Drug Information Journal, Vol. 35, pp. 1469–1478, 2001
Printed in the USA.
0092-8615/2001

METHODS FOR ASSESSING THE CREDIBILITY OF CLINICAL TRIAL OUTCOMES

ROBERT A. J. MATTHEWS
Visiting Research Fellow, Department of Information Engineering, Aston University, Birmingham, England

Credibility—the believability of new findings in the light of current knowledge—is a key issue in the assessment of clinical trial outcomes. Yet, despite the growth of evidence-based medicine, credibility is usually dealt with in a broad-brush and qualitative fashion. This paper describes how Bayesian methods lead to quantitative credibility assessments that take explicit account of prior insights and experience. A simple technique based on the concept of the critical prior interval (CPI) is presented, which allows rapid credibility assessment of trial outcomes reported in the standard format of odds ratios and 95% confidence intervals. The critical prior interval is easily determined via a graph, and provides clinicians with an explicit and objective baseline on which to base their assessment of credibility. The use of the critical prior interval is demonstrated through several working examples.

Key Words: Credibility; Critical prior interval; Bayesian methods

INTRODUCTION

THE OUTCOME OF MOST clinical research is now stated in quantitative terms, a trend encouraged by the emergence of evidence-based medicine. Typically a single number, such as an odds ratio (OR), is used to capture the essence of the research finding, for example, the OR for mortality after five years. This number is conventionally accompanied by a measure of the probability of the result emerging through the play of chance, such as a 95% confidence interval (CI). If this measure meets a preestablished criterion, for example, the 95% CI excludes an OR of 1.00, corresponding to no effect—the finding is held to be "statistically significant."

While specific statements of statistical significance are routinely included in reports of new findings, discussion of their *credibility*—that is, their believability in the light of existing knowledge—usually consists of broad-brush arguments based on the outcome of previous research. Such arguments are, however, all too easy to devise; Egger et al. (1) cite a case where the authors of two mutually contradictory studies were both able to supply apparently reasonable qualitative credibility arguments for their conflicting findings.

The desire to include some quantitative measure of credibility has prompted the illegitimate use of statistical significance as a surrogate for such a measure. However, the warnings of statisticians that "significance" and "confidence" have specific technical meanings with no direct connection with

Based on a presentation from the DIA Workshop "Statistical Methodology in Clinical R&D," April 2–4, 2001, Vienna, Austria.

Reprint address: Robert Matthews, 47 Victoria Road, Oxford, OX2 7QF, UK. E-mail: r.matthews@physics.org.

their everyday sense (2) seem largely to go unheeded.

Nevertheless, there remains a clear need for a quantitative measure of credibility that is statistically well-founded, easy to use, and capable of unambiguous interpretation by working clinicians. This paper describes how the methods of Bayesian inference lead to such a measure.

BAYESIAN METHODS

The statistical concepts now in most common use are based on the so-called frequentist interpretation of probability. As its name suggests, this assigns a probability $Pr(E)$ to an event E (say, a specific patient benefiting from a drug), on the assumption that E can be precisely repeated many times. In other words, frequentist methods treat all events as if they are coin-tosses or throws of a die: while one trial may not give the required result, in the long run the proportion of trials that do will tend toward $Pr(E)$. Of course, such a view of probability is hard to reconcile with the realities of clinical medicine, where, short of a visit to parallel universes, there is no hope of a large number of trials with identical patients under identical conditions. In contrast, Bayesian methods view probabilities not as idealized long-run frequencies, but as degrees of belief based on all the available evidence. This is increasingly being recognized as a much more relevant interpretation in many situations, and there is now a large literature explaining the origins, philosophy, and applications of Bayesian methods in statistical inference in general (3) and medicine in particular (4,5). For our purposes the key feature of Bayesian methods is that—again, in contrast to conventional ("frequentist") statistical methods—they allow new findings to be set in their proper context, and viewed in the light of other relevant sources of insight, such as previous trials, *in vitro* studies, and real-life experience on the wards.

