]. Schweizerische Zeitschrift für Strafrecht.

Development and Evaluation of a Contrastive Learning Framework with Applications to Micromorphometry of Aluminum Powder used in Explosives

Danica **Ommen** (Iowa State University), Christopher Saunders (South Dakota State University), JoAnn Buscaglia (Federal Bureau of Investigation Laboratory Division)

The identification of source framework can be used within forensic evidence interpretation to compare a pair of items and determine whether they have come from a common unknown source or from two different unknown sources. In comparing aluminum (Al) powder particles recovered from two pre-blast improvised explosive devices (IEDs), the goal is to determine whether the powder sources are associated, potentially providing investigative between-case linkages. These problems can be addressed using a variety of statistical techniques, including the Two-Stage, Likelihood Ratio and Bayes Factor approaches. Unfortunately, the complex nature of the evidence, such as replicate measurements taken on different levels of substructure within the source powder, creates difficulties in applying the usual approaches in a straightforward manner. For characterizing features of the Al powder, we take several subsamples from the bulk Al powder, several aliquots from each subsample, several fields of view are imaged on each aliquot, and then multiple micromorphometric parameters are measured for each particle in a field of view. The hierarchical nature of this type of data creates a complex dependency structure that is difficult to directly incorporate into the traditional statistical methods for source identification. In this presentation, a contrastive learning algorithm framework is developed for complex evidence types like Al powder micromorphometry. The contrastive learning methods consist of two major components: a method for quantifying the similarity (or dissimilarity) of pairs of evidential items and a method for determining the best separation between within-source or between-source comparisons.
During this presentation, we explore several different methods of quantifying pairwise similarity and several different methods of classifying pairs as within- or between-source comparisons. We will also present our approaches for evaluating the performance of the resulting score functions using micromorphometric data from Al powders.

Keywords: scores, machine learning, trace evidence, common source

Automated Segmentation of Breech Face and Firing Pin Images of Cartridge Cases and Firearm Identification using Deep Learning Methods

Roberto **Puch-Solis**, Muthu Rama Krishnan Mookiah, and Niamh Nic Daeid, Leverhulme Research Centre for Forensic Science, School of Science and Engineering, University of Dundee, Dundee, Scotland

Cartridge cases of fired bullets are essential in assessing whether a bullet was fired at a crime scene using a gun seized from a suspect. The firing pin of a gun leaves an impression on a cartridge case that is specific to the gun. Images of cartridge cases are used for firearm identification. In this work we present two aspects that aid in cartridge case evaluation: (a) automated determination of the area of the cartridge-case image that contains the breech face and firing pin while discarding the headstamp, and (b) automated classification of cartridge cases using deep neural networks. Both aspects were addressed using a dataset from The National Institute of Standards and Technology (NIST) of the USA, which consists of 1703 cartridge case images from 12 different guns: Ruger P89, Sig Sauer Model P226, Ruger P9PR15, Hi-Point C9, Smith and Wesson 10, Smith and Wesson 40VE, Colt VM, Glock VM, Ruger VM, Smith and Wesson VM, Sig Sauer VM, Ruger P95DC. The first aspect addressed is called image segmentation and it was performed using deep learning method DeepLabv3+, which achieved a high performance (accuracy of about 98%). The automated classification aspect tested nine deep learning models: DenseNet121, DenseNet201, EfficientNetB7, InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50, VGG19 and Xception. The best performing model obtained a very high classification performance (average F1-score of 0.9525 and average AUC of 0.9982). The Grad-CAM visualization technique was used to show the discriminative areas of the breech face and firing pin regions of the cartridge case image. The results reveal a great potential of deep learning models to accelerate firearm identification.

Keywords: Firearm identification; Cartridge case classification; Deep learning

Likelihood Estimation and Uncertainty Quantification

Michael **Puthawala,** Department of Mathematics and Statistics, South Dakota State University, Matti Lassas, Department of Mathematics and Statistics, University of Helsinki, Ivan Dokmanić, Department of Mathematics and Computer Science, University of Basel, Pekka Pankka, Department of Mathematics and Statistics, University of Helsinki, and Maarten V. de Hoop, Computational and Applied Mathematics and Earth Science, Rice University.

In this talk we will give a brief overview of some recent work on the theory and implementation of invertable and injective deep neural networks with a focus on their use in likelihood estimation and the role in uncertainty quantification. We will first do a general review of deep learning techniques with an emphasis on how they differ from traditional statistical inference. Then we will introduce the theory and application of some invertible networks, in particular the flow network and its variants, and their use in likelihood estimation and uncertainty quantification. Finally, we will conclude by discussing recent work on injective networks in manifold learning problems and their connections to existing bijective networks.

