ELSEVIER

Journal of Statistical Planning and Inference 94 (2001) 43–58

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Why *should* clinicians care about Bayesian methods?

Robert A.J. Matthews

Department of Information Engineering, Aston University, Birmingham B4 7ET, UK

Abstract

There is a growing awareness of Bayesian methods within the medical research community, and increasing discussion of their potential applications. This interest has, however, so far failed to convert into the routine use of such methods by working clinicians. I argue that attempts to encourage the use of Bayesian methods by highlighting the deficiencies of conventional (frequentist) inference have not succeeded because these deficiencies typically have minor practical consequences, while their more serious effects can usually be explained away by appeal to other statistical issues. As a result, Bayesian methods have not appeared to offer practical pay-offs big enough to justify the cost of acquiring the necessary expertise. In an attempt to remove this "cost–benefit" hurdle, I outline a simple Bayesian technique that can be used alongside frequentist methods to address an issue of routine practical concern to working clinicians: the credibility of new research findings.

1. Introduction

Medical researchers are heavy users of inferential methods. Leading medical journals are replete with clinical and epidemiological studies whose conclusions are backed by quantitative inferential concepts such as p-values and confidence intervals. Small wonder, then, that there is growing interest in Bayesian inference, which is said to have a number of key advantages over conventional methods (see, e.g. Lilford and Braunholtz, 1996; Bland and Altman, 1998; Spiegelhalter et al., 1999).

Yet despite the increasing awareness of their existence, the actual application of Bayesian methods remains very much the exception rather than the rule among working clinicians. There are many potential explanations for this, from concern about the much-discussed use of prior belief in Bayesian analysis, to the simple reluctance of old dogs to learn new (statistical) tricks.

I would argue, however, that the fundamental reason why Bayesian methods have yet to become widely adopted is that they have yet to pass the "cost–benefit analysis" any hard-pressed professional performs when hearing of some new technique: will the

E-mail address: rajm@compuserve.com (R.A.J. Matthews).

0378-3758/01/$ - see front matter
PII: S0378-3758(00)00232-9

cost of acquiring the necessary expertise be compensated for by real practical benefit? Many potential users of Bayesian methods, it would seem, have concluded that the effort of switching from the familiar old frequentist methods to Bayesian techniques is unlikely to give sufficient return on (intellectual) investment.

Of the two factors involved in such a cost–benefit analysis, the cost of adopting Bayesian techniques can certainly seem high to non-specialists: the typical introductory text on Bayesian inference has a mathematical content well above that of the equivalent texts for conventional statistics. Yet I would argue that this is not the principal reason why Bayesian techniques have yet to become widely adopted: their mathematical basis is within the reach of anyone — given the motivation. Rather, I suspect the chief reason is the failure of Bayesians to provide a ready answer to a rather obvious question: if the standard methods of inference are so awful, how come the whole scientific enterprise has not collapsed around our ears?

In what follows, I suggest that while the failings of conventional inference are real enough, they typically have either little practical impact on the assessment of evidence, or are easily explained away as the result of other statistical deficiencies. As such, avoiding them hardly provides much motivation for switching to Bayesian methods.

If the "stick" of supposedly dire consequences of using frequentist methods has not succeeded in promoting the widespread use of Bayesian methods, what might? In this paper, I suggest a possible "carrot", which focuses instead on one of the major practical advantages of Bayesian methods: their ability to take explicit, quantitative account of extant knowledge. Specifically, I outline a Bayesian method of assessing quantitatively the *credibility* of a new research finding. The technique is easy to apply, and produces a measure of credibility whose interpretation is both straightforward and transparent. As such, it may provide a gentle "entrée" to Bayesian modes of thought, which allows the non-specialist to appreciate their power without losing all contact with familiar frequentist methods.

I begin, however, by reviewing briefly the concerns raised over the years about frequentist inference, and why these concerns have had so little impact among working clinicians.

2. Five reasons to fret about frequentist methods

Frequentist concepts such as p-values have faced criticism almost since the advent of "significance testing" in the 1920s. The criticisms have ranged from qualms about the interpretation and arbitrariness of the well-known $p=0.05$ criterion for significance, through concern that p-values exaggerate real "significance", to claims that frequentist methods conceal the ineluctable presence of subjectivity in inference. Indeed, so sustained has this criticism been that one wonders how frequentist methods have survived for so long. Yet survive they have, which raises an awkward question: just how important can their supposed failings really be?

2.1. Frequentist methods are easily misinterpreted

Any textbook on statistical inference contains definitions of frequentist concepts such as p-values and confidence intervals (CIs). Yet it is doubtful whether one in ten non-specialists who routinely use these concepts could state their correct definitions cold. Indeed, so convoluted are they that not even textbooks or examination boards always get them right (Bourke et al., 1985; Heyes et al., 1993; MacRae, 1995). But does this matter? Take the case of the p-values used to determine "statistical significance". One might reasonably expect that these are trivially related to the probability that chance alone accounts for the results obtained. However, as the (correct) textbook definition makes clear, p-values are calculated on the *assumption* of the validity of the null hypothesis that chance alone is responsible for the observed result. As such, p-values cannot also measure the probability that the null hypothesis H_0 *really is* the right explanation. Yet this is precisely what p-values invite the unwary to do — and to conclude from a p-value of, say, 0.05, that there is just a 5% chance of the stated result being due to the play of chance. That is, p-values invite users to commit a variant of the well-known "transposition of conditioning fallacy" in which $\Pr(A|B)$ is mistaken for $\Pr(B|A)$. Most non-statisticians have difficulty appreciating the distinction between the simple question they thought was answered by p-values — $\Pr(H_0|\text{data})$ — and the strangely convoluted answer p-values actually supply, namely Pr(At least as impressive data as that seen $|H_0$). Certainly the re-iteration of this distinction in countless papers, courses and textbooks has failed to convince most users of statistical inference of the dangers of falling foul of this fallacy. And why should it? For where is the evidence that this (mis-)use of p-values has led to the collapse of modern science as we know it? The lack of such evidence suggests that, whatever the qualms of some statistics works, making this error cannot really be all that serious. Before considering the validity of this argument, let us look at another of the conceptual problems with p-values routinely raised by their critics.

