Facts *versus* Factions: the use and abuse of subjectivity in scientific research - PART 2

Working paper for the European Science and Environment Forum, September 1998

Robert A.J. Matthews

Bayesian inference

The classical frequentist techniques of inference are not, in fact, "classical" at all, but relative newcomers in the long history of statistical inference. Before the 1920s, another approach to statistical inference was in general use, based on a result that flows directly from the axioms of probability. As such, this approach has solid theoretical foundations, produces intuitive, readily-understood measures of "significance", and remains as valid today as it did before it was eclipsed by the flawed attempts of Fisher *et al.* to create an objective theory of statistical inference. It is known as Bayesian inference, after the 18th Century English cleric Thomas Bayes who first published the key theorem behind it: Bayes's theorem.

The power and importance of this theorem is immediately apparent in its solution to one of the central problems of standard statistical inference. As we have seen, frequentist methods do not tell us Prob(theory | data); that is, they do not tell us what our belief in a theory should be, given the data we actually saw. To answer that question, we must turn to the axioms of probability theory, from which we find that (see, e.g. Feller 1968 Ch 5):

$$Prob(A\,|\,B) = Prob(B\,|\,A).Prob(A)/Prob(B) \qquad (1)$$

This is Bayes's theorem, which becomes the basis of Bayesian inference when "A" is the event of a specific hypothesis being true, and "B" as the event of observing specific data. Bayesian inference was the standard means of performing statistical inference prior to Fisher's work in the 1920s, and it allows us to calculate a clear and unambiguous measure of support for a theory, Prob(theory | data) directly from experimental results via the relationship:

$$Prob(theory\,|\,data) = Prob(data\,|\,theory).Prob(theory)/Prob(data) \qquad (2)$$

This formulation of Bayes's theorem shows clearly that while we can calculate the quantity we are interested in, namely Prob(theory | data), this is not equivalent to Prob(data | theory), much less to a P-value. However, the formula also highlights the key stumbling-block to the application of Bayesian inference. To work out the value of Prob(theory | data), we must first establish Prob (theory); that is, we must be able to put some "prior probability" on the theory we are testing. As I shall show later, setting this prior probability is often far less problematic than some critics claim: it is rare that there are absolutely no previous findings or plausibility arguments available to constrain our estimate. It remains true, nevertheless, that in those cases where there is a complete absence of any previous results or insight, the prior probability of the correctness of the hypothesis will be based largely on opinion. In short, it will be *subjective*.

It is this unequivocal use of subjectivity that has made Bayesian inference so controversial, and has led to such determined attempts to find alternatives. As we have seen, working scientists may routinely use subjectivity when it suits them, but the idea of explicitly incorporating it into the very heart of data analysis remains anathema. But this attitude overlooks a striking fact about the scientific process: that all attempts to rid it of subjectivity have failed. By the usual standards of scientific research, the repeated failure of these attempts would be taken to imply that the basic thesis was flawed. And from (2) we now see that this would, indeed, be the correct conclusion to draw. For the axioms of probability, via Bayes's theorem, show that subjectivity cannot be wrung out of the scientific process for the simple reason that it is mathematically *ineluctable*. Much as we might want to, it is *impossible* to obtain the value of Prob(theory | data) without having some value for the prior probability Prob(theory).

The plain fact is that subjectivity in statistical inference is as unavoidable as uncertainty in quantum mechanics. Yet while we have all grown accustomed to the latter - not least because of the welter of theoretical and empirical support for its existence - there remains a deep-seated reluctance to embrace the presence of subjectivity in scientific research.

We have seen that this reluctance stems in part from concern about playing into the hands of the "enemies" of science, and also from past abuses in the application of subjectivity. Further barriers exist to the adoption of Bayesian methods in data analysis, however. Some of these are entirely pragmatic: it is undoubtedly harder to boil down Bayesian inference to the same "cook-book" approach used in standard frequentist methods. Except in simple cases, Bayesian inference is also more mathematically and computationally demanding than frequentist methods. The dearth of textbooks and software suitable for the non-specialist wanting to carry out real-life data analysis does nothing to helps (see, however, O'Hagan 1997).

None of this would matter, however, were the working scientist convinced that the effort involved in getting to grips with Bayesian methods was worthwhile. This leads one to suspect that there are other, more fundamental reasons for the failure of Bayesian inference to regain its primacy over frequentist methods.

First, advocates of Bayesian inference have failed to tackle the widely held belief that Bayesian prior probabilities are never more than wholly subjective guesses, "plucked out of the air" to suit some or other prejudice or preconception. It cannot be stressed too highly that only rarely will there be *absolutely nothing* on which to base a reasonable prior. In many cases, there will be sources of evidence on which to base a sensible prior probability: for example, results from previous studies of similar drugs and plausibility arguments concerning, say, cancer risks from radiation based on insights from physics. Even if there really is little solid evidence on which to base a prior probability, Bayesian inference can still provide insight by allowing one to study the effect of different levels of prior belief (see, e.g. Spiegelhalter *et al.* 1994). It is also possible to invert Bayes's theorem, and estimate what prior belief is needed for data to reach a given level of plausibility; I give examples of such "inverse Bayesian inference" below.

The second key feature of Bayesian inference that is not sufficiently appreciated is that initial prior beliefs in a specific hypothesis become progressively less important as data accumulate. It can be shown mathematically (see, e.g. O'Hagan 1994 p 74 et seq.) that whatever prior probability is used at the outset, Bayes's theorem ensures that everyone is driven towards the same conclusion as the data accumulate. Unless one's prior is precisely zero (which is not a rational stance), the only long-term effect of the prior belief is that a sceptic starting from a low prior probability will require more data to reach the same level of belief as an enthusiast for the theory - which is hardly an egregious feature of a theory of inference. Indeed, it is striking that this mathematical feature of Bayesian inference mirrors so well how science actually operates. Starting from a wide variety of opinions about, say, the link between some chemical and cases of cancer, the accumulation of experimental and epidemiological evidence drives the scientific community toward the same conclusion about the reality or otherwise of the link, with sceptics merely taking longer to be convinced.

In short, Bayesian inference provides a coherent, comprehensive and strikingly intuitive alternative to the flawed frequentist methods of statistical inference. It leads to results that are more easily interpreted, more useful, and which more accurately reflect the way science actually proceeds. It is, moreover, unique in its ability to deal explicitly and reliably with the provably ineluctable presence of subjectivity in science.

