in the period before, for a “relative risk ratio” of 6 (see Table 3). In other words, it now appears that suspicious deaths were six times more likely after the nurse joined the hospital staff than before. The *p*-value, indicating the probability of this difference occurring by chance is about 0.0068 (about 1 chance in 150), which sounds strongly incriminating for the nurse. But of course, the nurse is entirely innocent. The seemingly incriminating finding was generated by biased investigators who failed to take account of other factors that might have affected the death rate and interpreted the data in a manner that was inadvertently influenced by predictable human biases. An unbiased investigation would have shown a smaller (and therefore less incriminating) increase in deaths.

Table 3 Number of “Suspicious Deaths” Before and After Suspect Joined the Hospital Staff (as Reported by a Biased Investigation)

	Patients	Deaths	Deaths Deemed “Suspicious” in a Biased Investigation	Deaths Deemed “Suspicious” in an Unbiased Investigation
Before	1,000	100	5	10
After	1,000	200	30	20
Relative risk ratio			6	2
p-value[79]			0.0068	0.5876

Example 2

Another approach that investigators may take is to compare the number of “suspicious deaths” when the nurse was or was not on duty. Our second example illustrates how statistics produced by such comparisons can be distorted by (a) failure to take account of other causal factors that may correlate with the duty periods; and (b) investigative bias in determining which deaths are suspicious. It also allows the time periods over which the data are collected to be unequal in length.

Suppose that 16 patients die in circumstances assessed by investigators to be suspicious over a 15-day period on the ward in question, with ~~9~~ 11 of those deaths reported during the 7-hour morning shifts and the remaining ~~7~~ 5 during the afternoon and night shifts. The nurse under suspicion works 8 morning shifts, and 2 of the afternoon or night shifts. So the raw rate of suspicious deaths tends to be higher when the nurse is on duty than when not, simply by virtue of the nurse’s pattern of work. The first columns in Table 4 (under ‘Unbiased investigation’) tabulate these values. Compared to Example 1 it is now more difficult to interpret the data intuitively, but cross-classifying the deaths by shift and the nurse’s presence in this way suggests that the time of day is an important factor; an appropriate formal method of analysis is described in Section 5(c) below, and yields the *p*-values in the final line of the table. These show that allowing for the inherent differences between shifts transforms the strength of evidence against the nurse from statistically significant (*p*=0.~~017~~015) to very weak indeed (*p*=0.~~378~~301).

[79] *p*-values computed using (one-sided) Fisher’s exact test.

Table 4 Numbers of “suspicious deaths” when suspect was and was not on duty, under assumptions of both unbiased and biased investigations (see text)

		Deaths attributed to nurse on duty			
		Unbiased investigation		Biased investigation	
	Shifts	Ignoring morning effect	Allowing for morning effect	Ignoring morning effect	Allowing for morning effect
Nurse on duty, morning	8	~~10~~8	7	~~12~~10	8
Nurse on duty, other[80]	~~7~~2		~~3~~1		~~4~~2
Nurse off duty, morning	~~2~~7	~~6~~8	~~2~~4	~~4~~6	~~1~~3
Nurse off duty, other	28		4		3
p-value[81] for nurse effect		0.~~017~~015	0.~~378~~301	0.~~0007~~0006	0.~~031~~027

Finally, let us suppose that cognitive bias also influences the investigators’ assessments of whether each of the deaths was suspicious, in such a way that 2 additional deaths during the nurse’s shifts are now judged suspicious, one in the morning and one in the afternoon, while one fewer death was called suspicious in each of the counts where the nurse was not on duty. The final columns of Table 4 (under ‘biased investigation’) show these data, and the corresponding *p*-values, show that this small bias (which might also have been caused by simple mis-recording in duty records) is enough to make the evidence now statistically significant (*p*=0.~~031~~027) even when we assume there is no difference between morning and other shifts in mean rates of death, whilst if we do not allow for such differences the evidence is very highly significant (*p*=0.~~0007~~0006).

All of the analyses here assume that there are no other causal effects, such as seasonal factors or administrative changes, that need to be taken into account.

In Appendix 6, we describe analyses of the data in these two examples, and explain the logic and the calculations that lead to the *p*-values quoted above.

[80] i.e. afternoon, evening or night shift

[81] Using likelihood ratio test for equality of rates, with and without adjustment for morning effect. This is the chi-squared test conventionally used in the analysis of deviance; see Appendix 6.

Appendix 6: Patterns of occurrence of adverse events

Here we give some annotated examples of correct analyses of illustrative data on patterns of occurrence of adverse events. To simplify exposition we will write about unexpected deaths of patients in a section of a hospital, and take the “explanation” under consideration to be deliberate harm caused by a nurse. Of course, this exposition applies *mutatis mutandis* to many other scenarios, and any professional role, etc.

