The Cochrane lecture

Who's afraid Of Thomas Bayes?

R J Lilford, David Braunholtz

Abstract

Sometimes direct evidence is so strong that a prescription for practice is decreed. Usually, things are not that simple—leaving aside the possibility that important trade offs may be involved, direct comparative data may be imprecise (especially in crucial sub-groups) or subject to possible bias, or there may be no direct comparative evidence; but still decisions have to be made. In these circumstances, indirect evidence—the plausibility of effects—enters the frame. But how should we describe the extent of plausibility and, having done so, how can this be integrated with any direct evidence that might exist. Also, how can allowance be made in a transparent (that is, explicit) way for perceptions of the size of bias in the direct evidence. Enter the Reverend Thomas Bayes; plausibility (however derived—laboratory experiment, qualitative study or just “experience”) is captured numerically as degrees of belief (“prior” to the direct data) and updated (by the direct evidence) to yield “posterior” probabilities for use in decision making. The mathematical model used for this purpose must explicitly take account of assumptions about bias in the direct data. This paradigm bridges theory and practice, and provides the intellectual scaffold for those who recognise that (numerically definable) probabilities, and values (also numerically definable) underlie decisions, but who also realise that subjectivity is ineluctable in science.

(J Epidemiol Community Health 2000;54:731–739)

The end of history . . . Or just the beginning?

The onset of human labour remains an enigma but Mont Liggins, a New Zealand physiologist, speculated in the 1960s that a surge in cortisol might be the trigger for parturition in sheep. He tested this hypothesis by giving cortisol to pregnant ewes and confirmed that this resulted in premature birth. Serendipitously he noted that the lambs survived in far greater numbers than he would have expected, given the degree of their prematurity. Preterm lambs, like their human counterpart, tend to succumb to a disease called hyaline membrane disease—an affliction caused by collapse of the tiny air sacs or alveoli, where gas exchange takes place in the lung. Liggins hypothesised that the beneficial effect of cortico-

steroids on the neonatal lung would be constant between species. Luckily, corticosteroids do not cause pre-term birth in the human, in contrast with their action in sheep.

Liggins carried out a randomised trial in the human and this has since been replicated many times. Patricia Crowley, an obstetrician now living in Dublin, assembled the quantitative results in a meta-analysis¹ that confirms that the risk of both hyaline membrane disease and of perinatal mortality is reduced by the use of antenatal corticosteroids. A funnel plot (fig 1) showed that the results of individual trials distributed themselves symmetrically around the final meta-analysis result, providing reasonable reassurance, along with the comprehensive search strategy, that the result was not due to publication bias.

Corticosteroids were also noted to reduce the risk of intracerebral haemorrhage in the neonate—a potent cause of cerebral palsy—and long term follow up suggested that a short course of antenatal corticosteroids does not impede long term development of the brain.

This meta-analysis was widely published as part of a database of all trials in the maternity care field. Soon after that, as Chair of the Audit Committee of the Royal College of Obstetricians and Gynaecologists, the first author developed and promulgated a set of guidelines including, of course, an exhortation to use corticosteroids in people who threatened to give birth prematurely (and in whom there was no contraindication).

Initial observations of practice suggested that this treatment was grossly underused, as half (or less) of mothers giving pre-term birth

Figure 1 Funnel plot showing the effect of maternal corticosteroids in infant death (x axis) against the size of the trial (y axis).

NHS Executive,
Bartholomew House,
142 Hagley Road,
Birmingham B16 9PA

Correspondence to:
Professor Lilford

Accepted for publication
25 May 2000

had also been given corticosteroids. However, the majority of heart attack victims who had not been treated with the appropriate standard for this condition—intravenous administration of clot busting drugs—were ineligible if viewed prospectively; typically they presented too late for the treatment to be effective, or the ECG signs of a heart attack were not present, even though this turned out to be the correct diagnosis in retrospect.² Suspecting that similar factors might be at work with respect to antenatal corticosteroids, jointly with colleagues we carried out a NHS centrally commissioned R&D study to examine uptake of this effective standard among eligible patients. As shown in figure 2, we confirmed a massive increase in the use of corticosteroids among cases where there was an opportunity to give this treatment—in many cases the woman presented too late, and the figures would have been much less impressive had we not taken care to exclude these.

So here we have the paradigm of evidence-based medicine. An astute scientist carrying out basic research makes an interesting observation with implications for novel treatment. This is tested in randomised trials, in different parts of the world, and they are collated in a meta-analysis. In some cases, benefits must be weighed against the side effects by decision analysis. The results are then appraised by a professional group responsible for developing clinical policy (for example, a Medical Royal College or National Appraisal Centre) and in this way guidelines are developed. The impact of these guidelines is assessed by clinical audit and if barriers to implementation are identified, then these are tackled by some form of managerial action—clinical governance. This paradigm for development of clinical practice—so called evidence-based care—seems so self contained and well worked out, that it has been argued that as far as methodological development is concerned, we have reached what Francis Fukuyama dubbed “the end of history”.³ The purpose of this talk is to refute this notion, and to suggest that we are in the infancy of our methodological understanding.