2.2. Frequentist methods are arbitrary

Even the most statistically apathetic must have wondered precisely why a p-value of 0.049 is deemed "statistically significant", while one of 0.051 is not. As Jeffreys (1961) emphasises, the 0.05 cut-off was chosen by R.A. Fisher because of a handy mathematical coincidence: a conveniently low percentage of the total area under the Normal curve — 5% — lies beyond a conveniently (almost) round number of standard deviations either side of the mean: 1.96.

Arbitrary or not, the 0.05 criterion has proved extraordinarily resilient in the face of attempts to excise it from the theory of statistical inference. In these days of powerful PCs and statistics software, this resilience can hardly be attributed to computational convenience. A more plausible explanation is that the 0.05 criterion does seem to give a clear-cut, standardised and reasonable point of reference for the otherwise seemingly hopelessly subjective task of gauging "significance" — and clinicians (like most people)

like clear-cut answers. Furthermore, one can once again argue that there cannot be that much wrong with the 0.05 criterion, as it has been used for decades without the scientific sky falling in. Thus this second supposedly grave flaw in frequentist methods can all too easily be dismissed as of no practical importance for the working clinician.

2.3. Frequentist methods exaggerate significance

One criticism of frequentist methods that is not so easily dismissed is the charge that they routinely exaggerate the real "significance" of experimental findings. Again, the criticism has been made for decades: almost 40 years ago, Leonard Savage and colleagues warned that frequentist measures are "startlingly prone" to see significance in results properly ascribed to the play of chance (Edwards et al., 1963). A quarter-century later similar warnings were made in greater analytical detail by Berger and co-workers (Berger and Sellke, 1987; Berger and Delampady, 1987). How have p-values managed to survive so serious a challenge?

The answer lies in the quantitative implications of the issue already raised above: the difference between what p-values *seem* to tell us, and what they really do tell us. Suppose we want to test the (point-)null hypothesis H_0 that a particular set of data is the result of the play of chance alone. As we have seen, the p-value merely gives us $\Pr(\geqslant \text{data}|H_0)$, while the obvious quantity of inferential interest is the value of $\Pr(H_0|\text{data})$. This can be calculated directly (and non-controversially) from Bayes's Theorem:

$$\Pr(H_0|\text{data}) = \left(1 + \frac{1 - \Pr(H_0)}{\Pr(H_0)\text{BF}}\right)^{-1}, \tag{1}$$

where $\Pr(H_0)$ is the so-called *prior probability* that the null hypothesis explains the data, established in the absence of the new findings, and BF is the *Bayes factor*, which captures the weight of evidence provided by the data in favour of H_0. In a wide range of practical cases, the lower bound on BF is given by (Berger and Sellke, 1987)

$$\text{BF} \geqslant z \exp[(1 - z^2)/2] \quad (z > 1), \tag{2}$$

where z is the normal test statistic. As a p-value of 0.05 corresponds to $z = 1.96$, (2) leads to a Bayes factor of $\text{BF} \geqslant 0.47$. From (1) we thus see that to be justified in believing that a p-value of 0.05 is equivalent to $\Pr(H_0|\text{data}) = 0.05$ — as many users of p-values clearly do — they must hold a prior probability for the null hypothesis $\Pr(H_0)$ no higher than 0.1. In other words, only those who were already 90% certain that the null hypothesis is wrong are justified in taking results with $p = 0.05$ as implying there is now a 95% chance of the null hypothesis being wrong (which, incidentally, also highlights what little extra weight of evidence against the null hypothesis is provided by results with $p = 0.05$).

One might be tempted to argue that all this merely implies that the whole dispute over p-values versus $\Pr(H_0|\text{data})$ can be evaded by claiming that researchers sufficiently motivated to investigate a given hypothesis are indeed at least 90% certain that "there is

something in it". This will not do, however: if others — such as regulatory authorities — are to reach the same conclusion, they must also be convinced that such a level of prior belief is tenable, and evidence to substantiate $\Pr(H_0) \leqslant 0.1$ may not be easy to provide.

In medical research, there is a further difficulty in such casual evasion of this issue. The design of randomised controlled trials (RCTs) is such that, in principle at least, if the null hypothesis has been ruled out, then the only other explanation for a positive result is therapeutic action. Thus having a high prior probability against fluke necessarily implies one also has a high prior probability that the therapy would be efficacious. This immediately raises ethical issues over recruiting patients to any RCT whose final p-value is to be interpreted as $\Pr(H_0|\text{data})$, as this equivalence can only be justified by those who are already 90% certain that patients in the control arm will do less well.

