Bayes and health care research.

ALLMARK, P. J. <http://orcid.org/0000-0002-3314-8947>

Published version

ALLMARK, P. J. (2005). Bayes and health care research. Medicine, health care and philosophy., 7 (3), 321-332.

Running head: Bayes and health care research [This is a second resubmission of MHEP268]

Article type: Scientific contribution.

Title: Bayes and health care research.

Author: Peter Allmark, PhD, University of Sheffield.

Affiliation: Department of Acute and Critical Care Nursing, University of Sheffield.

Name and address of corresponding author: Peter Allmark PhD./ Samuel Fox House/ Northern General Hospital/ Sheffield S5 7AU.

Phone: 0114 226 6858

Fax: 0114 271 4944

Email: p.j.allmark@shef.ac.uk

Word count: 8817

Acknowledgements: This paper has benefited from the comments of Professor Richard Lilford, John Platt, Paul Ramcharan, the delegates at the *Nursing Philosophy* conference 2003 and the two anonymous referees for this journal.

Bayes and health care research

Abstract

Bayes' rule shows how one might rationally change one's beliefs in the light of evidence. It is the foundation of a statistical method called Bayesianism. In health care research, Bayesianism has its advocates but the dominant statistical method is frequentism.

There are at least two important philosophical differences between these methods. First, Bayesianism takes a subjectivist view of probability (i.e. that probability scores are statements of subjective belief, not objective fact) whilst frequentism takes an objectivist view. Second, Bayesianism is explicitly inductive (i.e. it shows how we may induce views about the world based on partial data from it) whereas frequentism is at least compatible with non-inductive views of scientific method, particularly the critical realism of Popper.

Popper and others detail significant problems with induction. Frequentism's apparent ability to avoid these, plus its ability to give a seemingly more scientific and objective take on probability, lies behind its philosophical appeal to health care researchers.

However, there are also significant problems with frequentism, particularly its inability to assign probability scores to single events. Popper thus proposed an alternative objectivist view of probability, called propensity theory, which he allies to a theory of corroboration; but this too has significant problems, in particular, it may not successfully avoid induction. If this is so then Bayesianism might be

philosophically the strongest of the statistical approaches. The article sets out a number of its philosophical and methodological attractions. Finally, it outlines a way in which critical realism and Bayesianism might work together.

Key words

Bayes, Bayesianism, frequentism, critical realism, statistics, Popper, induction, health care, research.

Bayes and health care research

Introduction

In this article I argue that Bayesianism has a positive contribution to make to the philosophy and methodology of health care research. The main philosophical contribution is that Bayesianism provides a plausible account of induction. The main methodological contribution is that the use of Bayes' rule in the interpretation of evidence may be preferable to the dominant method of statistics used in health care (that of frequentism). The methodological strengths of Bayesianism have already been set out in many articles aimed at health care researchers; this article focuses on the philosophical background.

I begin with an outline of Bayesianism method. I then contrast its philosophical basis with that of the dominant statistical method used in health care research, frequentism. I suggest that these philosophical differences are partly behind the dominance of frequentism. However, I argue that frequentism has problems of its own and that an alternative, related view, Popper's propensity theory, might retain at least one of these problems. Next I give an account of the philosophical and methodological attractions of Bayesianism. Finally, I outline an account of research method that attempts to combine insights from Popper's critical realist view of science with the Bayesian account of induction.

1. Bayes' rule and Bayesian conditionalization

The following account is drawn from various sources, particularly Hacking (2001), Earman (1992), Papineau (1995), Worral (1998), Goodman (1999b) and Spiegelhalter *et al* (2000).

At the heart of Bayesianism is Bayes' rule (or theorem). This is a simple and uncontroversial element of probability theory. It is based on the idea that many unknown quantities have a probability distribution. For example, (based on Bland and Altman [1998]), the prevalence of diabetes in a region of the UK might be an unknown quantity. However, we may know from other surveys that the prevalence in the UK as a whole is around 2% and that the prevalence within a number of areas of the UK lies between 1 and 3%. As a result, it is very unlikely that the prevalence in the region of interest would be 0% or 10%. Conversely, it is very likely to lie between, say, 0.5% and 4%. The probability of these various results could be plotted on a graph. This would resemble a normal, bell-curve distribution, with 2% at its apex. This graph would represent the probability distribution for the unknown quantity, prevalence of diabetes in the region.

Suppose that new evidence became available. 1.5% of a survey of 1000 people in the region is diabetic. Bayes' rule tells us how the probability distribution should alter in the light of this evidence. It states that the posterior probability (i.e. the distribution in the light of new evidence) is proportional to the prior probability (i.e. the distribution we had before) times the likelihood. Stated more formally,

$$\textit{Posterior probability of hypothesis} \propto \textit{Prior probability x Likelihood}$$

The likelihood is a function that tells us how probable the new evidence would be if our prior probability were correct. It can be used in Bayes' rule to help form the posterior probability. In the diabetes example, the most probable result of the survey was 2%. If this had been the result of the survey then the effect of the new evidence mediated through Bayes' rule would have been to leave the position of the apex of the probability distribution unaltered but to increase its height. In other words, the bell-curve would have become narrower around the same apex. As it is, the effect of the result showing 1.5% prevalence in the survey of 1000 will be to shift the apex to 1.7%, i.e. to shift the probability distribution to the left.