Keywords: Deep Learning, Manifold Learning, Likelihood Estimation, Uncertainty Quantification.

Q, R, S, T

Hierarchical Bayesian Models to Improve Likelihood Ratio Calibration in Forensic Glass Comparison

Pablo Ramirez-Hereza [a], Juan Maroñas [b], Daniel **Ramos** [a], Jose Almirall [c]

[a] AUDIAS Laboratory - Audio, Data Intelligence and Speech. Escuela Politecnica Superior. Universidad Aut´onoma de Madrid, Spain.

[b] Machine Learning group, Escuela Politecnica Superior. Universidad Aut´onoma de Madrid, Spain.

[c] Center for Advanced Research in Forensic Science. Department of Chemistry and Biochemistry. Florida International University, USA.

Likelihood Ratios (LRs) for multivariate trace evaluation is a prevalent metric in various source-level forensic disciplines. In this context, LRs compare the probability of two pieces of evidence if they originated from the same source versus the probability of these two pieces if they originated from different sources.

Traditional Bayesian approaches for the computation of these LRs statistically model the within-source and between-source variabilities of the multivariate traces, in such a way that parameters from the between-source variability are point-estimated from data. However, in forensic glass comparison, the uncertainty not modeled in the parameters, as well as the complexity of modeling multivariate distributions, the scarcity of data, and the curse of dimensionality, typically make these approaches to generate LRs with bad calibration.

In this work, we propose the use of hierarchical Bayesian models to incorporate uncertainty related to the model parameters and enhance the calibration of the resulting LRs. To achieve this, hierarchical Bayesian models define a prior distribution of their parameters and update it into a posterior distribution using Bayes' theorem and a given dataset. In this context, the Between-source variability is defined by this posterior distribution, avoiding to directly point estimate its parameters.

Our experiments with different LA-ICP-MS databases in the field of forensic glass comparison show the robustness of these models to data scarcity, as well as their ability to generate LRs more calibrated than previous approaches. For this reason, these hierarchical Bayesian model represent a more reliable and robust alternative than classical approaches.

Keywords: forensic glass comparison, la-icp-ms, likelihood ratio, hierarchical Bayesian models, Gaussianization, multivariate traces

Practical certainty and approximate probability - solutions for expressing uncertainty in scientific advice from food safety

Ulrika **Sahlin**, Associate Professor, Centre for Environmental and Climate Sciences, Lund University, Sweden

Food safety is an area relying on independent scientific experts for the synthesis of evidence to produce scientific advice. To meet demands from stakeholders and gain public trust, the European Food Safety Authority (EFSA) has adopted a policy for transparency about the assessment process. This implies to open about all steps of an assessment and communicate limitations in available knowledge. To do this, EFSAs Scientific Committee adopted a structured and flexible approach for uncertainty analysis (2018) and collected evidence-based recommendations on communication of uncertainty (2019). The underlying principle is that experts should at least try to express their uncertainty quantitatively using subjective probability. This is often done as a group judgement. For example, a final step of a scientific assessment might be to ask the group of experts to characterise their level of certainty in the conclusion by making judgement considering available evidence and identified sources of uncertainty. The scientific rigour of such judgements are contested, even among the experts themselves, and EFSA has made a huge effort in making expert judgements into a documentable and reviewable process, designed to reduce cognitive biases. I will summarise the emerging practice at EFSA on the characterisation of the level of certainty in a conclusion with formal Expert Knowledge Elicitation using behavioural aggregation. In particular, the possibility to express subjective probability in an approximate way and how to tackle situations where legislation requires a positive or negative conclusion without any expression of uncertainty.

Is Forensic Science in Crisis?

Michał **Sikorski,** Warsaw University of Technology

The results of forensic science are believed to be reliable (see e.g., Koehler 2016 or Murrie et al. 2019). At the same time, due to the lack of suitable empirical studies, we actually know very little about this reliability (see e.g., NRC 2009 or PCAST 2016).

I will attempt to re-assess the reliability of forensic results. I will argue that phenomena analogous to all of the main causes of the Replication Crisis in Psychology are present in forensic science and therefore forensic results are plausibly much less reliable than it is commonly believed.

I will start by sketching the methodological discussion surrounding the Replicability Crisis.

