Axiomathes
DOI 10.1007/s10516-009-9085-0

ORIGINAL PAPER

Questions of Evidence in Evidence-Based Policy

Eleonora Montuschi

Received: 19 February 2009 / Accepted: 15 May 2009

Abstract Evidence-based approaches to policy-making are growing in popularity. A generally embraced view is that with the appropriate evidence at hand, decision and policy making will be optimal, legitimate and publicly accountable. In practice, however, evidence-based policy making is constrained by a variety of problems of evidence. Some of these problems will be explored in this article, in the context of the debates on evidence from which they originate. It is argued that the source of much disagreement might be a failure to addressing crucial philosophical assumptions that inform, often silently, these debates. Three controversial questions will be raised which appear central to some of the challenges faced by evidence-based policy making: firstly, how do certain types of facts candidate themselves as evidence; secondly, how do we decide what evidence we have, and how much of it; and thirdly, can we combine evidence. In addressing these questions it will be shown how a philosophically informed debate might prove instrumental in clarifying and settling practical difficulties.

Keywords Evidence · Policy-making · Facts · Practical objectivity · Transparency

1 Introduction

Nowadays there is an increasing political emphasis about using evidence—in particular, scientific evidence—to inform and develop policy in a wide range of areas: health and social care, housing, transport, education, criminal justice, etc. As Cartwright (2007a, b) well explores, governments are more and more willing

E. Montuschi (✉)
Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK
e-mail: E.Montuschi@lse.ac.uk

Published online: 10 October 2009

to fund evidence-based approaches to policy-making; and government funding is increasingly tied to the demand for evidence.

In the United Kingdom evidence-based policy has become the way ahead for developing social programmes ever since the election of the Labour Government in 1997.[1] In June 2000, the UK Treasury established the Evidence-Based Policy Fund with a budget of £4 million over 2 years.

An example of this commitment is the so called "Sure Start" programme, started in 2001 with the aim of breaking the cycle of poverty by providing children and families with childcare, health, and educational support. This programme was conceived and carried out in evidence-based style: extensive reviews of research findings were compiled to show what approaches and early interventions are most likely to work. Also its execution and continuing evaluation and refinement have been evidence-based.[2]

We find another example of governmental commitment to evidence-based policy making in the decision of the UK Parliament's Select Committee on Science and Technology to establish an inquiry into "Scientific Advice, Risk and Evidence: How Government Handles Them" (November 2005). The inquiry focused on "the mechanisms in place for the use of scientific advice (including the social sciences) and the way in which the guidelines governing the use of such advice is being applied in practice across Government". Its aim was also specifically to "test the extent to which policies are "evidence-based""(Scientific advice… 2006).

The evidence-based movement has also gained importance in Europe. In its 2001 White Paper on governance, the European Union acknowledged that:

> Scientific and other experts play an increasingly significant role in preparing and monitoring decisions. From human and animal health to social legislation, the Institutions rely on specialist expertise to anticipate and identify the nature of the problems and uncertainties that the Union faces, to take decisions and to ensure that risks can be explained clearly and simply to the public. (Commission of the European Communities 2001)

So it seems that the generally embraced view is that with the appropriate evidence at hand, decision and policy making will be optimal, legitimate and publicly accountable; that with the appropriate evidence, bias and arbitrary decisions will be eliminated, or at least monitored and kept at bay.

Unfortunately, we know that things are far from being so straightforward and clear cut in practice. Evidence-based policy making is constrained by a variety of problems of evidence. The evidence may be uncertain (e.g. the long term impacts of radiation from mobile phones); it may be subject to differing interpretations (e.g. global warming); it may be misunderstood (e.g. misunderstandings of conditional probabilistic diagnostics in medicine); or it may be challenged by competing values

[1] A clear sign of this commitment can be found in the 1999 White Paper *Modernising Government,* which called for the "better use of evidence and research in policy making and better focus on policies that will deliver long term goals" and stipulates evidence as a key principle for policy making. See Cabinet Office (1999), p. 16.

[2] For example, proposals to expand the Sure Start programme led to a £16 million research project which intended to establish whether the programme was actually achieving results. See Hunter (2003).

(e.g. GM foods). Given these varying contexts in which evidence is problematic, evidence-based policy making might come to look more like an unduly optimistic generalization, if not a rhetorical statement, than a de facto description of a process of governance.

In an attempt to handle this complex host of problems a considerable amount of work has been put onto how to regulate the use of evidence. New institutional rules, structures and guidelines have been suggested with this aim in mind.

For example, the Campbell Collaboration is an "independent, international, non-profit organization that strives to provide decision-makers with evidence-based information to empower them to make well-informed decisions about the effects of interventions in the social, behavioural and educational arenas" (www.campbellcollaboration.org).

