Bayesian statistics in medical research: an intuitive alternative to conventional data analysis

Lyle C. Gurrin BSc (Hons), PhD, AStat,[1] Jennifer J. Kurinczuk MD, MSc, MFPHM, FAFPHM[2] and Paul R. Burton MD, MSc, MFPHM, CStat[3]

[1]Biostatistician, Women and Infants Research Foundation, King Edward Memorial Hospital, Subiaco, Perth, Australia
[2]Perinatal Epidemiologist, TVW Telethon Institute for Child Health Research, West Perth, Australia and Clinical Senior Lecturer, Department of Public Health, The University of Western Australia, Australia;
[3]Professor of Genetic Epidemiology, Department of Epidemiology and Public Health, University of Leicester, Leicester, UK and Head, Division of Biostatistics and Genetic Epidemiology, TVW Telethon Institute for Child Health Research, Department of Paediatrics, University of Western Australia, West Perth, Australia.

Correspondence
Dr Lyle C. Gurrin
Biostatistician
Women and Infants Research Foundation
King Edward Memorial Hospital
PO Box 134, Subiaco
Perth, W.A., 6008
Australia

Keywords: assisted reproduction technology, Bayesian statistics, confidence intervals, ICSI, medical statistics, *P* values, statistical inference

Accepted for publication:
21 September 1999

Summary

Statistical analysis of both experimental and observational data is central to medical research. Unfortunately, the process of conventional statistical analysis is poorly understood by many medical scientists. This is due, in part, to the counter-intuitive nature of the basic tools of traditional (frequency-based) statistical inference. For example, the proper definition of a conventional 95% confidence interval is quite confusing. It is based upon the imaginary results of a series of hypothetical repetitions of the data generation process and subsequent analysis. Not surprisingly, this formal definition is often ignored and a 95% confidence interval is widely taken to represent a range of values that is associated with a 95% probability of containing the true value of the parameter being estimated. Working within the traditional framework of frequency-based statistics, this interpretation is fundamentally incorrect. It is perfectly valid, however, if one works within the framework of *Bayesian* statistics and assumes a 'prior distribution' that is uniform on the scale of the main outcome variable. This reflects a limited equivalence between conventional and Bayesian statistics that can be used to facilitate a simple Bayesian interpretation based on the results of a standard analysis. Such inferences provide direct and understandable answers to many important types of question in medical research. For example, they can be used to assist decision making based upon studies with unavoidably low statistical power, where non-significant results are all too often, and wrongly, interpreted as implying 'no effect'. They can also be used to overcome the confusion that can result when statistically significant effects are too small to be clinically relevant. This paper describes the theoretical basis of the Bayesian-based approach and illustrates its application with a practical example that investigates the prevalence of major cardiac defects in a cohort of children born using the assisted reproduction technique known as ICSI (intracytoplasmic sperm injection).

Introduction

Many research questions in medical science can most naturally be answered by assessing the probability that a particular hypothesis is true, or false, having observed a relevant set of data. Medical researchers have long embraced statistical methods in determining how such data should impact on their belief about the plausibility of the hypothesis in question. There is, however, widespread misunderstanding as to the appropriate interpretation of the tools associated with statistical inference, such as *P*-values and confidence intervals. In this paper we will address some common misconceptions about the use of statistical methods in medical research, and suggest an alternative and more intuitive interpretation, based on the *Bayesian* theory of statistics (Lindley 1965a,b; Box & Tiao 1973; Lee 1989). The Bayesian methodology is particularly useful in both the clinical setting and the arena of public health policy when the results of a study must subsequently be used to facilitate a decision (Burton 1994; Lilford & Braunholtz 1996).

Frequentist statistics

Medical statistics is firmly founded upon the *frequentist* theory of statistics (Armitage & Berry 1994, p93–99), which is the best known and most widely used framework for statistical reasoning. In this framework, the process of inference requires us to consider every possible result that a study could potentially generate. This leads to the calculation of a *P-value*, defined as the probability of observing data at least as 'extreme' as, or more 'extreme' than, the data that was actually observed in the current study *given* that the particular hypothesis (usually the 'null' hypothesis of 'no difference' or 'no effect') is true. This is a statement of *frequency*-based probability since it involves the relative frequency of an outcome or event in a repeated series of identical, hypothetical experiments.

If the calculated *P*-value is small then the observed data are surprisingly extreme, in that they are improbable if the null hypothesis is true, and so it represents evidence against the null hypothesis. The converse is, however, not necessarily true; observing a set of data that is less extreme and is thus associated with a larger *P*-value is not necessarily evidence *for* the null hypothesis, since under some circumstances this can be quite likely to occur even if the null hypothesis is *false*. The design of a research study generating such data is said to have low statistical *power*, in that datasets that are apparently consistent with the null hypothesis can occur relatively frequently even if the null hypothesis is false and thus an alternative hypothesis is true (Armitage & Berry 1994, p195–206).