To gauge the real impact of confusing p-values with $\Pr(H_0|\text{data})$ one might therefore adopt a standard "agnostic" point-null prior of $\Pr(H_0)$ of 0.5, and study the consequences. This is what Berger and his co-workers have done, with fascinating — and somewhat disturbing — implications for results described as "significant at the $p < 0.05$ level". By convention, this term implies that the precise p-value lies in the range $0.01 < p < 0.05$. Using (1) and (2), one can show that for results with p-values in this range, a prior probability of 0.5 implies that $\Pr(H_0|\text{data})$ is *at least* 0.22. In other words, around at least a quarter of all positive findings with "$p < 0.05$" should, if viewed in a scientifically agnostic light, properly be regarded as nothing more than flukes.

So why has not this four-fold disparity between what p-values seem to mean and what they really mean revealed itself? Surely the plethora of failures to replicate predicted by this calculation should have made itself blatantly clear by now? The reason they have not, I suspect, again lies in the misconception that p-values rule out the null hypothesis with 95% reliability. This apparently impressive figure makes it seem only sensible to point the finger of blame for failures to replicate at one of the many other statistical issues — such as confounding or bias — that could be responsible. The alternative — that there must be something wrong in the meaning of an elementary statistical concept — is not a conclusion one would want to reach for on a regular basis.

2.4. Frequentist methods fail worst when you need them most

As (1) and (2) show, the disparity between p-values and $\Pr(H_0|\text{data})$ increases with $\Pr(H_0)$ for constant z. In consequence, p-values become increasingly misleading the more implausible the theory being investigated. For example, if the claim under study is so implausible as to justify a very high prior probability for the null hypothesis of $\Pr(H_0) = 0.99$, then a result said to be significant at the $p < 0.05$ level actually corresponds to $\Pr(H_0|\text{data})$ of *at least* 0.96, 20 times higher than the probability apparently implied by the p-value. Now the question posed in the previous sections becomes even more pressing: just how have p-values succeeded in lending their support to patently nonsensical phenomena while avoiding doubts being raised over their reliability?

Part of the answer, I would argue, is that most scientists are not very bothered about failures to replicate evidence for crazy ideas. Thus just when p-values are at their least reliable, most scientists are least interested in what they have to say anyway. Furthermore, any failure to replicate is again all too easily blamed on a plethora of other, apparently more likely, sources.

This is well illustrated by Terence Hines' fascinating recent review of evidence for one such crazy idea: the existence of biorhythms (Hines, 1998). The fundamental notion underpinning biorhythms is that humans are affected by a 23-day "physical" cycle, a 28-day "emotional" cycle and a 33-day "intellectual" cycle. This lacks any scientific basis, the various cycle lengths having simply appeared out of nowhere in the biorhythm literature between 1890 and 1930. The claim that these cycles are controlled by a biological "clock" that starts ticking at the moment of birth and maintains the requisite accuracy until death also lacks all biological plausibility. Despite this, Hines found that of 134 traceable scientific studies of biorhythms, over a quarter cited some support for biorhythms, often backed by "significant" p-values. Hines did not, however, see any need to question the reliability of frequentist methods to account for these false positives: his analysis revealed so many other statistical blunders that there was simply no need to invoke possible failings in the inferential method itself.

2.5. Frequentist methods are a poor basis for regulatory assessment

As we have seen, p-values are not a very stringent criterion by which to assess the significance of a finding. Despite this, medical regulatory bodies have long been happy to accept p-value evidence in support of claims of efficacy, such as two independent trials significant at the 0.05 level. This criterion has, not surprisingly, been misinterpreted as implying an apparently impressive overall significance level of $(0.05)^2 = 0.0025$. However, one can show that, using an agnostic point-null prior of $\Pr(H_0)$, this regulatory criterion is capable of approving therapies which, in fact, have *at least* a 7% probability of being based on nothing but fluke results.

While far less comforting than the 1-in-400 "false-positive" figure (falsely) implied by the p-value criterion, this 1-in-14 figure is still not especially worrying, and may well be too pessimistic: a case for using a prior less sceptical than agnostism can be made for some drugs, and many are approved on the basis of much more impressive p-values in any case. Those failures that do emerge can, moreover, readily be blamed on a host of other plausible explanations, yet again allowing the flawed frequentist inferential method off the hook.

3. Confidence intervals: better than *p*-values?

So far, I have focused primarily on the deficiencies of p-values, used to evaluate the significance of point-null hypotheses. However, criticism of p-values is not a uniquely Bayesian sport: frequentists have themselves argued against the use of p-values, on

the grounds that they fail to convey information about effect size or sample size. It is entirely possible for researchers to claim a "statistically significant" effect and yet for the effect to be so small as to have no significance in the everyday sense of the word.

It is also possible to obtain statistically significant results on the basis of very small samples. As even the most statistically naïve knows the dangers of basing grand conclusions on little data, this latter feature of p-values should again ring alarm bells about the precise meaning of "statistical significance". Yet it has done no such thing: p-values are still widely quoted without effect sizes or any qualms about sample size.

The reason appears to be that many users of p-values are indeed convinced that "statistical significance" implies real-life significance. As such, they see little need to quote effect size, or worry about inferring so much from so few data: the magic p-value formula has somehow "dealt with all that".

The upshot is that p-values continue to be wheeled out even in prestigious journals, whose editors seem convinced that they imbue findings with unquestionable import. One important exception has been the leading medical journals, where 95% confidence intervals (CIs) have gradually become the primary means of summing up findings. Thus it has become commonplace to see clinical studies summarised by a central figure giving the effect size, such as an odds ratio (OR), together with a "95% confidence interval", whose width reflects the statistical power of the study. Whether the result can be deemed "statistically significant" can still be quickly determined, simply by seeing if the 95% interval includes values corresponding to no effect (e.g. an OR of 1.00).