It is modelled on the Cochrane Collaboration for medical research and it plays a key role in the evidence-based practice (EBP) movement, trying to promote a culture which values empirical findings and rigorous research based on them. However, when it comes to advice on what research can be accepted for inclusion in systematic reviews, and on what methodologies are the best deliverers of evidence (and in what sense of "the best"), the debates turn controversial.

Arguably, the source of much disagreement often is a failure to addressing crucial philosophical assumptions that inform, often silently, these debates. Nonetheless, a current view among practitioners is that philosophical awareness has little bearing on how to handle real difficulties in practice.

In what follows I raise three questions which seem to me central to some of the challenges faced by evidence-based policy making, and hopefully will show how a philosophically informed debate might prove instrumental in clarifying and settling practical difficulties.

In the process of showing this, the choice of what type of philosophical analysis is to be used to reflect on the practical issues thrown up by policy-making will have to be assessed. In particular, the philosophical analysis will have to undertake the arduous task of proving itself to be sensitive to the type of problems it is to tackle. Making philosophy and societal issues interact is not a simple, nor an automatic process. An "applied philosophy of science" becomes then a new area of research worth its challenges. In other words, if carried out with due care and open-mindedness, an interaction between philosophy of science and societal issues will produce mutual readjustments in perspective and interpretation which will hopefully prove beneficial to both fields.

2 How Do Certain Types of Facts Candidate Themselves as Evidence?

The whole point of basing a policy on reliable evidence is to eliminate bias and decisions taken on arbitrary grounds. Conceiving policies on the basis of good, solid empirical findings seems just the way to achieve this.

What makes evidence solid? Even before that, what makes "evidence"?

Traditional philosophical theories do not seem to be of great help here. Cartwright, in the paper already referred to, rightly points this out by reference to

probabilistic theories of evidence. These theories focus on the probabilistic relations between evidence and hypothesis. For example many accounts demand that, for e to be evidence for a hypothesis h, e should increase the probability of h: $P(h/e) > P(h/\neg e)$.

Cartwright writes:

> These accounts are good at ensuring that evidence, as they define it, does what it is supposed to do, i.e. providing grounds for belief in h. But in policy making the question of evidence is not quite the same. (Cartwright 2007a, b, p. 4)

Probabilistic theories of evidence strive to attach degrees of certainty to a piece of evidence. However, this is only half of the story. In policy making we already know that evidence for a policy conclusion should make the conclusion probable. In practice, if evidence is to be used in order to make a policy conclusion probable, we need to figure out, for example, *what kinds* of facts make this conclusion probable and under what circumstances; or what makes them *relevant to* the conclusion they are meant to support. Besides, and as we will see later, we also want to be able to be open to new evidence, evidence which—though not conclusive or highly certain at a particular time—might indeed "vouch" in favour of a certain conclusion.

Therefore, Cartwright concludes, for purposes of policy deliberation, definitions of "evidence" in terms of probabilities, seem to put the cart before the horse.

So our question still stands: what makes "evidence"?

Practical guidelines are formulated, purportedly, to answer precisely this question. Nonetheless we encounter problems here too.

For example, there is a widespread recommendation in policy practice to evaluate evidence according to "evidence-ranking schemes".[3]

The idea behind these schemes is that a fixed rank can be given to *kinds* of evidence, and then depending on how the kinds are ranked, they are positioned in a hierarchy within the scheme. The scheme then "adjudicates" evidence from the very best to the second best, all the way down to the worst (less/least reliable).

By glancing through these hierarchical schemes, one cannot fail noticing that best evidence is almost always associated with one particular type of methodology: well conducted randomized control trials (RCTs).[4]

A typical example of a ranking scheme is as follows (SIGN 2004):

(1) ++ RCTs
(2) ++ case–control or cohort studies
(3) non analytic studies, e.g. case reports, case series
(4) expert opinion

[3] Examples of these "ranking schemes" can be found in SIGN (2004); or the Oxford Centre for Evidence-based Medicine Levels of Evidence (2007).

[4] A randomized controlled trial (RCT) is an experiment in which investigators randomly assign eligible subjects (or other units of study, e.g. classrooms, clinics, playgrounds) into groups. Each of the groups receives or does not receive one or more interventions (e.g. a particular treatment). Then the results are compared, and if the observed outcome is statistically significant, then it can be concluded that it has indeed been caused by the experimenters' manipulation, i.e. there is a high probability that the intervention actually works. Blind procedures (single, double, triple to even quadruple) are used to control bias.