We will assume that these events occur completely at random, but at a rate per unit of time (hour, shift, day, etc., as appropriate) that varies with time, and may be influenced by factors of the kind already discussed: seasonal and diurnal effects of disease, administrative changes, etc., and also, possibly, by wilful harm. We use the phrase “completely at random” in its proper mathematical sense, to mean that the occurrence of an event at a particular time has no direct influence on the time of any other event. (In other words, we consider only *exogenous* causes for the variation in rate of the adverse event, not *endogenous* ones). This rules out for example infections of a contagious disease, where there can be a direct causal link, but would cover heart attacks. The only other assumption we make is that when we consider two or more causal factors for the variation in rate, the effects of these are multiplicative: the percentage change in rate when one factor is present is the same whether or not other factors are also present.

As an artificially simple example, consider a hospital ward which is staffed either by nurse A or by nurse B. Numbers of deaths when each of the nurses is in charge are counted, and summarised here:

Table 5 Illustrative example: patient survival statistics under the care of two nurses

	Nurse A	Nurse B	Total
Died	15	9	24
Survived	25	31	56
Total	40	40	80

Could the apparent discrepancy in rates of death be attributed to chance, “just a coincidence”? We suppose that all circumstances of the Nurse A and Nurse B shifts are identical; there is no other conceivable reason for the apparent difference other than the presence of one nurse or the other.

This data structure is called a (2 by 2) contingency table: the standard way to analyse this, to test the hypothesis that there is no difference in the death rates attributable to the nurses, is “Pearson's chi-squared test”. This is an elementary technique, taught in the middle years of high school (eg GCSE level in England and Wales). This reveals that the probability of observing a difference in apparent death rates as large as, or larger than, that seen in the table*, if there was really no difference* is 14% (that is the *p*-value is 0.14). That means that if you were to repeatedly allocate 24 deaths and 56 survivals into two groups of 40 patients at random, a difference in apparent rates as large as that in Table 5 would be obtained about 1 time in 7. We have to conclude *there is no significant difference*. It would be misleading to the court to testify that there *was* a difference. A *p*-value less than 0.05 is a pre-requisite for publication in the scientific literature (and this is not a tough standard, very many scientific “findings” are never replicated by other scientists).

The *p*-value above is calculated as follows. Since 24 out of the 80 patients die, if there were no nurse differences, you would expect that (24/80) of 40, ie 12, of the deaths would occur on Nurse A’s shift. In the same way, for each of the other counts in the table (9, 25 and 31) you would expect respectively (12, 28 and 28). We denote the observed numbers (15, 9, 25, 31) by O_i and the expected numbers (12, 12, 28, 28) by E_i, then calculate

$$G = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i},$$

(that is, we take each of the cells of the table in turn and square the difference between the observed and expected numbers, and divide by the expected number; these fractions are added up over the four cells), which is 2.143. To convert this to the *p*-value quoted, we can use standard printed tables of the chi-squared distribution, or the function found on many calculators and all statistical software packages.

In contrast, if all of the numbers in Table 5 were exactly 10 times larger (150, 90, and so on), then the Pearson chi-squared statistic G would be 21.43 and the *p*-value turns out to be 0.000004, so there would be overwhelming evidence that the apparent different in death rate was *not* due to chance. (This is an example of the point made in Section 4(f) that "coincidental fluctuations from population means are more likely with small samples…").[82]

Contingency tables of any size, not just 2 by 2, can be analysed with Pearson's chi-squared test, but still very few criminal cases are simple enough to fit into this setting. Nevertheless, the analysis can be extended to deal with much more complex situations, allowing in particular more than one causal factor, and different durations of time. The more general framework is that of *Poisson log-linear models*, which are an example of *generalised linear models*. This is also a standard methodology, but one now taught not at high school but in undergraduate courses in mathematics and statistics. When applied to a 2-way contingency table, the results are the same.

These methods are provided in standard statistics packages, and will be part of the toolbox of all practicing professional statisticians. The assumptions underlying their use are simply those mentioned above, and courts should be able to accept results of such analyses in expert witness testimony, just as, for example, a scientist would expect to be able to present scientific evidence relying on data from electron microscopes or mass-spectrometers without needing to explain to judge and jury the physics needed to say how these complex machines function. In short, an expert witness, including a statistician, must be free to use adequate methodology for the task. In Appendix 8, we show computer code and output for the analyses in this section, using the well-regarded statistical system R, which is freely and universally available.

To illustrate appropriate methodology for analysing data on counts of deaths in different periods in the presence of other possible causal factors, consider the artificial example from Table 4 of Section 4. The deaths have been tabulated and summarised in the counts in four different categories of shifts. Note that these categories differ in various ways – they are of different durations; some are morning shifts, not all; and for some but not all the nurse in question is on duty. The rates of death vary between the extremes of 4 in 28 shifts and 2 in 2 shifts, a considerable difference, but can we attribute these differences in rate to the morning/other shift factor, or to the presence of the nurse, whilst allowing for the fact that among these small counts there will also be random variation?