There can be little question that 95% CIs are substantially better than p-values as a summary of findings. They are far from perfect, however: for example, their precise definition is even more subtle than that of p-values, and thus even easier to misinterpret. The perceived and correct interpretation of CIs do, however, coincide when there is an absence of any prior insight into the likely value of the parameter in question (that is, a so-called flat prior with infinite variance). A stance of such complete ignorance is, however, rarely justifiable: typically there will be some prior insight capable of imposing some bounds on the true value. In such cases, the frequentist CI will exhibit "shrinkage" in the direction of the prior insight. Thus in cases where extant knowledge gives grounds for doubting the reality of a finding, the frequentist 95% CI will — like p-values — tend to exaggerate the real "significance" of a finding.

As we have seen with p-values, however, none of these problems is likely to reveal itself in an inferential catastrophe which can unequivocally be laid at the door of the frequentist confidence interval. As such, there is little hope of their flaws providing enough incentive for working clinicians to embrace Bayesian methods.

4. What is required to change old habits?

Thus far I have outlined the principal reasons put forward by advocates of Bayesian methods for worrying about conventional statistical inference. I have also suggested why these failings have had so little impact outside the statistical community: while

they are real and of conceptual importance, they have not led to inferential meltdown in routine research.

If Cassandra-like warnings of the dire consequences of using frequentist methods have not convinced working clinicians of the merits of using Bayesian methods, then perhaps a more positive approach might work. I am hardly the first to suggest this: it was the stated objective behind the classic 1994 JRSS paper by Spiegelhalter, Freedman and Parmar, which in 31 pages covered the theory, practice and implications of the Bayesian analysis of randomised trials — and even included some worked examples (Spiegelhalter et al., 1994). Yet even this model of lucidity and practicality has patently not succeeded in sparking the widespread use of Bayesian methods in general medical journals.

The discussion section of that paper suggests some reasons, ranging from the perceived complexity of any "serious" Bayesian analysis to the usual qualms about the role of prior beliefs in Bayesian inference.

My own suspicions, however, are that all these would seem less of a barrier if a Bayesian technique were put forward that tackled an issue of routine practical concern to working clinicians. Ideally, this technique would be easy to apply and interpret, and be capable of being used alongside existing "conventional" statistical methods such as p-values and 95% CIs. If at all possible, it would also address the thorny issue of setting priors, which has proved a major barrier to the wider adoption of Bayesian methods.

A technique meeting this wish-list might help avert the stalemate that threatens to descend on the issue of the use of Bayesian methods in medical science. In what follows, I outline one such technique which addresses a key issue routinely faced in medical research: the credibility of new findings.

5. Credibility: a suitable case for Bayesian treatment

While the outcomes of medical research are now routinely stated quantitatively via such measures as odds ratios, this is rarely the case for their inherent credibility — that is, the level to which the finding is both plausible in the light of current knowledge, and backed by persuasive weight of evidence. Even where a finding is controversial or unexpected, assessments of its credibility almost always consist of broad-brush qualitative arguments based on previous research or experience. Such arguments are, however, all too easy to devise; Egger et al., 1998 cite a case where the authors of two mutually contradictory studies were both able to supply apparently reasonable qualitative arguments to back their opposing results.

The lack of a quantitative measure of credibility has led to the illegitimate use of statistical significance as a surrogate for such a measure. Such abuse of p-values and 95% CIs does at least demonstrate that there is an unfulfilled need for a statistically well-founded quantitative measure of credibility. Such a measure would ideally have a

number of features:

1. it should allow the new findings to be set in the context of existing knowledge, as befits any measure of credibility;
2. it should be transparent, allowing anyone to gauge the overall credibility of the finding in the light of their own perceptions of its plausibility;
3. it should reflect the effect of sample size, with findings based on small samples requiring more support from existing knowledge to achieve the same level of credibility as findings from large samples;
4. it should be easy to calculate from conventional statistical measures, and have a straightforward interpretation.

These desiderata are naturally met within the framework of Bayesian methods, which have at their heart the notion of updating belief in the light of new evidence. As with any potential application of such methods, however, the thorny issue of priors has to be addressed: that is, the requirement of Bayes's Theorem that we explicitly incorporate our prior knowledge in the inferential process.

Much effort has been focused on this controversial aspect of Bayesian inference, not least because of the suspicion that the reliance on prior knowledge threatens to undermine the whole notion of scientific objectivity, allowing different researchers to draw different conclusions from the same data. Some advocates of Bayesian methods see this dependency on prior knowledge as entirely natural: after all, researchers routinely "pick and choose" findings they find impressive in the light of what they know, despite the supposedly objective nature of standard statistical methods (see, e.g. Matthews, 2000). As such, Bayesian methods can be seen as merely capturing in a rigorous manner a feature of inference that is conventionally dealt with in an *ad hoc* way.

Even so, the standard means of dealing with prior knowledge do not seem to have allayed suspicions about Bayesian methods being overly vulnerable to subjectivity. For example, so-called elicitation methods for capturing prior knowledge from experts have been criticised on the grounds that such experts can be (and have been) overly optimistic or conservative in their views. Furthermore, those not polled for their expert views may justifiably feel excluded from the resulting assessment, yet unclear of how to replace the opinions of the experts with their own. Methods such as robustness analysis have been devised to tackle these problems, but their use is hardly elementary.