Figure 2 The compendium of effectiveness research “Effectiveness and Childbirth and Maternity Care” was issued in 1989 and the Royal College guidelines promulgated in the early 1990s. The figure shows compliance, in 20 randomly chosen hospitals in England and Wales, with the injunction to give antenatal corticosteroids to women who threaten to give birth prematurely before promulgation of the guidelines (1988) and later (1996).

A theory of decision making

As our subject is the relation between evidence and decision, we need a theory of decision making if we are to move forward. In circumstances where the desire is to maximise utility (well being), the optimum decision can be shown to be dependent on the probabilities of various outcomes, and the values placed on those outcomes. In the example given above, concerning a single short course of corticosteroids to promote fetal lung maturity, there are no known material disadvantages. When there are no “costs”, the main effect is “dominant” and there is no need for values, but more typically in modern health care, trade offs are involved. Expected utility can be calculated as a sum of the values, weighted by their probability of occurring. Decision analysis (and its economic variant, cost utility analysis) raises some fiendishly difficult issues, which we will not go into here, but which are discussed elsewhere.⁴ Here we wish only to make the key point that the underlying purpose of evaluative research is to provide the probabilities on which decisions turn. Simply stated, evaluative research seeks to answer the following question:

In scenario A, what are the probabilities of outcome M, or N, or ... given decision X, or Y, or ...

Here we come to a central point of the presentation—conventional statistics do not give probabilities in this form.

Two types of probability

Conventional “frequentist” statistics is based on “the probability of obtaining the observed result given a certain underlying state of the world”. However, as we have seen above, what is needed for decision making is “the probability of a given underlying state of the world, given the data that we have observed”. Put more simply, a decision maker wants to know the probability of an outcome (for example, the new treatment has lower mortality) given the data, not the probability of the data given an outcome (usually, that there is no difference between treatments.)

To get from the data to probabilities that can be used in decision analysis, Bayesian statistics are required.

Clinicians are very familiar with the use of Bayesian statistics to calculate the probability of a disease after, for example, a positive test result, given prevalence (prior probability) and test accuracy parameters (sensitivity divided by the false positive rate, also known as the likelihood ratio for a positive test LR^+).

The equation takes a general form:

Posterior odds (of disease) given a positive test = prior odds (of disease) $\times LR^+$

or

Posterior odds (of disease) given a negative test = prior odds (of disease) $\times LR^-$

(LR^- = probability test negative given disease/probability test negative given no disease)

However, Bayesian methods may be used not just where the outcome is binary (disease or not), but also where the outcome is continuous (for example, proportion of patients surviving).

This requires a prior belief “density” and likelihood curves in place of a single prior probability and single likelihood ratio. The likelihood curve represents the probability of the observed data at all possible levels of true effect. It is centred on the observed data and its width depends on the number of observations; it is often, but not always, closely related to traditional confidence limits.

As will be seen, posterior distributions calculated in this way may be very different to the likelihood (or confidence limits) derived solely from the data. Other things being equal, the tighter the distribution of the “prior” the closer will be the posterior to the “prior” (and consequently, in general, the further will be the posterior from the data), and the narrower the likelihood/confidence limits the closer will be the posterior to the data. Thus, the data will “win” over the prior when the dataset is large (in relation to the events that it is describing), as would be the case in, say, the results of the meta-analysis of the use of clot busting drugs in people with strong clinical evidence of a recent heart attack.⁵ In such circumstances, the practical distinction between the two methods all but disappears, and it is reasonable to use the data only to calculate “numbers needed to treat” (NNT)—a concept that, aimed as it is at directly informing decisions, should obviously be based in Bayesian, not frequentist, probabilities. However, use of NNT based solely on observed data is totally inappropriate when measurements are imprecise in relation to the “prior”.

The Bayesian way of thinking is described in much more detail elsewhere^{6–8} but the propositions that follow—some of them very radical—have this ontological stance at their heart. The general use of Bayesian methods was promoted by Savage and others in the early 1960s, but suffered both from a certain naivety regarding the real world (for example some exponents believed randomisation in clinical trials was irrelevant), and from lack of suitable software for doing the often difficult calculations. These problems are now much reduced, randomised Bayesian clinical trials exist (if not yet exactly run of the mill), and Bayesian software is becoming ever more powerful and user friendly.⁹ Philosophical problems remain, at least from a frequentist point of view: the loss of objectivity, the definition of (personal) probability—which in its commonest and most intuitive form involves gambling on different outcomes—seems distant from medical research, the question of what it is that the posterior represents (what a Bayesian believes, or should believe, or would believe given the prior and the model?)—for an early, intelligent but not particularly pro-Bayesian discussion of the issues see Mainland’s statistical ward round number 5.¹⁰

The non-dichotomous nature of knowledge and our changing view of the ethics of randomised trials