[82] See also Appendix 2.

To correctly assess the extent to which the deaths can be attributed to the presence of the nurse we must compare two hypotheses:

> (1) that the *only* cause of systematic difference in rates is the shift effect, and
>
> (2) that *both* the shift effect and the presence of the nurse have a systematic effect on the rates.

This way of posing the question accords with both sound scientific practice, and the criminal law principle of *in dubio pro reo*[83]. It is incorrect, and prejudicial, simply to examine whether the presence of the nurse affects the rates whilst ignoring the other potential causal factor.

This point is illustrated in the analyses of the Table 4 "unbiased investigation" data summarised in the final rows of that Table and in Table 6. If we ignore the shift effect, we are simply comparing the rates of ~~10~~ 8 per 15 shifts with ~~6~~ 8 per 30 shifts when the nurse is or is not on duty. The Poisson log-linear analysis (details explained in Appendix 6, with code in Appendix 8) gives a *p*-value of 1.~~7~~5% (0.~~017~~015) for the likelihood ratio test that the nurse's presence has no effect on the rates – and we would conventionally call this result significant, which would be incriminating. However, if we follow the correct practice of comparing hypotheses (1) and (2) above, the *p*-value becomes ~~37.8~~30.1% (0.~~378~~301). Because this higher p-value is not statistically significant, it provides no basis for rejecting hypothesis (1) and therefore cannot be incriminating. Table 6 also gives the expected numbers of deaths in each category of shifts, the maximum likelihood estimates according to the statistical models being fitted in the two approaches. It is easily verified by inspection that the values in the case of the correct analysis shown in the 6th column fit the observed data (4th column) much better than do the expected numbers under the incorrect analysis (5th column).

Table 6 Continuing the example in Table 4

Number of shifts	Nurse	Shift	Deaths	Expected deaths ignoring morning effect	Expected deaths allowing morning effect
8	on duty	morning	7	~~5.33~~6.4	7.~~87~~43
~~7~~2	on duty	other	~~3~~1	~~4.67~~1.6	~~2.13~~0.57
~~2~~7	off duty	morning	~~2~~4	~~0.40~~1.6	~~1.13~~3.57
28	off duty	other	4	~~5.60~~6.4	4.~~87~~43

[83] The principle that a defendant may not be convicted by a court when doubts about his or her guilt remain.

Appendix 7: Usual practice in medical statistics and epidemiology.

Two types of clusters of events are routinely investigated in medical statistics and epidemiology: outbreaks of food poisoning and clusters of severe adverse events or usually rare diseases. The difficulty of identifying possible causal factors is multi-faceted. The methods developed in medicine to bring together evidence from laboratory science, observational and experimental studies are important tools for investigation.[84]

The standard first approach is to design and conduct a case-control study, a rapid and relatively inexpensive method. For each precisely defined incident, one or more control incidents are found, and the antecedents investigated. The design stage establishes clear definitions of events, and consistent approaches to seeking evidence for cases and controls. The varieties of biases which can arise are well-studied, and methods to minimise the risks of such biases established. It is standard for those recording data on possible explanatory variables to 'be blind' to which people are cases or controls. For deaths, experts in the quality and coding of death certificates might provide a necessary complement to physicians or pathologists.

In the study of causes of disease, nine aspects of association, the Bradford Hill guidelines, are considered.[85] It would be sensible to consider these in other investigations of causes, as is happening in areas of civil litigation.[86] Detailed consideration of uncertainty is preferable to false confidence in a single explanation.

Comparisons of different authorities' methods of investigating clusters of events might well lead to mutual benefit.[87] The methods in Public Health England guidelines for investigating non-infectious disease clusters possibly due to environmental exposures are relevant to clusters of death.[88] As well as suggested membership, with roles and responsibilities, of an investigation team, the guidelines include a substantial list of useful data sources.

An example of an efficient investigation is that of a cluster of serious events in children with cystic fibrosis.

a) In 1993, doctors at Alder Hey Children's Hospital (AHCH), Liverpool, noticed that five children with cystic fibrosis (a condition in which the lungs and digestive system are clogged with thick sticky mucus) who needed surgery because of fibrosing colonopathy (obstruction of the intestine) presented between July and September, 1993. One response to this might have been to suggest that doctors at AHCH were failing in some way.

b) On 8 January 1994, a short report was published, which reported that "The only consistent change in management had occurred 12-15 months preciously when all five had switched to" high-strength pancreatic enzymes (high dose drugs). [89]

[84] ICCA & RSS, 2019, Statistics and probability for advocates, p.18.
[85] Hill, AB, 1965.
[86] ICCA & RSS guide, 2019, p.19.
[87] Stewart, Ghebrehewet & Jarvis (2016).
[88] Public Health England (2019).
[89] Smyth, et al, 1994.