Controversy surrounds the use of Bayesian methods when this external evidence takes the form of subjective opinion. While the informal use of such judgment is widespread, its formal acceptance into the assessment of clinical evidence has led to fears that Bayesian methods turn inference into a free-for-all, where anyone can reach any conclusion about any clinical trial finding.

Advocates of Bayesian methods have made strenuous efforts to counter this criticism, but suspicions remain. Nor are they wholly without foundation; for example, the use of a panel of "experts" as the source of subjective evidence raises concern that such experts can prove overly optimistic—or conservative—in their views. Furthermore, those clinicians not included among the "experts" may justifiably feel excluded from the resulting assessment, and yet have no clear idea of how to replace the "experts' " opinions with their own. Methods such as Bayesian elicitation and robustness analysis have been devised to tackle these problems, but their use is hardly elementary.

Despite these caveats, Bayesian methods cannot simply be dismissed as an unnecessary complication to statistical inference. "Conventional" statistical methods, for all their familiarity, do not address key questions in the assessment of data, and are all too easily misinterpreted (6,7). Bayesian methods, in contrast, can supply easily-interpreted answers to questions of direct relevance to researchers—including the credibility of new findings.

USING BAYESIAN METHODS TO ASSESS CREDIBILITY

In essence, Bayesian methods provide the means for combining existing insight with new findings to arrive at an updated level of insight; or, in the language of Bayesian inference,

$$\text{Prior insight} \oplus \text{data} \Rightarrow \text{Posterior insight.} \quad (1)$$

For example, if the outcome of a clinical trial is summed up by an OR and the associated 95% CI, Bayesian methods allow us to combine this result with prior insight expressed in the same format to produce an updated

level of insight, in the form of a "posterior" OR together with a "95% credible interval" (see Figure 1).

If this latter excludes no effect (OR = 1.00), then in the light of prior insight, the new result may be deemed *credible at the 95% level*. Crucially, such credibility is not in general equivalent to a result being significant at this level; one can show (see Appendix) that the two intervals are equivalent only in the special case of prior insight being so vague that any outcome OR is as plausible as any other. In all other cases, prior knowledge will lead to credible intervals that do not coincide with confidence intervals. Indeed, results with standard 95% CI excluding ORs of 1.00 may nevertheless have posterior credible intervals that fail to exclude OR = 1.00: in other words, results can be statistically significant, but not credible in the light of what is already known.

Such cases, which will be encountered below, highlight the impact of setting new data into their proper context. It is the ability of Bayesian methods to do this that makes them ideal for assessing the credibility of new findings.

Any practical technique for credibility assessment should have the following 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 all clinicians to gauge the credibility of the finding in the light of their own experiences,
3. It should reflect the effect of sample size, with a result obtained from a small trial possessing less credibility than the identical result from a large trial, and
4. It should be easy to calculate from conventional statistical measures, and have a straightforward interpretation.

These desiderata are all met within the framework of Bayesian methods, which capture the issue of credibility in the following formal form:

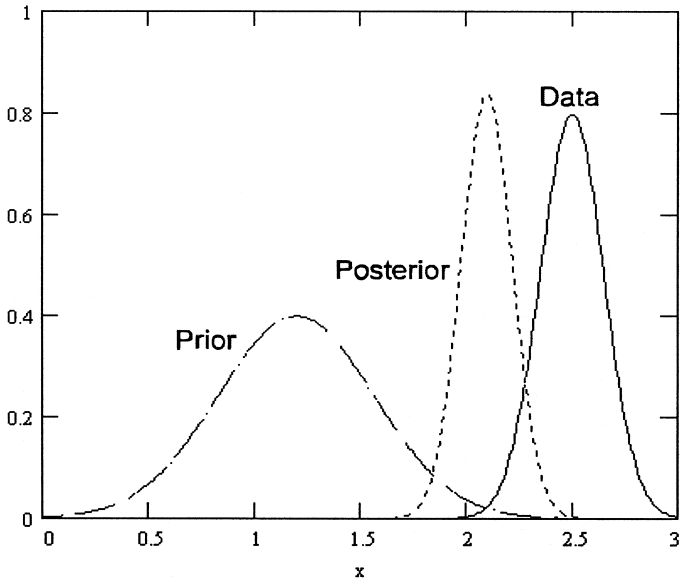

FIGURE 1. Relationship between Bayesian prior, posterior, and data. Prior knowledge and insight is captured by a prior distribution (here, a Normal distribution with relatively large variance, reflecting a certain level of uncertainty). Combined with data that have been acquired, the result is an updated "posterior" level of knowledge, with smaller variance, and a mean value determined by the relative evidential weight of the prior knowledge and the data.