Many evidence-based clinical standards were derived from trials of treatments that were otherwise freely and widely available in non-

trial practice. A recent systematic review of the ethics of trials¹¹ shows that until recently, the morality of this form of human experimentation was based on the “uncertainty principle”. Of course, “uncertainty” compared with “certainty” provides a very permissive paradigm within which to randomise patients—only very articulate patients on hearing that their doctor is “uncertain”, are likely to reply as follows: “Yes, but how uncertain are you?” or (better still) “what is your best prior guess about the effectiveness of treatment?” or (best of all) “what is your prior probability distribution?”. A patient who has a sense of her doctor’s “prior” (or who having looked into the subject, has developed her own “prior”), is in the best position to bring her values into play and determine which of the alternative treatments will maximise her expected utility. To develop this argument a step further, if she cannot say which treatment she prefers—then the alternatives provide the same expected utilities—she is, in decision analytic language, “indifferent” between them and in the ethics literature she would be referred to as being in “equipoise”.⁶

This topic is discussed in much more detail elsewhere,¹¹ but here we merely point out corollaries that contradict the prevailing wisdom. It is widely believed that small trials are unethical “in their use of subjects”.¹² They may be poor value for resources—assuming that these resources are transferable—but given equipoise they are not unethical in their use of subjects. As some patients who participate in clinical trials may include an element of altruism in their construction of equipoise, it is important that the scientific value of such studies is not “over sold”, but given the rise of meta-analysis, it is much more realistic to think of a given trial as a contribution to the world’s knowledge on a subject, rather than a one stop elevator to the truth, shorn of any need for replication or extension.

More fundamentally, the requirement that patients should be “indifferent” between treatment outcomes (not just that clinicians should be “uncertain” about treatment effects) is likely to result in slower recruitment to trials than has generally been typical hitherto. If this is so, then we may simply have to put up with slower recruitment to clinical trials in the future, and rely increasingly on other sources of evidence (even if less epistemologically sound), such as clinically rich databases. A recent report from the NHS Methodology Programme suggests, reassuringly, that the results of studies based on such databases do not always differ wildly from those of RCTs, and where they do there is often a rational explanation.¹³ The problem to be solved is predicting accurately the size and direction of biases in database studies, so they can be reliably subtracted. Although, as scientists, we appreciate the clarity of definitive clinical trials, we also think it is necessary to prepare for a time when trials might lose (for ethical and cost reasons) their late 20th century status as the dominant methodological paradigm for applied clinical research.

Inconclusive study results

Electronic fetal monitoring was introduced in the late 1960s and in the 1970s a number of small clinical trials were done involving a few hundred women each. These were obviously inconclusive, so a very famous study—The Dublin Randomised Trial of Foetal Monitoring—was established, involving no less than 12 000 women. The result, showing no statistically significant difference in stillbirth rates (between the monitored and unmonitored groups), was seized upon by the natural childbirth movement who castigated obstetricians for continuing to use this form of screening. What few realised was that although lots of women were randomised, only five stillbirths occurred among the group as a whole, and hence the trial provides little information: it was “under-powered”. Traditional statistical interpretations were simply unhelpful for decision makers, as least as far as stillbirth was concerned. However, the above trial did confirm that fewer babies in the monitored group had postnatal convulsions (of the type associated with hypoxia), and the introduction of fetal monitoring in developed countries was accompanied by a sharp reduction in intrapartum stillbirth.

On the basis of these indirect data, a “prior” centred on a reduction of, say, 50% in the risk of intrapartum stillbirth might have been constructed before the trial, with very low probabilities ascribed to an increase in stillbirth from fetal monitoring.

Such a “prior” is little changed by the stillbirth data actually observed in the Dublin Trial. The point we wish to emphasise is that imprecise information—or indeed no information at all—does not imply that we should prefer the null hypothesis over any other. “No direct evidence” neither favours nor disfavours the null hypothesis, it simply means we only have the indirect evidence (plausibility) to work with. The null hypothesis may or may not be favoured by the indirect evidence, but its routine use as a basis for statistical comparison is through convention. Of course, this does not imply that a person’s “prior” should be immune to challenge. For example, a “prior” as enthusiastic as that posited above might have been hard to defend, had it not been the case that there were less convulsions in the monitored group in the Dublin Trial, and that intra-

partum stillbirth rates had fallen *pari passu* with the introduction of fetal monitoring.

Medical history is full of examples of clinicians who formed excessively enthusiastic beliefs about the effectiveness of various treatments on the basis of biological plausibility,¹⁴ only to have these overturned by the results of direct comparative studies. While it is true that history teaches us that clinicians should seek to moderate their prior beliefs, this is not tantamount to saying that the null hypothesis is always to be preferred in the absence of statistically significant direct comparative data. Such a position would be downright ridiculous. To take an extreme example, should we assume, in the absence of direct comparative data, that warfarin does not aggravate a bleeding stomach ulcer? Further examples illustrating the point that plausibility cannot be expunged from our assessments of treatment effectiveness are given in table 1.