These features alone should motivate many working scientists to find out more about applying Bayesian inference in their own research. For those who still need to be convinced, however, I now demonstrate perhaps the most impressive reason for using Bayesian inference: its ability to provide a far greater level of protection than frequentist methods against seeing "significance" in entirely spurious research findings. For as we shall see, while frequentist methods are still widely used within the scientific community, they routinely exaggerate the real significance of implausible data, with results that can and do bring the scientific process into disrepute.

How P-values exaggerate significance

As we have seen, frequentist methods of inference provide measures of significance that are neither objective nor intuitive. More importantly, however, they give a fundamentally misleading view of the significance of data. To see this, take the simple case in which a hypothesis is to be tested via measurements of a specific parameter, θ ; for example, the hypothesis may be that a toxin is linked to some disorder in children, so that θ is the level of this toxin in children suffering from the disorder. Such an investigation would then consist of measuring values of θ in a group of affected children, θ_i, computing the data mean and variance, and comparing it with θ_0, the value of θ found among normal children. We would then test the "null" hypothesis that any difference we find is merely the result of chance by setting up a test-statistic, z, which takes into account the sample size, its mean and variance, and compares it to θ_0, the value expected if the null hypothesis is correct. Following the frequentist approach, one would typically convert this z-score to a P-value, the probability of obtaining at least as large a value of z, *assuming* the null hypothesis that chance alone is the cause. According to convention, if the P-value is less than 0.05, then the data are taken to be "significant".

However, as we have seen, a much more meaningful measure of "significance" is Prob(Null hypothesis | data), the probability that the difference in θ*really is* the product of chance alone. Just how big is the disparity between this measure of significance and the frequentist P-value ? To find out, we can use Bayes's theorem (2), which with a little algebra becomes

$$Prob(Null\ hypothesis\ |data) = \left\{1 + \frac{1 - Prob(Null)}{Prob(Null).BF}\right\}^{-1} \qquad (3)$$

where Prob(Null) is the prior probability for the null hypothesis that there is no real difference in the toxin level in the children, and BF is the so-called Bayes Factor, which measures how much we should alter our prior belief about the null hypothesis in the light of the new data, as captured by z. For the value of the Bayes Factor, one can show (see, e.g. Lee p131) that under very general conditions BF has a *lower* limit of

$$BF \geq exp(- z^2 / 2) \qquad (4)$$

As an example, suppose that past evidence concerning the toxin leads us to an agnostic view of the possibility that there are higher levels of the toxin in the children with the disorder; this is equivalent to setting Prob(Null) = 0.5. Inserting this and (4) into (3) we find that, for a given value of z, our initial agnosticism leads us to a probability that the null hypothesis of no real difference is indeed correct of *at least*

$$Prob(Null\ hypothesis\,|data) \geq \{1+\exp(z^2/2)\}^{-1} \qquad (5)$$

Suppose, for example, that the measurements of the toxin levels in the two groups revealed a difference with a z-value of 2.0. On the frequentist viewpoint, standard statistical tables shows that this implies a P-value of 0.044; as this is less than 0.05, the difference is deemed "significant at the P = 0.05 level". As we have stressed, however, this does *not* mean that the probability that the difference *really is* a fluke is also 0.044; we can only calculate this latter probability via Bayes's theorem. Plugging in z = 2 into (5), we find that our data actually imply that Prob(null | data), the probability the difference is just a fluke, is *at least* 0.12. In other words, while the frequentist methods led us to conclude that the difference was "significant", the Bayesian calculation pointed to a much higher probability of the finding being a mere fluke.

This conclusion, moreover, was based on an agnostic prior of Prob(Null) = 0.5. If there are no strong grounds for believing that the effect is genuine, then - in contrast to frequentist methods - Bayesian inference allows us to factor in this lack of plausibility explicitly into our analysis. This can have particularly dramatic effects in the assessment of "anomalous" phenomena (Matthews 1998), as the following example shows (Nelson 1997).

For over 250 years, Princeton students have attended Commencement on a Tuesday in late May or early June, an outdoor event for which good weather is vital. According to local folklore, good weather does usually prevail, prompting claims that those attending may "wish" good weather into existence. By analysing local weather records spanning many decades, Nelson found that Princeton's weather was generally no different from that of its surroundings. However, he did find some evidence that the town was less likely to be rained on during the outdoor events. The phenomenon gave z-scores as high as 1.996, which on a frequentist basis gives a "significant" P-value of 0.046. Properly mindful of the implausibility of the phenomenon, however, Nelson was reluctant to take this "objective" finding at face value, and instead reached a more subjective conclusion: "These intriguing results certainly aren't strong enough to compel belief, but the case presents a very challenging possibility".

A Bayesian analysis allows a far more concrete assessment of plausibility to be made. Clearly, with such a bizarre claim, there is little one can say about the precise value of a sensible prior probability for the null hypothesis of no real effect, other than to say that the probability is likely to be pretty high. In such cases, Bayesian inference still gives valuable insight, as it allows one to estimate the level of prior probability necessary to sustain a belief that the effect is illusory, even in the light of Nelson's data. Using (4) and (3) and z = 1.996, this inverse Bayesian inference shows that Prob(Null | data) > 0.5 for all Pr(Null) > 0.88 In other words, for anyone whose prior scepticism about the effectiveness of "wishful thinking" exceeds 90 per cent, the balance of probabilities is that the effect is illusory, despite Nelson's data.

As this example shows, frequentist methods greatly exaggerate the "significance" of intrinsically implausible data. However, as we shall now see, frequentist methods can also seriously exaggerate both the size and significance of effects in much more important mainstream areas of research, such as clinical trials.

Misleading "significance" of clinical trial results

Misleading P-values

The most common methods for investigating the efficacy of a new drug or therapy, or the impact of exposure to some risk-factor, are the so-called randomised clinical trials and case-control studies, in which a group of people given the new treatment or known to have the disease are compared with a "control" group. One common frequentist method of analysing the outcome is to reduce the results to a test-statistic (such as χ^2), which is then turned into a P-value; as before, if this is less than 0.05, then the difference between the two groups is deemed to be significant. Again, however, a Bayesian analysis reveals that the real "significance" of such a finding is typically much less impressive than the P-values imply.

As before, I shall demonstrate this by taking a real-life case. During the early 1990s, research emerged to suggest that the risk of coronary heart disease (CHD) is associated with childhood poverty (Elford *et al.* 1991). Following the discovery that infection with the bacterium *H. pylori* is also linked to poverty, some researchers suspected that the bacterium may form the "missing link" between the two. Precisely how a bacterium in the stomach might cause heart disease is less than clear - raising the key issue of plausibility, to which we shall return shortly. Nevertheless, a number of studies were undertaken to investigate the link between CHD and *H. pylori.* In one of the first such studies (Mendall *et al.* 1994), 60 per cent of patients who suffered CHD were found to be infected with *H. pylori*, compared with 39 per cent of normal controls. When the effects of age, CHD risk factors and current social class had been controlled for, the results led to a χ^2 value of 4.73. Using frequentist methods, this leads to a P-value of 0.03, implying that the rate of CHD among those infected with *H. pylori* is "significantly" higher than those without.