What we evince in terms of recommendation from these schemes seems to be the following:

(1) evidence is whatever appears on the list;
(2) the strength of evidence depends on the place a method has on the list;
(3) the recommended form of evidence amounts to something like "go with whatever appears at the top of the list" (i.e. well conducted RCTs).

There are several reasons why RCTs are normally chosen as best evidence providers. One of them is that they have inbuilt assumptions which ensure the results (their results can be directly deduced from the evidence provided). Another is that they calculate evidence in terms of probability, so the higher the probability, the better the evidence (and given that all concepts are operationalised, bias is under strict check).

However, Cartwright makes us aware that accepting RTCs as best evidence entails that we already have a theory of what evidence consists of, namely what RCTs provide (a further case of putting the cart before the horse). From here we then rightly feel entitled to eliminate all the rest. Nothing else quite matches with the ideal of evidence provided by RCTs, therefore it may simply be discarded.

At this point, at least two related orders of problem seem to arise:

Firstly, if we accept that RCTs are *best evidence*, this means that they can be applied universally and that the results they deliver are always the *best results* in terms of evidence, independently of where we use them. However, how do we know that RCTs are the best providers of evidence in any and every context?

Indeed, a well conducted RCT might be effective in one context but not necessarily in another.

Secondly, there are contexts in which other forms of evidence might be more effective in reaching certain conclusions; or the questions we are pursuing might not enter in the range of questions included in a RCT; etc. Why should we simply ignore all the evidence which comes from other sources?

By focusing on the quality of evidence (the better the evidence, i.e. the more probable, the more certain the results) it seems that two crucial related issues (at least in the context of policy-making) are neglected: the *relevance* of evidence (i.e. how to decide when a high quality result bears on the acceptability of a policy proposal, but also how other, though less certain, results might bear on it), and its *effectiveness* (i.e. how to decide that what is efficacious in an "ideal" or "closed" setting, e.g. an RCT, will also be so in a practical, "open" one, e.g. a school, or a family).

The social world is complex, and motley: why should we expect that a "one-size-fits-all" strategy can handle this complexity? Or that whatever might seem to work best in one situation should be the best in principle?

An example might help illustrating this.

In 1995 the Dutch Minister of Health authorised a randomised clinical trial of heroin-maintenance for heavy users. At the time there were 24,000 heroin users in the Netherlands, and about half of them participated in some or other methadone programme. However, the great majority of them did not improve. The aim of the

RCT was to find out whether additional provision of heroin would improve the quality of life (medical, psychological, social) of these drug users.

After years of fierce debates and opposition, the minister and her scientific advisers, arguing that RCT was the "truly scientific" approach which would lead to unambiguous, "objective", evidence-based results, won the battle and the experiment commenced.

Trudy Dehue studied the experiment, and followed it up between 1997 and 2001 to comment on its progress and results (Dehue 2002). Her analysis of the case is detailed and rich, but a few issues can be quickly and relevantly recalled to illustrate the points I am raising in this section.

The first aspect she delves into concerns the composition of the control groups for the experiment. The expectation was that the simple offer of free drugs would be enough to attract participants in large numbers. On the contrary, it was difficult from the start to meet even the minimum number required to make up the groups (and in fact the numbers were reconsidered when it became clear that the initial estimates were not going to be met). What the experiment did not take into account was that addiction is not just a physical affair. It is also a way of life, "part of an alternative culture, which is at least as important as the substance itself" (ib., p. 90).

In order to acquire their daily dose, the participants in the experiment had to go to a maintenance station three times per day, seven days a week, pass a metal detector door, be watched all the time (to avoid smuggling), be randomly selected for urine tests, undergo a thorough medical and psychological test every month… Heroin addicts started complaining that government's heroin "tasted differently", that there was probably something wrong with it, that it tasted like "nasty rotgut". No wonder, Dehue remarks:

> Which wine buff would enjoy the finest glass if it were handed out through a window in a maintenance station? […] Few people take pleasure in a free gourmet dinner followed by intrusive examinations and interviews urging them to break with the gastronomic community. (ibidem)

Heroin on prescription is nothing like free heroin: its recipients had to pay a price, by adhering to a regime of drug-taking which was totally foreign to them.

A second matter raised by Dehue concerns the use of classifications in the experiment. Drug users were defined as "patients", and this, we are made to realize, is not without its consequences. In particular, it partly explains the reason for the difficulty in recruiting participants in the experiment. It has been argued—for example by Nora Storm, spokeswoman of the Rotterdam Junkie Union—that only those capable and willing to adapt to the rules set out by the experiment in the end took part in it; and many left before it ended. Self-selecting strategies emerged within the selected groups, which cast doubt on the purported effectiveness of the results.