Definition: A result may be deemed credible at the 95% level if, when combined with prior knowledge, the new evidence leads to a posterior 95% credible interval that excludes no effect.

To demonstrate the use of this criterion, consider the outcome of a randomized controlled trial presented in terms of a central OR and associated 95% CI (L_D, U_D). Suppose that prior knowledge indicates that ORs above a level U_o or below L_o are implausible, that is, that there is less than a 5% probability of ORs that lie outside the interval (L_o, U_o). Using Bayes's Theorem to combine this prior interval with that obtained from the trial (see Appendix) then leads to an updated level of evidence of (L_P, U_P); if this posterior 95% credible interval excludes OR = 1, the finding may be deemed credible in the light of the stated prior knowledge.

An Example of Bayesian Credibility Analysis

An impressive demonstration of this approach is due to Pocock and Spiegelhalter (8), who used it in a credibility assessment of the controversial findings of the GREAT (9) study of early use of anistreplase. The study compared mortality among patients with suspected acute myocardial infarction given anistreplase at home with those given anistreplase in the hospital around two hours later. The initial results pointed to an impressively substantial and statistically significant decrease in three-month mortality among those treated early: OR = 0.47 (0.23, 0.97).

Pocock and Spiegelhalter argued that despite its apparent statistical significance, so dramatic a reduction was implausible in the light of prior evidence. They quantified this prior evidence by a 95% interval of (0.6, 1.0); that is, they argued that credible ORs most likely lay somewhere between the extremes of no benefit (OR = 1.00) and a $\sim$ 40% reduction in mortality. Combining this prior range with the 95% CI of the GREAT trial (using the method set out in the Appendix), they obtained a posterior OR of $\sim$ 0.73 (0.6, 0.9). Pocock and Spiegelhalter thus concluded that if the GREAT results were seen in the light of prior experience, the true impact of early anistreplase on mortality was likely to be much less impressive than suggested by the GREAT trial alone.

Subsequent research has confirmed this view. A recent meta-analysis by Morrison et al. (10) points to a reduction in mortality from early anistreplase equivalent to an OR of 0.83 (0.70, 0.98), in line with the relatively modest mortality reduction predicted eight years earlier on the basis of the Bayesian analysis.

This example highlights a number of important issues. First, it shows how the inclusion of prior knowledge can substantially change perceptions of apparently impressive trial results. Second, it underlines the warnings of statisticians that "statistical significance" is a poor means of assessing the credibility of a trial outcome. The GREAT study result was statistically significant, but this merely reflects the fact that the upper limit of its 95% CI excluded no effect. More important for the credibility of the result is the fact that the overall interval was relatively wide, reflecting the relatively small size of the trial (around 300 patients). In the context of conventional statistics, sample size usually raises concerns about inadequate power, that is, the risk of failing to detect small effects that are actually present. However, Bayesian methods (and, indeed, common sense) raise broader concerns about the reliability of any conclusion, positive or negative, based on small samples. Wide confidence intervals point to low evidential weight, and it was this that led to the GREAT result undergoing so much "shrinkage" toward no effect when combined with the prior evidence: it simply lacked the evidential weight to overwhelm the prior evidence that any reduction in mortality from early use of anistreplase was likely to be fairly modest.