On the face of it, this finding raises the intriguing prospect of being able to tackle one of the major killers of the western world using nothing more than antibiotics. Yet while the evidence that both CHD and *H. pylori* infection are more common among the poor is suggestive of a link between the two, it is hardly unequivocal. Such scepticism is underscored by the lack of any convincing mechanism by which a gastric bacterium could trigger heart disease. The frequentist P-value, however, cannot reflect any of these justifiable qualms; sceptics of the link have no option but to say that on this occasion they are just going to ignore the supposed "significance" of Mendall *et al.* 's finding.

In contrast, Bayesian inference requires no such arbitrary "moving of the goalposts": it allows explicit account to be taken of the plausibility of the findings. In the case of the supposed link between CHD to *H. pylori*, the lack of any convincing mechanism balanced against the socio-economic evidence of a link suggests that an agnostic prior probability of Prob(Null) = 0.5 would be a reasonable starting-point for assessing results like those found by Mendall *et al.* . Inserting this into (3) implies that the probability of the results being due to chance, given the observed data, is

$$Prob(Null\ hypothesis\ |data) \geq BF/(1+BF) \qquad (6)$$

where BF is the Bayes Factor for the null hypothesis of chance effect. One can show that for in a wide range of practical situations, including this type of case-control study, the *lower* bound on BF is given by (see, e.g. Berger & Sellke 1987)

$$BF \geq \sqrt{(\chi^2)}\exp[(1-\chi^2)/2] \qquad (7)$$

Inserting the value of $\chi^2 = 4.73$ found by Mendall *et al.* into (6) shows that the BF is *at least* 0.337. Putting this in (6) we find that Prob(Null | data), the probability that Mendall *et al.*'s results are due to nothing more than chance is *at least* 0.25. In other words, even using an agnostic prior, the frequentist P-value has over-estimated the real "significance" of the findings by almost an order of magnitude.

Those taking a more sceptical view of a link between a gastric bacterium and CHD would, of course, set Prob(Null) somewhat higher. Applying the concept of inverse Bayesian inference used earlier, it emerges that even a relatively modest sceptical prior of just Prob(Null) = 0.75 is enough to lead to a balance of probabilities that Mendall *et al.*'s findings are entirely illusory.

Misleading Confidence Intervals

Some defenders of frequentist methods regard criticism of P-values as an attack on a straw man, pointing out that P-values are increasingly being supplanted by 95 per cent confidence intervals (CIs), which convey more information about effect size than a single-figure P-value. Yet as we have seen, frequentist CIs share many of the same problems of interpretation as P-values. Most importantly, they also share an inability to take into account the plausibility of the hypothesis under test. As such, 95 per cent confidence intervals are also prone to exaggerate both the size and the "significance" of intrinsically implausible effects.

In contrast - and as one might expect by now - the Bayesian counterpart of CIs (known as Credible Intervals or Highest Density Regions), are more comprehensible, more meaningful and more reliable indicators of real significance. With frequentist CIs, the 95 per cent refers to the reliability of the statistical test; the Bayesian CI, in contrast, means precisely what it seems to mean: that there is a 95 per cent probability that the true value of the parameter lies within the stated range.

As already noted, Bayesian CIs are numerically identical to their frequentist counterpart if there is only very vague prior knowledge about plausible values of the parameter of interest (see e.g. Berger & Delampady 1987 p 328, and Appendix to this paper). However, such complete ignorance about the likely size of the effect under study is rarely defensible, and in general frequentist and Bayesian CIs will not coincide. In such cases, a Bayesian CI is always a more reliable guide to the true "significance" of a finding than its frequentist counterpart.

Again, let us illustrate this through a real-life example. In the early 1990s, the Grampian region early anistreplase trial study (GREAT Group, 1992) generated considerable interest in the medical community, as it seemed to show that heart-attack victims given this clot-busting drug at home had a 50 per cent higher chance of survival than those given the drug once they arrived in hospital. While there were good reasons for expecting that early intervention with the drug would produce some improvement, the size of the claimed benefit surprised many. Nevertheless, frequentist measures of significance appeared to give objective support to the finding: the team found a relative risk (RR) of death for those given the drug early of 0.52 - i.e. a 48 per cent risk reduction - with a 95 per cent CI of (0.23 0.97). As this excludes an RR of 1, this surprising result is also "significant" in frequentist terms, the equivalent P-value being 0.04.

However, as was pointed out shortly after the publication of the GREAT results (Pocock and Spiegelhalter 1992), a considerable amount of prior information existed with which to assess the plausibility of the GREAT finding; for example, a much larger European study involving the same drug pointed to a much smaller benefit. Drawing on this existing knowledge, Pocock and Spiegelhalter carried out a Bayesian re-assessment of the GREAT results; an outline of how such an analysis can be performed is given in the Appendix to this paper. The prior information was captured through a probability distribution which peaked at an RR of 0.83 while giving low probabilities to RRs greater than 1.0 (no benefit) or less than 0.6 (dramatic improvement). When combined with the GREAT data, the resulting ("posterior") probability distribution peaked at an RR of around 0.75, with a 95 per cent Bayesian CI of (0.57 1.0). While still pointing to a more impressive effect than that suggested by previous studies, the GREAT results emerge from the analysis as markedly less impressive than suggested by the frequentist methods.

At this point, it is natural to ask whether this Bayesian analysis really did give a more accurate picture of reality than the frequentist methods. The simple answer is yes. Six years after the publication of the GREAT findings, the overall picture emerging from international studies is that early use of clot-busters like anistreplase does indeed confer extra benefit, with RRs of around 0.75 to 0.8 (Fox, quoted in Matthews 1997). This is only half the improvement suggested by the frequentist analysis of the GREAT data, but in impressive agreement with Pocock and Spiegelhalter's Bayesian analysis.

In a similar vein, the current consensus concerning the supposed *H. pylori*-CHD link is that a plausible mechanism relating the two is lacking, and that a causal link remains dubious (Danesh *et al.* 1997). This suggests that the basis of the above Bayesian analysis of the supposed link remains valid - a conclusion supported by a recent large-scale study that failed to find any convincing evidence for an association (Wald *et al.* 1997).