Bayes' rule can be adapted to the more complex position where we are concerned with conflicting hypotheses. For example, we might have a null hypothesis that a drug will have no effect on mortality and a hypothesis that it will. Provided that these hypotheses are mutually exhaustive, that is, that no other hypothesis about the case is possible, Bayes rule states that,

Posterior probability of hypothesis $\propto$ *Prior probability x Bayes Factor.*

And the Bayes factor is,

Likelihood of hypothesis / Likelihood of null hypothesis

This can be adapted to a situation where the hypothesis is being set against a number of alternative hypotheses. Thus Bayes' rule is one that can be used in the analysis of

many types of quantitative data in health care research, from the search for a single figure to the comparison of the probability of alternative hypotheses.

This much is uncontroversial. The controversy lies in the application of Bayes' rule. In particular, there are at least three controversial assumptions underlying its application by Bayesians. The first of these assumptions is that probability is a measure of opinion; it is subjective rather than objective. Bayesians hold that all empirical statements about the world are beliefs that we hold to a greater or lesser extent. Therefore, we can assign probability scores to these statements: we assign a score close to 1 where we are almost certain that it is true, a score of 0 where we are almost certain it is false. All our beliefs will lie somewhere on this scale.

The other assumptions follow from the first. They are that statistics is an appropriate method for the revision of these subjective beliefs and that Bayes' rule is an appropriate consistent method for doing this. Different people will assign different scores to empirical statements. For example, some will be almost certain that the MMR vaccine is not implicated in autism, others almost certain that it is. Bayes' rule tells us how we should change our views in the light of evidence. It can also be adapted to situations where we are unsure whether or not the evidence has occurred (Howson, 1995). The process whereby we change our probability beliefs in the light of evidence is called conditionalization. Bayesian method has its advocates in health care research but is little used. The dominant statistical method is frequentism.

2. The philosophical background to Bayesianism and frequentism

This section explores the philosophical differences that are at the heart of the debate between Bayesians and frequentists. Briefly these are that i) Bayesianism is based on

a subjective view or probability, frequentism on an objective view; ii) Bayesianism is explicitly an inductive method whereas frequentism is, at least, compatible with the anti-inductive (critical realist) views of Karl Popper. Thus, this section also describes some of the problems with induction that led to the development of critical realism. I suggest that frequentism's apparent ability to avoid these problems (as Bayesianism cannot) is one of its chief appeals.

2i) Objectivism and subjectivism

Bayesianism takes a subjectivist view of probability. For subjectivists, a probability score is a statement of the strength of a belief. A statement beginning, "I am 90% sure that x" reflects a subjectivist view. For objectivists, a probability score is a statement of fact. For example, when we say an unbiased coin has a 50% chance of turning up heads we are stating a fact about the world.

In the subjectivist approach of Bayesianism the key methodological idea is conditionalization (described above). In the objectivist approach of frequentism the key methodological idea is the notion of the long run (or infinite run). Thus we would say that the unbiased coin has a 50% chance of turning up heads because in the long or infinite run it would turn up heads 50% of the time. It follows that if we are to make a statement of probability it must be about an event that, at least in principle, is repeatable in the long run. As we shall see, this causes some problems when it comes to single event probability.

This philosophical difference between Bayesianism (as subjectivism) and frequentism (as objectivism) is reflected in further methodological differences. Because

Bayesians are concerned with beliefs, the starting point of scientific method is the beliefs we have now and the finishing point is the beliefs we have in the light of new evidence. As such, prior beliefs are brought into the interpretation of any new data. By contrast, frequentists (as objectivists) are not interested in beliefs but only in what the data tells us about the world. As such, they do not use evidence from outside the trial in the interpretation of results, although meta-analysis does permit them to add data from different trials to get a more precise interpretation.

2ii) Induction and its problems

A second area of philosophical difference between frequentism and Bayesianism relates to the use of induction. Bayesianism is explicitly inductive whereas frequentism is compatible with attempts to avoid it. To explain this it is necessary first to set out what induction is and why it might be best avoided in scientific method.

Science attempts to gain knowledge through observation, including experiment, and through reasoning. The reasoning it uses can be categorised into two main types, induction and deduction.

In deductive reasoning, the conclusion follows from the premises without risk. In other words, provided the premises of the deductive argument are true, and provided the argument form is a valid one, the conclusion that follows will be true. For example, if I argue that the cat is on the mat and that the mat is in the house I can deduce that, therefore, the cat is in the house. Deductive reasoning, because it takes

no risks, appears to provide no new substantive content; all the information in the conclusion is already contained in the premises.

By contrast, inductive reasoning takes risks; it involves moving from data to opinions not entailed by the data. Hacking (2001) describes three types of inductive reasoning.

i) Sample to population. For example, we noted that drug x worked for 95% of a sample with a disease and we conclude that it will work on around 95% of the population with the disease.

ii) Population to sample. For example, we know that 5% of patients admitted to our hospital are Muslim and we conclude that a random sample of hospital patients will contain roughly 5% of Muslims.

iii) Sample x to sample y. For example, we sample 100 middle-aged men attending a screening clinic and note that 80% are middle-class. We conclude that the next hundred that we sample will contain a similar proportion.

This is probably not an exhaustive list. Richard Lilford (personal communication) describes a type of induction where we move from evidence in one area to a conclusion in a different one. For example, our basic knowledge of science allows us to induce that homeopathy is unlikely to work, or that vasodilators should not be given in cases of aortic stenosis. (He goes on to say that it is a strength of Bayesianism that it is able to accommodate this type of induction).

The information in the conclusions of all pieces of inductive reasoning is new; it does not simply follow from the premises. Furthermore, scientific reasoning appears to

make a great deal of use of such reasoning, as I hope the examples indicate. However, such reasoning is risky; it could go wrong.

The problems with using induction in science have been discussed very extensively (e.g. Chalmers, 1999; Papineau, 1995; Worral, 1998; O'Hear, 1995; and Ruben, 1998). I shall briefly describe four problems.