c) At the time the report was published, a case-control study to investigate the findings had been started: this is the appropriate method of reacting to the reports of new adverse events among patients. The Medicines Control Agency had been informed of the cases, and had issued appropriate warnings. There were about 7600 people known to have cystic fibrosis in the UK; 5/7600 is 0.07%.

d) On 11 November 1995, about 2 years later, the results of the case-control study were published.[90]

e) The study had 14 cases of fibrosing colonopathy, with each case matched to four controls. Data on these 70 patients showed a significant (at 5%) odds ratio of 1.45 per extra 1,000 high-strength capsules, and indicated which two particular proprietary formulations were associated with the highest odds ratios. That is, this association between particular formulations and fibrosing colonopathy could have arisen by chance one time in twenty.

f) Laxative use was also found to be associated with fibrosing colonopathy; odds ratio 2.42 (95% Conf. Int 1.20-4.94). From a case-control study, one cannot establish whether laxative use was a cause of fibrosing colonopathy, or a symptom of it.

g) Six of the 14 cases received care at AHCH. Care at Liverpool was associated with approximately a two-fold increase in risk of fibrosing colonopathy. If taken alone, this risk is statistically significant at the 4 percent level (p=0.04%), but adjusting for high-dose drugs removes the significance (p=0.3 or p=0.8).

h) In deciding whether to suggest that AHCH doctors were negligent, or actively harming children with cystic fibrosis, one must consider the competing explanations for fibrosing colonopathy.

[90] Smyth, et al, 1995.

Appendix 8: Annotated code and output

This appendix may be of limited interest to readers who are not statisticians, but is included for two main reasons. One, in the interests of full disclosure, is to verify the results in the illustrative numerical examples in Sections 4(f) and 5(c), and the calculations in Appendix 2, and make them completely reproducible. The other is that the codes may serve as templates for investigators and expert witnesses undertaking similar analyses in future, and a starting point for the more elaborate analyses that will be necessary in many real cases. These might entail additional explanatory factors, interactions between them, and possibly additional variables modelled as random effects. The last-mentioned here will necessitate use of generalised linear mixed models, available in R by using the lme4 package, for example.

Analysis in Table 3

Input

Create a 2 by 2 matrix containing the data for the biased investigation: deaths and survivals for the before and after cases, and display the data.

```
biased<-matrix(c(5,95,30,170),2,2)
biased
```

Conduct Fisher’s test for equality of the odds ratios before and after, against the alternative that the odds on survival is less.

```
fisher.test(biased,alternative='less')
```

Repeat for the unbiased investigation

```
unbiased<-matrix(c(10,20,90,180),2,2)
unbiased
fisher.test(unbiased,alternative='less')
```

Output

Biased case

```
> biased
     [,1] [,2]
[1,]    5   30
[2,]   95  170

> fisher.test(biased,alternative='less')

        Fisher's Exact Test for Count Data

data:  biased
p-value = 0.006818
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.0000000 0.7151649
sample estimates:
odds ratio
```

```
 0.2992371
```

Unbiased case

```
> unbiased<-matrix(c(10,20,90,180),2,2)
> unbiased
     [,1] [,2]
[1,]   10   90
[2,]   20  180

> fisher.test(unbiased,alternative='less')

        Fisher's Exact Test for Count Data

data:  unbiased
p-value = 0.5876
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.00000 2.08157
sample estimates:
odds ratio
         1
```

Analysis in Tables 4 and 6

Input

Create data frame containing the response variables `deaths`, two explanatory factors `nurse` and `morning`, and the variable `shifts`, the number of shifts for that row of the table, used on a log-scale as an offset since we are modelling rates of deaths per unit time.

```
shifts<-c(8,~~7,2~~2,7,28)
nurse<-as.factor(c('yes','yes','no','no'))
morning<-as.factor(c('yes','no','yes','no'))
deaths<-c(7,~~3,2~~1,4,4)
data<-data.frame(shifts,morning,nurse,deaths)
print(data)
```

Fit Poisson log-linear models for rates of death, both with just `nurse` included as an explanatory variable, and with `morning` also included. Print analysis of deviance table and fitted values in each case

```
fitN<-glm(deaths~nurse+offset(log(shifts)),
      family=poisson(),data)
print(anova(fitN,test='Chisq'))
print(fitted(fitN))
fitMN<-glm(deaths~morning+nurse+offset(log(shifts)),
      family=poisson(),data)
print(anova(fitMN,test='Chisq'))
print(fitted(fitMN))
```

Output

Display of data frame.

```
  shifts morning nurse deaths
1      8      yes    yes      7
2      ~~7~~      2      no    yes      ~~3~~1
3      ~~2~~      7      yes    no      ~~2~~4
4     28       no     no      4
```

Analysis of deviance table where only nurse is fitted. Note that the *p*-value for the nurse effect is 0.~~01728~~01509, ie 1.~~7~~5%, so apparently statistically significant.

```
Analysis of Deviance Table

Model: poisson, link: log

Response: deaths

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        3     9.7904
nurse  1   5.9056           2     3.8849  0.01509 *
~~NULL                        3     10.570~~
~~nurse  1   5.6678           2      4.902  0.01728 *~~
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fitted values for this model.