These cases are hardly the only examples of the tendency of frequentist methods to exaggerate both effect size and "significance" of clinical findings. Undoubtedly the most disturbing evidence comes from the continuing failure of many impressive drug trial results to produce similarly impressive results once approved for general release. It is widely recognised that most new therapies for cancer and heart disease have proved far less effective than initially believed (e.g. Fayers 1994, Yusuf *et al.* 1984). Very recently, a UK study uncovered evidence that the use of "clinically proved" drugs for myocardial infarction since the early 1980s has had no effect on mortality, with death-rates on the wards at least double those found in trials (Brown *et al* 1997).

Such a finding would come as no surprise to those familiar with the inherent ability of frequentist methods to exaggerate both effect sizes and "significance". It is of course perfectly possible that at least part of the explanation for such disappointing findings lies elsewhere: the greater care taken of all patients in clinical trials, for example, and the fact that trials tend to be conducted in centres of excellence. Brown *et al.* suggest that their disappointing findings may be due to a failure to optimise the use of the available treatments for myocardial infarction. This highlights another factor in the continuing failure of Bayesian methods to supplant frequentist methods: the existence of many other apparently plausible explanations capable of masking the failings of frequentist methods.

"Explaining away" frequentist failures

The most common explanation for studies whose spuriously "significant" findings fail to be confirmed is that the sample size was too small. This seems plausible enough: after all, everyone knows that the smaller a sample, the less reliable its conclusions. Yet the argument overlooks two key facts. First, the calculation of a P-value takes full account of sample size. On the frequentist viewpoint, we must regard a P-value of 0.03 as "significant" whether it is based on a sample of 10 or 10,000 people; larger samples are just more likely to detect "significance" in smaller effects. And this is related to the second flaw in the "sample size" defence of frequentist failures. Small samples are indeed more susceptible to statistical noise than large ones, but only in the sense that their lack of statistical power makes them more prone to missing real effects. For a given P-value, both small and large studies of the same quality are equally likely to see "significance" in results that are really due to chance. As such, blaming the failure of large studies to replicate "significant" positive findings from smaller studies purely on sample size is simply fallacious.

A more sophisticated, and plausible, defence of frequentist failures is that the original studies were undermined by biasing and confounding factors. Bias undermines the separation of subjects into cases and controls, due to, say, misdiagnosis of the disease whose cause is under investigation. Confounding undermines attempts to link a cause to its effects; for example, failure to take into account dietary differences can undermine attempts to link carcinogens to observed cases of cancer.

Both bias and confounding are exceptionally difficult to deal with, and undoubtedly explain many failures to replicate results. For example, when Mendall *et al.* applied further controls for the confounding effect of overcrowding and hot water supplies in childhood risk-factors for infection by *H. pylori*, the link between the bacterium and CHD remained, but its P-value was no longer significant.

The undoubted power of bias and confounding to undermine clinical research findings has provided defenders of frequentist methods with a further reason for shunning Bayesian inference. The argument is that while Bayesian methods may indeed deal more effectively with the risk of seeing significance in fluke results, it is no better at dealing with bias and confounding than the standard frequentist methods, and these are typically far more important.

This is also incorrect. Even relatively simple Bayesian analysis does allow concern about bias and confounding to be taken into account, via the form of the prior probability distribution, in the assessment of the posterior probability. Similar remarks apply to the supposed inability of Bayesian methods to take into account the many other potential influences on trial outcome, from poor randomisation to the better care received by patients in clinical trials. All these can be captured by a prior reflecting past real-life experience of just how successful drugs usually turn out to be.

Ultimately, however, all these supposed objections to the use of Bayesian methods serve only to conceal the key advantage of Bayesian inference: that it offers far greater protection against seeing significance in implausible results. The importance of this can best be seen through another real-life example, and one of great contemporary interest: the assessment of the risk of lung cancer faced by passive smoking of environmental tobacco smoke (ETS). The strongest evidence for this risk is generally held to be a recent meta-analysis of 37 published studies (Hackshaw *et al.*, 1997). This found a relative risk (RR) for lung cancer among life-long non-smokers living with smokers of 1.24 with a 95 per cent CI of (1.13, 1.36). A detailed assessment of both bias and confounding was carried out, but the central estimate for the RR remained essentially unchanged at 1.26 with a 95 per cent CI of (1.07, 1.47). On the basis of standard inference methods, this implies a highly "significant" link between passive smoking and lung cancer ($P < 0.005$). To underline the credibility of their results, Hackshaw *et al.* performed an informal plausibility assessment of their findings, using indirect measures of the likely intake of ETS by passive smokers. These suggest that passive smokers have about 1 per cent the exposure to cigarettes of their smoking partners. Assuming smokers typically consume 25 cigarettes a day, face an RR of 20 and that there is a linear dose-risk relation, Hackshaw *et al.* reached an estimate of RR ~ 1.19 for passive smokers.

While broadly similar to the RR found by the meta-analysis, this plausibility argument has itself been criticised as implausible (Lee 1998, Nilsson 1998 p 20). However, both Hackshaw *et al.* and their critics underestimate the crucial importance of a much more rigorous assessment of the plausibility of such weak results. Hackshaw *et al.* devoted about 10 times more of their paper to the assessment of bias and confounding than to plausibility; as I now show, however, a Bayesian analysis reveals that plausibility has a far more dramatic effect on the "significance" of the results.

Of the many criticisms that can be levelled at Hackshaw *et al.*'s plausibility argument, the most serious is their reliance on markers of ETS exposure which are both indirect and not linked to carcinogenicity. The use of such markers is especially hard to justify in the face of evidence from *direct* studies of ETS exposure that consistently point to much lower levels of exposure. An ongoing series of such studies (see e.g. Phillips *et al.* 1994, Phillips *et al.* 1998 and references therein) has found median exposures figures of ~ 0.02 cigarettes a day for the most exposed passive smokers. Even adopting the same linear dose-risk relation as Hackshaw *et al.* (which again is questionable, Nilsson 1998 pp21-22) this suggests a plausible RR for passive smoking of around 1.02, an excess risk 10 times lower than that estimated by Hackshaw *et al.* Only the top 10 per cent of the most exposed passive smokers in the studies by Phillips *et al.* were found to face anything like the risk predicted by Hackshaw *et al.*

Incorporating these results into a plausibility argument via a Bayesian prior distribution leads to an altogether different view of the risks of passive smoking. Specifically, it suggest that the excess lung-cancer risk is both 11 times smaller than that given by Hackshaw *et al.*, and has a 95 per cent CI of (1.00, 1.04). Bayesian inference thus strongly suggests that the growing consensus that ETS is a proven and major health risk is misplaced. Whether or not the outcome of this Bayesian analysis will be borne out is as yet unclear. What is clear is that there is a very real danger of the frequentist evidence for a "significant extra" risk from ETS becoming canonical. This, in turn, raises the possibility that Hackshaw *et al.*'s risk figure will be used routinely to subtract out the confounding effect of passive smoking in future studies of the causes of cancer. If this risk figure has been substantially over-estimated - as the above Bayesian analysis strongly suggests it has - attempts to assess the true risk posed by many other health hazards will be seriously undermined (Nilsson 1997 p140).