2ii – a) Hume's problem. Induction cannot be rationally justified because it is based on the assumption that the future will be like the past. For example, it assumes that because apples have fallen downwards in the past they will carry on doing so in the future. However, the only support we can find for this assumption is itself dependent upon induction. In other words, the only reason we have to believe that the future will be like the past is because in the past, futures have been like the past. Thus any attempt to justify induction falls foul of a vicious circle (Okasha, 2001).

2ii – b) Goodman's problem. Induction cannot be rationally limited: it is always possible to induce conflicting conclusions from the same data. For example, the repeated observation of green emeralds allows us to induce the conclusion that all emeralds are green. However, it also allows us to conclude that all emeralds are grue. Grue is the phenomenon of appearing green until, say, January 2010 and blue thereafter (Papineau, 1995). This is sometimes referred to as Goodman's "gruesome" problem or as the problem of projectability (Skyrms, 2000).

2ii – c) The Ravens paradox. Induction can draw upon an absurdly wide range of observations to support a hypothesis. For example, the hypothesis that all ravens are black is supported by the repeated observation of black ravens. However, it seems to gain equal support from the observation of things that are not black and not ravens. Thus the observation of a white swan supports the hypothesis (Ruben, 1998).

2ii – d) The tacking paradox. Induction can use a single observation to support an absurdly wide range of hypotheses. For example, elliptical orbits confirm Newton's theory of gravity. However, it is possible to tack on to Newton's theory a further hypothesis, for example, that the planet Pluto is pitted with green cheese. Induction implies that a hypothesis is supported by evidence if that evidence is a consequence of the hypothesis. Because elliptical orbits are consequential on the "gravity plus cheese" hypothesis, their existence supports that hypothesis (Papineau, 1995).

2iii) Critical realism, objectivism and the rejection of Bayesianism

A key starting point for Popper is his rejection of induction; he believes that it is not rationally justified and that science can and does avoid it (Popper, 1992, 1989). He says that science does not proceed by inducing theories and hypotheses from evidence. Rather it proceeds by the method of conjecture and refutation. From our observations of the world and our puzzlement we create explanatory hypotheses. We cannot prove these are true as induction implies, but we can prove that they are false. For example, the observation of a single white raven falsifies the hypothesis that all

ravens are black. Popper suggests that it is through making bold conjectures and attempting to falsify them that science progresses.

Popper's critical realism is the chief influence on the method of hypothetico-deduction that is the framework adopted, consciously or not, by most quantitative health-care researchers. Thus, for example,

> "In the 1990s we all believe that we reason under a Popperian hypothetico-deductive umbrella." (Vandenbroucke, 1998, p. 15).

Sklar (2000) suggests also that frequentism developed in tandem with critical realism in response to perceived problems with subjectivism and Bayesianism. As a result, frequentism and objectivism *prima facie* sit far better within a Popperian framework than does subjectivism. From this perspective, Bayesianism is subject to a number of criticisms.

In the first place, the subjectivism itself is disturbing to those who would like to perceive statistical probabilities as more "scientific". Bayesianism only tells someone how she ought to change her beliefs in the light of evidence, not what those beliefs should be. By contrast, objectivism, in the form of frequentism, seems able to take the data and deliver a probability judgement that is valid for all.

Furthermore, Bayesianism is explicitly an inductive method. As such it is prey to the induction problems outlined above. To these can be added, from a critical realist viewpoint, the fact that beliefs cannot be falsified. As such, Bayesian probabilities are

not scientific conjectures. By contrast, the probability claims of frequentism can be subjected to attempts at refutation e.g. by repeating an experiment. Popper points out that repetition of experiments is impossible from a Bayesian perspective. This is because the prior probabilities with which a researcher approaches a second experiment will be different to those with which he approached the first one (Popper 1983).

Thus frequentism's predominance in health care research is not just a result of the methodological problems of shifting to the Bayesian alternative, as Winkler (2001) suggests. Rather, frequentism has been adopted because it is perceived as better in the light of a critical realist view of science. Indeed, one of the pioneers of frequentism, R Fisher, took explicitly a critical realist view (Cox, 2001).

However, I shall now argue, first, that objectivism has substantial philosophical problems of its own and, second, that Bayesianism is not wholly at odds with an adapted form of critical realism (that is, one that is adapted in the light of the need to retain some form of induction).

3. Philosophical problems with objectivism in the form of frequentism

In this section I argue that there are significant philosophical problems with objectivism as frequentism. These problems lead Popper to develop a non-frequentist but objectivist theory of probability, which is explored in the next section.

One of the most important problems for frequentism is that it seems unable to give an account of objective, single event probability. These are such things as the

probability that a large meteorite will strike the earth, or that a 41 year old will suffer a heart attack this year. They play a large part in health care and the application of health care research (for example, the question of whether or not a particular person would benefit from a particular treatment). The problem is that such events do not, and cannot be imagined to, repeat in a long or infinite run.

Further problems arise from frequentism's use of the concept of long or infinite runs. If we use the idea of a long run then there is the problem of deciding how long is enough. If we use the idea of an infinite run then there is the problem that any repeating event in a long run will occur an infinite number of times (Papineau, 1995).

There are standard mathematical techniques for dealing with these problems but philosophically there remains unease here. In particular, this is because any probability score is ultimately derived from hypothetical data (i.e. that which belongs in the long or infinite run). Bayesianism is criticised for its use of data from outside an experiment in the interpretation of the data from within it. This is seen as prejudicial. However, frequentism is arguably worse in that it draws upon non-existent data to derive probabilities from its experimental data.