```
  1   2   3   4

6.4 1.6 1.6 6.4~~       1        2        3        4~~



~~5.333333 4.666667 0.400000 5.600000~~
```

Formatted: Space Before: 12 pt, After: 12 pt

Analysis of deviance table where morning and nurse are both fitted. Note that the *p*-value for the nurse effect is now 0.~~37849~~30145, i.e. ~~37.8~~30.1%, so is not statistically significant.

```
Analysis of Deviance Table

Model: poisson, link: log

Response: deaths

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        3     9.7904
morning  1   8.3494         2     1.4411 0.003858 **
nurse    1   1.0678         1     0.3733 0.301449  ~~NULL~~
~~3     10.5699~~
```

```
~~morning  1   8.6617           2      1.9081  0.00325 **~~
~~nurse    1   0.7756           1      1.1325  0.37849~~
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fitted values for this model.

```
        1         2         3         4
7.4254393 0.5745607 3.5745607 4.4254393~~        1        2        3        4~~
~~7.872829 2.127171 1.127171 4.872829~~
```

Biased investigation

This proceeds in exactly the same way, but using the biased data

```
deaths<-c(8,2,3,3) ~~deaths<-c(8,4,1,3)~~
```

in the input.

Analysis in Table 5

Input

Create data frame consisting of a numerical response variable `count` and two factors `nurse`, the explanatory variable, and `died`, the response category.

```
nurse<-as.factor(c('A','B','A','B'))
died<-as.factor(c('yes','yes','no','no'))
count<-c(15,9,25,31)
data<-data.frame(died,nurse,count)
print(data)
```

Fit a Poisson log-linear model, allowing for main effects `nurse` and `died`, and an interaction between them.

```
fit<-glm(count~died*nurse,data,family=poisson())
```

Output analysis of deviance table: the interaction term quantifies the differential effect of the two nurses on survival.

```
print(anova(fit,test='Chisq'))
```

The analysis of deviance table by convention uses the deviance as the test statistic: the following calculation demonstrates that it is numerically very similar to Pearson’s chi-squared statistic, as defined in the text.

```
E<-c(12,12,28,28)
print(sum((count-E)^2/E))
print(2*sum(count*log(count/E)))
```

Output

Display of data frame.

```
  died nurse count
1  yes     A    15
2  yes     B     9
3   no     A    25
4   no     B    31
```

Analysis of deviance table. Note that the *p*-value for the `died:nurse` interaction is 0.1416, ie 14.2%, so not statistically significant.

```
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           3    15.3254
died        1  13.1653         2     2.1601 0.0002852 ***
nurse       1   0.0000         1     2.1601 1.0000000
died:nurse  1   2.1601         0     0.0000 0.1416334
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Formatted: French (France)

Values of the two test statistics, the Pearson chi-squared statistic and the deviance statistic as used in the table above. They are very similar numerically, so it is immaterial which is used in calculating the *p*-value.

```
[1] 2.142857
[1] 2.160122
```

Calculations in Appendix 2

Input

Set up 4 illustrative data sets

```
casea<-matrix(c(10,30,5,35),2,2)
caseb<-matrix(c(100,300,50,350),2,2)
casec<-matrix(c(10,390,5,395),2,2)
cased<-matrix(c(55,345,5,395),2,2)
```

Define function to conduct chi-squared test, and calculate relative risk and absolute risk difference

```
comparerisks<-function(y)
{
```

```
ct<-chisq.test(y,,FALSE)
cat('statistic',ct$statistic,'  p-value',ct$p.value,'\n')

risks<-y[1,]/apply(y,2,sum); cat('risks',risks,'\n')
rr<-risks[1]/risks[2]
ard<-risks[1]-risks[2]
cat('RR',rr,'  AR difference',ard,'\n')
}
```

Formatted: French (France)

Apply function to data sets

```
casea
comparerisks(casea)
caseb
comparerisks(caseb)
casec
comparerisks(casec)
cased
comparerisks(cased)
```

Output