This raises this key issue of the use of prior evidence in credibility assessment. Critics of Bayesian methods frequently claim that the use of such evidence will lead to anarchy, with every clinician reaching different views about the same evidence. What

actually happens, however, is that as evidence accumulates, the role of prior belief on the Bayesian assessment of credibility becomes progressively less important: that is, "the truth will out." To see this, consider the impact of the evidence from Morrison et al.'s meta-analysis, summarized by (0.70, 0.98), on the state of knowledge in 1992 as summarized by Pocock and Spiegelhalter's prior of (0.6, 1.0). Using Bayes's Theorem to combine the two leads to a posterior interval of (0.71, 0.93), which is virtually identical to Morrison et al.'s original interval; that is, the evidential weight provided by the meta-analysis made the level of knowledge available in 1992 more or less redundant: the clinical trial data are starting to "speak for themselves." Furthermore, as the evidence accumulates, it will take a progressively narrow and skeptical view of the likely efficacy to counteract the weight of evidence, and undermine the credibility of the new findings.

Thus, far from creating anarchy, Bayesian assessments compel explicit statements to be made of why a specific trial outcome is, or is not, viewed as credible. Skeptics and enthusiasts alike must demonstrate that when their view of extant knowledge is combined with the new findings, the resulting posterior interval justifies their particular stance. By being compelled to state the precise prior that leads sceptics or enthusiasts to their conclusion, the plausibility of their stance becomes visible to all.

A SIMPLE BAYESIAN CREDIBILITY TEST

The above method of credibility assessment requires an explicit statement of prior knowledge. While this can often be relatively easy to give, the setting of prior intervals is undoubtedly the most controversial element of Bayesian methods. As we have seen, some of these criticisms, such as the claim that it leads to inferential anarchy, are not well-founded. Even so, it cannot be denied that credibility analysis based on a specific prior interval is a somewhat invidious process, in that it demands that clinicians seeking to base decisions on the analysis must "buy into" the same prior knowledge.

It is, of course, possible to use the methods in the Appendix to substitute any stated level of prior knowledge with one's own. However, the need for such recalculation can be obviated by using the concept of the CPI, a statistical parameter which can be calculated for any quoted 95% CI, and which provides a convenient standardized means of assessing the credibility of a trial outcome. The technical definition of the CPI and its derivation are given in the Appendix; in essence, the CPI shows what ORs one must already deem plausible in order to regard the outcome of a new trial as credible. As is shown in the Appendix, the CPI assumes a prior distribution symmetrical about an OR of 1.00, consistent with the ethical assumptions underpinning the randomization of patients to either arm of the trial. This, in turn, implies that the CPI corresponding to an OR 95% CI of (L_D, U_D) is given by $(L_o, 1/L_o)$ where

$$L_O = \exp\left\{-\frac{[\ln(U_D/L_D)]^2}{4\sqrt{\ln(U_D)\ln(L_D)}}\right\}. \qquad (2)$$

We can now formally state a standardized criterion for gauging credibility:

Credibility criterion: if prior knowledge indicates that plausible OR values exist outside the CPI of $(L_o, 1/L_o)$, then an OR with a 95% CI of (L_D, U_D) may be deemed credible at the 95% level.

In practice, for trial outcomes in which ORs < 1.00 indicate greater efficacy, the relevant CPI may be taken as $(L_o, 1.00)$; similarly, for outcomes where ORs > 1.00 indicate greater efficacy, the relevant CPI is $(1.00, 1/L_o)$, again calculated via (2).

While equation (2) allows direct calculation of the CPI from any 95% CI (L_D, U_D), an indication of the credibility of an OR can also be read off from the nomograph in Figure 2, which gives L_o for any 95% CI (L_D, U_D).