Subgroup analysis

Large trials (or meta-analyses of smaller studies) may give precise overall estimates on a certain outcome and yet yield imprecise measurements in certain important subgroups. What to do then, when the imprecise results in a subgroup differ appreciably from the overall result. If many subgroup effects have been examined, then the hypothesis testing (conventional) approach is fraught with problems. To avoid a high risk of “false positives” the current convention is to “correct” for multiple tests by raising the target “significance level” (that is, reducing the individual test *p* value) at which a subgroup effect is assumed to be “real”. This will result in over-estimates (in “significant” subgroups) and under-estimates (in “non-significant” subgroups) of effects. But more importantly we find the logic counter-intuitive: the estimate of a particular subgroup effect should not depend on how many other analyses were done. For example, we do not adjust confidence intervals on the main effect according to how many other research papers we have read that day, week, year . . .

On the other hand, to uncritically accept every subgroup estimate that manages to exceed a conventional (and uncorrected) *p* value runs a high risk of drawing spurious associations. For example, analysis of the results of aspirin for the prevention of stroke, in the early days, suggested that this treatment

Table 1 Theories having an impact on decisions emanating from empirical (quantitative) data

Issue	Comment
1 Astrological sign and duration of pain both influenced measured effectiveness of clotbusting drugs in stratified analysis of ISIS number 2 Trial	Only the effect of duration of pain taken seriously
2 Chorion villus sampling and apparent clustering of limb reduction defects	Chorion villus sampling seemed to be associated with increased risk of limb reduction defects when performed between 9 and 12 weeks gestational age (7 and 10 weeks embryonic age) Believability of these data was enhanced by the finding that: (i) the limb bud develops progressively into a limb over this period (ii) the later the intervention, the more peripheral the limb defect
3 Short versus long acting calcium antagonist in hypertension; inconclusive RCT	Harmful effect of short acting drugs widely accepted because they cause re-bound high blood pressure and separate studies show that intermittent hypertension is particularly damaging to the heart
4 β Blockers less effective than diuretics among elderly; inconclusive RCT — no difference of effect in previous RCT's of younger people	β Blockers impede cardiac reactivity, further reducing homeostasis, which is already compromised in elderly people
5 Angiotensin converting enzyme (ACE) inhibitors prolong life in heart failure	Not used when heart failure is caused by aortic stenosis because obstruction is proximal to site of drug action, despite no direct evidence on this point

was effective only in men and not in women. Some rather tendentious biological explanations were invoked retrospectively to explain this phenomenon, but we now know that the treatment works equally well in both sexes. An example from the ISIS 2 study of clot busting drugs for probable heart attack serves to further illustrate the point.⁵ Here, the protective effect of clot busting drugs was not seen in patients whose treatment was delayed for a considerable period of time after the onset of symptoms. The treatment also turned out to be ineffective in those born under the astrological star sign Gemini. Clearly these two observations are of altogether different epistemological significance; it is highly plausible that clot busting drugs would be ineffective once the clot had “organised”, while hardly credible that astrological star signs would affect the action of these drugs. To argue that the difference between these two observations lies in the notion that one was a prior hypothesis, and the other not, can be challenged by observing that the results concerning duration of symptoms would be equally impressive (and believable), even if the investigators had forgotten to specify duration of symptoms as a prior stratification variable. It is biological plausibility that makes the essential difference to our interpretation. Put another way, such plausibility is an essential and inevitable ingredient in our “prior” beliefs—in this case beliefs about the existence of a subgroup difference. The question, however, is how to integrate these prior beliefs into statistical analysis. We can think of one way only and that is to transform indirect data (that is, in this case data relating to the biological plausibility) into a “prior” expressing our beliefs in what a perfect set of direct data would show. Note that plausibility may be derived from laboratory experiment, epidemiology, or qualitative studies/experience¹⁵—a point to which we return.

Statistical analysis and public policy

Statisticians frequently emphasise the point that there is nothing magical about the p value. However, the (near universal) use of p values inevitably results in a dichotomy. The use of confidence limits does not resolve this problem—they either do or don’t cross the line of no effect. It is not clear what a decision maker is to make of a result where the confidence limits fail to exclude “no effect” by a greater or lesser amount. It is very hard for a policy maker not to act in the face of a “statistically significant” result. Such was the case with the most recent “pill scare”, when observational studies purported to show a doubling of the rate of venous thrombosis in those taking third generation oral contraceptive.¹⁶ Reanalysis of the results, using a Bayesian perspective, showed that the extent of the risk may well have been overestimated but, more important, a possible countervailing protective effect on myocardial infarction was not brought into consideration, because this was “not statistically significant”. However, this generation of pill is known to improve the profile of blood fats, so that prior belief and the data were both

KEY POINTS

- The probabilities used in clinical and managerial decisions should be based on prior probabilities and direct comparative data.
- Prior probabilities are subjective; they involve mentally transforming indirect evidence into a distribution representing belief.
- Subjectivity is an ineluctable component of science, but quantification makes this transparent.
- Qualitative research is a source of indirect evidence on which prior belief may be based and it is especially relevant to social and managerial decisions.
- Bayesian approaches are useful in the design of studies and flexible monitoring of accumulating data during a study.

orientated in the same direction. A reanalysis of the results using a Bayesian and decision analytic framework (which takes account of the fact that heart attack, although rarer than venous thrombosis, has a higher mortality) suggested that it was still not at all clear whether third generation pills cause more or less harm.⁷