This situation is compounded when the results of a study are declared to be 'statistically significant' if the *P*-value is observed to fall below an *arbitrary* threshold, typically 0.05. Such 'tests of statistical significance' or 'hypothesis tests' represent an unnecessary dichotomization of the set of all possible results of a study into an over-simplified 'accept/reject' decision analysis. The continuum of evidence across the range of potential data is completely ignored by such significance testing, which is often inappropriately viewed by clinicians and medical researchers as the statistical equivalent of a diagnostic test in medicine (Burton 1994; Burton *et al*. 1998). In a typical test of significance given some null hypothesis, $P < 0.05$ is often interpreted to mean 'there is a difference (the null hypothesis is false)', while $P \geq 0.05$ is understood to mean that 'there is no difference (the null hypothesis is true)'. These common but incorrect interpretations both express results in terms of the null hypothesis being true or false, and suggest that the *P*-value provides a direct quantitative measure of the plausibility of the null hypothesis. Taken to its ultimate conclusion, this results in the fundamental misconception that the *P*-value measures the probability that a given null hypothesis is true having observed a particular set of data.

A *P*-value actually reflects the probability of obtaining a particular pattern of results, or one more extreme, on the basis of an hypothesis *that is assumed to be true*. The probability that an hypothesis is true or false is *not* the long-run probability of an event and cannot even be expressed in the framework of frequency-based probability. In any particular case the hypothesis must either be true or false, and no frequency-based probability should be attached to it (Armitage & Berry 1994, p76–77). Any formal inferences in this vein must at best be indirect. As a minimum, any reasonable assessment of the viability

of the null hypothesis requires simultaneous consideration of the relative plausibility of a variety of competing hypotheses that are also consistent with the data and cannot, therefore, be based on the calculation of a single *P*-value, assuming that the null hypothesis is true!

Degrees of belief and subjective probability

In order to overcome the problems discussed in the last paragraph, we need to subscribe to a more general notion of probability. While we would wish to maintain the simple frequentist interpretation of probability as the long-run frequency of events in circumstances where it is appropriate, we would also like to make probabilistic statements and judgements about statistical parameters and, ultimately, scientific hypotheses.

Most statisticians now accept the concept of subjective probability, where statements involving the use of probability are taken to represent a 'degree of personal belief' about the quantity or event of interest (Lindley 1965a). This removes the need to associate probability with observable events (however hypothetical such events may be) and allows us to make quantitative judgements about the likelihood of an assertion being correct in circumstances where there is no reasonable long-run frequency interpretation.

A typical example occurs when we attach a probability to an event in public affairs, such as the statement 'there is a 10% chance that Australia will become a republic before 31 December 2005'. Clearly Australia will not debate the transition from a constitutional monarchy to a republic a large number of times under identical conditions and so a relative frequency interpretation of the probability of this event is simply not possible. Frequentist statisticians should refrain from attaching probabilities to such one off events, though many events (for example, sporting contests) can be similar enough for the associated probabilities to warrant a frequentist interpretation. Although it is not strictly necessary for subjective probabilities to be based on data, they should change, in a rational manner, as new data accrue. More formally, a sequence of subjective probability statements must be internally consistent or *coherent* in the sense of Walley (1991) in order to avoid irrational behaviour in response to them (Walley 1991; Walley *et al.* 1996).

Bayesian statistics

Suppose that we plan to conduct an observational or experimental study to further our knowledge about some quantity of interest (called a *statistical parameter*), and thus to collect information that will provide evidence to support or refute a current hypothesis about the quantity of interest. The *Bayesian* approach to statistical inference (named after the 18th century English clergyman the Reverend Thomas Bayes) initially asks the researcher to collate all pre-existing information, reflecting both evidence based on past studies and current beliefs, before prospectively collecting any new data. This information is then expressed in mathematical form as a *prior* probability distribution. The prior distribution is simply a quantification of the current state of understanding about the unknown quantity of interest and can be thought of as attaching a weight, expressed as a probability, to each possible value of the quantity of interest before additional data are recorded and examined. Values of the quantity of interest that are viewed as being *a priori* fairly likely to represent the true quantity are assigned a high prior probability and those that are viewed as less likely receive a correspondingly lower prior probability.

The prior distribution by definition allows investigators to incorporate pre-existing information into their analysis, something that is more difficult to do in the frequentist theory of statistics. Lilford *et al.* (1995) comment that 'Bayesian methods utilise all available data'. This provides a distinct advantage over conventional methods of analysis, which Lilford & Braunholtz (1996) rightly observe 'do not allow decision makers to take explicit account of additional evidence'. The choice of an appropriate prior distribution is usually based on a combination of the following three sources of information:

(i) evidence from previous studies via the inspection of historical data;

(ii) consultation with experts in the field to elicit their clinical opinion, which potentially involves a degree of subjective judgement;

(iii) the development of theoretical physical or biological models.

New evidence from the data collected during the current study is summarized by the *likelihood function* (Edwards 1992; Berger & Wolpert 1988). This is a mathematical object that describes how the probability distribution of the observed data depends on the particular values of the statistical parameters that govern the chosen class of statistical models.

The last step in the Bayesian process is to combine the prior distribution with the likelihood function using a mathematical routine derived from *Bayes' Theorem* (Lindley 1965a,b; Armitage & Berry 1994, p71–77). The result of this process, called the *posterior* probability distribution, is an updated reflection of our beliefs about the statistical parameters and has a probabilistic interpretation analogous to the prior distribution. From a mathematical point of view, weighting the prior distribution by the likelihood function forms the posterior probability distribution. It is the posterior distribution that is used to draw inferences and thus form conclusions about the relevant quantity of interest.