The assessment of credibility using CPIs

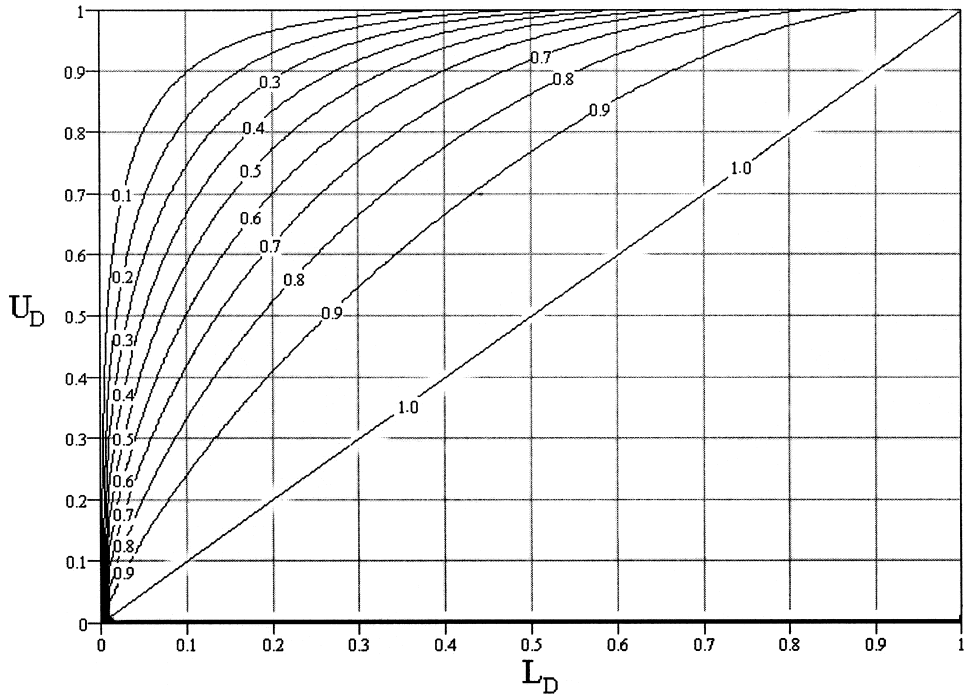

FIGURE 2. CPI contours for an OR with 95% CI of (L_D, U_D). The nomograph gives the bounds for the CPI for ORs where U_D is less than 1.00. It may also be used with ORs for which L_D is greater than 1.00 by replacing L_D and U_D by their reciprocals, and calculating $U_o = 1/L_o$. For example, an OR 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_o of 0.9, and so the CPI for the results as stated is (1.00, U_o) where $U_o = (0.9)^{-1} = 1.1$.

has obvious similarities to the standard method for assessing statistical significance. As such, the two may be combined in the following two-step procedure:

1. The OR is *significant* at the 95% level if an OR of 1.00 lies outside the calculated 95% confidence interval (CI), and
2. The OR is also *credible* at the 95% level if prior experience indicates that plausible ORs lie outside the CPI.

Some features of step 2 should be noted. First, the CPI is calculated directly and objectively from results quoted in standard (frequentist) terms. Second, the credibility assessment is entirely transparent, allowing all clinicians to make their own judgment as to the credibility of a finding. Finally, by taking explicit account of the width of confidence intervals, credibility assessments based on CPIs deal explicitly with the issue of sample size. The larger the trial, the tighter the resulting 95% CI, and as Figure 2 shows, the tighter the bounds of the CPIs become, until only a very limited range of ORs are capable of rendering the new finding not credible. Thus, as sample size increases, skeptics are progressively "forced into a corner" if they are to maintain their stance in the face of the accumulated evidence. On the other hand, findings based on small samples with relatively broad 95% CIs will have broad CPIs encompassing a wide range of prior ORs, thus making the existence of plausible values outside this range harder for enthusiasts to justify.

Examples of Credibility Assessment Using CPIs

GREAT Study of Early Anistreplase. To illustrate these points, consider again the results of the GREAT study. The stated 95% CI of (0.23, 0.97) excludes an OR of 1.00, and is thus *statistically significant* at the 95% level. However, the same interval leads via either equation (2) or the nomograph in Figure 1 to a CPI of (0.1, 1.00). Thus, while the outcome may be statistically significant, it can only be regarded as *credible* if prior experience indicates that early use of anistreplase can produce ORs for mortality of 0.1 or better; that is, only those already convinced that giving anistreplase two hours earlier can produce a 90% mortality reduction can consider the GREAT study result to be credible. This credibility assessment would thus lead many, if not most, clinicians to regard the GREAT study result with considerable skepticism. Such a conclusion again reflects the lack of evidential weight of the GREAT study, as witnessed by a very broad CPI, which extends down to implausibly impressive ORs. Furthermore, as we have seen, such skepticism has since proved justified in the light of subsequent trial results.