The crisis consists of the fact that most of the psychological (or even scientific) results are not reliable. I will focus on the causes (e.g., questionable research practices or publication bias) of this low reliability and large-scale replication projects (e.g., Collaboration 2015) that were instrumental in uncovering the Crisis. I will also discuss what is known about the reliability of forensic science.

The main source of our evidence concerning it are black-box studies. In such studies, the error rate is estimated by testing the performance of forensic experts in experimental setups. Then, I will discuss some of the reliability-reducing factors identified in the forensic literature.

I will try to show that problems analogous to all of the main causes of the Replicability Crisis are present in forensic science. For example, practices similar to the Questionable Research Practices are present in forensic science (e.g., Thompson 2009). Similarly, a well-documented tendency of prosecutors to hide exculpatory forensic evidence (see e.g., Jones 2010) has an effect analogous to publication bias. Finally, the biasing effects of the contextual information and motivations of scientists are present in both academic and forensic science (see Cooper and Meterko 2019 and Wilholt 2008). Moreover, I will argue that black-box studies are not effective in detecting the effects of those factors and therefore the results of such studies plausibly overestimate real-world reliability.

In conclusion, I will argue that we have strong reasons to believe that forensic science is now in a state of crisis similar in its severity to the Replicability Crisis, and in order to obtain reliable estimates of the reliability of forensic results we need to move beyond the black box studies by conducting studies analogous to the large-scale replication projects conducted in psychology.

References:

Collaboration, Open Science. 2015. “Estimating the reproducibility of psychological science.” Science 349 (6251).

Cooper, G., and Vanessa Meterko. 2019. “Cognitive bias research in forensic science: A systematic review.” Forensic science international 297:35–46.

Jones, C. 2010. “A Reason to Doubt: The Suppression of Evidence and the Inference of Innocence.” Journal of Criminal Law & Criminology 100:415–474.

Koehler, J. 2016. “Intuitive Error Rate Estimates for the Forensic Sciences.”

Murrie, D., et al. 2019. “Perceptions and estimates of error rates in forensic science: A survey of forensic analysts.” Forensic science international 302:109887.

National Research Council. 2009. Strengthening forensic science in the United States: A path forward. 1–328.

President’s Council of Advisors on Science and Technology. 2016. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.

Thompson, W. 2009. “Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation.” Law, Probability and Risk 8:257–276.

Wilholt, Torsten. 2008. “Bias and Values in Scientific Research.” Studies in History and Philosophy of Science Part A 40 (1): 92–101.

Finite Mixture Modeling for Hierarchically Structured Data with Application to Keystroke Dynamics

Andrew **Simpson** and Dr. Semhar Michael, South Dakota State University

Keystroke dynamics has been used to both authenticate users of computer systems and detect unauthorized users who attempt to access the system. Monitoring keystroke dynamics adds another level to computer security as passwords are often compromised. Keystrokes can also be continuously monitored long after a password has been entered. Many of the current methods that have been proposed are supervised methods in that they assume that the true user of each keystroke is known a priori. This is not always true for example with businesses and government agencies which have internal systems that multiple people have access to. This implies that unsupervised methods must be employed for these situations. One may propose using finite mixture models to model the keystroke dynamics but we show that there is often not a one-to-one relationship between the number of mixture components and the number of users. Also, users usually type numerous times during the session or block of time while using the system which means the keystroke dynamics from the session can be assumed to have arisen from the same user. We propose a novel method that accounts for the lack of a one-to-one relationship between the number of users and the number of components as well as accounts for known information based on when keystrokes were typed. Based on simulation studies and the motivating real-data example the proposed model shows good performance.

Keywords: multilayer mixture model, keystroke dynamics, source identification, system security, semi-supervised mixture model

Incident series: can this be just coincidence?

Marjan **Sjerps** [1,2], Leen van der Ham [1], Peter Vergeer [1], Ivo Alberink [1], Patricia de Bruin [3]
1 Netherlands Forensic Institute
2 Korteweg-de Vries Institute for mathematics, University of Amsterdam
3 Utrecht University

When a series of incidents happens involving the same person(s), and these incidents normally are quite rare, people intuitively suspect that there may be some underlying common (criminal) cause. Examples are people being suspected of killing several patients, killing their children, setting fire to several properties, or causing multiple car accidents. As some notorious cases (Sally Clark, Lucia de Berk, Daniela Poggiali) have shown, assessing the evidential value of the occurrence of an incident series is rather difficult, and mistakes can have severe consequences.
The Royal Statistical Society published a report in 2022 concerning statistical issues in the investigation of suspected medical misconduct. This is an important guideline for future cases. Furthermore, we consider the paper by Gill et al. (2018) very useful in this context.
At the Netherlands Forensic Institute, we reported in a number of cases involving possible arson and car accidents (possible insurance fraud). Furthermore, a student project resulted in a master thesis on the topic. This presentation focusses on the statistical methods used, and discusses some of the difficulties we encountered.