From a critical realist perspective there is a further concern. Frequentism appears to be inductive. Induction, as we have seen, uses the assumption that the future will be like the past. The long or infinite run also makes this assumption; it postulates that the recurring event seen in our sample will recur in an approximately similar way in the future. This is induction. This point of itself does not undermine frequentism, but it does undermine its claim to a special place in a critical realist approach.

Popper perceived the single event problem as devastating to frequentism; thus, whilst he was a frequentist at the time of *Logic of Scientific Discovery* (first published in 1934) he had rejected it by 1956 when *Realism and the Aim of Science* was written. He proposed instead a propensity theory of probability. This is an objectivist but non-frequentist approach and is described and criticised in the next section.

4. Popper's propensity theory of probability and its problems

The propensity theory attempts to establish more firmly the idea that probability is a fact rooted in the world, not a belief about the world. Frequentism does this by reference to the long or infinite run, with the resultant problems described above. Propensity does it by suggesting that a given set of conditions has a propensity to produce an event.

Popper gives an example of two dice; one is biased (towards, say, 6), the other fair. He imagines an infinite run of the biased die interpolated with a limited number of throws of the fair one. In that long run of throws, he asks, what should we take the probability of a 6 to be for any throw of the fair die? He suggests that on the frequentist approach data from the biased die will swamp that from the fair one. Thus the fair-die probability will be the same as for the biased die. He also suggests that this seems wrong; we should want to say that the probability of a 6 would be 1/6.

Popper says this problem can be overcome if we understand probability to be a limiting relative frequency only in a sequence of runs under the same experimental conditions. Setting this criterion reveals frequency to be a propensity of the

experimental conditions to produce a result. For example, we can say that the probability of a particular 40-year-old smoker dying from a heart attack within a year is, say, 1 in 1000. This is based on the experimental conditions (being male, 40 and a smoker) having the propensity to produce that result.

This overcomes at least two of the problems for Popper. One is that this approach can make sense of single event probabilities from an objectivist viewpoint. A second is that such single event probabilities do constitute scientific conjectures in that the experimental conditions can be restated as a testable conjecture (e.g. that a sample of 40-year-old male smokers will die at approximately the rate stated). However, the approach does run into other problems.

One is termed the "reference class problem". The probability of our 41-year-old smoker dying is a product of putting him into various reference classes based on age, sex and smoking habit. This gives us the particular figure stated. However, were we to add another class, say, that he belongs to a genetically "strong" family of smokers none of whom have ever had heart disease, the figure would alter completely. It could be that, given enough information, we would be able to say for certain whether or not this man will have a heart attack in the next year. This "reference class problem" suggests that a Bayesian approach to single event probability is, at present, the only credible one (see Gillies, 2000 for discussion).

As well as this, there is some doubt as to whether Popper's propensity approach does without induction, as he intends it to. In particular, a propensity to produce an event

under the same experimental conditions is a propensity to do so in the long or infinite run – something that I have suggested is an inductive concept.

Popper has been accused of using induction in at least one other area connected to probability. This relates to our tendency to believe that a well-tested hypothesis is more probably true than one which whilst not falsified has not yet been subjected to testing. Popper (1983) acknowledges the widespread feeling that a well-tested theory is more likely to be true than one that is untested. He adds that a well-tested theory is not necessarily one that has been tested often; rather it is the rigour of the test that is important.

The problem for Popper is that this seems to be inductive. Induction is involved when we induce a conclusion wider than that contained in the data (such as a universal theory from particular observations). What seems to happen with well-tested theories is that we begin with two particular facts and induce a universal conclusion. The two particular facts are, first, that a particular theory has survived rigorous testing and, second, that well-tested theories in the past have been much more reliable than untested ones. The universal conclusion is that well-tested theories are more reliable, more probable, than untested ones.

Popper attempts to avoid this inductive manoeuvre through an idea he terms "corroboration" (Popper 1983). He argues that we are easily confused in this area because it is idiomatically correct to say that a well-tested theory is more probable than an untested one. As a result we tend to think that testing increases probability and that, therefore, induction can be said to have a logic that is set down by the rules

and axioms of probability. This is precisely what Bayesianism does. However, Popper claims this is a mistake. He identifies a number of problems. Some of these are set out above as problems with induction. However, Popper suggests there is a further problem in that this idiomatic probability does not follow the standard rules of probability in a number of ways. I shall describe two of these.

The first is that, given two well-tested and unfalsified theories we tend to prefer the one that has the most content, the one that explains the most. If the first theory explains all that the second one does and more we tend to think it is closer to the truth; idiomatically it is more probable. However, in terms of the rules of probability the one with most content is always going to be less probably true. Take the situation where we have two theories, A and B, that explain the same phenomena but where B also explains additional phenomena. Here, A is more probable than B because there are more facts that could render B false. Thus, when we say that theories with greater content are more "probable" we are using the term idiomatically, but not in a way that follows the rules of probability.

A second way in which idiomatic notions of probability do not follow the standard rules of probability is that an unfalsified theory is always going to have a probability of 1, or close to 1, because it is verified by unfalsifying evidence all around. For example, the theory that all swans are white is verified by everything that is not white and not a swan (this is a version of the Ravens paradox set out above).

Whilst Popper is keen to preserve idiom wherever possible, here he believes the confusion is so great that a new term, corroboration, is required. The key point is that

severe testing of theories increases their level of corroboration, not probability. Popper claims corroboration can be derived without induction or Bayesian updating of probabilities in the light of evidence.

