
Commentary on Meester et al. “On the (ab)use of statistics in the legal case against nurse Lucia de B.”

David Lucy*

What has been dubbed the “nurse/roster” problem has yet again reared its ugly head. Not this time in the United Kingdom, nor the United States of America, both nations being ones in which there have been many instances of wayward members of medical staff and carers being accused of, and in many cases convicted of, harming those in their care; but this time in the Netherlands. Again, as in recent cases in the United Kingdom involving instances where parents are accused of harming their children, and where some component of the evidence revolves around the number of instances of adverse outcome, there has been a widespread public interest in the value of these type of observations and their interpretation as evidence.

Despite the relatively high public profile of this sort of problem, there is a dearth of available literature upon it. So any paper which concerns the “nurse/roster” problem, particularly from a group with recent exposure to it, is a welcome addition to what little has been written.

Meester et al. devote a portion of their paper to a detailed discussion of a recent analysis by Elffers. I shall not discuss the specific case as I know little of it, and little of Elffers analysis except what has been described by Meester et al.. However, they (Meester et al.)

*Department of Mathematics & Statistics, Lancaster University, Lancaster, UK, LA1 4YF.
go on to discuss various approaches which have been used in the past to the statistical evaluation of these types of evidence.

To some extent Meester et al.’s discussion is, understandably, coloured by their background in a European inquisitorial jurisdiction. They discuss the notion of hypothesis testing type statistics, and the calculation of \( p \)-values in relation to legal evidence. This may be a perfectly legitimate analysis in the context of the criminal justice system of the Netherlands, but in the United Kingdom would be very much discouraged as it carries the implication of the “failure to reject the null hypothesis”, or the converse. The null hypothesis relates to what a United Kingdom court might think of as the “ultimate issue”, that is the question of guilt, or innocence of any individual, and that is solely the province of other agents in both the criminal justice systems of the United Kingdom. In the United Kingdom it the role of the expert witness to comment upon the evidence, not upon whether an individual is guilty or not guilty. As such, not only is there no single agreed optimum method for dealing with these types of observations, but it is also unlikely that were any optimal agreed method devised, that that method could be applied across legal jurisdictions, as each jurisdiction has its own unique set of characteristics and rules of evidence which necessitate a similarly unique set of analyses by statisticians working within that framework.

I shall say little about the actual methods of evidence evaluation themselves. Meester et al. make a fine job of analysing the assumptions, implications, strengths and drawbacks of Elffers \( p \)-value approach, the Bayesian approach of de Vos, and “epidemiological” approaches. However, I should like to clarify the following points:

1. The paper by Lucy and Aitken, referred to by Meester et al. as publication [3] was not published. It was withdrawn by the authors from publication precisely because of major weaknesses in the argument.

2. One of those weaknesses, mentioned by Meester et al., was the treatment of \( \mu \).
This weakness could be addressed by regarding $\mu$ as an unknown quantity in a fuller Bayesian treatment.

3. Lucy and Aitken did use the relative risk but this is not a novel use. It had been used in the United States of America by investigators at the Centres for Disease Control for exactly this purpose, that is where health workers and carers were suspected of harming patients, and it was used by Lucy and Aitken as a basis for comparison with the likelihood ratio approach.”

Meester et al. do not restrict their arguments to the realm of the statistical sciences. Rather, they rightly emphasise the philosophical and epistemological aspects of evidence evaluation. An important area of discussion is the in their section entitled “using data twice: the post hoc correction”. Here I believe Meester et al. begin to identify and explore the area of real weakness underlying each and every statistical approach to the nurse/roster problem, and is the matter to which the remainder of this commentary will address.

“Using data twice”, or “double counting” of observations is where the observation of $n$ adverse outcomes associated with a particular individual, or individuals, is used as evidence that the $n$ adverse outcomes are the result of criminal activity, and additionally, that the criminal activity has been conducted by the specified individual, or individuals. The only observation is of $n$ adverse outcomes, no other observations being available.

These misgivings of those who are wary of “double counting” are understandable. Two sequential propositions are considered, first, that a criminal offence has been committed, and secondly, that a specific individual is responsible for that criminal offence. There is one, and only one, piece of information being available to bear on these two propositions.

The archetypal forensic problem is a “matching” problem. The simple forensic matching problem arises when, prior to any observations being made, it is known that some crim-
inal offence has been committed. Investigators look for evidence, and some evidence is found which defines some shared value of the same property between the offender and a potential suspect. A suspect is found from some other source, and the suspect is found to share the value of the linking property. Some measure, the current paradigm is a likelihood ratio, can be defined which can lead to a strength of belief that the suspect is truly the offender. Here, in this archetypal problem, the observation of the shared property is used solely to increase, or decrease, the belief that the suspect is the offender. An example might be where some specimen of DNA has been unequally attributed to the execution of a criminal offence, a suspect identified through some independent means, and the DNA profile of the suspect is an exact match for that DNA profile associated with the criminal offence. Provided suitable estimates can be made for the distribution of the DNA profile in question a statistically based solution for this problem is unproblematic.