Keywords: naked statistical evidence, serial crime, Likelihood Ratio, evidence evaluation

References:

Royal Statistical Society (2022) Healthcare serial killer or coincidence? Statistical issues in investigation of suspected medical misconduct (https://rss.org.uk/RSS/media/File-library/News/2022/Report_Healthcare_serial_killer_or_coincidence_statistical_issues_in_investigation_of_suspected_medical_misconduct_Sept_2022_FINAL.pdf)
R.D. Gill, P. Groeneboom, and P. de Jong (2018) Elementary statistics on trial - the case of Lucia de Berk. Chance, pages 9–15, December 3, 2018. doi: 10.1080/09332480.2018.1549809.
P. de Bruin (2022) The forensic statistical analysis of incident series (https://studenttheses.uu.nl/handle/20.500.12932/41688)

Differentiating between monozygotic twins on the basis of a mixed offender-victim DNA profile

Klaas **Slooten**, Netherlands Forensic Institute and VU University, Amsterdam

In a recent Dutch sexual assault case, a suspect denied all involvement; he had a monozygotic twin brother. Based on standard autosomal DNA profiles, no distinction between the twins was of course possible.
In order to make this distinction possible, mutations in their DNA were identified. Most of these were present in the DNA of both brothers, meaning that the mutations were of pre-twinning origin.
However, while they both had the usual and the mutated variant, they had these with different proportions. The trace profile consisted of DNA of one of the twins and of the victim.
I will discuss the statistical model, incorporating various sources of uncertainty, that was defined in order to arrive at a likelihood ratio for the hypotheses that brother one, versus brother two, contributed to the trace DNA.

Recent methodological advances in Bayes factors for use in forensic analysis and reporting

Dan **Spitzner**, Associate Professor of Statistics, University of Virginia

This paper covers two recent methodological advances in forensic source problems, both connected to Bayes factors.

The first advance is the establishment of an evidence-reporting strategy derived from the concept of pool reduction. In response to studies of the perceptions of statistics used to present forensic evidence, the pool-reduction strategy is equipped with conceptual grounding through the argument that it reflects an unconventional interpretation of a Bayes factor.

The second advance addresses a prominent criticism of the use of Bayes factors, which is that they are oversensitive to the prior distribution. To address this criticism, the paper puts forward a robust Bayes factor, derived from a calibration technique that makes use of a certain concept of neutral data. Steps in the development of this technique are demonstrated toward its use with potentially complex articulations of background knowledge. Both of these advances are made with forensic applications in mind, and other new tools for the forensic analyst to consider.

Experimentation with common evidence-reporting strategies - stated as random match probabilities, likelihood ratios, and their verbal equivalents - has uncovered fallacious interpretation of forensic evidence, suggesting a priority to examine additional strategies. The pool-reduction strategy takes as its blueprint an illustration of the prosecutor's fallacy as it played out in the 1996 Doheny appeals case, wherein the source of a DNA sample was a key factor. Rather than accept a statement such as “there is a high probability the DNA is from the defendant”, the judge instructs the jury to interpret the forensic analysis to imply that the pool of potential sources has been reduced to a small pool of men, among which includes the defendant. This strategy highlights uncertainty while conveying the perception of discriminative strength, a characteristic that appears to benefit evidence-reporting strategies.

The paper formalizes the pool-reduction strategy through its connection to Bayes factors, and proposes a number of conventions for its implementation in practice.The sensitivity of Bayes factors to the prior distribution is well known among Bayesian methodologists, and has long been criticized in forensic science. A widely-studied solution is to substitute a default-prior setting within the formulation of a Bayes factor; this solution, however, is insufficient in forensic analysis, where background information expressed via theprior distribution is to be retained. The robust Bayes factor examined in this paper is shown to exhibit desirable sensitivity properties, and to show promise for adaptation to elaborate data-analysis scenarios.

Keywords: Bayes factor; evidence reporting; source attribution; prosecutor's fallacy; robust Bayes factor.