This example of passive smoking and lung cancer provides the final strand in the case for the widespread and routine use of Bayesian inference in the analysis of data. This can be summed up as follows:

- It allows both previous knowledge and the inherent plausibility of a hypothesis to be explicitly taken into account;
- It gives measures of "significance" that are more meaningful than those generated by frequentist methods;
- These measures have more intuitive and straightforward definitions than their frequentist counterparts, and are thus much less prone to misinterpretation;
- Bayesian inference is less likely to see "significance" in entirely spurious findings, especially in poorly-motivated research of low inherent plausibility. As such, it provides more protection against seriously - even dangerously - misleading findings whose attempted replication or extension will ultimately prove futile.

Conclusions

In this paper, I have shown that the scientific community has a deeply ambiguous attitude towards the presence of subjectivity in research. While both desiring and proclaiming objectivity, working scientists routinely use subjective criteria in their everyday research. The justification is pragmatic, and entirely reasonable: it is impossible for working scientists to deal with the plethora of new results and theories that constantly present themselves in any other way. However, mindful of past abuses in the history of science, the scientific community remains committed to keeping the presence of subjectivity in the research enterprise to a minimum.

This commitment has led to the widespread adoption of techniques for statistical inference that appear to be "objective". Known as frequentist methods, they have become central to the research enterprise, with their outcomes - P-values and 95 per cent confidence intervals - becoming a *sine qua non* for acceptance by leading science journals. As I have shown, however, these textbook methods are neither objective nor reliable indicators of either effect size or statistical significance of research findings. By failing to take into account the intrinsic plausibility of the hypothesis under test, frequentist methods are capable of greatly exaggerating both the size and the significance of effects which are in reality the product of mere chance.

The implicit recognition of these failings by scientific community is evidenced by the way in which essentially identical results from the supposedly "objective" frequentist methods are interpreted in entirely different ways, according to the subjective belief of researchers. Thus, a large and "highly statistically significant" result in parapsychology will be ignored, while a small and statistically non-significant link between passive smoking and cancer will be deemed to "add considerably" to the case against environmental tobacco smoke.

The persistent failure of scientists to rid the research process of subjectivity, and the failings of frequentist techniques, can both be traced to the same fundamental source: the axioms of probability. These show that in the assessment of hypotheses, subjectivity is mathematically *ineluctable.* All attempts to banish subjectivity from the research process are thus ultimately futile, and are at best no more than exercises in sweeping subjectivity "under the carpet".

The vexed problem of subjectivity in science has its solution in those same axioms, however. Bayes's theorem provides the underpinning for an entire theory of statistical inference which takes explicit account of plausibility, and supplies measures of statistical "significance" that are more relevant, more comprehensible and more reliable than those of frequentist methods. As such, the wider adoption of Bayesian inference will undoubtedly save substantial amounts of time, resources and public money currently spent on futile attempts to replicate "significant" support for intrinsically implausible hypotheses.

Some idea of the extent of this waste can be obtained by noting that each month journals covering disciplines from sociology and psychology to geology and genetics carry many papers claiming to have results "significant at the 0.05 level" with P-values in the range $0.01 \leq P < 0.05$. Even assuming that these claims are all sufficiently well-motivated to merit an agnostic prior, it can be shown that (6) and (7) point to *at least* a quarter of such claims are meaningless flukes (Matthews 1998). For research meriting even a very moderate level of scepticism, this proportion rapidly rises to over 50 per cent. This is a finding that should worry anyone concerned with the reliability and funding of scientific research.

The fact that just two independent clinical trials with results "significant at the 0.05 level" are sufficient for new therapies to win approval from national regulatory bodies is hardly less worrying. As so often with frequentist concepts, this P-value standard can and is misinterpreted as implying that the probability of the therapy being ineffective is less than 1 in 400 (see, e.g. Buyse 1994). The true proportion will be far higher, especially among therapies whose claims of efficacy are poorly motivated - a fact reflected in the many cases of where initial euphoria over some new "breakthrough" turns into disappointment (Yusuf *et al.* 1984, Pocock & Spiegelhalter 1992, Fayers 1994, Brown *et al.* 1997). Bayesian assessment of trial results give regulatory bodies a formal means of incorporating this crucial "reality check" into their deliberations. In contrast, the frequentist methods currently used by regulatory bodies have no means of incorporating such key knowledge: given the same raw data, they cannot distinguish between streptokinase or snake-oil. A number of regulatory bodies will accept Bayesian assessments of drug trials; in the light of the above, the use of such methods should not be optional but mandatory.

Lack of theoretical underpinning has an especially large impact on areas of research such as parapsychology and alternative medicine. Bayesian inference applied here would certainly cast grave doubt on claims that appear impressive from a frequentist viewpoint. It is important to stress, however, that this does not imply that all research into "alternative" or "anomalous" fields should be abandoned. Bayesian inference merely implies that the standard frequentist criteria for judging statistical "significance" in these areas are especially inadequate. It can be shown that in such fields of research, there are few grounds for viewing as "significant" anyresult whose two-tailed P-value exceeds 0.003 (Matthews 1998). This value assumes an agnostic prior of Pr(Null) = 0.5, which is undoubtedly generous for most claims for the existence of anomalous phenomena; even so, the resulting P-value is 17 times more demanding than the conventional 0.05 criterion used for gauging significance, and it is clear that many current claims for anomalous phenomena fail to meet it.

Reputable researchers would no doubt feel more confident defending evidence for an anomalous phenomenon by applying at least a mild level of scepticism in their assessment of significance. In this case, a P-value of no more than around $2\text{x}10^{-4}$ is appropriate, a value 250 times more demanding than the conventional 0.05 criterion. These technical results can be stated much more succinctly, however: extraordinary claims require extraordinary evidence. This is a well-attested and widely-accepted principle, yet it is noticeable by its absence in the mathematics of frequentist inference.