Interestingly for us, Dehue does not take issue with whether the methodology of RCT in the case of the Dutch heroin experiment was strictly adhered to, or whether infringements in the protocols of experimental research were responsible for its shortcomings. "Even in perfect RCTs—she argues—in which each collaborator or participant fully keeps to the experimental protocol, the results cannot *represent*

reality as it is".[5] We could rephrase this by saying that the "reality" portrayed by the experiment goes hand in hand with the methodological constraints set out by the experiment itself. So the experiment does not provide "objective evidence" in the sense argued for by both Dutch scientists and politicians, that is evidence which is impartial, unambiguous and credible. What the experiment guarantees is that, given the constraints and assumptions of the RCT set up for this particular case, results follow according to the controlled style of the experiment (what Cartwright would call a "clinched" methodology). But what control do we have on those very constraints and assumptions?

This illustrates our first problem above: there is no reason in principle why RCTs are the best providers of evidence in any and every context—let alone the only providers. The American National Institute of Drug Abuse published a volume on "Qualitative Methods in Drug Abuse and HIV Research" where it is pointed out how only few social scientists pay enough methodological attention to the complex types of behaviour (individual as well as social) related to drug use and HIV infection (ib., p. 91). It is then recommended that an *ad hoc* ethnography could be used to develop such a methodology. This would entail studying natural groups rather than artificial ones, or interpreting drug-culture by participating in it rather than observing, testing and recording it. This is not to be taken as a suggestion that ethnography is more objective than RCTs in analysing the reality of drug abuse (ethnographic methods have indeed their own problems and riddles). It should rather lead us to acknowledge on one side, that social phenomena can (and ought to) be approached by different methods, each of which displays different degrees of effectiveness; and on the other, that ignoring evidence coming from different sources might be not only wasteful, but harmful.[6]

Of course the problem then becomes that of how to compare and combine the different types of evidence, as I will discuss in the final section. Before coming to this, a preliminary and equally important issue ought to be addressed.

3 How Do We Decide What Evidence We Have, and How Much of It?

The way in which scientific information/data are presented and communicated bears on how those data might be used as crucial or relevant or effective evidence for or against certain conclusions. Gerd Gigerenzer in his *Reckoning with Risk* discusses some interesting examples.[7]

In 1995 a "contraceptive pill scare" occurred in Britain. According to the official statement,

[5] Ib., p. 86. Dehue claims that the Dutch experiment was indeed designed in the respect of the highest standards.

[6] On this more general issue see also Cartwright (2007a, b), Seckinelgin (2007).

[7] Gigerenzer (2002). Gigerenzer's examples are discussed in the context of dealing with risk and the uncertainties of daily life. Nonetheless the way they are set out become instructive vis a vis some of the features we are discussing here concerning evidence.

> the contraceptive pill is associated with a 100% increase in the risk of thromboembolism.

This can indeed be taken as strong evidence against its use. However, Gigerenzer asks, how is such a risk calculated? The figure in the statement is one of what is called "relative risk", that is a measure of the efficacy of a treatment in terms of the relative number of people saved.

This though is not the only way to represent risk. A different calculation takes the form of "absolute risk", by measuring the efficacy of a treatment in terms of the absolute number of people saved.

Gigerenzer shows that the statement reporting the figure of 100% increase in risk of thromboembolism can be restated in terms of absolute risk as follows:

> the risk of thromboembolism increases about 1 to 2 in 14,000 women.

Indeed, the new figure makes us look at the official statement with different eyes.

Different ways of calculating the risk associated with mammography screening offer similar results, and prompt similar reflections.

There are here at least three ways of presenting the benefits of screening:

(a) *relative risk reduction*: mammography screening reduces risk of dying by 25%;
(b) *absolute risk reduction*: mammography screening reduces the number of women who die of breast cancer by 1 out of 1,000 (0.1%);
(c) *number of women needed to be screened*: in order to prevent 1 death, the number of women who need to be screened for 10 years is 1,000.[8]

Very rarely, Gigerenzer argues, these types of risk are presented according to the figures provided by absolute-risk calculations. Why? The general public is normally impressed by large figures, the figures which hit the headlines—even though going along with them might create extensive social and individual damage (regarding the two cases just analysed: an increase in unwanted pregnancies and abortions; or physical and psychological consequences for about a million women per year in the US which get false positives as a result of mammography screening).

This, as I take it, is not to argue against screening programmes, or in favour of a reckless use or underuse of the contraceptive pill, but to make a point about the fact that relying on scientific evidence should be approached with a rather nuanced attitude. There are different techniques to calculate, and then to present, communicate and use, the evidence relevant to making certain decisions (or, in Gigerenzer's context, of reckoning with certain risks), to the point that "favourable" or strong evidence might appear not as favourable and strong by using, say, a different method of calculation. Awareness of the pros and cons of these methods should be publicly accessible and criteria for making informed decisions should be made part of the general debate.