However, the theory has been criticised on a number of fronts (e.g. Rosenkrantz, 1994). One problem is that it does still appear inductive. Popper's argument, outlined in the previous paragraph, shows only that induction may not obey the rules of probability in the ways Bayesians claim, not that it does not occur. Corroboration still ultimately seems to rest on some idea that the future will be like the past (i.e. that well-tested theories will carry on being reliable).

A second problem is that if we do not invoke induction, the probability of any universal theory being true is always zero (a point Popper seems to accept). However, many balk at this conclusion. As Rosenkrantz (1994, p. 473) puts it,

> "Unless we track the changes in our confidence by using Bayes' rule to update inductive probabilities, all unrefuted hypotheses remain equally trustworthy and equally testworthy [i.e. there is a] … need to go on testing hypotheses no matter how much (putative) inductive support there is to their credit."

Thus, the theory that blood circulates is just as worthy of testing as a far more speculative theory, say, that BSE can be passed on through a blood transfusion. This is problematic in itself and suggests a further problem for Popper. His propensity theory of probability required the use of "experimental conditions" to permit us to

give single case probability; but if those experimental conditions are invocations of unrefuted hypotheses with zero probability then the single case probability will always be zero. Take the case of the 40-year-old smoker discussed above. The probability of his dying in the next year was based upon hypotheses concerning 40-year-old male smokers as a universal group. As universal hypotheses, the probability of these being true is zero. Hence, the probability of our 40-year-old smoker dying of heart disease is also zero.

I suggest, therefore, that Popper's theories of propensity probability and of corroboration remain problematic. If the criticisms prove to be correct then some form of induction may be inevitable in scientific method. In such circumstances, Bayesianism has a number of attractions.

5. The attractions of Bayesianism

There are both philosophical and methodological attractions to Bayesianism. The philosophical ones include the following.

5i) Single event probability. Bayesianism has no problem with single event probability. Asked to judge, say, the probability of a 40-year-old male smoker having a heart attack within the next year the Bayesian can invoke prior probabilities to deliver a result. Given further evidence, such as the genetic "strength" of the man, the Bayesian can modify that result.

5ii) Good account of induction. In so far as induction is necessary in the development and assessment of science (*pace* Popper) then, as Earman (1992, p.2) puts it,

"Bayesianism is the only view presently in the offing that holds out the hope for a comprehensive and unified treatment of inductive reasoning."

Bayesianism explicitly uses induction and has several counters to the problems of induction. I shall briefly describe some of these.

1. Hume's problem. Hume tells us that inductive inference cannot be justified because it depends upon the assumption that the future will be like the past, an assumption that is itself an inductive inference. However, a Bayesian does not depend on this assumption (Okasha, 2001). Bayes' rule tells us how we should change our beliefs on the basis of data: it does not depend on the world being arranged in a certain way. In particular, it permits fallibility: the beliefs we hold are conjectural and subject to change. The future may not turn out to be like the past, but Bayes' rule will allow our beliefs to shift accordingly in the light of the evidence. Okasha (2001) suggests that not only does this evade Hume's problem, it also provides a more accurate account of how our beliefs develop in the light of evidence.

2. Goodman's problem. Goodman himself suggested one solution to the problem on to which Sklar (2000) puts a Bayesian interpretation. Earman (1992) posits his own Bayesian solution, as does Good (1975). Essentially, Earman's solution draws on the notion of the "washing out" of a hypothesis by the accumulation of evidence. As evidence accumulates, the probability of the hypothesis "emeralds are grue" diminishes (for example, as other

time-dependent notions like grue fall by the wayside when they fail to eventuate). This idea of "washing out" would work equally with less bizarre hypotheses.

3. The Ravens paradox. This is fairly easily dealt with using Bayes' rule. Evidence confirms a hypothesis to the extent that it is likely given the hypothesis and unlikely otherwise. Thus observing a white swan does not confirm the black ravens hypothesis because it is evidence that is likely whether or not the hypothesis is true (Earman, 1992).

4. The tacking paradox. The problem here is that evidence that confirms one theory can also confirm that theory plus an element tacked onto it. For example, elliptical orbits confirm Newton's gravity theory but also that theory plus the hypothesis that Pluto is pitted with cheese. Papineau (1995) says that a Bayesian can challenge the inductive assumption that occurrence of the consequences of a theory confirms the theory itself. Theories do not have to be understood holistically. A Bayesian approach allows for evidence supporting elements of a theory to different degrees. In the example here, the Pluto and cheese element is not confirmed at all by elliptical orbits.

Bayesianism's ability to deal with some of these induction problems lies behind Earman's statement at the beginning of this section. It is not without problems, including some of those developed by Popper, but it is a philosophically promising

account of induction. There are also methodological reasons for a health care researcher to find Bayesianism attractive; I turn to these next.

5iii) Explicit use of external data. Bayesians are explicit in the use of data from outside a trial in the interpretation of results; frequentists use it implicitly and, therefore, less clearly. All research brings with it a huge amount of background belief when results are interpreted. For example, faced with a statistically significant relationship between aspirin and deep vein thrombosis the frequentist will be likely to conclude there is reason to believe the hypothesis of a relationship. By contrast, faced with a statistically significant relationship between a homeopathic remedy and deep vein thrombosis the conclusion is likely to be far more cautious; that this is probably statistical artefact and that further research is required (Linde *et al*, 1997). Thus in interpreting the results, the researcher has brought in prior beliefs about biological plausibility.

There are many other ways in which prior beliefs find their way in to frequentist analysis. One simple one is in setting the level at which results are taken to be significant. The selection of, for example, $P < 0.05$ is based on a value judgment that this level of error is acceptable (Goodman, 1991a). Thus, frequentism's desire to base probability statements on data alone does not succeed.