An important advantage of the Bayesian approach to statistical analysis is that it provides probability distributions for the quantities of interest. This makes it possible to make genuine probability statements about the magnitude of such parameters, such as the probability that a clinical effect lies within a particular range (e.g. 'the probability that the odds ratio is between 0.2 and 0.5 is 95%'). Furthermore, one consequence of this is the intuitively appealing opportunity to attach a probability to a statistical hypothesis of interest, since such a hypothesis is merely a statement about the value or nature of such parameters (e.g. 'there is a 5% chance that the treatment effect is greater than 0'). This provides direct and explicit answers to the sorts of questions that are usually posed by clinicians and medical researchers. Such an interpretation can be extrapolated immediately to clinical practice and could form the basis of decisions about policy in public health medicine.

Example 1

Figure 1 illustrates the Bayesian posterior distribution generated from a hypothetical phase II clinical trial investigating the fall in diastolic blood pressure (DBP) following commencement of a new drug 'X' in a group of 40 patients with mild untreated hypertension. A pharmaceutical company wishes to determine whether preliminary results are good enough to warrant proceeding to a full phase III trial. At the outset it is stated that the drug will be regarded as 'potentially useful' if it reduces DBP by at least 5 mm Hg. An appropriate prior distribution was chosen by seeking the advice of experts.

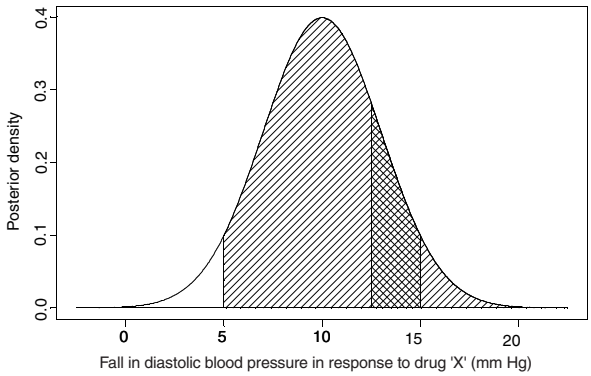

Figure 1 A Bayesian posterior distribution for the fall in diastolic blood pressure.

As an example of the type of inference which may be drawn from the posterior distribution, let us use the figure to estimate the probability that the true fall in DBP in response to 'X' lies somewhere between 12.5 mm Hg and 15 mm Hg. The required posterior probability is equal to the area under that part of the curve which falls between 12.5 mm Hg and 15 mm Hg (the area which is cross hatched) as a proportion of the total area under the curve. The total area under the curve is 1.0, the crosshatched area is 0.155 and the required probability is therefore 15.5%. In other words, having observed the current data, and assuming that the prior distribution was chosen appropriately, there is a probability of 15.5% that drug 'X' reduces DBP by between 12.5 mm Hg and 15 mm Hg in patients with mild untreated hypertension.

Equivalently we can answer the research question that was of primary interest to the pharmaceutical company: '*what is the probability that the true fall in DBP is at least 5 mm Hg?*'. The relevant probability is represented by the full shaded area (single line shading and cross hatching) on Figure 1, which encompasses 95.2% of the total area under the curve.

Thus there is a 95.2% posterior probability that the true fall in DBP in response to drug 'X' is at least 5 mm Hg.

Choosing a prior distribution

Specification of the prior distribution is a matter of ongoing concern for those contemplating the use of Bayesian methods in medical research. Clearly any conclusions drawn from a Bayesian analysis will potentially be sensitive to the choice of prior distribution. Some authors have devoted considerable thought to the process of formalizing the choice of a prior probability distribution. Freedman & Spiegelhalter (1983), Spiegelhalter & Freedman (1986), Chaloner *et al.* (1993) and Kadane *et al.* (1980) have all made some suggestions as to eliciting and quantifying the prior opinions of clinicians, but this remains a difficult task. It is sometimes fancifully suggested that clinicians and 'consumers' should come equipped with their own prior distribution which they can then combine with the likelihood function provided by the statistician!

If there is important pre-existing information that needs to be taken into account then it can be incorporated into a subsequent analysis by formulating a suitably descriptive prior distribution. This is a crucial step in the Bayesian process, despite the fact that it is often treated with scepticism by traditionally minded statisticians and clinicians. Nevertheless, although we do not wish to downplay the importance of choosing an appropriate prior distribution in situations where there is considerable prior knowledge, there are, to be realistic, many circumstances where little or no relevant pre-existing information is available. It is perhaps a reasonable criticism of the Bayesian approach to statistical analysis that, in this situation, attempting to specify a prior distribution is effectively trying to quantify something that does not exist. Alternatively, we may wish to restrict attention to the current data so that we can, in some sense, let the data 'speak for themselves', or, in the words of Lilford *et al.* (1995) 'represent the information arising just from the data'. Lindley (1965a) comments 'even when one has some appreciable prior knowledge of theta [a quantity of interest] one may like to express the posterior beliefs about theta without reference to them [i.e. the prior distribution]'. Equivalently, it is not unusual to hear researchers in a scientific context say that they need to draw an 'objective' inference that is untainted by the personal opinions and prejudices of those participating in the project.