Rapidly evolving technologies

It is difficult to know when to start a clinical trial when technology is developing rapidly; for example, when a new type of device such as that used for endovascular stenting of aneurysms is undergoing frequent refinement. If a trial is started too early, then comparisons may involve a method that has become obsolete by the time the results are published. However, wait and equipoise may be largely dissipated. Furthermore, the considerable time required to first obtain the necessary funding and then launch a trial is such that large amounts of potentially useful data are lost between the point at which a technology settles down and the start of a trial. We therefore, advocate early randomisation between “family members” of a new technology and conventional care. In some cases, it may be possible to factor in the different examples within a family of new treatments, but more often skills and experience are peculiar to a particular method. Thus, while comparisons between the new family of treatments and the standard are randomised, intra family comparisons are typically observational, but can be expressed not just in direct terms but also relative to control treatments. Furthermore, the greater the number of comparators, the less evidence is required before an unpromising method should be abandoned—an intuition confirmed by theoretical studies.^{17 18} We have argued elsewhere that analysis within a Bayesian paradigm is ideal under such circumstances.¹⁹ We argue further, that perhaps results should be made publicly available at regular intervals, so that anyone who may be affected can make a personal decision according to their own priors and values.

Feedback trials: an alternative to data monitoring and ethics committees

Conventional trials are scrutinised at regular intervals by a data monitoring and ethics committee (DMEC). The results, however, are sequestered until the end of the study. Data monitoring committees thus have some horribly difficult issues to tackle, but at heart, they are required to make a trade off between the welfare of “near patients” (patients who will go into the trial if it continues) and “far patients” (patients who may benefit from more precise results in the long term). The rules that guide this awesome trade off are completely opaque and the public, who are affected by these decisions, are kept at arms length in the debate. We argue that DMEC decisions should not be held “in camera” or, at least, more empirical evidence of the acceptability of the practice to the public should be sought. An alternative, transparent system would allow people to make up their own minds about whether or not they wished to participate in an ongoing study, in the light of prevailing data (both from within and external to the index trial). Such a policy would increase public understanding of what is going on and in some circumstances may actually increase recruitment, as different patient/doctor pairs would be equipoised at different stages as data accumulate.²⁰

The need for flexible research commissioning

Tracker and feedback trials require a more flexible approach on behalf of those who commission research. In addition, there are other reasons, apart from rapid evolution of technology, which may make it difficult to specify an entire protocol in advance. In some cases, piloting may be required to prove that recruitment is feasible. In others, scoping may be necessary to establish that sufficient variety of practice exists to justify a large scale study. Moving from a general area of concern to a crisp research question may itself involve a fairly extended process of inquiry in its own right. If those various “preliminary” stages are to progress smoothly and expeditiously, then those who commission research need to be much more heavily involved in projects as they unfold, and need to have the intellectual skills to steer these projects from their inception onwards. They also need to be able to take rapid “executive” decisions as projects unfold. Such decisions, if they are to be soundly made, need to be based on explicit models wherever possible. If commissioners are to move to a more flexible—iterative—mode of commissioning, then they will need to justify their decisions by reference to such models, rather than exclusively by reference to committees. This is described in more detail elsewhere.²¹

While the traditional method of commissioning research, stereotyped in the introduction to this paper, has produced some spectacular successes, it has also opened up a yawning chasm between the worlds of research and practice—particularly managerial practice. The arcane and stylised statistical processes of modern Health Technology Assessment offer

little to hard pressed health services managers. From time to time mainstream research in its current form produces a clear cut answer but it leaves many service delivery issues almost untouched. For R&D to rely solely on a form of statistical analysis originally devised for investigation of crop yields, risks alienating researchers from health service managers. No wonder managers fall back on management consultants and qualitative research in which outcomes are not measured, but people are simply asked what they thought of this or that innovation—a form of inquiry that has been justly parodied as “how was it for you research”. This takes me to the issue of qualitative research itself.

Qualitative research and the decision maker

Qualitative researchers interview people in open format, they observe them in real time or they study documents and other archived material. While their results may subsequently be rendered in a form suitable for statistical analysis, the data are not collected in a pre-specified numerical or ordinal form. This has the enormous advantage that it enables the research to produce findings the existence of which (not just the direction or size of which) the investigator had not considered in advance. This advantage does not, of course, come “free”, as the interpretation of results collected in open format is more subjective, and hence open to implicit biases, than is the case with a quantitative study. However, we will not delve here into such biases and ways of protecting against them.

Qualitative research has three uses that are not controversial:

- 1 By exploring a topic or domain in a relatively unstructured way, new hypotheses (for subsequent formal testing) may be generated.
- 2 By finding out what is important to people who are affected by decisions better methods of measuring outcomes may be devised.
- 3 By helping to explain results obtained in a piece of research through the meanings that people place on events, further hypotheses may be generated.

In summary, qualitative methods may produce better tools for the research process, generate altogether new hypotheses, and explain why results were as they were. But what about decisions—can qualitative research inform decisions directly and if so, how?