[8] There is also a fourth way to present the benefit: "increase in life expectancy" (women between 50 and 69 who participate in screening increase their life expectancy by an average of 12 days). See Gigerenzer (2002, p. 59).

4 Can We Combine Evidence?

As we have seen, evidence can come from different sources, be provided by different methodologies, and be communicated by different means. If we believe that such a variety should not be ignored, how can we compare and combine evidence in view of assessing (often "on balance") the so-called "strength of evidence"—or, at a more basic level, in view of acquiring at least sufficient and yet relevant evidential support?

Two examples will allow us to reflect on this important issue, and give us some clues as to how to address it.

The first comes from the legal field.

DNA sampling, when first introduced as forensic evidence for the identification of the culprit of a crime, was treated by the courts of law as a conclusive proof of identity. It is well known that it is impossible for two individuals to have exactly the same genome (with the exception of clones and absolutely identical siblings). However, in reality things are not so clear cut. Forensic scientists could never have the time and large resources needed to produce a full genetic map from a DNA sample. Instead they use, by routine, a limited number of markers, and at this lower-scale level it is no longer impossible that the same DNA sample could be shared by distinct individuals. However, if this is the case, a DNA match (and its impeccable scientific credentials) cannot count any longer as a "proof" of identity.

Does this then mean that DNA evidence should be inadmissible (due to its unreliability) evidence in the courts of law? Statistical calculation comes to the rescue: we are able to calculate the strength of DNA matches in terms of probabilities, and assess rather accurately how much the evidence speaks in favour or against a person having committed a certain criminal offence. Nonetheless, in the course of a trial different bits of evidence come into play which, because of their nature and origin, might not so obviously be treated statistically. How can science-informed evidence be compared with and weighed against other non-scientific evidence, in view of reaching a balanced, and as far as possible complete and "objective" assessment of the case under scrutiny?

An interesting case illustrates how to handle these questions.[9] In January 1995 Denis John Adams, who lived in an area where a crime of rape was committed, was accused and tried for sexual assault on the evidence of a DNA match between Adams and a sample of semen extracted from the victim. No other incriminating evidence was put forward by the prosecution. The defence presented two bits of counter-evidence: the victim did not identify Adams as her assailant, and Adams's girlfriend testified that they were together at the time when the crime was committed.

Figures attached to the DNA match probability went from 1 in 200 million (favoured by the prosecution) to 1 in 2 million (not excluded by the defence). Still, even accepting the 1 in 2 million scenario, the likelihood ratio:

[9] For a discussion of this case I refer to Dawid (2008) and Lynch and McNally (2003).

$$\frac{\text{The probability of obtaining the DNA match, if Adams is guilty}}{\text{The probability of obtaining the DNA match, if Adams is not guilty}}$$

is 2 million, that is very strong evidence of guilt.

However, the problem in the Adams case is that all the other evidence submitted to the court was evidence against guilt. Should all this extra evidence be ignored, given that the DNA evidence appears to be so overwhelming? Were we to put the DNA evidence in a ranking scheme, it would position itself at the top of the hierarchy by being associated with seemingly undisputable (i.e. quantitatively measured) credentials, unlike other types of evidence, engulfed in highly disputable questions (how do we know that Adams' girlfriend is not lying? Under what and how many circumstances is a victim unable to recognize his/her assailant? etc.). We might then be tempted to follow the recommendation: "stick with the best".

However, a jury, in order to be fair, cannot afford following such a recommendation. But how can it make all evidence bear on the case?

The clever move of the defence lawyer was to try also to put the evidence in favour of innocence in terms of probabilities, with a resulting overall likelihood ratio of 1 in 18 in favour of guilt. If there was cause for reasonable doubt before the defence evidence, after it there can be absolutely no case for conviction.[10]

Dawid's conclusion is that without the principles of probability theory for guidance, forensic scientists would be very unclear as to how to interpret evidence. However, even setting aside the thorny issue of how scientific and statistical evidence can be competently handled by a lay jury,[11] we should indeed question.

- to what extent can/should non scientific evidence be put in scientific form (e.g. statistical form);
- whether the aim of combining scientific and non scientific evidence can only be pursued by "translating" the latter in the language of the former.