Forensic Scientists' Decision Thresholds and the Accuracy of Verdicts

William C. **Thompson**, University of California, Irvine

Forensic pattern analysts often report conclusions categorically (e.g., identification; inconclusive; or exclusion). This talk will use signal detection theory to model examiners' reported conclusions, focusing on the connection between the decision threshold and the probative value of the forensic evidence. It will use a Bayesian network model to explore how shifts in forensic examiners' decision thresholds may affect rates and ratios of true and false convictions in a hypothetical legal system. It demonstrates that small shifts in decision thresholds, which may arise from contextual bias, may dramatically affect the value of forensic pattern matching evidence and its utility in the legal system.

U, V, W, X, Y, Z

Comparing a trace to a known reference: the applicability of score-based and common-source likelihood ratios

Peter **Vergeer**, Netherlands Forensic Institute

In recent scientific literature, forensic scientists have argued against the use of score-based and/or common-source likelihood-ratio models when comparing a trace to a known reference. Opponents of score-based likelihood ratios (LRs) argue that score-based LRs lack a measure of rarity, while opponents of common-source LRs for this type of problem argue that these models only apply to a question whether two traces have a common source or not. This paper argues the opposite: these LR-approaches are applicable to the trace-reference problem.

Following the sampling models underlying score-based, feature-based, common- and specific-source LRs (when applied to the trace-reference problem), I will first show that they are all intimately related, and in particular a specific-source LR model can be seen as a nested version of a common-source LR model. Using this perspective, it becomes clear that from a Bayesian perspective common-source LR models applied to the trace-reference problem can be formalized by using an empirically motivated prior for the relevant characteristics describing the known reference.

Next, I will address the question whether score-based LRs are applicable (again from a Bayesian perspective), using well-founded performance measures (strictly proper scoring rules). I will show that when using a score-based LR system repeatedly in casework to update prior odds, its long-term performance is never less than using prior odds only. This makes score-based LR systems applicable, although in principle they perform worse than their feature-based counterparts do.

Keywords: likelihood ratio, score based, common source, specific source

Probabilistic foundations for the use of the logistic regression Bayes factor in forensic source identification

Peter **Vergeer***, Dylan Borchert#, Christopher P. Saunders#
* Netherlands Forensic Institute
# South Dakota State University

In comparison to likelihood ratios (LRs), Bayes factors (BFs) have the advantage that uncertainty in model parameter values is taken into account in a logical and coherent way. From a practical point of view, the BF is generally shrunken (i.e. value of evidence closer to 1) compared to its (maximum likelihood) LR counterpart. This makes Bayes factors in principle more suitable for automated value-of-evidence calculation than likelihood ratios. In forensic literature, it is common to calculate Bayes factors for generative models. It is also common to calculate likelihood ratios for discriminative models, for example using ML estimates of logistic regression parameters. In this poster, we present an approach to calculate Bayes factors when using logistic regression as a model to discriminate between two classes. Logistic regression can be used to discriminate between two populations where the log of the likelihood ratio follows a linear model. This is equivalent to the log of the posterior odds of group membership following a linear model. Using a database of observations known to be generated under two different models, we can obtain a posterior distribution for the parameters of the logistic regression, and use this distribution to obtain the posterior odds of group membership for a latent observation with unknown membership. This posterior odds ratio can then be divided by the prior odds ratio to obtain the corresponding Bayes factor. An important note is that by constructing the database with a prespecified number of observations under each model, we are fixing the base rates. This removes the Bernoulli sampling process of the labels used to construct the likelihood function for the logistic regression; by using logistic regression with our designed database, we are assuming a slightly different generative model, which will be discussed in the context of McLachlan (1992).

Keywords: Value of Evidence, Logistic Regression, Bayes Factor, Discriminative Model

Reference:

McLachlan, G.J. (1992) Discriminant Analysis and Statistical Pattern Recognition. John Wiley, New York.
http://dx.doi.org/10.1002/0471725293

A Collection of Idioms for Modeling Activity Level Evaluations in Forensic Science

Marouschka **Vink** MSc[a,b], Prof. Dr. Marjan Sjerps[a,b]
[a] Netherlands Forensic Institute, The Hague, Netherlands.
[b] University of Amsterdam, Amsterdam, Netherlands.

Bayesian networks are a powerful tool in forensic science for modeling activity level evaluations. However, constructing them can be complex, and the desire for a more standardized and structured approach to the modeling process is growing. Research suggests using an idiom-based approach for modeling complex evaluations using Bayesian networks. The idiom-based approach is a bottom-up modeling approach that begins with small reasoning patterns – we call them "idioms" - and combines them into a larger network. The ability to combine several idioms to create more comprehensive templates allows for greater flexibility in modeling and enables analysts to represent complex scenarios more accurately. In literature, collections of idioms have been made for several disciplines. We would like to add our own collection to the existing literature.