```
> casea
     [,1] [,2]
[1,]   10    5
[2,]   30   35
> comparerisks(casea)
statistic 2.051282   p-value 0.1520781
risks 0.25 0.125
RR 2   AR difference 0.125
> caseb
     [,1] [,2]
[1,]  100   50
[2,]  300  350
> comparerisks(caseb)
statistic 20.51282   p-value 5.923318e-06
risks 0.25 0.125
RR 2   AR difference 0.125
> casec
     [,1] [,2]
[1,]   10    5
[2,]  390  395
> comparerisks(casec)
statistic 1.698514   p-value 0.1924825
risks 0.025 0.0125
RR 2   AR difference 0.0125
> cased
     [,1] [,2]
[1,]   55    5
[2,]  345  395
> comparerisks(cased)
statistic 45.04505   p-value 1.925539e-11
risks 0.1375 0.0125
RR 11   AR difference 0.125
```

Appendix 9: Members of the working party drawing up this report

Professor Peter Green FRS, Emeritus Professor of Statistics, University of Bristol, and Distinguished Professor, University of Technology, Sydney.

Professor Richard Gill, Emeritus Professor of Statistics, Leiden University.

Neil Mackenzie QC, Arnot Manderson Advocates, Edinburgh.

Professor Julia Mortera, Professor of Statistics, Università Roma Tre.

Professor William Thompson, Professor Emeritus of Criminology, Law, and Society; Psychology and Social Behavior; and Law, University of California, Irvine.

In addition, we are grateful to Professor Jane Hutton, Professor of Medical Statistics, University of Warwick, for providing Appendix 7.

References

Bacon F. (1620) *Novum Organum, Book I*, 109, point 46, reprinted in Hutchins RM (Editor) *Great Books of the Western World* (vol 30). New York: Britannica Publishing 1952.

Balding DJ & Donnelly P. (1994) The prosecutor's fallacy and DNA evidence. *Criminal Law Review,* 711–721.

Burger JM. (1981) Motivational biases in the attribution of responsibility for an accident: A meta-analysis of the defensive-attribution hypothesis. *Psychological Bulletin. **90** (3): 496–512. doi:10.1037/0033-2909.90.3.496. S2CID 51912839*

Centers for Disease Control (1985) *Cluster of unexplained deaths at a hospital – Maryland*. Report EPI-86-17-1. Centers for Disease Control, Atlanta.Chatfield, Tom. *Critical Thinking* (2017) Sage Publishing.

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

de Keijser, J & Elffers H. (2012) Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. *Psychology, Crime & Law, 18,* 191–207. http://dx.doi.org/10.1080/10683161003736744

Dotto F, Gill RD & Mortera J. (2022). Statistical analyses in the case of an Italian nurse accused of murdering patients. *Law, Probability and Risk*, https://doi.org/10.1093/lpr/mgac007.

Dror IE & Charlton D. (2006) Why experts make errors. *J Forensic Identificatio*n, 56:600–16.

Dror IE, Charlton D & Peron A. (2006) Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Sci Int* 2006;156:174–8.

Dror IE & Hampikian G. (2011) Subjectivity and bias in forensic DNA mixture interpretation. *Sci Justice* 51(4):204–8.

Dror I, Melinek J, Arden JL, Kukucka J, Hawkins S, Carter J & Atherton DS. (2021a) Cognitive bias in forensic pathology decisions. *J Forens Sci* 66(5): 1751-57.

Dror I, Melinek J, Arden JL, Kukucka J, Hawkins S, Carter J & Atherton DS. (2021b) Authors' response to Peterson et al. commentary. *J. Forens Sci* 66:2545-48.

Dror IE & Rosenthal R. (2008) Meta-analytically quantifying the reliability and biasability of forensic experts. *J Forensic Sci* 53(4):900–3.

Dror IE, Thompson WC, Meissner CA, Kornfield I, Krane, D, Saks MJ & Risinger DM. (2015) Context management toolbox: A Linear Sequential Unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *Journal of Forensic Science* 2015;60(4):1111-1112. http://dx.doi.org/10.1111/1556-4029.12805

Evett IW. (1995) Avoiding the transposed conditional. *Science & Justice, 35,* 127–131. http://dx.doi.org/10.1016/S1355-0306(95)72645-4

Fienberg SE & Kaye DH. (1991). Legal and statistical aspects of some mysterious clusters. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154(1):61—74.

Findley KA & Scott MS. (2006) The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review* 291-395

Forensic Science Regulator (2015). Cognitive bias effects relevant to forensic science examinations.

Forensic Science Regulator (2021a). Development of evaluative opinions.

Forensic Science Regulator (2021b). Codes of Practice and Conduct for Forensic Science Providers and Practitioners in the Criminal Justice System, Issue 7.

Forensic Science Regulator (2022). Draft ‘Statutory’ Code of Practice

Forrest, ARW. (1995) Nurses who systematically harm their patients. *Medical Law International* 1(4):411-421. doi:10.1177/096853329500100404

Galef J. (2021) *The Scout Mindset: Why Some People See Things Clearly and Others Don't.* Penguin Books.

General Electric v. Joiner, 522 U.S.136 (1997).