Morrison et al. Meta-analysis for Early Anistreplase. Now consider again Morrison et al.'s meta-analysis, with its 95% CI of (0.70, 0.98). Like the GREAT finding, this result is statistically significant, but its corresponding CPI is much tighter: (0.72, 1.00), which reflects the much greater evidential weight contained within the meta-analysis. The CPI also implies that the meta-analysis result may be regarded as credible by those whose prior experience leads them to consider that early use of anistreplase can produce at least a 28% reduction in mortality. This is much lower than the > 90% reduction demanded in order to make the GREAT study outcome credible. However, it may still be regarded as implausible; this is an issue that could conveniently be addressed in a discussion section following on from an explicit calculation of the CPI. The contribution of the CPI is to set such a discussion on a firm, quantitative basis.

Subcutaneous Sumatriptan for Migraine. The previous examples have shown how sample size can dramatically affect the credibility of new findings. This may prompt concern that small studies pointing to dramatic effects are likely to suffer at the hands of a CPI credibility analysis. To show that this is not the case, consider the results of an early study of sumatriptan (11), the 5-HT agonist which has proved a highly effective treatment against migraine. The results of this early study were dramatic, with 79% of patients given subcutaneous injections of 8 mg of sumatriptan reporting an improvement in symptoms, compared with only 25% of those given placebo; the overall OR on reporting improvement was 11.4, with a 95% CI of (6.00, 21.5). With only around 100 patients in each arm, however, the study carries only modest evidential weight: witness the relatively broad 95% CI. Nevertheless, so impressive was the study's overall result that it still produces a CPI of (1.00, 1.20); that is, the result is credible at the 95% level unless modest levels of efficacy above ~20% are deemed implausible. Thus, despite being relatively small, this study presents surprisingly credible evidence for the effectiveness of sumatriptan. Subsequent trials have confirmed the credibility of this early study, with sumatriptan now regarded as a major breakthrough in migraine relief.

ACEi Treatment for Acute Myocardial Infarction. As the previous example shows, the CPI does not automatically penalize the results from small studies. Nor does it automatically penalize modest effect sizes. Consider a recent systematic overview of data on the effectiveness of angiotensin-converting enzyme inhibitors (ACEi) in reducing mortality following acute myocardial infarction (12). The overall finding was an OR for 30-day mortality of 0.93, that is, a modest 7% reduction in mortality. Based on data from almost 100000 individuals, however, this OR was accompanied by a

relatively tight 95% CI of (0.89, 0.98), which thus leads to a relatively undemanding CPI of (0.95, 1.00), that is, although the study indicates that ACEi has only a relatively small impact on mortality, the finding remains credible at the 95% level unless a level of efficacy above just 5% is deemed implausible.

CONCLUSION

As evidence-based medicine continues to make in-roads into standard clinical practice, the need for quantitative measures of the credibility of study findings becomes ever more pressing. The concept of statistical significance cannot be used in this way: as statisticians have repeatedly (though unavailingly) emphasized, significance is a highly misleading concept of relatively meagre inferential interest. The aim of this paper has been to show how Bayesian methods provide a statistically valid framework for assessing the credibility of clinical trial outcomes. Specifically, the paper has demonstrated how the concept of the CPI provides a simple, standardized assessment of the credibility of new findings stated in the standard format of odds ratios and 95% confidence intervals.

As such, the CPI shows how Bayesian methods can "add value" to conventional statistical measures such as confidence intervals. This may help overcome the prevailing image of Bayesian methods as some weird technique wholly incompatible with standard statistical methods.

The examples of the CPI given in this paper will, it is hoped, also go some way to dispel the notion that Bayesian techniques lead to inferential anarchy, where the use of prior insights allows clinicians to reach any conclusion from the same set of data. Strong evidence for reasonable levels of efficacy produces relatively undemanding CPIs that will be deemed credible by all but the most skeptical clinician. In contrast, weak evidence for a dramatic effect leads to a CPI rendering the outcome credible only by the most ardent enthusiast. In short, the CPI compels enthusiasts and skeptics alike to justify their stance in quantitative terms.