It must also be emphasised that many of the concerns about frequentist inference expressed here have been recognised by leading statisticians for decades (see e.g. Jeffreys 1961, Edwards *et al.* 1963, Lindley 1970). This inevitably raises the question of why Bayesian inference is still failing to (re)gain its central role in the scientific enterprise. This is, I believe, due largely to the failure of its advocates to convey three key facts to working scientists:

- That while subjectivity may be an unwelcome feature of the scientific process, the axioms of probability show that it is unavoidable, and that Bayes's theorem is the correct way to deal with it;
- That while Bayesian inference does allow subjective prior knowledge to be incorporated into the assessment of data, such knowledge is not "plucked out of thin air". Rather, it allows an entirely reasonable yet crucial assessment of plausibility to be factored into the analysis.
- That, in any case, the effect of the choice of prior becomes increasingly irrelevant as data accumulates, with the only persistent effect of priors being the entirely natural one that sceptics of a specific claim require stronger evidence to reach the same level of belief than its advocates.

There is a dangerous irony in the continuing reluctance of the scientific community to adopt Bayesian inference. For this reluctance stems largely from a deep-rooted fear that adopting methods that embrace subjectivity is tantamount to conceding that the scientific enterprise really is a social construct, as claimed by the post-modern advocates of the "anti-science" movement. The central lesson of Bayes's theorem is, however, quite the opposite. It shows, with full mathematical rigour, that while evidence for a specific theory may indeed start out vague and subjective, the accumulation of data progressively drives the evidence towards a single, objective reality about which all can agree.

It is ironic indeed that by failing to recognise this, the scientific community continues to use techniques of inference whose unreliability undermines confidence in the scientific process, and which thus threatens to deliver science into the hands of its enemies.

Appendix: Bayesian inference using confidence intervals

A growing proportion of research findings are reported via confidence intervals, in which a central parameter value, M, is accompanied by a range of values of the form (L , U), which form the so-called 95 per cent confidence interval (CI) for the results. As discussed in the main article, the frequentist interpretation of a CI is not as straightforward as it may appear: the 95 per cent figure refers to the reliability of the statistical test applied, and not to the probability that the true parameter value lies in the stated range. In contrast, a Bayesian 95 per cent CI (often also called a Credible Interval) means precisely what it seems to mean: there is a 95 per cent probability that the true value lies within the stated range.

We now outline the procedure for calculating Bayesian CIs for a given set of data. For both frequentist and Bayesian CIs, the range (L, U) is calculated from the mean parameter value, M and its standard deviation SD, via the formulas

$$L = M - 1.96.SD \qquad \text{(A1)}$$

$$U = M + 1.96.SD \qquad \text{(A2)}$$

In the textbook frequentist approach, M and SD are calculated directly from the raw data. In the Bayesian approach, however, the M and SD are the so-called "posterior" mean and standard deviation, formed by combining the raw values extracted from the data with "prior" values based on extant knowledge and insight about the effect under study. The resulting posterior mean and standard deviation thus sets the new findings into their proper context, taking explicit account of their intrinsic plausibility.

The first step in a Bayesian analysis is thus to capture this prior knowledge and insight. In many real-life cases, this can be achieved by specifying a Normal distribution which peaks at the most plausible value for the parameter of interest, M_o, and whose 95 per cent "tails" (L_o, U_o) reflect the plausible range of that parameter. The standard deviation of this prior distribution, SD_o can then be calculated from (A1), (A2):

$$SD_0 = (U_0 - L_0)/3.92 \qquad \text{(A3)}$$

The next step is to combine this prior distribution with the experimental data, whose mean is M_d and standard deviation is SD_d; the resulting "posterior" distribution will have a mean M_p and standard deviation SD_p. It can be shown that Bayes's theorem leads to a posterior distribution with parameters given by (see e.g. Lee 1997 Ch 2)

$$SD_p = 1/\sqrt{(1/SD_0)^2 + (1/SD_d)^2} \qquad \text{(A4)}$$

$$M_p = (SD_p)^2 \left[(M_0/SD_0^{\,2}) + (M_d/SD_d^{\,2})\right] \qquad \text{(A5)}$$

The Bayesian 95 per cent CI then follows putting A4 and A5 into A1 and A2; the result is a range of values for the parameter of value in which the true value will lie with 95 per cent probability. Two key implications of equations A4 and A5 should be noted. First, they show that the frequentist and Bayesian definitions of the CI are equivalent only when SD_o is infinite, corresponding to a stance of complete ignorance about the plausible range of values for the parameter of interest. This is rarely justifiable, and in general the frequentist and Bayesian CIs will not coincide. Equations A4 and A5 also show that the inclusion of prior information has the effect of moving the posterior probability distribution in the direction of the prior. Thus if results from, say, a clinical trial are strikingly more impressive than seems plausible, failure to account for this lack of plausibility via a prior distribution will exaggerate both the size of the effect, and its statistical significance. As we have seen, frequentist methods cannot explicitly incorporate such plausibility arguments, and are thus especially prone to lend unjustified credibility to remarkable data.

The growing tendency to state results in terms of frequentist 95 per cent CIs does at least summarise results in a form that can easily be combined with prior knowledge using the techniques given above, as I now show.

Example: In their analysis of the GREAT study, Pocock and Spiegelhalter captured the implications of previous studies via a "prior" relative risk (RR) of death of 0.825, with a 95 per cent CI of (0.6, 1.0). To apply the above formulas, a logarithmic transformation has to be applied to the central estimate and range, and the RRs should be transformed into a so-called Odds Ratio (OR), but in this particular case the difference between OR and RR can be ignored to a first approximation. Thus we take the prior distribution to be Normal, with a peak at ln(RR_o), with its standard deviation SD_o being calculated from A3 using the natural logarithms of the upper and lower ranges of the CI, ln (U_o) and ln(L_o). This leads to a prior distribution that peaks at $M_o = -0.19$, with a standard deviation of 0.13.

To calculate M_d and SD_d, we note that the GREAT study found a mean RR of 0.515, with a (frequentist) 95 per cent CI of (0.23, 0.97). We can convert this into the mean and standard deviation required by A4 and A5 by taking natural logarithms and using A2: this gives $M_d = -0.664$, and $SD_d = 0.367$. Using A4 and A5 we can now work out the posterior probability distribution; we find $M_p = -0.245$ and $SD_p = 0.123$.