The nurse/roster problem shares few of the characteristics of the “matching” archetype. The only observation allowed is that observation which enumerates the adverse outcomes, and some form of list of individuals within the same spatial location, both coeval with the adverse outcomes, and at other times. There is no prior knowledge that a criminal offence has taken place, and the only linking property is that of spatio-temporal coincidence between any individual and an adverse outcome. It is not a case of the observations simply having to serve to inform the investigator as to which individual the evidence points, but also whether the evidence should be pointing to any specific individual at all. Consequently this evidence type shares many characteristics with forensic evidence types such as drug residues upon bank notes, being used as evidence against those who are allegedly involved in the supply of illicit substances. This is probative because any quantities of compounds associated with illicit substances detected upon the banknotes under examination are presumed to have come from, and indeed given the specificity of the detected compound must have come from, some direct or indirect association with the supply of illicit substances. However, the link with any putative offence is not direct in the same
way that a shoe print, or any other matching type evidence might be. The argument is made through some sort of “causal” necessity. In the case of drugs on banknotes the causal necessity is reasonably strong in that the trace compounds must have come from somewhere, and that somewhere must be connected with an offence at some level. However, with some other types of possible evidence, such as the social profiles of possible offenders, were they to be used as evidence, the argument is one of: the offender has a particular social profile, the suspect has the same very particular social profile, hence the possession of that particular social profile indicates to some extent that the suspect was the offender. The nurse/roster problem has in some ways an even more tenuous claim to probative value than that of social profiling. In cases where roster data might be employed as evidence there is some uncertainty as to whether an offence has even been undertaken. Consequently the data taken from the roster, or whatever record of presence or absence exists for the particular institution in question, has to be used to support the proposition of an offence being committed, and a proposition as to whom the offender might be.

One’s intuition suggests that a single piece of evidence should not be used to support more than a single proposition, however, the notion that evidence can be somehow “used up”, or depleted, by it’s use is not logically justified. For instance, were a specific individual to be observed emerging from a residence carrying a felling axe covered in what appears to be blood, one could argue reasonably that the observation is evidence that a violent assault had taken place, and that it points to a main suspect for that violent assault. Here it is debatable as to how many observations one is making, but on the face of it it is a single piece of evidence, and here is being used without problem to support two propositions: first, that a violent assault had taken place, and secondly, that the individual seen carrying the axe is indicated as having perpetrated the assault. One would be considered foolish to observe the individual emerging from the residence carrying a felling axe, conclude a violent assault had taken place, then say that one could have no knowledge of who the perpetrator might be. Conversely, it would be equally foolish to
observe the individual with the blood covered felling axe, but say that one could have no knowledge that a violent assault had taken place because that same observation had already been used to suggest who had participated in the violent assault.

The instance above where a single observation has been employed to inform several propositions may be an extreme example, but it is illustrative of the fact that there is no necessary principle which suggests “one observation, one proposition”. There are evidence types where a single observation has the dual role of identifying a suspect, and forming evidence against that suspect. An example, albeit a controversial example, is the use of a DNA profile located in a database of known individuals. Here a DNA profile might have been used to pick out a suspect, and the same profile may also be employed as evidence that the suspect was the offender. For many working in the field of statistical evidence evaluation this seeming dual use of a DNA profile is without any particular epistemological difficulty, although some have indicated that there may be statistical issues arising from this process.

There is no particular general philosophical problem with a single observation shedding light on two propositions. However, Meester et al. point to the fact that there may be problems with the particular pair of propositions concerning whether a criminal offence has been committed, and, which individual committed the offence. A discussion, and solution, of this epistemological problem may hold the key to the statistical evaluation of many of these evidence types; and it is only a statistical evaluation which holds any realistic potential for making sense of this type of evidence.

It is difficult not to have some sympathy for Elffers who has tried to find some sort of rational evaluation of the observed evidence. The task, as it is currently formulated, is nearly impossible. Not just because the necessary statistical methods have not been agreed upon, nor even because the required philosophical framework does not yet exist to allow a coherent statistical analysis of this type of evidence, but because whatever
statisticians evaluating evidence have to say is often misunderstood, or misused by the court. Apparently, according to Meester et al., the court in the first instance transposed the conditional from Elffers work from a likelihood form, to a probability form, and despite the having at least two more statisticians present, so did the court of appeal. If courts are going to insist upon transposing conditional probabilities then there is little which any epistemological basis or statistical framework, no matter how sophisticated, can do to effect a reasonable and accurate evaluation of the evidence.

This paper serves to remind us just how far we have to go, and how much work there is to do, before we as a statistical and scientific community can approach this type of problem in a consistent and agreed way. The problems are not so much statistical, but philosophical and epistemological. We need to be certain about just what it is we are expecting our evidence to do, and how much any particular set of observations can be used as evidence for multiple propositions. Finally, there is unlikely to be any single optimal solution to these sorts of problems, because what may be considered optimal under one jurisdiction cannot be considered transferable across jurisdictions without careful thought.