There is, however, an approach to the issue of priors that has received relatively little attention, despite its simplicity and efficacy being recognised by Jack Good half a century ago (Good, 1950). It consists of turning Bayes's Theorem around, and reversing the usual inferential sequence of prior belief being combined with new evidence to produce an updated ("posterior") level of belief. That is, it asks instead what prior level of belief is required in order to regard the hypothesis under test as persuasive, in the light of the new findings. As Good has stressed, this reversal is both logically and mathematically legitimate, despite the impression given by the conventional terms "prior" and "posterior".

In what follows, I make this approach the basis of a Bayesian method for assessing credibility. Specifically, this method takes the new data bearing on the reality of an effect, and uses Bayes's Theorem to calculate from them the *critical prior interval* (CPI) which renders the reality of the claimed effect not credible (i.e. which leads to a posterior level of belief which fails to exclude no effect at the 95% level). In practical terms, the criterion for the assessment of credibility is thus as follows:

If previous evidence indicates that plausible values for the parameter in question exist outside the critical prior interval (*CPI*), *the reality of the stated effect may be deemed credible at the* 95% *level.*

For example, the credibility of a finding based on an odds ratio and 95% CI would be assessed by calculating the CPI from the stated confidence interval, and assessing whether current knowledge is consistent with the existence of plausible odds ratios beyond this critical interval. If plausible ORs do exist beyond the CPI, then the overall findings may be regarded as "credible at the 95% level".

6. Derivation of the critical prior interval (CPI)

To fix ideas, I shall henceforth focus on log-normally distributed ORs as the means of summarising both the new data and the assessment of credibility. In deriving the Critical Prior Interval, I shall also assume this interval to be symmetrical in log-odds about a mean value of zero. The credibility assessment is thus being made on the basis of "cautious reasonable scepticism" (Kass and Greenhouse, 1989), consistent with the judiciously sceptical stance conventionally adopted in scientific research.

With this assumption, it can be shown (see Appendix) that the CPI corresponding to the standard 95% CI for a stated OR of (L_D, U_D) has a range of $(L_0, 1/L_0)$, where L_0 is given by

$$L_0 = \exp\left\{ \frac{-0.5[\ln(U_D/L_D)]^2}{\sqrt{[\ln(U_D L_D)]^2 - [\ln(U_D/L_D)]^2}} \right\}. \tag{3}$$

We can now formally state the criterion for gauging credibility: an odds ratio with a standard 95% CI of (L_D, U_D) may be deemed credible at the 95% level if prior knowledge indicates that plausible values for the OR do exist outside the critical prior interval $(L_0, 1/L_0)$.

In practice, this criterion may be simplified by noting that we need consider only whether plausible values for the OR exist *below* the calculated L_0 for CIs < 1.00 or *above* $1/L_0$ for CIs > 1.00. In either case, if the ORs meet these criteria, the finding may be deemed credible at the 95% level, with the CPI being stated as $(L_0, 1.00)$ or $(1.00, U_0)$.

While (3) allows direct calculation of the CPI range from any OR range (L_D, U_D), an indication of the credibility of an OR can also be read off from the nomograph in Fig. 1, which gives L_0 for any 95% CI range (L_D, U_D).

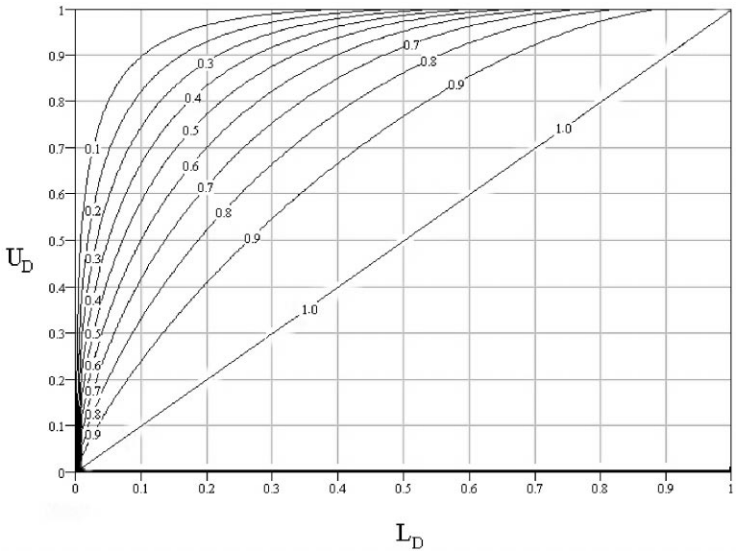

Fig. 1. CPI contours for an OR with 95% CI of (L_D, U_D). The nomograph gives the bounds for the CPI for ORs where U_D is less than 1.00. It may also be used with ORs for which L_D is greater than 1.00 by replacing L_D and U_D by their reciprocals, and calculating $U_0 = 1/L_0$. For example, an OR of 1.9 (1.5, 2.5) has a CPI corresponding to the reciprocal of the CPI for 0.53 (0.40, 0.67). From the above nomograph, the latter CI gives an L_0 of 0.9, and so the CPI for the results as stated is $(1.00, U_0)$ where $U_0 = (0.9)^{-1} = 1.1$.

