ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Healthcare serial killer or coincidence?

Statistical issues in investigation of suspected medical misconduct

Summary Report

by the RSS Statistics and the Law Section

September 2022

1. Introduction

Justice systems are sometimes called upon to evaluate cases in which healthcare professionals are suspected of killing their patients illegally. These cases are difficult to evaluate because they involve at least two levels of uncertainty. Commonly in a murder case it is clear that a homicide has occurred, and investigators must resolve uncertainty about who is responsible. In the cases we examine here there is also uncertainty about whether homicide has occurred. Investigators need to consider whether the deaths that prompted the investigation could plausibly have occurred for reasons other than homicide, in addition to considering whether, if homicide was indeed the cause, the person under suspicion is responsible.

The RSS's report, which this summarises, provides advice and guidance on the investigation and evaluation of such cases. Our report was prompted by concerns about the statistical challenges such cases pose for the legal system. The cases often turn, in part, on statistical evidence that is difficult for lay people and even legal professionals to evaluate. Furthermore, the statistical evidence may be distorted by biases, hidden or apparent, in the investigative process that render it misleading. In providing advice on how to conduct investigations in such cases, this report particularly focuses on minimising the kinds of biases that could distort statistical evidence arising from the investigation. This report also provides guidance on how to recognise and take account of such biases when evaluating statistical evidence and more broadly on how to understand the strengths and limitations of such evidence and give it proper weight.

This document is intended to summarise the issues covered in the full report and present our recommendations. The full report is designed specifically to help all professionals involved in investigating such cases and those who evaluate such cases in the legal system, including expert witnesses. It will also be of interest to

scholars and legal professionals who are interested in the role of statistics in evidentiary proof, and more generally to anyone interested in improving criminal investigations. With such a wide range of audiences, it is inevitable that for some readers certain sections may seem more relevant, and some less so, but we believe it is important not to aim particular sections at particular kinds of reader. We want, for example, the barrister to see what advice we give to the expert statistical witness – and we hope understand it, at least in broad terms – and vice versa; we believe that is important in helping all parties to appreciate the contributions of others in reaching just outcomes.

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 are serious statistical challenges in distinguishing event clusters that arise from criminal acts from those that arise coincidentally from other causes. Our analysis shows that seemingly improbable patterns of events (eg apparent clusters, rising trends, etc) can often arise without criminal behaviour and may therefore have less probative value than people assume for distinguishing criminality from coincidence. Full details are in **Section 2** of the full report.

When a medical professional faces criminal charges for killing patients, competing theories are often advanced by the prosecution and defence. The prosecution's theory is typically that a medical professional, previously trusted to perform critical life-saving functions, has unexpectedly (and sometimes inexplicably), chosen to murder patients in his or her care. While history has shown that humans are capable of such behaviour, and there have indeed been cases in which, for example, physicians have murdered multiple patients, nevertheless proven instances are thankfully extraordinarily rare – a mere handful of documented cases, perhaps a dozen or so per year, among the many millions of healthcare professionals worldwide. So the prosecution's theory in such cases is often one

that appears, *a priori*, to be improbable. Alternative theories – i.e., that some unknown factors, or mere chance, caused deaths to occur in apparently extraordinary numbers among patients under the care of a particular professional – often also appear improbable. So the assessment of the case invariably turns, at least in part, on a weighing or balancing of the probabilities of seemingly extraordinary events. Such assessments are challenging under the best of circumstances but become especially difficult when the evidence adduced to distinguish between the competing theories may be biased or presented in a misleading manner. We set out these challenges in detail in **Section 3** of the full report.

We discuss the kinds of investigative biases that can arise in these cases (in **Section 4** of the full report). We focus on ways that investigators' desires and expectations may ***unintentionally and even unconsciously*** influence what they look for, how they characterise and classify what they find, what they deem to be relevant and irrelevant, and what they choose to disclose. Examiner bias is a well-known phenomenon in both scientific and forensic investigations. It arises in large part from what are known as observer effects, a tendency for human beings to look for data confirming their expectations (confirmation bias) and to interpret data in ways that are subtly (and often unconsciously) influenced by their expectations and desires. Statisticians have long studied the ways in which examiner bias can distort statistical evidence emerging from scientific and forensic investigations. We apply insights from this scientific literature to an analysis of the investigative process in the types of cases discussed in the report. We also draw examples from investigations of actual cases that illustrate what we believe to have been biased investigative processes and discuss how such biases can generate misleading statistical findings. It bears repeating that our focus in this section is on processes that can *unintentionally and unconsciously* influence the investigative process. We are not questioning the general honesty, integrity or good intentions of those involved in

investigating such cases. We focus instead on investigative procedures that can distort statistical findings in ways that, while entirely unintentional, may nevertheless be important.