Qualitative research may contribute to decisions by helping work out the values and preference weights attached to each outcome. But what about probabilities? If we go back to the earlier section on decision analysis, we see that logically decisions require probabilistic information—remember we said that: given scenario A, what are the probabilities of outcome M, or N, or . . . given decision X, or Y, or . . . One cannot escape from the constraint that decisions require probabilistic information; for example, that M is so much more likely to happen with X, and N is so much more likely to occur with Y, etc. If qualitative

research is to be directly useful in providing probabilities for decisions, then it must be able to influence our estimates of the effects of X and Y on the probabilities of M and N. You will also remember that in the section on quantitative clinical studies, we alluded to biological plausibility, and indicated that this could be quantified by producing a Bayesian “prior”. Note that in this process, data of one sort (clot busters work by dissolving clots, and clots organise after a number of hours) are transformed (by someone) into data of another sort (in which the reduced effectiveness of clot busting drugs in heart attack is quantified). We argued first that such transformation was logically necessary to produce the kind of probabilities that a decision maker needs (such “posterior” probabilities could be calculated only when a “prior” was given), and also that taking such indirect knowledge into account is both natural and desirable when decisions have to be made. It seems to us that precisely the same paradigm is necessary when interpreting the effectiveness of social interventions. Quantitative social experiments (including those involving service delivery) should also be interpreted against a Bayesian “prior” and in many cases, qualitative work will form the basis for such a “prior”. The degree of scepticism and enthusiasm manifest in such priors will vary from case to case.

In the most extreme instances, we may conclude that a formal qualitative study is not even necessary—some things may be taken, as the American Constitution would have it, as self evident truths. In these extreme cases the qualitative “work” consists simply of our own life experiences. For example, few of us would wish to dispute the notion that patients who cannot speak the language of their caregivers will do better if they can avail themselves of the services of an interpreter. One would not wish to do a formal study simply to try and prove that people who cannot communicate with professional staff should be eligible for interpretation services.

Other topics addressed by formal qualitative work may produce results so convincing and/or that resonate so strongly with our intuitions, as to provide a basis for action without the need for further quantitative work. For example, Dearden and colleagues showed that children do not fraternise on paediatric wards,²² and this result was so powerful that without any further direct comparative studies, barriers to family visits were reduced or removed. Again, qualitative work (this time of a rather more formal nature), interacted with a “prior” hunch to generate a sufficiently strong “posterior” to lead people to decide that children would fare better under a more permissive visiting regime.

Another example comes from a study of social workers and their clients showing a mismatch of expectations—the social workers wanted to give counselling therapy while their clients wanted mainly practical help.²³ This seemed strong enough to at least slant practice in the latter direction (a prescription for action) but it also begs questions about when, and in

what form, counselling may be most effective (a prescription for research).

In yet other cases, qualitative work may influence our beliefs, but rather cautiously. In figure 3 we give an invented worked example of a situation where the qualitative work would contribute to a very cautious “prior” belief (the topic is an intervention aimed at reducing unwanted teenage pregnancies). Here, qualitative work suggesting that children were not knowledgeable about conception, but would welcome any educational intervention, leads to a rather cautious prior about the effects that would materialise in practice.

In the case of biological sciences, we have argued that clinicians should be exhorted to adopt a somewhat sceptical posture and allow the lessons of history to influence “priors” developed on the basis of laboratory data. It would seem that a similar scepticism should be invoked when drawing a Bayesian prior on the basis of qualitative data. Many service delivery interventions have turned out to be useless or harmful, including: helicopters to transport accident victims²⁴; provision of staff with paramedic skills in ambulances²⁵; routine counselling for survivors of disasters²⁶; and use of “care management” (also referred to as “standard” or “brokerage” case management) for patients with severe mental illness.²⁷

Action research

Action research gained currency in a number of fields, after its inception in 1940s; especially those of teaching in school, social care and business management. Action research has also gained ground in several branches of health science; principally in health promotion, in nursing and to a limited extent in primary care. According to some of its proponents, action research constitutes an approach to research radically different to that hitherto accepted as a requisite in the health services field. Thus, some proponents of action research in relation to health services are keen to position their enterprise as a specific challenge to customary ways of conducting mainstream research. A theoretical paper by Susman and Evered,²⁸ often referenced in the writings of proponents of action research, makes detailed claims purporting to show that action research can legitimately anchor itself in a distinctive, anti-rationalist philosophical tradition as least as legitimate as that invoked by the rationalist scientific approach. Our colleague Brian Morrison and ourselves have examined this claim.

Action research has five tenets:

1 The flexible planning tenet

The detailed content and direction of a research subject are not to be determined at the outset. These only take on a definitive shape as the work progresses and are kept continuously under review.

2 The iterative cycle tenet

Research activity is to proceed by a cycle of considering the problem, proposing action, taking action, learning lessons from that action

and then reconsidering the problem in the light of those lessons. Each of these phases is carried out in consultation with stakeholders.