Some features of (3) are worth highlighting. First, the CPI can be calculated directly and objectively from results quoted in standard (frequentist) terms. The element of judgement thus comes not in setting the prior (as in "direct" Bayesian inference), but in assessing whether prior knowledge indicates that plausible values of the OR lie outside the calculated CPI. The credibility assessment is also entirely transparent, allowing anyone to make their own judgement as to the credibility of a finding. Such judgements are, of course, routinely performed by all researchers: the CPI approach merely formalises it.

Finally, by taking explicit account of the *width* of confidence intervals, credibility assessments based on CPIs deal explicitly with a widespread concern about many studies: insufficient sample size. Implausible yet statistically significant findings based on small samples are often criticised by arguing that their significance is based on insufficient numbers. This is, however, a fallacy: the calculation of statistical significance takes account of sample size, and thus cannot be undermined by appeal to small sample size. Nevertheless, the intuitive notion that sample size affects the *credibility* of a finding is well founded, and is clearly reflected in CPIs. A finding based on a large sample, leading to a tight CPI, will be harder for sceptics to dismiss, as they must then justify their belief that this tight CPI contains all plausible values for the OR. On the other hand, a finding based on a small sample, with a relatively broad CI, leads to a correspondingly broad CPI encompassing a relatively wide range of ORs — making the existence of plausible values beyond this range harder to justify.

7. Credibility assessment: a worked example

As an example of credibility assessment in action, consider a recent case-control study of the impact of simple lifestyle modifications on cardiovascular risk (Spencer et al., 1999). This found that among men aged 27–64, taking non-vigorous exercise and avoidance of added salt were both associated with statistically significant reduced risks of acute myocardial infarction (AMI). For non-vigorous exercise, an OR of 0.5 with a standard 95% CI of $(0.4, 0.7)$ was found, while avoidance of added salt led to an OR of 0.6 and 95% CI of $(0.4, 0.9)$.

Thus, at first glance, both these lifestyle modifications seem to have produced impressive and statistically significant risk reductions. However, calculation of their corresponding CPIs highlights a crucial difference between these two findings (see Fig. 2).

For non-vigorous exercise, Eq. (1) — or Fig. 1 — shows that the 95% CI of (0.4, 0.7) leads to a CPI of $(0.87, 1.00)$. Thus the stated impact of non-vigorous exercise can be regarded as statistically significant and credible if existing experience suggest that such exercise could reasonably be expected to produce at least a 13% reduction in the odds of AMI. Given the wealth of evidence that mild to moderate exercise confers a substantial level of protection against heart disease (Berlin and Colditz, 1990; Manson et al., 1999; Lemaitre et al., 1999) the modest 13% improvement required for credibility by the CPI seems eminently reasonable. We may thus reasonably conclude that, in the

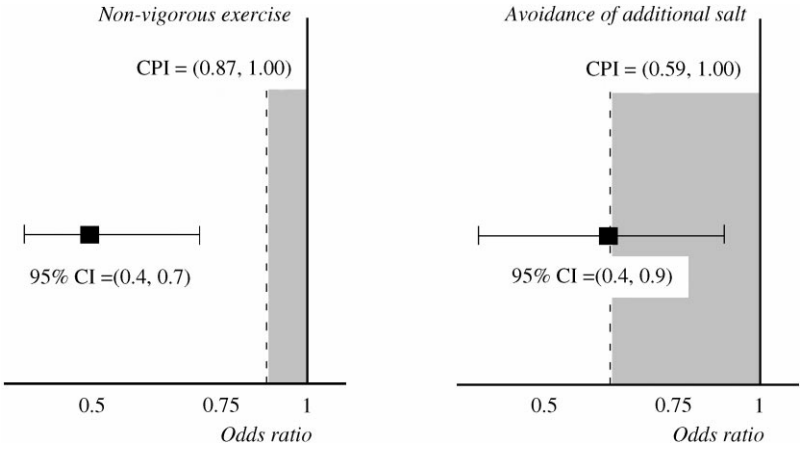

Fig. 2. Comparison of the CPIs for the results of Spencer et al. (1999) concerning the impact on acute myocardial infarction risk of (a) non-vigorous exercise and (b) avoidance of additional salt. While the odds ratios for both findings are statistically significant and apparently impressive, their CPIs are markedly different, with implications for the credibility of the two findings. Specifically, the result for non-vigorous exercise may be regarded as *credible* as well as statistically significant at the 95% level if existing evidence supports the possibility of such exercise cutting AMI risk to an OR of 0.87 or lower; there is ample evidence to supports such a view. In contrast, the findings regarding salt avoidance are credible only if existing evidence points to this measure being capable of producing an OR as low as 0.59. So great a risk reduction lacks evidential support.

light of existing knowledge, the study's positive finding concerning non-vigorous exercise is both statistically significant and credible at the 95% level.

A quite different conclusion emerges in the case of salt avoidance, however. Here the stated 95% CI leads via (7) — or Fig. 1 — to a CPI $(0.59, 1.00)$. Thus, this finding may only be regarded as credible if existing knowledge indicates that avoidance of added salt can plausibly produce at least a 41% reduction in AMI risk. While there is considerable evidence that salt reduction does confer some reduction in blood pressure (He et al., 1999; MacGregor and de Wardener, 1998) so large a reduction in AMI risk merely from avoiding added salt lacks any evidential support. Thus the study's finding concerning the apparently dramatic effect on AMI risk of this health measure is statistically significant, but it is not credible.

This worked example highlights a number of features about CPIs and their use in assessing credibility. First, it shows that CPIs allow a clear distinction to be made between *statistical significance* and *credibility*. In the above example, both results were statistically significant; however, only the CPI for moderate exercise was sufficiently tight to make the finding credible in the light of current knowledge.