Section 5 of the report provides advice on how to improve investigative procedures in order to minimise investigative biases. While it is impossible to eliminate all human biases from a criminal investigation, there are a number of procedures that can reduce bias and thereby improve the quality and objectivity of the evidence emerging from the types of investigations we discuss here. We focus particularly on the advantages of blinding and masking procedures, which involve temporarily withholding potentially biasing facts from some of those involved in the investigation. We go on to discuss ways to reduce “tunnel vision” in which the investigation becomes a search for evidence confirming a particular investigative theory while ignoring or dismissing evidence inconsistent with that theory. We provide and explain advice on appropriate correct analyses of data, and discuss two worked examples.

We provide advice on evidence evaluation and fact-finding in these cases in **Section 6** of the report. We expect our report to be relevant and useful anywhere such cases may arise; hence we do not limit our discussion to the needs of a particular legal system, and expect our advice to be useful both in inquisitorial and adversarial legal processes. We believe the statistical issues in these cases pose challenges to legal fact-finders in every jurisdiction, whether they are professional judges or lay jurors, and are challenging for lawyers as well. Our advice focuses on identifying and appreciating ways in which statistical evidence may be misleading, and assuring (to the extent possible) that presentations of evidence are balanced in order to help triers-of-fact appreciate both the strengths and limitations of the evidence, and give it only the weight it deserves. We provide examples of presentations and arguments that we consider to be misleading or inappropriate. We will discuss cautionary instructions that may be helpful to lay fact-finders. Ultimately, we hope our

comments will help lawyers and judges, and statisticians and other experts, refine their presentation and evaluation of evidence in these difficult cases in order to better serve the interests of justice.

2. Case studies

The ideas, analysis and recommendations in the report are illustrated by three very different real case studies, summarised below.

In the Case of Jane Bolding, there was extremely compelling statistical evidence associating this US nurse's duty periods with times when unexpected cardiac arrests occurred in her ward. After an initial confession was retracted, the court decided that the statistical evidence alone was not enough for a conviction, and she was acquitted.

In the Case of Harold Shipman, anecdotal observations about apparently unusually many deaths among his patients followed by a police investigation found (non-statistical) evidence incriminating this English family doctor, and he was convicted on many counts of murder. A subsequent inquiry suggested that statistical monitoring techniques might have raised the alarm earlier and saved many lives.

In the Case of Lucia de Berk, this Dutch nurse was initially convicted of 4 murders and 3 attempted murders of children under her care, based largely on a statistical analysis, supported by anecdotal observations of character and behaviour. Subsequently serious flaws in the statistical evidence were exposed, and new medical evidence became available, together casting great doubt on the convictions, and she was exonerated.

Jane Bolding

Statistical evidence often plays a prominent role in the investigation of suspected healthcare serial killers. In 1988, Jane Bolding, an American nurse who worked in the intensive care unit of Prince George's Medical Center in Maryland, was prosecuted for serial murder of patients, allegedly by administering lethal doses of potassium chloride. The key evidence against Bolding was the high incidence of cardiac arrest during periods when Bolding was on duty. Evidence suggested that she had been the primary nurse on duty when 57 heart attacks occurred, while the number during comparable periods when other nurses were on duty had never exceeded five. An analysis by epidemiologists from the U.S. Centers for Disease Control (CDC) concluded that Bolding's patients were 47.5 times more likely to experience cardiac arrest than those of other nurses and that "an epidemic" of cardiac arrests ceased when Bolding left the hospital unit where it occurred (Sacks et al., 1988; CDC, 1985). Sacks testified at Bolding's trial that "[t]he chances of [this large number of cardiac arrests] happening by chance is about one in 100 trillion."

Other than the statistical evidence, the key evidence against Bolding was an alleged confession. During a 23-hour interrogation, Bolding reportedly confessed to killing two patients and agreed to write letters of apology to their families. She later retracted this confession, however, and it was excluded from the trial after a judge found that it had been obtained through illegally coercive methods that violated Bolding's constitutional rights. Consequently, prosecutors had little to rely upon during the trial other than the statistical evidence. No one testified to seeing Bolding inject any patients with potassium chloride, and although post-mortem examinations showed that the patients had higher than normal potassium levels, it was impossible to determine whether potassium chloride poisoning was the cause of the deaths. Defence lawyers offered alternative theories for the elevated rate of deaths during periods when Bolding was present.

A judge, who decided the case without a jury, found the prosecution's statistical evidence insufficient to warrant a conviction, saying "the state at most has placed [Bolding] at the scene of the offenses…but that is insufficient to sustain a conviction." (Washington Post, June 21, 1988). According to the judge, the statistical evidence "failed to supply the missing link that would connect the defendant with the alleged criminal act," and consequently "the state's reach hopelessly exceeded its grasp" (AP News, June 20, 1988).