Under these circumstances some statisticians have proposed what might loosely be called an objective Bayesian theory of statistical inference. They advocate the use of 'vague', 'flat' or 'non-informative' prior distributions that in some sense emphasize the role of the current experimental data and obviate the need for specific reference to prior beliefs (Lindley 1965b; Hughes 1993). One such distribution is the *uniform* probability distribution, which assigns equal prior weight to each possible value of the quantity of interest on the scale of the chosen outcome measure. Each value of the quantity of interest is viewed as 'equally likely' before the new data are observed, which seems intrinsically reasonable. The use of a uniform prior probability distribution focuses attention on current rather than pre-existing data (Lindley 1965b), in that the shape of the posterior distribution depends entirely on the likelihood function. Although the use of the prior distribution has a number of shortcomings and does not truly represent a formal mathematical expression of the state of 'prior ignorance' (Walley 1991; Walley *et al.* 1996; see also the discussion), it provides an *ad hoc* standard or reference analysis, from a common starting point, that aids comparison between current experimental or observational data, and that obtained from other sources. Furthermore, the uniform prior distribution provides an important link between frequentist and Bayesian theories of statistical analysis, which can be conveniently illustrated by exploring the role of the confidence interval in statistical inference.

The interpretation of confidence intervals

The concept of a confidence interval was developed by frequentist statisticians in order to represent the precision of a parameter estimate as the size of an interval of values that necessarily includes the estimate itself. Confidence intervals are generated by inverting a probability statement about the data *given* the value of the parameters, in order to come up with a range of values for the true parameter to which we can attach a probabilistic interpretation. In order to remain faithful to the frequency based

definition of probability, however, a conventional 95% confidence interval is properly defined in a somewhat subtle manner, in terms of hypothetical repetitions of the study and analysis under consideration. To paraphrase: if new data were to be repeatedly sampled, the same analysis carried out and a series of 95% confidence intervals calculated, 19 out of 20 such intervals would in the long run include the true value of the quantity being estimated (see Armitage & Berry 1994, p93–99).

Most researchers, however, interpret a 95% confidence interval in a rather different manner. They infer that the confidence interval contains the unknown quantity with 95% probability (Burton 1994). Within the frequentist framework, this interpretation of the confidence interval is fundamentally incorrect (Armitage & Berry 1994, p93–99). In any particular case, the unknown true value either does lie within the bounds of the confidence interval or it does not; there is no appropriate frequency-based interpretation of the probability of this 'single' event. If, however, an objective Bayesian analysis is carried out using a prior distribution which is uniform on the principal scale of analysis, it can be shown that a conventional C% confidence interval encloses a range of values that also encompasses C% of the area under the posterior distribution (Lindley 1965b). A 95% confidence interval is therefore equivalent to a posterior subjective probability of 95% that the true value lies between the lower and upper bounds of the confidence interval – an interpretation that corresponds precisely to that stated above. Some statisticians would prefer to call a confidence interval interpreted in this manner a *Bayesian 95% credible interval* (Winkler 1972). The Bayesian interpretation we have suggested is acceptable provided that one acknowledges that we are dealing with subjective probability, not frequency probability, and that one is assuming that any prior information is to be viewed as 'vague'. This congruence between conventional frequentist confidence intervals and the corresponding Bayesian credible interval associated with a uniform prior probability distribution provides a straightforward and intuitive interpretation of the results of a conventional statistical analysis, and affords a simple introduction to the Bayesian approach to statistical inference (Burton 1994).

Example 2

Intracytoplasmic sperm injection (ICSI) involves the selection and injection of a single spermatozoon into an oocyte. The procedure is an extension of standard *in-vitro* fertilization treatment and represents the most significant development in the field of assisted reproduction since the birth of the first 'test tube baby' in 1978. ICSI offers, for the first time, the prospects of genetic parenthood for men with profound oligozoospermia (low sperm count) and, with the use of testicular biopsy and epididymal aspiration techniques, even for those men with azoospermia (no sperm present in the ejaculate). The success of ICSI has led to its use throughout the world. There are, however, several theoretical concerns about the safety of ICSI and a series of potential risks for the offspring have been identified (Patrizio 1995; Cummins & Jequier 1994; De Kretser 1995).

The ICSI technique was developed by a group at the Brussels Free University (Palermo *et al*. 1992). From the outset, this group had the foresight to set in place follow-up of the infants born after ICSI treatment at their centre. As their cohort of infants has increased in number, they have published their findings in an overlapping series of papers (Bonduelle *et al*. 1994; Van Steirteghem *et al*. 1994; Tournaye *et al*. 1995; Liebaers *et al*. 1995; Bonduelle *et al*. 1996). By 1995 they had assessed 420 live born infants conceived following ICSI treatment at their centre and had identified a series of birth defects (Bonduelle *et al*. 1996). They used a definition of major birth defects for which population comparison data were not available. They determined that 14 of 420 live born infants (3.3%) had a major birth defect and concluded that there was no increase in the prevalence of birth defects in infants born after ICSI (Bonduelle *et al*. 1996). However, by reclassifying the reported defects using the classification system used by the Western Australian Birth Defects Registry, researchers were able to compare the birth prevalence of defects with the population prevalence estimates from Western Australia for live births during the same time period (Kurinczuk & Bower 1997). Following the reclassification 31 of the 420 children (7.38%) were defined as having a major birth defect, compared to 3.8% of the general Western Australian population of live births. Of particular interest were