A frequentist could respond that the problem arises only when people misuse or misunderstand statistics. For example, faced with 95% confidence intervals, or $P < 0.05$, they falsely conclude that they can be 95% sure that the null hypothesis is false. In reality, frequentists know that it is necessary to draw on other background

knowledge, such as the plausibility of the mechanism, before deciding what they should believe in the light of results. A good frequentist will not draw conclusions only from the results of one study. The Bayesian response to this is that their method is superior because it closes this gap between results and belief. Faced with the results of a study a Bayesian can say what your posterior belief should be given your prior one. This links to the next point, that Bayesianism is intuitively better.

5iv) Bayesianism is intuitively better. Bayesians point out that frequentism is often misunderstood (Goodman, 1999a). Whilst frequentism can respond in the way just described, the tendency to misunderstand suggests that the Bayesian reading of statistics fits the way people usually do read them. For example, each piece of epidemiological evidence on MMR and autism is taken to ratchet up our belief of no association rather than to be conclusive in itself. Bayesians believe this is the right way to view probabilities and evidence, but it is not frequentism's way. Thus Bayesianism fits our "natural" way of understanding probability better than does frequentism (Gurrin, *et al* 2000; Winkler, 2001).

There is a further way in which Bayesianism "fits" thinking in health care. Bayesian reasoning is commonplace in diagnostic thinking (see Wulff and Gøtzche, 2000: chapter 4). The clinician is used to taking each diagnostic test result, its specificity and sensitivity, and assessing the probability that a given patient has a given disease. This is Bayesian in at least two ways. First, the probability is a subjective one that rises and falls with each piece of evidence. Second, the probability is a single event probability, that for a particular person at a particular time.

5v) Frequentism wrongly dichotomises results into the significant and the non-significant. This leads to two possible errors (Goodman, 1999a). The first is that results that are non-significant are completely ignored or discarded. Thus a result might indicate a strong trend that Bayesian analysis would cause to change our (posterior) beliefs a little but which has no effect from a frequentist point of view. The second is that statistically significant results are taken to be true and we are required to act upon them. Lilford and Braunholtz (1996) show how this can have adverse consequences in health policy. The finding of significant relationships between, for example, the pill and thrombosis in one piece of research is taken to be conclusive and something that requires immediate action. In Bayesian terms it would be taken as something that alters our posterior beliefs without necessarily doing so enough to require immediate action.

These two points, concerning intuition and the false dichotomies of frequentism, show an overall advantage of Bayesianism; that it does justice to uncertainty. Usually one piece of research is not conclusive. It would be far better if we understood evidence to ratchet our beliefs in a certain direction.

5vi) P-values greatly overestimate the strength of evidence against the null hypothesis. Given fairly plausible assumptions a set of results with $P = 0.05$ is compatible with a situation where nearly half the studies have a true null hypothesis (Sterne and Smith, 2001). Goodman (2001b) shows that the Bayes factor gives a much better indication of the strength of evidence against the null hypothesis.

5vii) Bayesianism can draw on a wider range of research. Frequentism can only be applied with quantitative research. It is to frequentism that we owe the notion that the randomised controlled trial is the gold standard in research because it attends exclusively to the data and is unaffected by prejudicial prior beliefs. Qualitative research is seen as being at best a different, unrelated and inferior discipline. The findings of the two types of research cannot be combined.

By contrast, Bayesianism can draw on a wide range of data. Lilford and Braunholtz (2003) give an example. It involves research into the question of whether school-based training in the management our social encounters can reduce unwanted pregnancies and sexually transmitted diseases. In this context, beliefs can be ratcheted up by a series of qualitative and quantitative studies. A "real life" example of the two types of research being combined is that of Roberts et al (2003). This ability to combine the two types of research will be appealing to health care researchers in the qualitative arena.

A related strength of Bayesianism is that it is able to extract meaning from incomplete data, such as a small trial. This is particularly important where one is researching rare diseases. In such cases, frequentist methods require the use of sample sizes, to detect evidence of effect, that are unobtainable. By contrast, Bayesian methods are able to show the shift in posterior belief that is appropriate given data from a small trial for a range of prior beliefs (Lilford and Braunholtz, 1995)

Winkler (2001) suggests that this apparent strength of Bayesianism is also one of its problems in terms of acceptance in health care research. He suggests that the need to

think carefully about inputs into prior probabilities makes the Bayesian approach harder to use than frequentism. Whilst he goes on to suggest ways of making this as easy as possible, he says that ultimately, "Hard decision-making problems … deserve serious thought" (p. 61).

A frequentist might respond that in the practical choice situation it is possible to draw upon other factors, such as background knowledge. Thus, although frequentism does not explicitly factor in the findings of qualitative research into the interpretation of its results, it is possible for them to use such things in the application of results. For example, a frequentist study's result on the likely success rate of radical mastectomy versus lumpectomy could be combined with qualitative research in deciding which should be provided. However, this response does not address the key point at issue, which is not what should we do but what we should believe. Bayesians are able to draw on the results of different types of study in the interpretation of the results of a specific study; this is something frequentism cannot explicitly do.

Some of the methodological problems with frequentism can be corrected; Sterne and Smith (2001) offer some very useful guidelines to interpreting research in a way that is frequentist but which avoids some of the problems. Nonetheless, there seem to be strong methodological grounds to add to the philosophical grounds for considering a move to the wider use of Bayesian methods in health care research.

However, one barrier to this is the influence of critical realism in this area, particularly in the form of the hypothetico-deductive method and the key role played by attempts to disprove one's hypothesis (and prove the null one). A number of

writers have suggested that some form of combination of Bayesianism and critical realism might be possible (e.g. O'Hear, 1980; Good, 1975). In the next and final section I give an outline of how this might be done.