Harold Shipman

There are documented cases in which medical professionals have intentionally engaged in misconduct that put their patients at risk. A well-known example is that of Harold Fredrick Shipman, an English physician in general practice. In 2000, Shipman was found guilty of the murder of 15 patients under his care. Investigators suspected he was responsible for the deaths of many others, perhaps as many as 250, making him one of the most prolific serial killers in modern history.

Concerns about Shipman were first raised by other medical practitioners, who noted what appeared to be an unusually high rate of deaths among Shipman's patients. An initial police investigation in 1998 found insufficient evidence to bring charges, but police subsequently learned that the wills of some of Shipman's former patients had been altered under suspicious circumstances to leave assets to Shipman, rather than family members of the deceased. Further investigation found evidence that Shipman had administered lethal doses of sedatives to healthy patients, and had then altered medical records to indicate falsely the patients had been in poor health. Based on this evidence Shipman was prosecuted and convicted.

In light of this grim episode, there were calls for improved monitoring of adverse medical outcomes, to allow dangerous medical misconduct to be detected earlier. For example, statistician David Spiegelhalter and colleagues suggested that statistical monitoring of patient death rates would have raised red flags about Shipman's misconduct years earlier, thereby saving lives (see Spiegelhalter, D. et al., 2003).

Lucia de Berk

In 2003, Lucia de Berk, a Dutch paediatric nurse, was convicted of four murders and three attempted murders of children under her care. In 2004, after an appeal, she was convicted of seven murders and three attempted murders. Thereafter, several academic commentators questioned the quality of the evidence used to support the conviction, particularly statistical testimony.

De Berk had been under suspicion in her hospital for some months as a result of gossip about her tough, disturbed childhood and striking personality. When a child in her care died suddenly, the death was immediately announced to be completely unexpected and, by implication, suspicious. Hospital officials identified eight further deaths or resuscitations that had occurred while she had been on duty as medically suspicious. Additional suspicious deaths were identified and linked to de Berk at two other hospitals where she had worked. For two of the patients, investigators found toxicological evidence supporting the claim that de Berk had poisoned them, although the probative value of this evidence was weak. Statements in de Berk's diary about "a very great secret" and a "compulsion" on a day that a patient had died were given a sinister interpretation.

During her original trial, a criminologist (who had years earlier graduated in mathematics) presented statistical evidence according to which the probability of so many deaths occurring while de Berk was on duty was only 1 in 342 million. This number was the product of three p-values, one for each hospital. Prominent statisticians came forward to argue that the incriminating statistic was based on an over-simplified and unrealistic model, biased data collection, and a serious methodological error in combining p-values from independent statistical tests. The probability of so many deaths occurring by chance may have been as high as one in 25.

In light of these doubts, and further medical evidence that came to light in post-conviction investigations, the case was re-tried in 2010 and de Berk was acquitted. The original convictions are widely viewed as miscarriages of justice that were prompted, in part, by an inadequate investigation and misuse of statistical evidence. They led to various reforms in the Dutch legal system.

3. Recommendations

We make a series of recommendations to help manage the difficulties that arise in cases of alleged medical misconduct involving statistical evidence. Because the statistical aspects of these cases are often nontrivial, fraught with difficulties, challenging to laypeople (jurors, media reporters, the public) and to lawyers, and indeed are not entirely straightforward to the specialists:

- ***Recommendation 1****:* It is therefore important that all parties involved in investigation and prosecution in such cases consult with professional statisticians, and use only such appropriately qualified individuals as expert witnesses. [Section 5(c)][1]

There are two kinds of error in drawing inferences about effects from data: inferring an effect that is not real, or missing one that is. Both have grave effects in the judicial setting. It has been argued that if one decreases the error rate of one of the two kinds, the error rate of the other kind will go up; thus any change in practice shifts the balance between prosecutor and defence, shifting the errors from Type 1 to Type 2 or vice versa. That is only the case if nothing is changed in statistical methodology, apart from merely shifting a decision threshold. But one can reduce both error rates by increasing the amount of information extracted from the already available data, using superior statistical methods, and of course by acquiring more and different kinds of data.

- ***Recommendation 2****:* In presenting the results of statistical tests, both the level of statistical significance (*p*-value) and the estimated effect size should be stated. One addresses

[1] The references in square brackets in this section refer to the full report.

the question of whether an effect is truly *detected*, the other quantifies the *size* of that effect, if it exists. These are different concepts and both are important; neither should be confused with subjective judgements about the credibility of the expert witness. [Section 4(c), Section 5, and Appendix 2]

Special care is needed to assure that *p*-values, when presented in reports and testimony, are understood and used properly. While *p*-values are an important statistical and scientific tool, they are difficult for people to understand and are frequently misinterpreted. They may, for example, be misunderstood as statements about the probability that a coincidence occurred, rather that the probability of observing a given number of deaths (or more) by chance, and this kind of misinterpretation can be extremely unfair to individuals suspected of misconduct.