the 14 (3.33%) infants with cardiac malformations defined as major by Kurinczuk & Bower (1997). There was some concern, however, that because of the unusually close surveillance of the Belgian cohort, the increased risk of cardiac birth defects described by Kurinczuk & Bower (1997) may have been due to the over-diagnosis of defects that would otherwise never have come to medical attention (Kurinczuk & Bower 1997; Bonduelle *et al*. 1997). Having excluded all cardiac defects that may (even remotely) have fallen into this category, 5 of the 420 (1.19%) infants were deemed to have at least one major cardiac defect that would *definitely* have been identified under routine surveillance. This was then compared to the corresponding prevalence of major cardiac defects in the population of Western Australian live births, that is, 0.67%.

Researchers in Western Australia wished to determine whether these results warranted the submission of a grant application to investigate this issue further using local ICSI data. They wished to know how likely it was that ICSI was associated with an increase in the birth prevalence of major birth defects, particularly cardiac defects, and if so, how likely it was that such an increase in cardiac defects was large, for example, greater than two-fold.

Let us initially consider how these data might be analysed in a conventional setting. The 'null' hypothesis will be that 'the birth prevalence of major cardiac defects in the ICSI birth cohort is the same (0.0067) as in the general Western Australian population'. A conventional test of the null hypothesis based upon the standard Normal approximation to the binomial distribution (Armitage & Berry 1994, pp70–71, 118–125) would utilize a standard error for the proportion of $((0.0067 \times 0.9933)/420)^{1/2} = 0.00398$. The observed proportion in the ICSI cohort is 5/420 = 0.0119 and so the standardized Normal deviate (Z) is ((0.0119 – 0.0067)/0.00398) = 1.31, which (from the usual statistical tables) is equivalent to a 2-tailed *P*-value of 0.191. In calculating the 95% confidence interval for the proportion, we now ignore the null hypothesis and use the observed proportion to calculate its standard error (Armitage & Berry 1994, sections 4.7, 4.9): ((0.0119 × 0.9881)/(420) = 0.00529. This produces a 95% confidence interval of 0.0119 ± 1.96 ×0.00529 = 0.00153 to 0.0223.

Using these standard results the data are likely to be interpreted in one of three ways. First, it may be noted that $P > 0.05$ and this may be interpreted as suggesting that the null hypothesis should be 'accepted' and the conclusion drawn that there is no evidence of an increased prevalence of major cardiac defects in children born following ICSI. This interpretation is, of course, fundamentally incorrect. Second, it may be noted that there appears to be a potentially important increase in the prevalence of major cardiac defects in the ICSI cohort that is close to twice the corresponding proportion in the general population. However, because the result based on the five cases of cardiac defects was not statistically significant, it might be argued that this data set is too small to draw any meaningful inferences. This interpretation is safer than the first, but fails to use the data to their full potential. A third alternative is to interpret the 95% confidence interval. This interval (calculated above as 0.00153 to 0.0223) is wide and encompasses values that would lead to quite different inferences. For example, a birth prevalence of 0.002 would suggest that ICSI infants had a prevalence of major cardiac defects that was only 30% of that in the general population, whereas a prevalence of 0.0201 would suggest that it was three times as high. Both of these values are contained in the confidence interval and are therefore, in some sense, consistent with the observed data. This confirms that the sample size is too small and suggests that further study is important. This interpretation is both valid and informative and there is no question that if a standard approach to analysis is to be adopted it should be based upon confidence intervals. This approach does not, however, allow us to express some of these qualitative impressions in a quantitative manner. For example, although a prevalence of 0.0201 falls within the 95% confidence interval and is therefore 'consistent' with the data, it is unclear how likely it is, on the basis of this preliminary analysis, that the true birth prevalence really is this high or maybe even higher.

As an alternative, we would propose that a Bayesian analysis be carried out using a 'non-informative' prior distribution that is uniform on the scale of proportions. Having made this assumption we can now make use of the equivalence of a standard C% confidence interval and a Bayesian C%

credible interval. In order to generalize the ensuing calculations, let us consider what may be called a *critical confidence interval*. This is defined as the confidence interval with a midpoint at the observed value (in this example, at a proportion of 0.0119) and a lower limit at the value of a threshold of interest (in this example, at a proportion of 0.0067 corresponding to the 'null value' associated with the rate in the general population), and an upper limit that is by symmetry the same distance above the observed value as the lower limit is below. Using Z tables it is straightforward to determine the percentage coverage of this critical confidence interval. In this example, such a confidence interval on the proportion scale extends from 0.0067 to 0.0119 + (0.0119 – 0.0067) = 0.0171. This is symmetric about the observed proportion, that is 0.0119, and extends 0.0052/0.00529 = 0.983 standard errors in either direction. Reference to a table of the Z distribution indicates that 83.71% of the area under the curve lies below Z = 0.983, thus 16.29% lies above this point and, by symmetry, 16.29% of the area lies below Z = –0.983. This particular confidence interval is therefore a (100 – [2 × (16.29]) = 67.42% confidence interval which means that, having adopted a prior distribution that is uniform on the scale of proportions, the range 0.0067 to 0.0171 is a Bayesian 67.42% credible interval. This means that there is 67.42% posterior probability that the true proportion lies between 0.0067 and 0.0171 and a 16.29% posterior probability that it is greater than 0.0171. There is therefore a 67.42% + 16.29% = 83.71% posterior probability that the true proportion is greater than 0.0067 and thus a relatively high probability that the risk of a major cardiac defect in a baby conceived using ICSI is higher than the risk in the general population. Readers should note that, in this particular case, the posterior probability of 83.71% could have been obtained directly from the table of the Z distribution: '83.71% of the area under the curve lies below Z = + 0.983'. However, we explain the calculation in terms of a two-sided confidence interval, because we believe that this clarifies the full procedure and it is appropriate under all circumstances.