It must be stressed, however, that the numerical value of the CPIs 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 CPI must be taken on its merits: a narrow CPI for one finding may be easier to justify in some cases than a much broader CPI in others. As befits any measure of credibility, it is only when viewed in the light of existing knowledge that the implications of the CPI can be assessed.

Do CPIs provide a panacea for the problem of misleading clinical study results? Of course not. While credibility in the light of existing knowledge is an important safeguard, it is still possible for studies to produce seemingly plausible results and yet ultimately prove unreliable. For example, a meta-analysis may be skewed by publication bias, producing an overly optimistic indication of efficacy. Nevertheless, just as the so-called funnel plot technique can highlight the presence of such bias (13), CPIs can alert clinicians to a lack of credibility in an otherwise "significant" study outcome. In specific cases, one may also argue over the use of a prior distribution symmetric about no effect in defining CPIs. As stated earlier, such a prior is well justified on ethical grounds in randomized clinical trials in general. There is, of course, nothing to prevent clinicians using a more general Bayesian approach, again outlined here, to reflect their own beliefs. However, in most cases CPIs should prove useful to working clinicians wanting a simple means of making sense of new clinical trial findings.

Acknowledgments—I am very grateful to Stephen Senn for prompting me to develop the above ideas, and to David Spiegelhalter, Mark Selinger, Dennis Lindley, and Catherine Elsworth for valuable discussions.

APPENDIX
THE MATHEMATICS OF BAYESIAN CREDIBILITY ASSESSMENT

Suppose the outcome of a study is represented by a parameter Normally distributed with mean μ_D and vari-

ance φ_D. This can be converted to a 95% CI of (L_D, U_D) via

$$L_D = \mu_D - 1.96\sqrt{\varphi_D} \quad \text{(A1)}$$

$$U_D = \mu_D + 1.96\sqrt{\varphi_D}. \quad \text{(A2)}$$

Bayes's Theorem provides the means of combining evidence in this form with prior knowledge, captured as a 95% CI of (L_o, U_o). The result is a posterior distribution corresponding to a *95% credible interval of* (L_P, U_P) in which

$$L_P = \mu_P - 1.96\sqrt{\varphi_P} \quad \text{(A3)}$$

$$U_P = \mu_p + 1.96\sqrt{\varphi_P}. \quad \text{(A4)}$$

with μ_P and φ_P calculated from (14)

$$1/\varphi_P = 1/\varphi_o + 1/\varphi_D \quad \text{(A5)}$$

$$\mu_P = \varphi_P[(\mu_o/\varphi_o) + (\mu_D\varphi_D)]. \quad \text{(A6)}$$

In the special case of odds ratios and their associated 95% CIs (L, U), one can calculate the various means and variances from

$$\mu = [ln(U) + ln(L)]/2 \quad \text{(A7)}$$

$$\varphi = [0.255 \; . \; ln(U/L)]^2. \quad \text{(A8)}$$

Equations A5 and A6 show that a 95% posterior credible interval coincides with the conventional 95% CI only in the special case where $\varphi_o \to \infty$; this corresponds to a stance of vague prior knowledge, where any OR is as plausible as any other; such a stance is rarely justifiable. Wherever a finite value of φ_o is used, the trial outcome OR is "pulled" toward the prior OR to an extent determined by the relative strength of the prior evidence and the evidence gathered during the trial.

In their analysis of the GREAT trial results (9), Pocock and Spiegelhalter (8) summarized prior knowledge concerning the early use of anistreplase via a prior OR interval of (0.6, 1.0), which from A7 and A8 above leads to $\mu_0 = -0.255$ and $\varphi_0 = 0.017$. Similarly, the GREAT 95% CI of (0.23, 0.97) gives $\mu_D = -0.755$ and $\varphi_D = 0.135$. From A5 we can now calculate the posterior variance: $\varphi_P = 0.015$; the posterior mean follows from A6: $\mu_P = -0.311$. Converting from natural logarithms and using A3 and A4 we finally obtain the posterior OR and 95% credible interval produced by combining the prior knowledge with the trial results: 0.73 (0.6, 0.9).