Gill RD, Groeneboom P & de Jong P. (2018) Elementary statistics on trial (the case of Lucia de Berk). *Chance* 31 (4) 9–15. https://doi.org/10.1080/09332480.2018.1549809

Gill RD. (2021) Aart de Vos commentary, posted in translation at: Richard’s blog: https://gill1109.com/2021/05/24/condemned-by-statisticians

Gill RD, Fenton N & Lagnado D. (2022) Statistical issues in serial killer nurse cases, *Laws* (MDPI: Basel), 11(5):65. https://doi.org/10.3390/laws11050065

Hamilton G. (2011) *The Nurses Are Innocent: The Digoxin Poisoning Fallacy*. Dundurn. See also https://en.wikipedia.org/wiki/Toronto_hospital_baby_deaths

Herman AE. (2017) *Visual Intelligence: Sharpen Your Perception, Change Your Life Paperback.* Eamon Dolan/Mariner Books.

Hill AB. (1965). The environment and disease: association or causation? *Proc R Soc Med*. 58(5): 295–300.

Inns of Court College of Advocacy & Royal Statistical Society (2019) *Statistics and probability for advocates: Understanding the use of statistical evidence in courts and tribunals.*

Jeng M (2006). A selected history of expectation bias in physics. *American Journal of Physics*, 74, 578. https://doi.org/10.1119.1.2186333.

Kahneman D, Sibony O & Sunstein CR. (2021) *Noise: a Flaw in Human Judgement*. Little, Brown Spark, New York.

Kassin SM, Dror IE & Kukucka J. (2013) The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition* 2(1):42-52. doi.org/10.1016/j.jarmac.2013.01.001

Kennedy v Cordia Services LLP 2016 SC (UKSC) 59

Koehler, JJ. (1993) Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics Journal*, 34, 21–39.

~~The Law Commission (2015) *The admissibility of expert evidence in criminal proceedings in England and Wales: A new approach to the determination of evidentiary reliability*. Consultation Paper No 190.~~

Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

The Law Commission (2015) *The admissibility of expert evidence in criminal proceedings in England and Wales: A new approach to the determination of evidentiary reliability*. Consultation Paper No 190~~.~~

Lucy D & Aitken C. (2002) A review of the role of roster data and evidence of attendance in cases of suspected excess deaths in a medical context. *Law, Probability and Risk*, 1(2):141—160.

Meester R, Collins M, Gill RD & van Lambalgen M (2007) On the (ab)use of statistics in the legal case against the nurse Lucia de B. *Law, Probability and Risk* 5 233–250. With discussion by David Lucy. https://doi.org/10.1093/lpr/mgm003

Miller LS. (1984) Bias among forensic document examiners: a need for procedural change. *J Police Sci Admin*, 12(4):407–11.

Mitler MM, Hajdukovic RM, Shafor R, Hahn PM & Kripke DF (1987) When people die: Cause of death versus time of death. *The American Journal of Medicine* 82, 266–274. https://doi.org/10.1016/0002-9343(87)90067-2

Murrie DC, Boccaccini MT, Guarnera LA & Rufino, KA. (2013) Are forensic experts biased by the side that retained them? *Psychological Science*, 24(10): 1889-1897.

Nakhaeizadeh S, Dror IE & Morgan RM. (2013) Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. *Sci Justice*, 54(3):208–14.

Nickerson RS. (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2(2):175-220.

Nisbett RE & Ross L. (1980) Human *inference: strategies and shortcomings of social judgment*. Prentice-Hall: Englewood Cliffs, NJ.

Nisbett RE & Wilson TD. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231–259.

Osborne NK, Woods S, Kieser J & Zajac R. (2014) Does contextual information bias bitemark comparisons? *Sci Justice*, 54(4):267–73.

Peterson BL, Arnall M, Avedschmidt S, et al. (2021) Commentary on: Dror et al (2021a) *J. Forens Sci* 66:2541-44

Public Health England (2019). Non-infectious disease clusters: investigation guidelines..

Pires AM & Branco JA. (2010) A statistical model to explain the Mendel-Fisher Controversy. *Statistical Science* 25(4): 545-65.

Risinger DM, Saks MJ, Thompson WC & Rosenthal R. (2002) The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review* 90*:*1–55.

Ross L. (1977) The intuitive psychologist & his shortcomings: distortions in the attribution process. *Advances in Experimental Social Psychology*, 10: 173-220.

Royal Society and Royal Society of Edinburgh (2020) *The use of statistics in legal proceedings: A primer for courts*.

Sackett DL. (1979) Bias in analytic research in *The Case-Control Study Consensus and Controversy* Michel A. Ibrahim (ed.), Pergamon, pp 51-63.

Sacks JJ, Stroup DF, Will ML, Harris EL, Israel E., Anderson D., Aung KH, Nelson B, Quinley J, Hathcock AL, Irwin K, Sniezek J & Tyler, CW. (1988). A nurse-associated epidemic of cardiac arrests in an intensive care unit. *Journal of the American Medical Association*, 259(5):689—695.