In general, if the percentage coverage of the confidence interval is C%, the posterior probability that the true value of the quantity of interest exceeds the stated threshold is C% + (100% – C%)/2. This is because any value which falls *inside* the critical confidence interval (posterior probability = C%) must by definition exceed the threshold of interest and symmetry dictates that one half of all values which fall *outside* the confidence interval (posterior probability = (100% – C%)/2) will also exceed the threshold. In order to calculate the probability that the true value of a quantity of interest is *less* than a given threshold, one may carry out a series of analogous calculations using the critical confidence interval whose *upper* limit falls at the threshold.

Returning to the example, let us calculate the probability that the true proportion exceeds 0.0134, which is twice the rate in the general population. Since this value exceeds the observed value of 0.0119, we set the upper bound of the critical confidence interval to the threshold of interest, namely 0.0134, and calculate the lower bound to be as far below the observed value of 0.0119 as 0.0134 is above, giving a value of 0.0119 – (0.0134 – 0.0119) = 0.0104. This confidence interval, extending from 0.0104 to 0.0134, is ± 0.2835 standard errors around the estimated proportion of 0.0119. This is a 22.32% confidence interval and the posterior probability that the true proportion exceeds 0.0134 is half of the probability lying outside this interval, or (100% – 22.32%)/2 = 38.84%.

These results tell the researcher that it is very likely (approximately 84%) that the true prevalence of major cardiac defects is greater in the ICSI cohort than in the general population and that there is close to a 40% probability that it exceeds twice the background rate. Similar calculations demonstrate that the chance that the true proportion in the ICSI cohort is as high as three times the rate in the general population is only 6.06%. To extend the characterization further, Table 1 details the posterior probability that the true proportion exceeds a series of thresholds of interest.

Analyses such as those illustrated above proved to be of considerable value to the medical scientists in Western Australia investigating the risks associated with ICSI therapy. The investigators were subsequently successful in obtaining a research grant (from the March of Dimes Birth Defects Foundation in New York) to continue their work in this area.

Table 1 The posterior probability that the prevalence of major cardiac defects in the ICSI cohort exceeds a series of thresholds based on the prevalence in the general population

Threshold of interest	*Threshold as a multiple of the prevalence in general population*	*Posterior probability that the true rate exceeds the stated threshold*	*Posterior probability that the true rate is less than the stated threshold*
0.0067	1.0	83.71%	16.29%
0.01005	1.5	63.67%	36.33%
0.0134	2.0	38.84%	61.16%
0.01675	2.5	17.97%	82.03%
0.0201	3.0	6.06%	93.94%
0.02345	3.5	1.45%	98.55%
0.0268	4.0	0.24%	99.76%

Discussion

Conventional statistical analyses based upon the 'frequency-based' view of probability and statistical inference often fail to make the best and most complete use of available data when assessing the evidence for an hypothesis under investigation (Burton 1994; Burton *et al*. 1998; Lilford *et al*. 1995). This results from having to restrict attention to the consideration of just one hypothesis (usually the 'null' hypothesis) and then to the comparison of the resultant *P*-value with an arbitrarily chosen threshold (usually $P = 0.05$) instead of viewing it on a continuous scale as an (indirect) measure of evidence. This reduces a potentially powerful inferential tool to a simplistic, mechanistic and ultimately very poor form of decision analysis known as the statistical significance test. The widely held belief that all studies and experiments that result in a non-significant *P*-value provide the *same* support for the specified hypothesis is just one example of the type of misunderstanding that can easily arise from a failure to appreciate the subtleties of interpretation associated with conventional frequentist statistical analysis (Freeman 1993).

We have proposed an alternative approach that views the problem of statistical inference from a Bayesian perspective. Such an approach allows one to make full use of the available data. A genuine probabilistic interpretation, based on the concept of subjective probability, provides direct answers to questions about the probable magnitude of the effects of interest and hence permits one to compare competing hypotheses in a straightforward and understandable manner. The use of subjective probability within a Bayesian framework is particularly useful in circumstances where a conventional approach to statistical analysis may be difficult or misleading. These include circumstances where: (i) a statistically non-significant result may be large enough to be clinically relevant (small sample size); (ii) a statistically significant result may be too small to be of clinical relevance (large sample size); or (iii) where one wishes to draw quantitative conclusions regarding the probability that two or more outcomes are sufficiently similar that any difference is unlikely to be clinically relevant.