- ***Recommendation 3****:* In reports and testimony, experts should take care to explain the proper interpretation of p-values and should avoid drawing fallacious inferences from them. In jurisdictions that rely on lay jurors, judges should consider providing instructions about the proper use of *p*-values. Lawyers, judges and investigators should educate themselves to the dangers of fallacious statistical interpretation. Lawyers should endeavour to present the case in a manner conducive to correct understanding, avoiding to the extent possible testimony or arguments conducive to misinterpretations.

It is important to take a broad and informed view of all the circumstances in which a cluster of adverse outcomes is observed, to ensure that all potential causal factors are identified, and the problem that those best-informed may be implicated in alternative explanations for the data, with a consequent risk of bias. We therefore advocate that

- ***Recommendation 4****:* Investigations should be guided by panels representing all relevant areas of expertise but independent of both the suspect and the employing institution. [Section 5(a)]

Statistical investigations of the kind we discuss are not controlled experiments, but observational studies directed by humans, with all the inherent unconscious biases pervading all human reasoning. It is impossible to eliminate completely the role of human judgement in organising and conducting statistical data acquisition and analysis, but:

- ***Recommendation 5****:* To the maximum extent practicable, experts informing an investigation, such as DNA specialists, fingerprint examiners, toxicologists, and pathologists should be kept "blind" to all aspects of the case irrelevant to the question they are being asked to answer. Blinding is a key tool in minimising prejudicial subjective effects such as unconscious bias. [Section 5(b)]

Guidelines of this nature for evidence of other kinds already exist in some jurisdictions. For example, organisations that issue practice guidelines for matters such as DNA evidence include SWGDAM (USA), FSR (England and Wales), ENFSI (Europe), and the International society of Forensic Genetics (ISFG; International).[2] Our recommendation for blinding is more comprehensive than what is currently required in most jurisdictions.

[2] In England and Wales, the Forensic Science Regulator has issued Codes of Practice and Conduct for Forensic Science Providers and Practitioners, with recent proposed updates; see Forensic Science Regulator (2021b, 2022).

A second universal consequence of basing decisions about causes of effects on observational studies is captured by the well-known aphorism “correlation is not causation”.

- ***Recommendation 6****:* It is vital that investigators appreciate the truth of this, and the fact that the connection between them is well-studied, and that in fields such as medical diagnosis there are accepted criteria to guide the valid drawing of conclusions in observational studies [Section 6 and Appendix 5]. Possible confounding factors must be identified, and their effect quantified, before attributing causes to observed effects. [Sections 2, 4(a,c)]

Our work is designed to promote stronger, more scientifically rigorous investigations of alleged medical misconduct. While that is the ideal, courts may still occasionally be called upon to evaluate evidence generated by poorly conducted investigations that produce problematic results. In jurisdictions that rely on lay juries as triers of fact, judges should consider whether the results of such an investigation are sufficiently reliable and trustworthy to meet legal standards for admissibility.

- ***Recommendation 7****:* When courts must evaluate the results of problematic investigations, it is particularly important that they consider reports and expert testimony from independent statisticians. If investigative bias is a significant concern, lawyers and courts should also consider seeking evaluations from experts of cognitive bias and factors associated with the accuracy of expert judgment.

Understandably, most participants in the legal world have little training in matters of statistics and the scientific evaluation of uncertainty. In some countries, organisations in parts of the legal community ensure that training is available to those who would like it on probabilistic reasoning, statistical modelling, and statistical

inference. In our opinion, defence lawyers first of all need to know that there is a whole scientific field out there which can help them serve their clients better. They need to be able to learn about the possibilities and to know how to find the professional community which can help them. Similarly, prosecution lawyers will need to learn about these matters – and if they do not, can expect cases built around inadequate statistical analysis to be successfully challenged by defence lawyers with aid of expert testimony. Judges too will need to be sufficiently informed to be able to determine admissibility and guide jurors accordingly. Not every legal professional needs to know everything: obviously, they cannot. However, within the different parts of the legal community, there do need to be people who do understand enough to know when professional support and further education is necessary. Our final, strong recommendation is therefore that

- ***Recommendation 8****:* Further interaction between legal and statistical communities should be fostered by the leaders of the legal and statistical communities, with a view to promoting joint educational activities.

From past to present...

The image of the wheatsheaf first appeared in our original seal. Being the end product of the harvesting and bundling of wheat, it was a pictorial way of expressing the gathering and analysis of data: the foundations of statistical work.

It also implied that statistical practice comprises more than the collection of data: it consists of active interpretation and application as well (threshed for others, if the rural analogy is sustained). Rigorous data gathering is still at the heart of modern statistics, but as statisticians we also interpret, explain and present the data we collect.