We are excited to share with you a collection of idioms that are essential for modeling activity level evaluations in forensic science using Bayesian networks. Firstly, we identified four degrees of specificity of Bayesian networks: idioms, instantiated idioms, template models and case models. We therefore believe that the idiom-based approach is a four-step method to advance the ladder of specificity: going from widely applicable models (idioms) to specific case models. Secondly, we present a collection of existing idioms and our own suggestions of forensic idioms useful for activity level evaluations. We have divided these idioms into five groups, each with a specific modelling objective: cause-consequence idioms, narrative idioms, synthesis idioms, hypothesis-conditioning idioms, and evidence-conditioning idioms. We strongly support the use of an idiom-based approach and want to emphasize the significance of our collection by demonstrating the ability to combine several of the presented idioms to create a more comprehensive template model.

Keywords: Bayesian networks; Forensic Science; Activity level; Evidence; Idioms; Probability

Do verbal probability equivalents have any meaning?

Gerhard **Wevers**, Science Leader, ESR Ltd, New Zealand

ESR, the New Zealand forensic laboratory has been using a Bayesian framework to interpret evidence since the late 1980s. This evidence includes all trace evidence (glass, fibres, paint etc) as well as comparative evidence such as toolmarks, firearms and shoeprint impressions. With the exception of glass evidence, the other evidence types are subjectively interpreted, and a LR subsequently assigned.
In the absence of a number-based probability, verbal equivalents are used when considering the probability of the examination findings (or evidence, E) given a particular proposition (usually the prosecutor's proposition (Hp) and the alternate proposition (Ha)). These verbal equivalents consist of an eight step range from negligible to very high. The difference in the number of steps for these verbal equivalents forms the basis of the assigned verbal LR. For example, if the probability of the examination findings for either proposition was the same, then the LR would be considered neutral. If there are five or more steps difference; for example, P(E|Hp) = very high and P(E|Ha) = very low, the resulting LR would be assigned as extremely strong support for Hp.

In an on-going attempt to improve this process, examiners that use this procedure were surveyed and asked to assign a numerical probability for each of the probability verbal equivalents. This survey was also sent to DNA analysts, examiners who do not use the Bayesian framework for evidence interpreting and (non-scientific) administration staff. An explanation of the process used was given to the DNA analysts, whereas the other two groups were not given any information and were only asked to assign a numerical probability to each of the verbal equivalents.
The results of the survey show that examiners who use this process think in terms of percentages and not probability. There was also no significant difference in the probabilities assigned to the verbal equivalents between examiners who use this process and the two groups that were only asked to assign probabilities without any other information.

Using the extreme values of the probabilities assigned in this survey from those who use this procedure, a P(E|Hp) = very high and P(E|Ha) = very low (extremely strong support for Hp) would have a calculated LR of between 5 and 100. However, a conclusion of extremely strong support would have a LR of greater than 1,000,000, suggesting that the use of probability verbal equivalents is meaningless.
This presentation will discuss the benefits and drawbacks of using the Bayesian framework to subjectively interpret forensic evidence.

Keywords: Bayesian, firearms, toolmarks, impressions, subjective

Bringing an LR system from concept to practice - mRNA analysis

RJF **Ypma***[1], P Maaskant[2], S van Soest[2], M Sjerps[1], M van den Berge[2]
[1]Division of digital and biometric traces, Netherlands Forensic Institute
[2]Division of biological traces, Netherlands Forensic Institute
*presenting author

Messenger RNA (mRNA) profiling can identify cellular origin such as the body fluids present in a stain, yielding information on what activities could have taken place at a crime scene. In current reporting practice at our institute, statements of the form '(no) indication for the presence of body fluid X' or 'no reliable statement possible' are made [1,2]. To progress to a probabilistic statement, several statistical models have been proposed in the scientific literature [3–7]. Publishing a model is not the end goal, however, and more work has to be done before such models are used in practice (i.e. getting from "foundational validity" to "validity as applied" [8]). Here, we present our experience in the broader process of getting from idea to practical implementation of a Likelihood Ratio (LR) system for mRNA body fluid identification, from a statistician's perspective. Relevant questions addressed include: when is a model 'good enough' to replace current practice, how to convince and instruct reporting officers, and how to work together to ensure their use of model output in evaluation is correct and formulation of conclusions is both correct and understandable.