Figure 3 (A) A prior probability density for (log) relative risk of pregnancy of a 15 year old girl in one year, with or without receiving a sex education pack. Note that the originator of this prior is near certain the relative risk is between 2.0 (a doubling of risk) and 0.1 (a 10-fold reduction in risk) and judges there to be a 75% chance that there will be a reduction in risk. There is a 70% chance that the true relative risk lies between 1.33 (a $\frac{1}{3}$ increase) and 0.5 (half the risk). (B) The researchers' beliefs change on consideration of a (new) comparison between Belgium and the Netherlands where the children are more knowledgeable. The change is equivalent, in this researcher, to having observed 7/1000 versus 12/1000 pregnancies in the intervention/control arms of an RCT—indeed this is one way of quantifying how much weight a person wants to give to a “qualitative” or indirectly relevant piece of evidence. (C) Results from a RCT become available, in which 30/3000 15 year old girls in the control arm become pregnant, compared with 18/3000 in the intervention arm. The likelihood curve shows how the probability of the (fixed) observed RR varies as the assumed value of the true RR is varied. It is a maximum when the assumed true RR is equal to the observed RR. The posterior is simply calculated by weighting the prior by the likelihood (or vice versa).

3 The subjective meaning tenet

The situational definitions and subjective meanings that those directly affected attach to the problem being researched must be investigated and must also be allowed to determine the content and direction of the research project.

4 The simultaneous improvement tenet

The research project must set out to change the problem situation for the better in the very process of researching it.

5 The unique context tenet

A research project must explicitly take into account the complex, ever changing, and hence unique nature of the social context in which the project is carried out.

We have seen from this paper that with the exception of the last of these tenets, all the remaining tenets may, and have been, incorporated in mainstream research. Firstly, it is now accepted dogma that the perspectives of service users and clinicians should be taken into account in the design of research and psychological outcomes should be measured; such factors are formally and explicitly included in decisions that are modelled using expected utility theory (decision analysis). The concept that study design should be flexible is central to the tracker trial concept. Iteration is, as the name would suggest, intrinsic to the iterative commissioning process. Lastly the notion that the research itself might change outcomes is fundamental to orthodox scientific thinking and finds expression in, for example, the use of placebos and the blinding of measurement of treatment outcome. Clinical researchers have made much of a putative generalised (beneficial) “trial effect” that is alleged to operate across both arms of a clinical study. Therefore, we strongly dispute the notion that there is some philosophical distinction to be drawn between action research and mainstream research—the latter has borrowed many ideas from action research, even if it has done so inadvertently. The exception, of course, is the final—the unique context—tenet. Mainstream researchers have no difficulty with the notion that each person is unique, or indeed that each situation is unique—they deal with this problem in part by inflating sample size in proportion to the variance in their samples and in part by using judgement when extrapolating findings from one place and time to another.

Where they may part company with action researchers is if the latter adopt a “hard-line” insistence that different people in different situations are not only unique, but so unique that nothing can be usefully extrapolated from one circumstance to another. However, such a position is tantamount to arguing that there is no value in the publication of action research findings, and that the research is nothing more than a management tool of interest only in its local context. This extreme position would seem to preclude action research from research budgets, at least in so far as they are directed to producing generalisable knowledge. Consequently, we suspect that this “fundamentalist”

position on action research is not widely held, even by those who consider themselves “action researchers”—put another way, we suspect that most action researchers are prepared to relax the fifth tenet above. As the remaining four tenets are fully acceptable within mainstream research, we conclude that there is no fundamental difference at a “deep” philosophical level, between mainstream and action research. What we fully accept, however, is that inquiry may take different forms along a number of different dimensions. One of these is statistical power, something with which we believe we need be less obsessed, given the rise of meta-analysis. Another, is the extent to which the scope and nature of a piece of research are set in stone at the outset.

Conclusion

Much applied research is funded and carried out with the principal objective of informing decisions. Decisions turn on probabilities and values—it follows that statistical analysis should produce probabilistic information that can be used in such decisions. Decision analysis makes explicit the assumptions about probabilities and values contributing to a decision. Bayesian statistics is a component of decision analysis, and this is the source of the advantages of the Bayesian approach:

- probabilities in a form directly useful for decision taking
- the need to make explicit all assumptions contributing to the decision

Assumptions are inevitable when “extrapolating” from an observed frequency to the probabilities that underlie decisions at the bedside or in the boardroom. Some of these underlying “assumptions” derive from indirect evidence and are captured in a “prior” belief. Much rational behaviour can be predicated on the logic inherent in updating such priors with the direct comparative data—a small “significant” trial may not change behaviour because it would not have moved a sceptical “prior”, while a ‘non-significant’ result, in the traditional sense, may be the rational basis for a change in policy if indirect evidence also points in this direction. Qualitative research can impact directly on decisions through its effect on prior belief—the Bayesian paradigm does not consign it to a purely hypothesis generating (and tool providing) role. When the results are context-dependent, or subject to possible bias, Bayesian methods provide a framework for describing how much weight we ascribe to the evidence in a new setting. Like decision analysis itself, it provides the intellectual scaffold for focused discussion by pinpointing why people might wish to make different decisions—and if consensus remains elusive, the prescription for research is provided by a calculation of the amount of data required to produce conversions. And this is the charm of Bayes; unbiased comparative data win through in the end but a rational basis for action is available in the mean time.