Even only a sketchy answer to both of these queries should take into account firstly, that not any sort of evidence can indeed be translated into scientific or formal terms without any loss in meaning or relevance. What about, for example, the demeanour of witnesses, or the credibility of testimony? As has been pointed out: "They involve elements of trust, which are fallible, difficult to justify, and impossible to quantify". Besides, the language of quantification, which has become the emblem of late-modern society, seems not only to carry with it an "intimidating sense of objectivity", but also to project a "misleading or potentially confusing *appearance of objectivity* when applied to "non scientific evidence"".[12] Secondly, even assuming that all the available evidence can be made comparable by adopting

[10] By means of Bayesian calculus, Phil Dawid shows us that what we get at the very end is that there are five chances of guilt in a total of 14, which means in terms of guilt probability $5/14 = 35\%$.

[11] It is interesting to note that the jury, despite struggling with the complex statistical argument which was presented to them, and accepted without objection at trial, reached a verdict of guilt. Clearly, the immense odds of the DNA evidence had an overwhelming effect on the jurors' assessment of the evidence.

[12] See Lynch and McNally (2003, p. 96). On the use and role of numbers in modern society see Hacking (1975), Porter (1995) and Gigerenzer et al. (1989).

a common language of communication, the combination of various bits of evidence in view of reaching an objective verdict or judgment is neither automatic nor straightforward, and is deeply dependent on the method/s of combination adopted—as we learn from a second case study.

The field of this second case is vaccine research.

A storm of controversy in England arose in 1998 around a paper which claimed evidence in favour of the existence of some causal link between MMR vaccine (Measles, Mumps and Rubella) and child autism. The paper, written by Dr. Simon Wakefield (accredited doctor in the field), was published in *The Lancet* (prestigious medical journal) (Wakefield et al. 1998) in a receptive milieu of public apprehension towards the use of vaccines, especially on children. The result was a dramatic decrease in children vaccination, with a consequent decrease in children immunization.

Due to public and media pressure, Wakefield's paper was subjected to accurate examination. A whole host of ethical and procedural violations were raised. For example, a problem of sample bias was pointed out, i.e. the subjects involved in the study comprised a subset of children whose parents were already aware of Wakefield's interests. The findings quoted as evidence had not been reproduced to such an extent which could have been convincing for the experts. The uncertainty of the laboratory evidence used to support the causal claim also raised concern, in particular since evidence against the claim was supported by epidemiological studies. In the end, all this led to a retraction. However, and interestingly for us, none of the objections put forward as "evidence to the contrary" provided *on its own* the crucial clue to the resolution of the controversy.

The question then becomes: how is it possible in situations as controversial as this to reach a conclusion which is *objective*—that is, able to take into account the different types of evidence and combine them is such a way that a well balanced judgment is ultimately reached?

A good starting point for answering this question is to establish what "objective" entails in practical domains. There are three features which are normally referred to in talks of objectivity. Rather than positive features, they are features which point out what ought to be excluded from objective assessment.[13]

The first is the feature that excludes subjective or individual opinion from the process of assessment. Transparency is by routine invoked in the evaluation of the evidential findings claimed to bear on a controversial case. To be transparent entails making the conditions of evaluation not only explicit, but more specifically independent of subjective judgement. Often this is said to be achievable by relying on quantitative methods, rather than qualitative (see Suter 1993). The assumption is that qualitative methods are inherently subjective, in a way that quantitative ones are not.

A second feature which seems to be associated with objectivity is the exclusion of uncertainty. Diversity of opinion, for example, might be a source of uncertainty in judgment. Therefore, it might be claimed, it needs to be reduced in favour of consensus.

[13] In what follows I make reference to the three features of objectivity as discussed in Martin (2006).

A third feature importantly associated with objectivity is the exclusion of values, ideologies and political interference. Such an exclusion would achieve that ideal of objectivity which Daston and Galison, in a different context, have defined in terms of "mechanical",[14] that is that type of objectivity entirely based on methods and instruments which reduce human intervention close to zero.

Interestingly, all three features of exclusion seem to portray objectivity more as an idealized concept, as it would have only few chances to succeed in practice.

As to the first feature, it is not necessarily the case that quantification is ipso facto a guarantee for objectivity. A quantitative interpretation of data can be flawed, it could rely on bad methods, and arguably it is not necessarily the case that such an interpretation does not entail judgements of any sort (let alone value judgments). Besides, as pointed out in relation to the Adams case above, there might be features or aspects which might be untranslatable in quantitative terms, or at least at the cost of loosing specificity.

As to the second feature, it is not necessarily the case that if the findings are uncertain then because of this they are necessarily negligible. Cartwright for example makes a useful distinction between evidence which is conclusive, as it is provided by methods which "clinch" their conclusions, and evidence which is not conclusive, as it is offered by methods which only "vouch" for their conclusions. Examples of the latter include ethnographic methods, qualitative comparative analysis and the hypothetico-deductive method (which, incidentally, is at the heart of successful physics). Clinching methods (such as RCTs) are indeed appealing, though their range of application is quite narrow, and they leave us with the problem of how to decide what to retain and what to discard of all the non conclusive evidence provided by non clinchers (Cartwright 1999, 2007a, b).