Second, the differences in the credibility of the two health measures can be traced to the different levels of evidential weight concerning the impact of moderate exercise and salt avoidance: the 95% CI for the former is much tighter, and thus impressive, than that of the latter. All too often, new research findings are assessed just in terms of whether their 95% CIs exclude no effect, while ignoring the width of the CI. Use of the CPI allows the crucial issue of weight of evidence to be included in the assessment.

Finally, the example shows how CPIs provide a clear *quantitative* focus for judging the credibility of a research finding. While it is easy to provide broad-brush qualitative arguments for why both moderate exercise and salt avoidance might produce reduction in AMI risk, defending the plausibility of specific ORs is clearly much more demanding.

CPIs thus require that those reporting new findings do more than merely demonstrate statistical significance and walk away. They provide sceptics and advocates alike with an explicit and transparent base-line around which to base their case. As such, CPIs would seem to have an obvious role to play in the discussion section of papers describing new clinical findings.

8. Conclusion

The relatively minor inroads that Bayesian analysis has made into mainstream medical journals is pretty disheartening for those who (like the author) believe it has much to offer working clinicians. I have argued here that the blame lies principally with the strategy adopted by many Bayesians (including, again, myself) for convincing others to take up Bayesian methods: Cassandra-like warnings of the terrible fate that awaits those who persist in their frequentist habits. In this paper, I have tried to show why these dire predictions have not manifested themselves in the form most likely to impress the typical working scientist: as inferential scandals incontrovertibly tied to the

failings of frequentism. The fact is that there are always other statistical scapegoats on hand to deflect attention from the worst effects of these failings.

Given the apparent futility of this somewhat negative strategy, I have here explored another, which is more pragmatic and positive: that of showing how Bayesian methods can "add value" to existing frequentist methods. In particular, I have outlined a method of assessing the credibility of new findings, based on the attractive ability of Bayesian techniques to allow extant knowledge to be factored in to the assessment of new evidence.

The central concept involved — the Critical Prior Interval — allows a new finding stated in conventional frequentist terms to be compared with extant knowledge, and its credibility assessed. Strong evidence for reasonable effects leads to CPIs easily justified on the basis of existing knowledge; weak evidence for a dramatic effect, in contrast, gives CPIs that make clear the inherent lack of credibility of the claim.

It cannot be too strongly stressed, however, that the numerical value of the CPI per se does not determine the credibility of a finding: there is no sense in which a particular CPI value could become the equivalent of the ubiquitous "$p < 0.05$" hurdle over which research findings must jump. Each case has to be taken on its merits: a narrow CPI for one finding may actually be easier to justify in some cases than a much broader CPI in others. It is only when the CPI is viewed in the light of existing knowledge for each claim that credibility can properly be assessed.

Even so, there may be concern that CPIs could become a(nother) stick with which cynical referees can beat authors. The basis of CPIs is such, however, that only authors whose claims fly in the face of extant knowledge and lack persuasive evidential weight have much to worry about — and few would lament the demise of such claims in any case.

In short, the use of CPIs should seem unreasonable only to those making unreasonable claims: either advocates claiming extraordinary effects on the basis of weak evidence, or sceptics refusing to accept strong evidence for a plausible hypothesis.

Do CPIs provide a panacea to the problem of assessing clinical evidence? Of course not: research findings whose CPIs imply credibility at the 95% level may still be rendered unreliable by such effects as confounding or bias. But to reject CPIs on this basis is like refusing to take paracetamol for a headache because it doesn't cure all known diseases. Assessing the credibility of research findings is a regular headache for working clinicians, and CPIs are a statistically well-founded aid for distinguishing clinical reality from statistical artefact.

Perhaps the most valuable role of CPIs, however, would be in allowing clinicians to feel less apprehensive about taking that crucial step from being interested in Bayesian inference to actually exploiting its power in their everyday work.

Acknowledgements

It is a pleasure to thank David Spiegelhalter, Dennis Lindley and Mark Selinger for their constructive comments on an early draft, and Iain Chalmers, Richard Lilford and Catherine Elsworth for valuable discussions.

Appendix derivation of the critical prior interval for a 95% CI

Suppose the outcome of research study is represented by a parameter distributed as $N(\mu_D, \varphi_D)$, corresponding to a 95% CI of (L_D, U_D) where

$$L_D = \mu_D - 1.96\sqrt{\varphi_D}, \tag{A.1}$$

$$U_D = \mu_D + 1.96\sqrt{\varphi_D}. \tag{A.2}$$

Bayes's Theorem provides the means of combining evidence in this form with prior knowledge, captured via a 95% CI of (L_0, U_0). The result is a posterior distribution corresponding to a 95% CI of (L_c, U_c) in which

$$L_c = \mu_c - 1.96\sqrt{\varphi_c}, \tag{A.3}$$

$$U_c = \mu_c + 1.96\sqrt{\varphi_c}, \tag{A.4}$$

where μ_c and φ_c are calculated from (see e.g. Lee, 1997, Chapter 2)

$$1/\varphi_c = 1/\varphi_0 + 1/\varphi_D, \tag{A.5}$$

$$\mu_c = \varphi_c[(\mu_0/\varphi_0) + (\mu_D/\varphi_D)]. \tag{A.6}$$

Conventionally, both the study outcome and prior knowledge are known, allowing the calculation of the posterior "credible range" (L_c, U_c); if this posterior range includes parameter values corresponding to no effect, one may then regard the conclusion that a real effect has been detected as "not credible at the 95% level".