Derivation of the Critical Prior Interval

The above method inevitably leads to a credibility assessment based solely on the prior knowledge of one set of authors, and some effort is required to substitute a different prior interval. This lack of transparency can be overcome by following Good (15) and inverting the standard Bayesian procedure; that is, *calculating* what prior convictions are needed to render a finding credible at the 95% level; this is the basis of the CPI. To maintain consistency with the ethical assumption underpinning random clinical trials, the CPI assumes that patients in either arm are equally likely to benefit; mathematically, this implies that the interval representing prior knowledge is symmetrical about an OR of 1.00.

The CPI is then defined as the prior interval (L_o, U_o), such that when combined with the 95% CI from the clinical trial, the resulting posterior interval encompasses an OR of 1.00. Thus, if prior knowledge indicates that plausible ORs lie *outside* the CPI, the trial findings may be regarded as *credible at the 95% level*.

Supposing that no effect corresponds to a parameter value of zero, we are thus seeking a prior interval (L_o, U_o) such that $\mu_o = \pm 1.96\sqrt{\varphi_o}$. With the assumption that the prior distribution is N(0, φ_o), the above equations fix φ_o in terms of the μ_D and φ_D as extracted from the 95% CI of the trial outcome:

$$\varphi_o = [(\mu_D/\varphi_D)^2/3.84 - (1/\varphi_D)]^{-1}. \quad \text{(A9)}$$

This variance, together with $\mu_0 = 0$, completely specifies the CPI; for an odds ratio with 95% CI of (L_D, U_D), we find that the CPI is given by (L_o, $1/L_o$) where

$$L_o = \exp\left\{\frac{-[\ln(U_D/L_D)]^2}{4\sqrt{\ln(U_D)\,\ln(L_D)}}\right\}. \quad \text{(A10)}$$

REFERENCES

1. Egger M, Schneider M, Davey Smith G. Meta-analysis: Spurious precision? *Br Med J.* 1998;316:140–144.
2. Freedman D, Pisani R, Purves R, *Statistics.* (3rd Ed) New York, NY: Norton; 1998: Chapter 29.
3. O'Hagan A. *Kendall's Advanced Theory of Statistics.* Vol 2B: Bayesian Inference. London: Arnold; 1994.
4. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *Br Med J.* 1996;313:603–607.
5. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to Bayesian methods in health technology assessment. *Br Med J.* 319:508–512
6. Matthews RAJ. *Facts versus Factions: the use and abuse of subjectivity in scientific research.* Cambridge: European Science and Environment Forum; 1998. Reprinted in *Rethinking Risk and the Precautionary Principle.* Morris, J, Ed. Oxford: Butterworth; 2000: 247–282. Available online at: http://ourworld.compuserve.com/homepages/rajm/openesef.htm.
7. Matthews RAJ. Why *should* clinicians care about Bayesian methods? *J Stat Inf Plan.* 2001;94:43–58. See also discussion, 59–71.

8. Pocock SJ, Spiegelhalter DJ. Letter (untitled). *Br Med J.* 1992;305:1015.
9. GREAT Group. Feasibility, safety and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial. *Br Med J.* 1992;305:548.
10. Morrison LJ, Verbeek R, McDonald AC, Sawadsky BV, Cook DJ. Mortality and prehospital thrombolysis for acute myocardial infarction: a meta-analysis. *JAMA.* 2000;283:2686–2692.
11. The Subcutaneous Sumatriptan International Study Group. Treatment of migraine attacks with sumatriptan. *N Engl J Med.* 1991;325:316–321.
12. Roberto Latini R, et al. Clinical effects of early angiotensin-converting enzyme inhibitor treatment for acute myocardial infarction are similar in the presence and absence of aspirin: Systematic overview of individual data from 96,712 randomized patients. *J Am Coll Cardio.* 2000;35:1801–1807.
13. Egger M, Davey Smith G, Schneider M, Minder CE. Bias in meta-analysis detected by a simple graphical test. *Br Med J.* 1997;315:629–634.
14. Lee PM. *Bayesian Statistics: An Introduction.* 2nd Ed. London: Arnold; 1997.
15. Good IJ. *Probability and the Weighing of Evidence.* London: Griffin; 1950.