As to the third feature, again it seems that it is not necessarily the case that if findings are value-laden then they are necessarily biased. Not all values are alike, and not all values constitute bias. Values might inform and solicit good science (as for example in the case of climate science) to the point that the naturalistic fallacy might arguably cease to be a fallacy. A value-sensitive science might be a better option, besides being a more realistic one, than a value-free one.

All in all, what all these features of exclusion fail to show us is how objectivity should itself be looked at as a practical achievement, and as the result of practical procedures, which take necessarily on board the questions, the methodological assumptions and the empirical findings made available by a context of investigation.[15]

We can now go back to our question above: how is it possible in situations as controversial as the MMR vaccine to reach a conclusion which is *objective*—at least "on balance"? A practical concept of objectivity tells us that the way ahead entails combination. To be objective is not to exclude aspects from the final assessment, should they appear too subjective, too uncertain, too value laden. To be objective is to make use of the evidence available, in a fruitful combinatorial framework.[16] An

[14] Daston and Galison (1992, 2007). For some, the way to achieve this task consists of a proper use of statistical analysis. See for e.g. Mayo (1988).

[15] On how to describe a model of objectivity with these characteristics see my (2003).

[16] Haack (2003) uses the image of a crossword puzzle.

objective conclusion is a conclusion which takes into account different types and sources of evidence (for example, in the vaccine case, epidemiological, laboratory, clinical). Such a combination is not achieved at random. It must rather rely on a variety of methods of combination. A mixture of practical strategies and methods are already in place, for example:

(1) *leave it to the experts*—where somehow the authority of individual scientists comes to play a prominent role, and is taken almost for granted;
(2) simply *pool all the evidence together*—whereby the experts will have a "reference point" that summarises the state of current research;
(3) *literature review*—which offer surveys of relevant studies and then draws a conclusion about what these studies have in common, or what they point out.[17]

More sophisticated methods for combining evidence might include meta-analysis (statistical combination of the results of several studies); or Bayesian nets (complex diagrams that organize a body of knowledge by mapping out cause-and-effect relationships among key variables and by encoding them with numbers that represent the extent to which one variable is likely to affect another); or the already discussed ranking schemes (although these do not seem so much to "combine" evidence, but rather to select certain types of evidence to the exclusion of others).

Each of these methods has limitations, and poses further methodological challenges, as Martin points out in his paper. For example, the "leave it to the expert" strategy immediately raises a question as to who the experts are, and what type of expertise is invoked (as well as how inclusive or exclusive the expertise in question is). Or, in the case of "pool all the evidence together", we might well concede that data can be pooled into a sort of "library of evidence",[18] but the evidence these data elicits is not the sort of thing which can easily be aggregated in view of reaching a univocal conclusion. Besides, we should not expect that these methods miraculously combine all the evidence needed. Finally, and most importantly, any methodological standard fixed apriori puts serious restrictions on our need to respond to new evidence. If we fix standards in advance, and we only allow ourselves to proceed according to them, we might find it hard even to recognize that new evidence has been achieved, let alone put forward.

Nonetheless, all these difficulties should not deter us from pursuing the goal of combination, if one of the aims of policy making is accountability, that is favouring procedures which are on one side both accessible and trustworthy, and on the other successful in achieving objective outputs.

5 Conclusion

Practical objectivity is then a form of objectivity "on balance". It is the only kind of objectivity which fits the domain of policy making, a domain where it is never the case that just one piece of evidence speaks in favour or against a hypothesis.

[17] These are listed in Martin (2006).

[18] The expression is Thomas Jefferson's from Jefferson (2003); quoted by Martin (2008, p. 13).

Evidence for policy always includes conflicting claims, and these claims need then be evaluated and combined in view of reaching a policy decision. It often also includes uncertain claims, which ought not to be discarded simply because of their uncertainty. Uncertain evidence is not necessarily bad, or false evidence; and it is often better that no evidence at all.[19]

The central question then becomes how to arrive at rational, effective and objective decisions when different relevant (or candidate) voices speak differently for different outcomes, or when it is not clear what specific result a voice speaks for. Though such a question is at the core of evidence-based policy making, it might not find adequate answers by relying on policy makers alone; nor on the experts called in to provide the required findings to informing those voices.