The use of a uniform prior probability distribution promotes a confluence between the Bayesian and conventional frequentist approaches, since 95% confidence intervals can be viewed legitimately as containing the true value of interest with 95% probability. Many researchers already interpret confidence intervals in precisely this manner and thus our proposal does not require a radical modification of the way in which many researchers approach statistical analyses. It is important, however, that one acknowledges that this interpretation of confidence intervals is *only* valid if one works within a Bayesian framework using a uniform prior distribution. We have suggested reporting, where appropriate, the posterior probability that the quantity of interest exceeds a series of clinically relevant thresholds rather than just a single 95% confidence interval.

Although the use of a uniform prior probability distribution provides a neat introduction to the Bayesian process, there are a number of reasons why the uniform prior distribution does not provide the

foundation on which to base a bold new theory of statistical analysis!

First, the uniform prior probability distribution does not provide a formal mathematical representation of 'prior ignorance'. No *single* prior distribution is appropriate when one is faced with a complete lack of information (Walley 1991). Walley *et al.* (1996) note that '. . . any [single] Bayesian prior distribution assigns precise probabilities to hypotheses and therefore has strong behavioural implications, e.g. it precisely determines "fair" betting rates on the truth of the hypotheses.' Bayesian statisticians would endorse repeating the analysis using many different prior distributions in the hope of encapsulating a wide range of prior beliefs about the values of the relevant parameters. This is known as a *Robust* Bayesian approach to analysis (Berger 1984, 1990, 1994; Greenhouse & Wasserman 1995). Such a *sensitivity analysis* is clearly important if we are to ascertain how the posterior distribution is affected by changes to the prior probability distribution or by changes to the model used to create the likelihood.

A second difficulty with the uniform prior distribution is its sensitivity to transformation; the uniform distribution may in fact be very non-uniform when transformed to another scale of analysis. A prior distribution that is uniform on the scale of proportions, for example, cannot simultaneously be uniform on the scale of odds and *vice versa*, and yet in many cases either scale would be appropriate for analysis. We would argue that if two scales really are equally appropriate, and the use of a prior which is uniform on one scale leads to a qualitatively different conclusion to an analysis based upon a prior which is uniform on the other scale, then inferences must, of course, be viewed as uncertain. One hopes that in situations where more than one analytical scale is appropriate, the choice of scale would result in relatively small quantitative changes rather than large qualitative alterations to the principal conclusions. To illustrate, if Example 2 had been worked assuming uniformity on the scale of $\log_e$(odds) rather than on the scale of proportions, the estimated posterior probability that the true rate of cardiovascular birth defects in ICSI baby exceeded the general population rate, or twice that rate, would have been 90.1% and 39.5%, respectively. Because the sample size is *so* small (only five cases), these probabilities are noticeably different to the original values of 83.7% and 38.8%, respectively. Nevertheless, this change would make little or no difference to the principal conclusion of the analysis.

In most settings in medical statistics, confidence intervals are calculated in a way that assumes that the distribution of the data, or at least a relevant summary statistic, can be approximated by a suitable Normal distribution. The correspondence of a C% confidence interval calculated using such an approximation to a C% credible interval is therefore only exact when the data are Normally distributed. For most of the standard probability distributions used to analyse and model medical data, the approximation is, in general, quite close even when the sample size is relatively small (for an example, see Burton (1994)).

One of the problems with Bayesian analysis is that it is often a non-trivial problem to combine the prior information and the current data to produce the posterior distribution. Despite the increasing availability of purpose-designed software for Bayesian analysis (BUGS, Spiegelhalter *et al.* 1995), specialist advice and software is generally required in order to bring Bayesian statistics into the medical research workplace. The congruence between conventional confidence intervals and Bayesian credible intervals generated using a uniform prior distribution does, however, provide a simple way to obtain inferences in Bayesian form which can be implemented using standard software based on the results and output of a conventional statistical analysis.

The use of Bayesian methods is growing amongst clinical scientists and clinicians. The congruence between a Bayesian analysis using a uniform prior and a conventional analysis provides a non-threatening introduction to Bayesian methods and means that analyses of the type we describe can be carried out on standard software. Our approach is straightforward to implement, offers the potential to describe the results of conventional analyses in a manner that is more easily understood, and leads naturally to rational decisions. We do not suggest that this approach should be used all the time, nor should it be used is an excuse for designing studies which are too small or a fallback position when a conventional analysis fails to produce a statistically significant result. However, when it is used appropriately, we

believe that this approach is a useful addition to conventional methods.

Acknowledgements

Jennifer Kurinczuk gratefully acknowledges receipt of a two year project grant from the March of Dimes Birth Defects Foundation, New York (#6-FY98–497; #6-FY99–683).

This work was funded in part by the National Health and Medical Research Council of Australia as one component of Program Grant #96\3209.

References

Armitage P. & Berry G. (1994) *Statistical Methods in Medical Research*. 3rd edn. Blackwell Scientific Publications, Oxford.

Berger J. (1984) The robust Bayesian viewpoint (with discussion). In: *Robustness in Bayesian Analyses*. (Ed. J. Kadane). North-Holland, Amsterdam. pp. 63–144.

Berger J. (1990) Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference* **25**, 303–328.

Berger J.O. (1994) An overview of robust Bayesian analysis (with discussion). *Test* **3**, 5–124.

Berger J.O. & Wolpert R.L. (1988) *The Likelihood Principle*. 2nd edn. Institute of Mathematical Statistics, Hayward, California.