Using A1 and A2 and then transforming back out of natural logarithms, we finally arrive at a posterior RR figure of 0.78 with a Bayesian 95 per cent CI of (0.6, 1.0). This central risk figure is substantially less impressive than the value that emerges from the raw data; this reflects the impact of the inclusion of a prior reflecting the implausibility of gaining so large a risk reduction. Furthermore, the Bayesian 95 per cent CI encompasses an RR of 1.0, which implies that the possibility that there is no benefit is not entirely ruled out by this (small) study. As discussed in the main article, the results of Pocock and Spiegelhalter's Bayesian analysis ultimately proved more realistic than those suggested by the raw GREAT data alone.

REFERENCES & ACKNOWLEDGEMENTS

References

Aronson, J. L. 1984 *A realist philosophy of science* (London: Macmillan)

Asimov, I. 1975 *Asimov's Biographical Encyclopaedia of Science and Technology* (London : Pan Books)

Barber, B. 1961 Resistance by Scientists to Scientific Discovery *Science* **134** 596

Berger, J. & Berry, D. Statistical Analysis and the Illusion of Objectivity 1988 *American Scientist* **76** 159

Berger, J. & Delampady, M. 1987 Testing Precise Hypotheses *Stat. Sci.* **2** 317

Berger, J. & Sellke, T. Testing a point null hypothesis: the irreconcilability of P-values and evidence *J. Amer. Statist. Ass.* **82** 112 (1987)

Berry, D. A. 1996 *Statistics: A Bayesian Perspective* (Belmont, Ca: Duxbury)

Bofetta, P., Brennan, P., Lea, S. Ferro, G. 1997 Lung cancer and exposure to environmental tobacco smoke. *Biennial Report 1996/7* (Lyon: IARC/WHO)

Bourke, G. J., Daly, L.E. McGilvray, J. 1985 *Interpretation and uses of Medical Statistics* (3rd Edn). (St Louis : Mosby)

Breslow, N., Day, N. E. 1980 Statistical methods in cancer research vol 1: The analysis of case-control studies *IARC Scientific Publication No. 32* (Lyon : IARC)

Brown, N., Young, T., Gray, D., Skene, A., Hampton, J.R. 1997 Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register *Brit Med J* **315** 159

Buyse, M.E. 1994. Remarks in response to Spiegelhalter *et al.* 1994 (below); p 399

Collins, H. 1998 What's Wrong with Relativism? *Physics World* **11** (4) 19

Cooper, C. Dennison, E. 1998 Do silicone breast implants cause connective tissue disorder ? *Brit Med J* **316** 403

Crease, R.P., Mann, C.C. 1996 *The Second Creation* (London: Quartet)

Danesh, J. Collins, R. Peto, R. 1997 Chronic infections and coronary heart disease: is there a link ? *Lancet* **350** 430

Dunstan, D. 1998 Letter *Physics World* **11** (6) 15

Edwards, W. Lindman, H. & Savage, L. J. 1963 Bayesian statistical inference for psychological research. *Psychol. Rev.* **70** 193

Elford, J. Whincup, P. Shaper, A.G. 1991 Early life experience and adult cardiovascular disease *Intl J Epid* **20** 833

Fayers, P. 1994 Remarks in response to Spiegelhalter *et al.* 1994 (below); p 402

Feller, W. 1968 *An Introduction to probability theory and its applications* 3rd Edn. (New York: Wiley)

Feynman, R.P. 1985 *Surely you're joking Mr Feynman* (London: Unwin)

Fletcher, H. 1982 *Physics Today* June 43

Freedman, D. Pisani, R. & Purves, R. 1998 *Statistics* (3rd Edn.) (New York : Norton)

Gell-Mann, M. 1964 A schematic model of baryons and mesons *Physics Letters* **3** 214

Gell-Mann, M. 1994 *The Quark and the Jaguar* (London: Little, Brown)

Grayson, L. 1995 *Scientific Deception* (London: The British Library)

Grayson, L. 1997 *Scientific Deception - An Update* (London: The British Library)

GREAT Group 1992 Feasibility, safety and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial *BMJ* **305** 548

Greenstein, G. 1998 *Portraits of Discovery* (New York: Wiley).

Hackshaw, A.K. Law, M.R., Wald, N.J. 1997 The accumulated evidence on lung cancer and environmental tobacco smoke *Brit Med J* **315** 980

Hoffmann, B. 1975 *Albert Einstein* (London: Paladin)

Holton, G. 1978 Sub-electrons, Presuppositions and the Millikan-Ehrenhaft Dispute *Historical Studies in the Physical Sciences* **9** 161

Hellman, H. 1998 *Great Feuds in Science* (New York: Wiley)

Heyes, S., Hardy, M., Humphreys, P., Rookes, P. 1993 *Starting Statistics in Psychology and Education* 2nd Edn (London : Weidenfeld & Nicolson)

Howson, C. Urbach, P. 1993 *Scientific Reasoning* 2nd Edn (Chicago: Open Court)

Jeffreys, H. *Theory of Probability* 1961 (3rd Edn), (Oxford : University Press)

Kempthorne, O. 1971 "Probability, statistics and the knowledge business" *in Foundations of Statistical Inference* (Ed. Godambe & Sprott) (Toronto: Holt, Rinehart & Winston)

Lakatos, I. 1978 *Philosophical Papers* (Worrall & Currie, eds) vol 1 (Cambridge: the University Press) .

Lee, P. M. 1997 *Bayesian Statistics: An Introduction* 2nd Ed. (London : Arnold)

Lee, P.N. 1998 Difficulties in assessing the relationship between passive smoking and lung cancer *Stat Meth Med Res* **7** 137

Lindley, D. V. *Introduction to Probability & Statistics Part 2: Inference* 1970 (Cambridge: University Press)

Linet, M, *et al.* 1997 Residential exposure to magnetic fields and acute lymphoblastic leukaemia in children *New Eng J Med* **337** 1

Macdonald, V. 1998 Official: passive smoking does not cause cancer *The Sunday Telegraph* 8 March p 1

Matthews, R.A.J. 1992 *Unravelling the Mind of God* (London: Virgin)

Matthews, R.A.J. 1997 Faith, Hope and Statistics *New Scientist* **156** 36

Matthews, R.A.J. 1998 The statistical assessment of anomalous phenomena *J Sci Expl* (accepted)

Medawar, P. 1978 *Advice to a Young Scientist* (New York : Harper & Row)

Mendall, M.A. *et al.* 1994 Relation between *Helicobacter pylori* infection and coronary heart disease *B Heart J* **71** 437

Milton, R. 1994 *Forbidden Science* (London: Fourth Estate)

Nelson, R. 1997 Wishing for good weather: a natural experiment in group consciousness *J. Sci. Expl.* **11** 47