Confronting this question firstly entails critical awareness of the principles, the standards, the methods and the epistemological justifications of the methods used to reach certain conclusions. Secondly, making use of evidence in domains other than the scientific ones (e.g. the practical contexts of policy making) does not only raise an issue over the reliability of that very evidence (how scientifically "sound" it is). Most importantly, it raises an issue over the "applicability" of that evidence, where the contexts of application have features and causal powers of their own, able to affect the strength, the legitimacy and the relevance of the scientific findings themselves.

Both fields (meta-methodological awareness and critical applicability) can, I believe, be conquered and mastered by philosophical expertise, for the benefit of all players (including non philosophers). Of course this type of philosophy in practice should be up to the challenges and the sui generis problems coming from practice itself.

Acknowledgments This paper presents some of the issues and questions pursued in the research project "Evidence for Use", hosted by the Centre for Philosophy of Natural and Social Science at the London School of Economics. I am grateful to Nancy Cartwright and the other members of the research group for enlightening discussions over the topic.

References

Cabinet Office (1999) Modernising government, white paper Cm 4310, HMSO

Campbell Collaboration. http://www.campbellcollaboration.org

Cartwright N (1999) The vanity of rigour in economics. Discussion paper series, CPNSS. Expanded version in P. Fontaine and R. Leonard (eds) (2005) The experiment in the history of economics. Routledge, London-New York, pp 135–153

Cartwright N (2007a) Are RCTs the gold standard? BioSocieties 2(2):11–20

Cartwright N (2007b) Evidence based policy and its ranking schemes: so, where's ethnography? (mimeo)

Cartwright N et al (2007) Evidence-based policy: where is our theory of evidence? CPNSS/Contingency and Dissent DP, London. Also published in Beckermann A, Tetens H, Walter S (eds) (2008) Philosophy: foundations and applications. Main lectures and colloquia talks of the German analytic philosophy conference GAP. 6. Mentis-Verlag, Paderborn

Cochrane Collaboration. http://www.cochrane.org

Commission of the European Communities (2001) European governance: a white paper. Commission of the European communities: Brussels. COM

[19] See for example in the case of the precautionary principle.

Daston L, Galison P (1992) The image of objectivity. Representation 40:135–156
Daston L, Galison P (2007) Objectivity. Zone Books, New York
Dawid AP (2008) Statistics and the law. In: Bell A, Swenson J, Tybjerg W-K (eds) Evidence. Cambridge University Press, Cambridge, pp 119–148
Dehue T (2002) A Dutch treat. Randomized controlled experimentation and the case of heroin-maintenance in the Netherlands. Hist Human Sci 15:2
Gigerenzer G (2002) Reckoning with risk. Penguin Press, London
Gigerenzer G et al (1989) Empire of chance: how probability changed science and everyday life. Cambridge University Press, Cambridge
Haack S (ed) (2003) Clues to the puzzle of scientific evidence: a more-so story. In: Defending science—within reason. Prometheus Books, New York
Hacking I (1975) The emergence of probability. Cambridge University Press, Cambridge
Hunter DJ (2003) Evidence-based policy and practice: riding for a fall? J R Soc Med 96(4):194–196
Jefferson T (2003) Unintended events following immunization with MMR: a systematic review. Vaccine 21:3954–3960
Lynch M, McNally R (2003) Science, "common sense", and DNA evidence: a legal controversy about the public understanding of science. Public Underst Sci 12:83–103
Martin E (2006) Evidence, objectivity and public policy: methodological perspectives on the vaccine controversy. APA Proc Address 81(3) (mimeo)
Mayo D (1988) Towards a more objective understanding of carcinogenic risk. PSA Proc 2:489–503
Montuschi E (2003) The objects of social science. Continuum Press, London/New York
SIGN (Scottish Intercollegiate Guideline Network) (2004). http://www.sign.ac.uk/guidelines/fulltext/50/compevidence.html
Oxford Centre for Evidence-based Medicine Levels of Evidence (2007). http://www.cebm.jr2.ox.ac.uk/docs/level.html
Porter T (1995) Trust in numbers: the pursuit of objectivity in science and public life. Princeton University Press, Princeton
Scientific advice, risk and evidence: how government handles them (2006) Evidence Report 15 Feb 2006. http://www.parliament.uk/parliamentary_committees/science_and_technology_committee/sag.cfm
Seckinelgin H (2007) Evidence based policy for HIV/AIDS interventions: questions of external validity, or relevance for use. Dev Change 38(6):1219–1234
Suter G (1993) Ecological risk assesment. Lewis Publ, Chelsea
Wakefield A et al (1998) Ideal lymphoid-nodular hyper-plasia, non specific colitis, and pervasive developmental disorder in children. Lancet 351:637–642