Bonduelle M., Desmyttere S., Buysse A., Van Assche E., Schietecatte J., Devroey P. *et al.* (1994) Prospective follow-up study of 55 children born after subzonal insemination and intracytoplasmic sperm injection. *Human Reproduction* **9**, 1765–1769.

Bonduelle M., Legein J., Buysse A., Van Assche E., Wisanto A., Devroey P. *et al.* (1996) Prospective follow-up study of 423 children born after intracytoplasmic sperm injection. *Human Reproduction* **11**, 1558–1564.

Bonduelle M., Devroey P., Liebaers I. & Van Steirteghem A. (1997) Commentary: Major defects are overestimated. *British Medical Journal* **315**, 1265–1266.

Box G.E.P. & Tiao G.C. (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading, Massachusetts.

Burton P.R. (1994) Helping doctors to draw appropriate inferences from the analysis of medical studies. *Statistics in Medicine* **13**, 1699–1713.

Burton P.R., Gurrin L.C. & Campbell M.J. (1998) Clinical significance not statistical significance: a simple Bayesian alternative to *p* values. *Journal of Epidemiology and Community Health* **52**, 318–323.

Chaloner K., Church T., Louis T.A. & Matts J.P. (1993) Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* **42**, 341–353.

Cummins J.M. & Jequier A.M. (1994) Treating male infertility needs more clinical andrology, not less. *Human Reproduction* **9**, 1214–1219.

De Kretser D.M. (1995) The potential of intracytoplasmic sperm injection (ICSI) to transmit genetic defects causing male infertility. *Reproduction Fertility and Development* **7**, 137–142.

Edwards A.W.F. (1992) *Likelihood*. Johns Hopkins University Press, Baltimore.

Freedman L.S. & Spiegelhalter D.J. (1983) The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician* **32**, 153–160.

Freeman P.R. (1993) The role of p-values in analysing trials results. *Statistics in Medicine* **12**, 1443–1452.

Greenhouse J.B. & Wasserman L. (1995) Robust Bayesian methods for monitoring clinical trials. *Statistics in Medicine* **14**, 1379–1391.

Hughes M.D. (1993) Reporting Bayesian analyses of clinical trials. *Statistics in Medicine* **12**, 1651–1663.

Kadane J.B., Dickey J.M., Winkler R.L., Smith W.S. & Peters S.C. (1980) Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association* **75**, 845–854.

Kurinczuk J.J. & Bower C. (1997) Birth defects in infants conceived by intracytoplasmic sperm injection: an alternative interpretation. *British Medical Journal* **315**, 1260–1265.

Lee P.M. (1989) *Bayesian statistics: an introduction.* Arnold, London.

Liebaers I., Bonduelle M., Legein J., Wilikens E., Van Assche E., Buysse A. *et al.* (1995) Follow-up of children born after intracytoplasmic sperm injection. In: *Fertility and Sterility: a Current Overview.* (Hedon B., Bringer J., Mares P. eds.) Parthenon, New York.

Lilford R.J., Thornton J.G. & Braunholtz D. (1995) Clinical trials and rare diseases: a way out of a conundrum. *British Medical Journal* **311**, 1621–1625.

Lilford R.J. & Braunholtz D. (1996) The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal* **313**, 603–607.

Lindley D.V. (1965a) *Introduction to probability and statistics from a Bayesian viewpoint. Part 1 Probability.* Cambridge University Press, Cambridge. pp. 19–25, 29–42, 50, 58.

Lindley D.V. (1965b) *Introduction to probability and statistics from a Bayesian viewpoint. Part 2 Inference.* Cam-

bridge University Press, Cambridge. pp. 1–13, 15, 18, 19.

Palermo G., Joris H., Devroey P. & Van Steirteghem A.C. (1992) Pregnancies after intracytoplasmic injection of a single spermatozoon into an oocyte. *Lancet* **340**, 17–18.

Patrizio P. (1995) Intracytoplasmic sperm injection (ICSI): potential genetic concerns. *Human Reproduction* **10**, 2520–2523.

Spiegelhalter D.J. & Freedman L.S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective opinion. *Statistics in Medicine* **5**, 1–13.

Spiegelhalter D., Thomas A., Best N. & Gilks W. (1995) BUGS. Bayesian inference using Gibbs sampling, Version 0.60. MRC Biostatistics Unit, Cambridge. http://www.mrc-bsu.cam.ac.uk/bugs.

Tournaye H., Liu J., Nagy Z., Joris H., Wisanto A., Bonduelle M. *et al.* (1995) Intracytoplasmic sperm injection (ICSI): the Brussels Experience. *Reproduction Fertility and Development* **7**, 269–279.

Van Steirteghem A.C., Nagy P., Liu J., Joris H., Smitz J., Camus M. et al. (1994) Intracytoplasmic sperm injection – ICSI. *Reproductive Medical Review* **3**, 199–207.

Walley P. (1991) *Statistical reasoning with imprecise probabilities.* Chapman & Hall, London.

Walley P., Gurrin L. & Burton P. (1996) Analysis of clinical data using imprecise prior probabilities. *The Statistician* **45**, 457–486.

Winkler R.L. (1972) *An introduction to Bayesian inference and decision.* Holt, Rinehart and Winston Inc., New York. pp. 395–396.