6. Bayesianism plus critical realism

Bayesianism and critical realism share at least two important beliefs. The first is that humans are fallible and that scientific theories are fallible conjectures. In Bayesianism this is expressed through the idea that hypotheses or conjectures have a level of probability based on the evidence, and that this is never, or very rarely, equal to 1. In critical realism it is expressed through the idea that all (universal) hypotheses or conjectures have a probability of zero, but that well-tested theories are corroborated to a greater or lesser extent.

The second shared belief is that the starting point of scientific inquiry is theory-laden. Popper's refutation of logical positivism used this point. The positivists believed that a non-analytic statement would be meaningful only if there were some (pure) observation statement by which ultimately it could be justified. Popper argued that there are no such pure observation statements. What we observe is the result not just of what is "out there" in the world but also of our beliefs and conjectures about the world. Even a very simple observation statement such as "The sky is blue" has underlying it hypotheses such as that human eyes reliably report the colour of the sky, and the metaphysical hypothesis that there exists an external world.

This argument proved fatal to logical positivism. Popper's account of scientific method (set out above) thus discards the positivist idea that science progresses from

observation to the induction of universal theories and then through to the application of those theories in, for example, predicting future events. Instead the scientist proceeds from an initial theoretical position, through conjecture and testing to a new theoretical position. In Bayesianism an almost identical route is followed. The scientist begins from an initial theoretical position reflected in his prior probabilities, through experimentation and observation to a new theoretical position reflected in his posterior probabilities.

That critical realism shares these central beliefs with Bayesianism lies behind O'Hear's (1980, p.43) statement that,

> "[An] inductive Popperian might be rather similar to a Bayesian."

There are, as we have seen, disjunctions between the two theories and Popper was irredeemably hostile to Bayesianism. The Bayesian, Okasha (2001) suggests that Popper's non-inductive account of science contains two elements, one plausible and the other not. The plausible element is that scientific theories are the product of conjecture rather than inductive inference. The implausible element is that scientists can only attempt to falsify these conjectures rather than prove them, and that this is what (good) scientists try to do: science consists only of unfalsified hypotheses; no theories are taken to be true, or very probably true. Okasha disagrees: scientists clearly try to prove their conjectures rather than falsify them. Furthermore, it is clear that some theories are taken to be highly probable (e.g. the theory that the heart pumps blood round the body).

On this basis, the following is a tentative suggestion of the view of scientific methodology that would result from putting together Popper and Bayes. I shall suggest four main stages.

Stage one. The scientist begins with problems and questions. These may be problems of inadequate knowledge – for example, we see apples fall, or pressure sores form, but don't know why. Or they may be problems of conflicting theories – for example, we don't know whether human behaviour is due primarily to our genetic or our social inheritance.

Stage two. The scientist conjectures trial solutions or hypotheses that explain the phenomena or resolve the conflict.

Stage three. The scientist subjects these conjectures to rigorous testing. For Popper, this testing is aimed solely at falsifying the conjectures. For a Bayesian, the aim is to find the evidence most likely to increase our posterior confidence in the probability of the conjecture being true. However, in practice, the Popperian and Bayesian will be looking for the same thing. This is because the most confirmatory evidence from a Bayesian standpoint is that which is most likely if the hypothesis is true and unlikely if it is false. Thus, if the hypothesis is false, looking for the most confirmatory evidence is also the best way to falsify it (Jeffrey, 1975).

Stage four. A new theoretical position is reached. For Popper, we shall either have falsified the hypothesis or not. If we have not falsified it, we may add it to all the other unfalsified conjectures in our theoretical framework. But note that we cannot

say we have increased likelihood of the hypothesis being true. No matter how many attempts at falsification a conjecture survives all we can say is that it is not falsified, not that it is probably true.

For a Bayesian it is unlikely that we can reach the stage where we are certain that a hypothesis is true or false, but we will be able to say that we have increased or decreased our belief in its probability. If the test has been rigorous and the hypothesis has survived then our shift in belief will be substantial.

This is a simple pen portrait of scientific method. What it suggests is that Bayesianism can be introduced into a Popperian framework, adding the element of induction that might be necessary if Popper's propensity theory of probability and his theory of corroboration were not successful.

One practical point needs addressing: what should the researcher do when s/he wishes to present results to, say, the readers of the *BMJ*? In the first place, there have already been a number of papers published that have used Bayesian analysis alone (e.g. Linde *et al,* 1997; Roberts *et al*, 2002). Furthermore, at least one journal has made attempts to attract Bayesian papers (Davidoff, 1999). There are also numerous articles that explain Bayesianism (e.g. Goodman, 1999b). As such, the use of Bayesian analysis alone in papers submitted to medical journals would seem a reasonable course of action for those convinced of its efficacy. Conventional analysis alongside might be appropriate, however, in this period when we have, in effect, two systems running. This would be helpful both as an aid to those unfamiliar with one or the other system and also as an illustration of the practicability of Bayesianism.

Conclusion

I have tried to show both methodological and philosophical reasons for considering the use of Bayesianism in health care research. Philosophically, Bayesianism is a subjectivist account of probability that can be set against the objectivist accounts of frequentism and of propensity theory. Section 3 set out some objections to objectivism in the form of frequentism.

Section 4 considered Popper's alternative objectivist theory of probability, propensity theory and identified problems with it. To these can be added the general question of whether Popper is successful in ridding science of the need for induction (through both his propensity theory and his theory of corroboration). I suggested that if he is not successful then Bayesianism offers the best hope for an account of the use of induction in science that is compatible with the major insights of critical realism.