Nilsson, R. 1997 Is environmental tobacco smoke a risk factor for lung cancer ? In *What Risk: science, politics and public health ed. Bate, R* (Cambridge: The European Science and Environment Forum)

Nilsson, R. 1998 *Environmental Tobacco Smoke Revisted: The reliability of the evidence for risk of lung cancer and cardiovascular disease* (Cambridge: The European Science and Environment Forum)

Nyren, O. *et al.* 1998 Risk of connective tissue disease and related disorders among women with breast implants: a nationwide retrospective cohort study in Sweden *Brit Med J* **316** 417

O'Hagan, A. 1994 *Kendall's Advanced Theory of Statistics Vol 2B: Bayesian Inference* (London: Arnold)

O'Hagan, A. *FirstBayes -freeware bayesian inference software.* Available from the Mathematics Dept homepage based here

Oppenheimer, J. R. 1955 *The Open Mind* (New York : Simon & Schuster)

Pais, A. 1982 *Subtle is the Lord* (Oxford: University Press)

Pais, A. 1991 *Niels Bohr's Times* (Oxford: Clarendon Press).

Phillips, K. Howard, D.A., Browne, D. Lewsley, J.M. 1994 Assessment of personal exposures to environmental tobacco smoke in British non-smokers *Environment International* **20** 693

Phillips, K. Howard, D.A., Bentley, M. C. Alvan, G. 1998 Measured exposures by personal monitoring for respirable suspended particles and environmental tobacco smoke of housewives and office workers resident in Bremen, Germany *Int Arch Occup Environ Health* **71** 201

Pocock, S. J., Spiegelhalter D. J. 1992 Letter *Brit Med J* **305** 1015

Popper, K. 1963 *Conjectures and Refutations* (London : Routledge)

Radin, D. 1997 *The Conscious Universe: the scientific truth of psychic phenomena* (San Francisco: Harper)

Sivia, D. S. 1996 *Data Analysis: A Bayesian Tutorial* (Oxford: University Press)

Spiegelhalter, D. J., Freedman, L.S., Parmar, M.K.B., 1994 Bayesian approaches to randomised trials (with discussion) *J Roy Stat Soc A* **157** 357

Theocharis, T. & Psimopolous, M. 1987 Where science has gone wrong *Nature* **329** 595

Vallance, A. K. 1998 Can Biological Activity be Maintained at Ultra-High Dilution? An Overview of Homeopathy, Evidence, and Bayesian Philosophy *J Alt Comp Med* **4** 49

Vandenbroucke, J.P. 1997 Homoeopathy trials: going nowhere *The Lancet* **350** 824

Wald, N. Law, M.R. Morris, J.K. Bagnall, A.M. 1997 *Helicobacter pylori* infection and mortality from ischaemic heart disease: negative result from a large prospective study *Brit Med J* **315** 1199

Weinberg, S. 1993 *The Discovery of Sub-atomic particles* (London: Penguin)

Williams, T. 1994 *Biographical Dictionary of Scientists* (London: Collins)

Wolpert, L. 1992 *The Unnatural Nature of Science* (London: Faber)

Wolpert, L., Richards, A. 1989 *A Passion for Science* (Oxford: the university press)

Yusuf, S., Collins, R., Peto, R. 1984 Why do we need some large, simple randomized trials ? *Statistics in Medicine* **3** 409

Acknowledgements

In arriving at the arguments presented in this Working Paper, I have benefited enormously from the assistance, advice and comments of many researchers. I should particularly like to thank David Balding of Reading University, James Berger of Purdue University, Colin Howson of the London School of Economics, Robert Nilsson of Stockholm University, and Stuart Pocock and Ian White of the London School of Hygiene and Tropical Medicine. I am also most grateful to Roger Bate of ESEF for inviting me to collect together my thoughts in this Working Paper, and to three anonymous referees who made useful comments. My biggest debt, however, is to Dennis Lindley, from whom - like so many others working in this field - I have learned so much.

The author

Robert Matthews is Visiting Fellow in the Neural Computing Research Group at Aston University, Birmingham. A graduate in physics from Oxford University, he has published many research papers in fields ranging from astrodynamics and probability theory to the statistical analysis of anomalous phenomena and the mathematical basis of "urban myths". A Fellow of the Royal Statistical Society and Royal Astronomical Society, he also acts as science correspondent for *The Sunday Telegraph*, London.

David Williams

From: Robert Matthews [r.matthews@physics.org]
Sent: Friday, 26 September 2003 6:44 PM
To: David Williams
Cc: J Enstrom
Subject: credibility of air pollution

Dear Mr Williams,

Thanks for the note. I'm not aware of any work looking at the inherent credibility of the epidemiology on air pollution; it would be fascinating to see what it throws up. The technique of Credibility Analysis put forward in my papers may well help.

Essentially the question is whether the 95% confidence intervals found by epidemiological studies are not merely statistically significant but also "credible", in the sense that plausible relative risk/odds ratio values lie outside the CPI corresponding to these quoted 95% CIs.

This is certainly questionable in the case of environmental tobacco smoke (ETS, or "passive smoking"). The 1997 meta-analysis of case-control studies by Hackshaw et al found a relative risk figure of 1.26, with a 95% CI corrected for bias and confounding of (1.07, 1.47). Plugging the latter two figures into the formula given in my paper produces a CPI of (0, 1.17). Thus the Hackshaw et al finding may be statistically significant, but it is only credible if plausible values for the risks posed by ETS greater than 1.17 can be found. Hackshaw et al made a claim for RRs ~ 1.19 in their paper, which would render their meta-analytic finding credible. However, this was based on some very questionable assumptions. If instead one uses the results of *direct* measures of ETS intake (by personal air monitoring devices) by Phillips et al, they found RRs of around 1.02. These are well within the CPI for Hackshaw et al's results, which are thus rendered not credible.

A similar approach would doubtless work for other areas of pollution.

I hope this is of some use.
best wishes
Robert Matthews

----- Original Message -----

> **From:** David Williams
> **To:** r.matthews@physics.org
> **Sent:** Friday, September 26, 2003 3:16 AM
>
> Dear Mr Matthews,
>
> I have just revisited your article in the New Scientist (March 8) and read your paper on the same topic on your website. I am neither a statistician nor a medical researcher, my interest being in air pollution. Some fairly expensive strategies have been put in place to reduce air pollution. Much of these have been based on health effects and $ savings from avoided pollution-related injury. I have always been sceptical of much of the epidemiological and clinical data (gut feeling), and the reasons may well lie in the statistics used. Before I go to the original documents, have you had chance to look at this or do you know of someone who has?
>
> Sincerely