We would like to thank the following people who have been formative in our thinking: James Thornton, Reader in Obstetrics in Leeds, with whom we worked on Bayesian approaches to

breech delivery and from whom we have adapted the fetal monitoring example. Sarah Edwards, Research Fellow at the University of Birmingham, with whom we worked on the ethical implications of Bayesian thought. Brian Morrison, Research Assistant at the University of Birmingham, with whom we worked on action research. John Swales, University of Leicester, who provided many of the examples in table 1.

Funding: while writing this paper the authors were financed by the NHS Executive of England, but the views and opinions are our own and do not necessarily effect those of the NHS Executive.

- 1 Crowley PA. Antenatal corticosteroid therapy: a meta-analysis of the randomized trials, 1972 to 1994. *Am J Obstet Gynecol* 1995;173:322–35.
- 2 Anonymous. Acute myocardial infarction: pre-hospital and in-hospital management. The Task Force on the Management of Acute Myocardial Infarction of the European Society of Cardiology. *Eur Heart J* 1996;17:43–63.
- 3 Hampton JR. The end of medical history? *J R Coll Physicians Lond* 1998;32:366–75.
- 4 Thornton JG, Lilford RJ, Johnson N. Decision analysis in medicine. *BMJ* 1992;304:1099–103.
- 5 Anonymous. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Lancet* 1988;ii:349–60.
- 6 Lilford RJ, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med* 1995;88:552–9.
- 7 Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;313:603–7.
- 8 Matthews RAJ. *Facts versus fictions: the use and abuse of subjectivity in scientific research*. Working paper for The European Science and Environmental Forum, 4 Church Lane, Barton, Cambridge, CB3 7BE, UK, 1998.
- 9 Gilks WR, Thomas A, Spiegelhalter DJ. A language and programme for complex Bayesian modelling. *Statistician* 1994;43:169–78.
- 10 Mainland D. Statistical ward rounds—5. *Clin Pharmacol Ther* 1967;8:738–48.
- 11 Edwards SJL, Lilford RJ, Braunholtz DA, et al. Ethical issues in the design and conduct of randomised clinical trials. In: Black N, Reeves B, Fitzpatrick R, et al, eds. *Health services research methods*. London: BMJ, 1998:98–107.
- 12 Edwards SJ, Lilford RJ, Braunholtz D, et al. Why “underpowered” trials are not necessarily unethical. *Lancet* 1997;350:804–7.
- 13 Britton A, McPherson K, McKee M, et al. Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment (Southampton, UK)* 1998;2.
- 14 Vandembroucke JP. 175th anniversary lecture. Medical journals and the shaping of medical knowledge [published erratum appears in *Lancet* 1999;353:848]. *Lancet* 1998;352:2001–6.
- 15 Porta M. Re: “Biologic plausibility in causal inference: current method and practice” [letter]. *Am J Epidemiol* 1999;150:217–19.
- 16 McPherson K. Type 3 errors, pill scares, and the epidemiology of oral contraception and health. *J Epidemiol Community Health* 1999;53:258–60.
- 17 Strauss N, Simon R. Investigating a sequence of randomized phase II trials to discover promising treatments. *Stat Med* 1995;14:1479–89.
- 18 Yao TJ, Begg CB, Livingston PO. Optimal sample size for a series of pilot trials of new agents. *Biometrics* 1996;52:992–1001.
- 19 Lilford RJ, Braunholtz DA, Greenhalgh R, et al. Trials and fast changing technologies: the case for tracker studies. *BMJ* 2000;320:43–6.
- 20 Thornton JG, Lilford RJ. Preterm breech babies and randomised trials of rare conditions. *Br J Obstet Gynaecol* 1996;103:611–13.
- 21 Lilford R, Jecock R, Shaw H, et al. Commissioning Health Services Research: an iterative method. *Journal of Health Services Research and Policy* 1999;4:164–7.
- 22 Dearden R, Pill R, Robinson D. *Hospitals, children and their families; The report of a pilot study*. London: Routledge and Keegan Paul, 1970.
- 23 Mayer J, Timms M. Clashes in perspectives between worker and client. In: Tuckett D, Kauffer J, eds. *Basic readings in sociology*. London: Tavistock, 1978.
- 24 Snooks HA, Nicholl JP, Brazier JE, et al. The costs and benefits of helicopter emergency ambulance services in England and Wales. *J Public Health Med* 1996;18:67–77.
- 25 Nicholl J, Hughes S, Dixon S, et al. The costs and benefits of paramedic skills in pre-hospital trauma care. *Health Technology Assessment (Southampton, UK)* 1998;2.
- 26 Wessely S, Rose S, Bisson J. *Brief psychological interventions (“debriefing”) for trauma-related symptoms and the prevention of post traumatic stress disorder (Cochrane Review)*. Oxford: Update Software, The Cochrane Library: 2000:issue 1.
- 27 Marshall M. Case management: a dubious practice. *BMJ* 1996;312:523–4.
- 28 Susman G, Evered I. An assessment of the scientific merits of action research. *Administrative Science Quarterly* 1978;23:582–603.