Much of the argument in this article has focused on the philosophical reasons to consider Bayesianism in health care research. I have done this, at least in part, because this area seems to have been neglected in the health care literature advocating Bayesianism. Thus, I have described some methodological reasons to preferring Bayesianism over frequentism in health care research, but these are given in far greater detail elsewhere (e.g. Spiegelhalter *et al*, 2000; O'Hagan, Luce 2003). It is to these texts also that the reader should look for discussion of the practicality of the use of Bayesianism in health care research. My general conclusion is that taking together the methodological and philosophical argument, Bayesianism can make a positive contribution to health care research.

References

Bland J and Altman D.: 1998, 'Bayesians and frequentists', *BMJ* 317, 1151.

Chalmers A.: 1999, *What is This Thing Called Science?* 3rd edition. Buckingham: Open University Press.

Cox D.: 2001, 'Another comment on the role of statistical methods', *BMJ* 322, 230-1.

Davidoff F.: 1999, 'Standing statistics right side up: editorial'. *Ann Intern Med* 130, 1019-21.

Earman J.: 1992, *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* Cambridge MS: MIT Press.

Gillies D.: 2000, 'Varieties of propensity', *British Journal for the Philosophy of Science* 51, 807-35.

Good J.: 1975, 'Explicativity, corroboration, and the relative odds of hypotheses', *Synthese* 30, 39-73.

Goodman S.: 1999a, 'Towards evidence based medical statistics. 1: The P-value fallacy', *Ann Intern Med*, 130**,** 995-1004.

Goodman S.: 1999b, 'Towards evidence based medical statistics. 2: The Bayes factor', *Ann Intern Med,* 130, 1005-13.

Gurrin L, Kurinczuk J, Burton P.: 2000 'Bayesian statistics in medical research: an intuitive alternative to conventional data analysis', *Journal of Evaluation in Clinical Practice* 6, 193-204.

Hacking I.: 2001, *An Introduction to Probability and Inductive Logic.* Cambridge: Cambridge University Press.

Howson C.: 1995, 'Theories of probability', *British Journal for the Philosophy of Science* 46, 1-32.

Jeffrey R.: 1975, 'Probability and falsification: critique of the Popper programme', *Synthese* 30, 95-117.

Lilford R, Braunholtz D.: 1995, 'Clinical trials and rare diseases: a way out of the conundrum', *BMJ* 311, 1621-5.

Lilford R, Braunholtz D.: 1996, 'The statistical basis of public policy: a paradigm shift is overdue', *BMJ* 313, 603-7.

Lilford R, Braunholtz D.: 2003, 'Reconciling the quantitative and qualitative traditions - The Bayesian approach', *Public Money & Management* 23, 203-7

Linde K, Clausius N, Ramirez G *et al*.: 1997, ' Are the clinical effects of homeopathy placebo effects? A meta-analysis of placebo controlled trials', *Lancet* 350, 834-43.

O'Hagan A, Luce B.: 2003, '*A Primer on Bayesian Statistics in Health Economics and Outcomes Research*', Chebs/MEDTAP URL: http://www.shef.ac.uk/chebs/ [accessed 5.9.03].

O'Hear A.: 1980, *Karl Popper.* London: RKP.

O'Hear A (ed.).: 1995, *Karl Popper: Philosophy and Problems*. Cambridge: Cambridge University Press.

Okasha S.: 2001, 'What did Hume really show about induction?', *Philosophical Quarterly*, 51 (204), 307-27.

Papineau D.: 1995, 'Methodology: The elements of the philosophy of science', in: A. Grayling (ed.), *Philosophy.* Oxford: OUP, pp. 123-80.

Popper K.: 1983, *Realism and the Aim of Science.* London: Hutchinson.

Popper K.: 1989, *Conjectures and Refutations* 5th edition. London: Routledge.

Popper K.: 1992, *The Logic of Scientific Discovery*. London: Routledge.

Roberts K, Dixon-Woods M, Fitzpatrick R, Abrams K, Jones D.: 2002, 'Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence', *Lancet* 360, 1596-99.

Rosenkrantz R.: 1994, 'Bayesian confirmation: paradise regained', *British Journal for the Philosophy of Science* 45, 467-76.

Ruben D.: 1998, 'The philosophy of the social sciences', in: A. Grayling (ed.) *Philosophy 2.* Oxford: OUP, pp. 420-69.

Sklar L.: 2000, 'Introduction', in: L. Sklar (ed.), *Bayesian and Non-Inductive Methods.* New York: Garland, pp. vii-x.

Skyrms B.: 2000, 'Bayesian projectability', in: L. Sklar (ed.), *Bayesian and Non-Inductive Methods.* New York: Garland, pp. 1-22.

Spiegelhalter D, Myles J, Jones D, Abrams K.: 2000, 'Bayesian methods in health technology assessment: a review', *Health Technology Assessment* 4 (38).

Sterne J, Davey-Smith G.: 2001, 'Sifting the evidence – what's wrong with significance tests?', *BMJ* 322, 266-31.

Vandenbroucke J.: 1998, 'Clinical investigation in the 20th century: the ascendancy of numerical reasoning', *Lancet* 352 (SII), 12-16.

Winkler R.: 2001, 'Why Bayesian analysis hasn't caught on in healthcare decision making', *International Journal of Technology Assessment in Health Care* 17(1), 56-66.

Worral J.: 1998, 'Philosophy and the natural sciences', in: A Grayling (ed.), *Philosophy 2.* Oxford: OUP, pp. 197-266.

Wulff H, Gøtzche P.: 2000, *Rational Diagnosis and Treatment* 3rd edition. Oxford: Blackwell Science.