Keywords: mRNA, likelihood ratio, validation, forensic practice

References:

1. van den Berge M, Bhoelai B, Harteveld J, Matai A, Sijen T. Advancing forensic RNA typing: On non-target secretions, a nasal mucosa marker, a differential co-extraction protocol and the sensitivity of DNA and RNA profiling. Forensic Sci Int Genet. 2016;20: 119–129.
2. Lindenbergh A, Maaskant P, Sijen T. Implementation of RNA profiling in forensic casework. Forensic Sci Int Genet. 2013;7: 159–166.
3. de Zoete J, Curran J, Sjerps M. A probabilistic approach for the interpretation of RNA profiles as cell type evidence. Forensic Sci Int Genet. 2016;20: 30–44.
4. Dørum G, Ingold S, Hanson E, Ballantyne J, Snipen L, Haas C. Predicting the origin of stains from next generation sequencing mRNA data. Forensic Sci Int Genet. 2018;34: 37–48.
5. Fujimoto S, Manabe S, Morimoto C, Ozeki M, Hamano Y, Hirai E, et al. Distinct spectrum of microRNA expression in forensically relevant body fluids and probabilistic discriminant approach.
Sci Rep. 2019;9: 14332.
6. Ypma RJF, Maaskant-van Wijk PA, Gill R, Sjerps M, van den Berge M. Calculating LRs for presence of body fluids from mRNA assay data in mixtures. Forensic Sci Int Genet. 2021;52:102455.
7. Li Z, Lv M, Peng D, Xiao X, Fang Z, Wang Q, et al. Feasibility of using probabilistic methods to
analyse microRNA quantitative data in forensically relevant body fluids: a proof-of-principle study. Int J Legal Med. 2021;135: 2247–2261.
8. President's Council of Advisor on Science and Technology. Report to the president on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016.

Linking theory to forensic practice: A likelihood ratio-based approach for bloodstains dating

Grzegorz **Zadora**[1,2], Alicja Menżyk[1,2], Agnieszka Martyna[1], Alessandro Damin[3], Marco Vincenti [3,4] and
1 Forensic Chemistry Research Group, Institute of Chemistry,University of Silesia in Katowice
2 Institute of Forensic Research, Krakow, Poland
3 Department of Chemistry, University of Torino, Torino, Italy
4 Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Orbassano (Torino), Italy

In the era of DNA testing, it often gets forgotten that verifying who was the donor of the bloodstain is not always the most important issue. In order to establish an evidential value of the trace, it is necessary to demonstrate an intrinsic link between the evidence and the investigated crime, providing the final piece of the forensic puzzle – information about the time of bloodstain deposition. Unfortunately, in spite of the nearly century-old efforts [1, 2], a reliable method for bloodstains dating is still missing.
Having looked into research struggles [1, 2], it seems that the difficulty lies not so much in developing new analytical tools capable of monitoring the time-dependent behaviour of bloodstains, as in the inability of translating these methods to practical proceedings. Researchers – often inexperienced in the forensic caseworks – neglect the multifactorial blood decomposition process and, above all, its dependency on external factors, such as storage conditions or the chemical make-up of blood. For this reason, the present study will focus on the development of novel strategy for estimating the time elapsed since trace deposition, substituting the conventional dating approach that relies on establishing calibration models. The pivotal point of this concept is to bring the question of bloodstains age down to the comparison problem considered within likelihood ratio (LR) approach, where the state of evidence degradation would be confronted with a set of reference materials. These control bloodstains would be obtained through the process of supervised ageing, conducted under conditions simulating – as closely as possible – the actual conditions of evidence decomposition at the crime scene, for the period corresponding to prosecutor's scenario. Eventually, the comparison of the time-dependent characteristics of bloodstains, reflected herein in their Raman spectra [3], would allow to verify whether the state of evidence degradation is in accordance with the reference materials created in accordance with the prosecutor's hypothesis, and thereby establish the relevance of the preserved evidence to the case in question.
The LR models proposed herein for comparative dating of bloodstains were built for the first latent variable from regularised MANOVA, which has been proven fairly effective in separation of the sources (considered herein as sets of bloodstains of given age) [3]. As it could have been expected, this discrimination capability has been successively deteriorating throughout the entire degradation period, nevertheless obtained LR models were still characterised by acceptable rates of false positive and false negative answers, as well as the empirical cross entropy (ECE) plots.