Schneps, L & Colmez, C (2013) *Math on Trial: How Numbers Get Used and Abused in the Courtroom*. Chapter 7 – Math Error Number 7 –The Incredible Coincidence – The Case of Lucia de Berk: Carer or Killer? Basic Books, New York.

Scottish Government (2018) *Guidance for doctors completing medical certificate of cause of death (MCCD) and its quality assurance*. Advice from the Chief Medical Officer and National Records of Scotland.

The Shipman Inquiry (2003). Third Report: Death Certification and the Investigation of Deaths by Coroners.

Sidebotham P, Atkins B & Hutton JL. (2012) Changes in rates of violent child deaths in England and Wales between 1974 and 2008: an analysis of national mortality data. *Archives of disease in childhood*, *97*(3), 193-199.

Simon D. (2019) Minimizing error and bias in death Investigations. *Seton Hall Law Review*. 49: 255-305. Online at: https://scholarship.shu.edu/shlr/vol49/iss2/1

Sienkiewicz v Greif (UK) Ltd [2011] 2 AC 229 at p299, paragraph 206, per Lord Kerr of Tonaghmore JSC.

Smyth RL, van Velzen, D, Smyth AR, Loud DA & Heaf DP. (1994) Strictures of ascending colon in cystic fibrosis and high-strength pancreatic enzymes. *Lancet*, 343:85–86.

Smyth RL, Ashby D, O'Hea U, Burrows E, Lewis P, van Velzen D & Dodge JA. (1995) Fibrosing colonopathy in cystic fibrosis: results of a case-control study. *Lancet*, 346:1247–1251, 1995.

Spiegelhalter D. (2017) Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4:31—60.

Spiegelhalter D, Grigg O, Kinsman R. & Treasure T. (2003) Risk-adjusted sequential probability ratio tests: application to Bristol, Shipman and adult cardiac surgery. *Int J Qual Health Care*, vol. 15, pp 7–13 (2003).

Stewart AG, Ghebrehewet S, & Jarvis R (2016). Cancer and chronic disease clusters. In: Ghebrehewet S, Stewart AG, Baxter D, et al. [eds]. *Health Protection, Principles and practice*. Oxford University Press, Oxford, UK,. pp 163-173

Stoel RD, Berger CE, Kerkhoff W, Mattijssen E & Dror I. (2015) Minimizing contextual bias in forensic casework. In: Strom K, Hickman MJ, editors. *Forensic science and the administration of justice*. New York, NY: Sage, 67–86.

Stoel RD, Dror IE & Miller LS. (2014) Bias among forensic document examiners: still a need for procedural changes. *Australian Journal of Forensic Sciences*, 46(1):91–7.

Syed, M. (2015) *Black Box Thinking: The Surprising Truth About Success*. Penguin Books.

Taylor MC, Laber TL, Kish PE, Owens G & Osborne NK. The reliability of pattern classification in bloodstain pattern analysis, part 1: bloodstain patterns on rigid, non-absorbent surfaces. *J Forensic Sci* 2016;61(4):922–7.

Thompson WC (2009a) Interpretation: observer effects, in *Wiley Encyclopedia of Forensic Science*, Jamieson, A., Moenssens, A. (eds). John Wiley & Sons Ltd., Chichester, UK, pp 1575-1579.

Thompson WC. (2009b) Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation. *Law Probabability and Risk* 2009;8:257–76.

~~Thompson WC (2009) Interpretation: observer effects, in *Wiley Encyclopedia of Forensic Science*, Jamieson, A., Moenssens, A. (eds). John Wiley & Sons Ltd., Chichester, UK, pp 1575-1579.~~

Thompson, WC. (2011). What role should investigative facts play in the evaluation of scientific evidence? *Australian Journal of Forensic Sciences. 43(2-3*): 123-134.

Thompson, W.C. (2015). Determining the proper evidentiary basis for an expert opinion: What do experts need to know and when do they know too much? In C. Robertson & A. Kesselheim (Eds.) *Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science, and Law.* Elsevier, Inc. pp 133-150.

Thompson WC & Newman EJ. (2015) Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior*, 39(4): 332-349.

Thompson WC & Schumann E. (1987) Interpretation of statistical evidence in criminal trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy. *Law and Human Behavior, 11,* 167–187. http://dx.doi.org/10.1007/BF01044641

Thompson WC & Scurich N. (2018) When does absence of evidence constitute evidence of absence? *Forensic Science International*, 291 e18-e19.

Thompson WC & Scurich N. (2019) How cross-examination on subjectivity and bias affects jurors' evaluations of forensic science evidence. *Journal of Forensic Science*s, 2019. https://doi.org/10.1111/1556-4029.14031

Tressoldi PE. (2011). Extraordinary claims require extraordinary evidence: the case of non-local perception, a classical and Bayesian review of evidence. *Frontiers in Psychology*, 2, 117. https://doi.org/10.3389/fpsyg.2011.00117

Wartenberg, D. (2001) Investigating disease clusters: why, when and how? *J.R. Statist.Soc.A,* 164, Part 1, pp 13-22.

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.

FOUNDED
1834