Following Good, however, one may turn this process around, and *calculate* the prior level of belief — the critical prior interval (CPI) — required to render a finding not credible at the 95% level. This is the basis of the method of plausibility assessment put forward in this paper.

Specifically, we seek the prior interval (L_0, U_0) such that study findings of the form (L_D, U_D) lead to a posterior interval whose 95% CI bounds encompass no effect. Supposing that no effect corresponds to a parameter value of zero, we are thus seeking a prior interval (L_0, U_0) such that $\mu_c = \pm 1.96\sqrt{\varphi_c}$. Taking this prior distribution to be $N(0, \varphi_0)$, i.e. centred on no effect according to the "cautious reasonable scepticism" principle of Kass and Greenhouse (1989), the above equations allow us to fix φ_0 in terms of the μ_D and φ_D extracted from the stated 95% CI:

$$\varphi_0 = [(\mu_D/\varphi_D)^2/3.84 - (1/\varphi_D)]^{-1}. \tag{A.7}$$

This variance, together with $\mu_0 = 0$, completely specifies the CPI capable of rendering a given 95% CI not credible at the 95% level. In the special case of odds ratios and their 95% CIs (L_D, U_D), taken to be log-normally distributed, we have

$$\mu_D = [\ln(U_D) + \ln(L_D)]/2, \tag{A.8}$$

$$\varphi_D = [0.255 \ln(U_D/L_D)]^2. \tag{A.9}$$

Inserting these into (A.7) and converting back into a 95% CI, we find that the CPI for an Odds Ratio with 95% CI of (L_D, U_D) is given by $(L_0, 1/L_0)$ where

$$L_0 = \exp\left\{ \frac{-0.5[\ln(U_D/L_D)]^2}{\sqrt{[\ln(U_D L_D)]^2 - [\ln(U_D/L_D)]^2}} \right\}. \tag{A.10}$$

If prior knowledge is such that OR values lying beyond this CPI are plausible, then the overall finding may also be deemed "credible at the 95% level".

References

Berger, J., Delampady, M., 1987. Testing precise hypotheses. Statist. Sci. 2, 317.

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

Berlin, J.A., Colditz, G.A., 1990. A meta-analysis of physical activity in the prevention of coronary heart disease. Amer. J. Epidemiol. 132, 612.

Bland, J.M., Altman, D.G., 1998. Bayesians and frequentists. British Med. J. 317, 1151.

Bourke, G.J., Daly, L.E., McGilvray, J., 1985. Interpretation and Uses of Medical Statistics., Third Edition. Mosby, St Louis.

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

Egger, M., Schneider, M., Davey Smith, G., 1998. Meta-analysis: spurious precision? Meta-analysis of observational studies. British Med. J. 316, 140.

Good, I.J., 1950. Probability and the Weighing of Evidence. Griffin, London.

He, J., Ogden, L.G., Vupputuri, S., Bazzano, L.A., Loria, C., Whelton, P.K., 1999. Dietary sodium intake and subsequent risk of cardiovascular disease in overweight adults. J. Amer. Med. Assoc. 282, 2027.

Heyes, S., Hardy, M., Humphreys, P., Rookes, P., 1993. Starting Statistics in Psychology and Education., Second Edition. Weidenfeld & Nicolson, London.

Hines, T.M., 1998. Comprehensive review of biorhythm theory. Psychol. Rep. 83, 19.

Jeffreys, H., 1961. Theory of Probability., Third Edition. University Press, Oxford.

Kass, R.E., Greenhouse, J.B., 1989. Comments on "Investigating therapies of potentially great benefit: ECMO" (by Ware, J.H.). Statist. Sci. 4, 310.

Lee, P.M., 1997. Bayesian Statistics: An Introduction., Second Edition. Arnold, London.

Lemaitre, R.N., Siscovick, D.S., Raghunathan, T.E., Weinmann, S., Arbogast, P., Lin, D.Y., 1999. Leisure-time physical activity and the risk of primary cardiac arrest. Arch. Internat. Med. 159, 686.

Lilford, R.J., Braunholtz, D., 1996. The statistical basis of public policy: a paradigm shift is overdue. British Med. J. 313, 603.

MacGregor, G.A., de Wardener, H.E., 1998. Salt, Diet & Health. University Press, Cambridge.

MacRae, A.W., 1995. Statistics in A-level psychology: a suitable case for treatment? Psychologist 8, 363.

Manson, J.E., Hu, F.B., Rich-Edwards, J.W., Colditz, G.A., Stampfer, M.J., Willett, W.C., Speizer, F.E., Hennekens, C.H., 1999. A prospective study of walking as compared with vigorous exercise in the prevention of coronary heart disease in women. N. Engl. J. Med. 341, 650.

Matthews, R.A.J., 2000. Facts versus factions: the use and abuse of subjectivity in scientific research. Working Paper, European Science and Environment Forum, Cambridge, reprinted in: Morris, J. (Ed.), Rethinking Risk and the Precautionary Principle. Butterworth-Heinemann, Oxford, pp. 247–282.

Spencer, C.A., Jamrozik, K., Lambert, L., 1999. Do simple prudent health behaviours protect men from myocardial infarction? Internat. J. Epidemiol. 28, 846.

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

Spiegelhalter, D.J., Myles, J.P., Jones, D.R., Abrams, K.R., 1999. An introduction to Bayesian methods in health technology assessment. British Med. J. 319, 508.