Keywords: Bloodstains; Forensic dating; Likelihood ratio; Case-suited approach; Raman spectroscopy

References:
1. R.H. Bremmer, K.G. de Bruin, M.J.C. van Gemert, T.G. van Leeuwen, M.C.G. Aalders, Forensic Science International, 216 (2012) 1–11.
2. G. Zadora, A. Menżyk, Trends in Analytical Chemistry, 105 (2018) 137–165.
3. A. Menżyk, A. Damin, A. Martyna, E. Alladio, M. Vincenti, G. Martra, G. Zadora, Talanta, 209 (2020) 120565.

Verification of (un)common source of capsaicinoid profiles of oleoresin capsicum sprays

Grzegorz **Zadora**[1,2,] Rafał Borusiewicz[1], Agnieszka Martyna[2]
[2]Institute of Forensic Research, Krakow, Poland
[1]Forensic Chemistry Research Group, Institute of Chemistry, University of Silesia in Katowice

Pepper sprays contain a solution of capsaicinoids, obtained by extraction from peppers. Quantitative relations of natural capsaicinoids depend on the plant material they were extracted from [1]. Pepper spray is a non-lethal weapon that should only be used for self-defense, but is often used by criminals to attack and incapacitate victims. Evidence related to these types of incidents, such as containers, clothes of victims or suspects, as well as traces of substances found at the scene, are submitted to the forensic laboratory. The purpose of the analysis is to identify the ingredients of the preparation (especially active components) and compare the traces found on objects from the victim or the scene with the preparation from the can or traces found within objects related to the suspect.
The study aimed to investigate the possibility of differentiating OC gases based on capsaicinoid profiles recorded in GC-MS analyses.
Sixty-four gases from 12 different manufacturers were purchased and tested. The likelihood ratio (LR) approach was applied to the data expressing the relative capsaicinoids contents computed by integrating GC-MS signals. Two hypotheses were assumed that stated either common or different origins of the samples. Several LR models have been developed, and their performance has been controlled by the number of false positives and false negatives as well as empirical cross entropy.
The results obtained show that OC sprays may be distinguished, even if they were produced by the same producer presumably if produced using different batches of pepper extract [2].

Keywords: comparison problem, likelihood ratio, capsaicinoids, pepper sprays, gas chromatography–mass spectrometry

References:
1 A. Garcés-Claver, M.S. Arnedo-Andrés, J. Abadía, R. Gil-Ortega, A. Álvarez-Fernández, Journal of Agricultural and Food Chemistry, 54 (2006) 9303–9311.
2. R. Borusiewicz, A. Martyna, G. Zadora, A. Zehrebelna, Forensic Science International, 328 (2021) 111

Automatic Cartridge Evidence Scoring

Joseph **Zemmels**, Iowa State University; Susan van der Plas, University of Nebraska – Lincoln and Heike Hofmann, Iowa State University.

Automatic comparison algorithms have grown in prevalence in a number of forensic disciplines following the reports from National Research Council (2009) and PCAST (2016). Many of these algorithms objectively measure the similarity between evidence, such as two fired cartridge cases, based on markings left on their surface, such as impressions left by a firearm's breech face during the firing process. We introduce the Automatic Cartridge Evidence Scoring (ACES) algorithm to compare pairs of three-dimensional topographical surface scans of breech face impressions. The ACES algorithm pre-processes the scans, extracts numeric features, and returns a similarity score indicating whether two cartridge cases were fired from the same firearm. The numeric features are computed based on a cell-by-cell registration procedure, results from a density-based unsupervised clustering algorithm, and derived from visual diagnostic tools we developed to investigate the performance of cartridge case comparison algorithms. We use scans taken at the Roy J Carver High Resolution Microscopy Facility of cartridge cases collected by Baldwin et al. (2014) to train and test the ACES algorithm. The performance of ACES compares favorably to several other methods, such as random forests on smaller feature sets, logistic regressions, decision trees, and some variants of previous Congruent Matching Cells methods (Song 2013; Zhang et al. 2021). The ACES algorithm is implemented in a free, open-source interactive web application called cartridgeInvestigatR.

Keywords: impression evidence, firearms and toolmarks, cartridge cases, pattern recognition,comparison algorithms

References:

Baldwin, David P, Stanley J Bajic, Max Morris, and Daniel Zamzow. 2014. "A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons." FortBelvoir, VA: Ames Lab IA, Performing; Defense Technical Information Center. https://doi.org/10.21236/ADA611807.