Developing Confirmation Theory for Medicine: (i) Classification of Types of Evidence

by Donald Gillies, University College London

Talk in Prague, 17 October 2018

Contents

1. Introduction

I will begin by making a few remarks about the general project of developing a confirmation theory. Scientists, including medical scientists, do use informally a notion of empirical confirmation. I regard the community of natural sciences as a largely rational community. Its members consider various hypotheses and try to judge how well they are empirically confirmed by the data available. If a particular hypothesis comes to be very strongly confirmed, it becomes accepted and is used as the basis of practical applications. However, this acceptance is always provisional. If new evidence undermines the hypothesis, then it is modified, or sometimes rejected completely. It should be stressed that changes of opinion are not instantaneous. Sometimes years elapse before the community changes its mind in the light of new evidence, but they do eventually change if that evidence is strong. What I have said here applies to natural science, and to medicine in so far as it is a natural science. However, it does not apply to the social sciences. In the social sciences, in my opinion, ideological factors are always present, and they are often as strong, if not stronger, than considerations of empirical confirmation. However, I won't here be considering the problems of the social sciences.

Returning to the natural sciences, research in this area does involve an informal notion of empirical confirmation. Confirmation theory, as it has developed in philosophy of science, is an attempt to formalise this informal notion, to explicate it, as Carnap said. Now is this project worthwhile? Why should we not be content with an informal notion? I think that the formalisation project is worthwhile for at least two reasons, but, at the same time, I would like to add a qualification. The two reasons are the following.

First of all an attempt to formalise a notion is at the same time an attempt to make it more precise, and this often brings to light complexities in the notion whose existence might not otherwise have been appreciated. My second reason is more practical. Machine learning is now one of the major techniques in artificial intelligence (AI). Now machine learning, like any other form of learning, does involve the notion of empirical confirmation, which therefore has to be formalised in

order to be used in machine learning programs. Before taking an interest in philosophy of medicine, I worked for about 15 years in the philosophy of AI, and wrote a book on the subject, which was published in 1996. As part of this interest, I investigated how confirmation theory might be formalised in order to incorporate measures of confirmation into machine learning programs. In my 1998 paper, I proposed a measure of confirmation, which proved successful in a particular application of machine learning. Since then, however, I have not studied confirmation theory at all, but in these two seminars I want to look at whether and how my 20 year old measure of confirmation might be modified in order to apply to evidence in medicine.

Thus I do support, and have even engaged in, formal confirmation theory, but now comes my qualification. I think it is important, before starting to formalise, to conduct an informal analysis of how the concept of confirmation is used in the field in question, and then to make to make sure that the formal analysis is largely in agreement with the informal one. Unless this is done, there is a risk that a formal theory of confirmation is produced, which bears little relation to the way in which scientists estimate the empirical confirmation of a theory in practice.

The idea of a confirmation theory for medicine is obviously in the air just at the moment, because in January of this year (2018), the European Journal for Philosophy of Science published a paper by Landes, Osimani, and Poellinger on exactly this topic. These authors adopt a Bayesian approach to confirmation theory, which they augment, following Bovens and Hartmann (2003) by including the new theory of Bayesian networks. Actually, as I will explain in the next seminar, I favour a neo-Popperian rather than Bayesian approach to confirmation theory. However, now I want to stress that Landes, Osimani, and Poellinger do satisfy the qualification just given. They begin by giving an informal analysis of evidence for causal claims in medicine, which is based on Bradford Hill's 1965 paper, where he presents nine viewpoints which he regards as useful in assessing the validity of causal claims.

Now I am a great admirer of Bradford Hill's contributions to the use of statistical methods in medicine, and will say something brief about them in a moment. Nevertheless, I don't think his 9 viewpoints present a very coherent and satisfactory classification of the types of evidence in medicine. They constitute more of a checklist than a good theoretical account. I want next to argue for a simpler, and I believe more insightful, classification of the types of evidence in medicine. My argument is historical. I will show how ideas about evidence in medicine evolved historically; and this will lead naturally to a view, which combines the various historical stages.

2. Initial Predominance of Evidence of Mechanism

Scientific medicine may be said to begin about the middle of the 19th century. Of course anatomy and physiology had been put on a scientific basis before then, but, although these sciences are relevant to medicine, medicine is significantly different. Medicine is concerned at the theoretical level to find out the causes of diseases, and at the practical level to cure and prevent these diseases. A scientific approach to these tasks is hardly to be found much before 1850. Scientific medicine is, in many ways, surprisingly late in appearing.

One of the early pioneers of scientific medicine was Claude Bernard, author of the famous and influential essay of 1865: *An Introduction to the Study of*

Experimental Medicine. Bernard's aim was to turn medicine into a science by basing it on a scientific physiology. This in turn would be based on animal experiments of the kind which Bernard himself carried out. Thus for Bernard, the key type of evidence in medicine was evidence of physiological mechanisms established by animal experiments. While stressing evidence of mechanism in this way, Bernard regarded statistical evidence as of no value whatever. He states this view very strongly and clearly as follows (1865, p. 138):

"… never have statistics taught anything, and never can they teach anything about the nature of phenomena. I shall further apply what I have just said to all the statistics compiled with the object of learning the efficacy of certain remedies in curing diseases. Aside from our inability to enumerate the sick who recover of themselves in spite of a remedy, statistics teach absolutely nothing about the mode of action of medicine nor the mechanics of cure in those in whom the remedy may have taken effect."

Bernard held statistics in low esteem because he believed that absolute determinism was the basis of science. As he says (1865, p. 136);

"I acknowledge my inability to understand why results taken from statistics are called *laws*; for in my opinion scientific law can be based only on certainty, on absolute determinism, not on probability."

As his overall project was to turn medicine into a science, this meant eschewing the use of statistics. As he says (1865, p. 139):

"… if based on statistics medicine can never be anything but a conjectural science; only by basing itself on experimental determinism can it become a true science, i.e., a sure science. I think of this idea as the pivot of experimental medicine …"

It should be remembered that the use of statistics in medicine was a relatively recent innovation when Bernard was writing. Although there had been some suggestions in Britain in the 18th and early 19th century that statistics should be used in medicine, as far as France is concerned, Pierre-Charles-Alexandre Louis's research on the efficacy of blood-letting is probably the first significant use of statistics in medicine. Louis published a paper on this subject in 1828 and a revised and extended version as a book in 1835. This was only three decades before Bernard's essay. Clearly statistics was a new technique in medicine in 1865, and one which Bernard rejected in favour of determinism and animal experiments.

It is interesting to look at Koch's postulates in the light of Bernard's views of 1865. The postulates were formulated in the years 1878 and 1882, 13 and 17 years after the appearance of Bernard's influential essay. Moreover Koch always employed deterministic causality.

Koch's first two postulates in the version given in Gillies (2016) are as follows:

1. The micro-organisms must be shown to be present in all cases of the disease.
2. Their presence must be in such numbers and distribution that all the symptoms of the disease can be explained.

Here postulate 1 gives statistical evidence. All patients suffering from the disease must be shown to have the micro-organisms, presumably by taking a blood sample and examining its contents. This gives statistical evidence in a human population. Postulate 2 gives evidence of mechanism, presumably obtained by autopsies, and animal experiments, where available. However, a point should be noted. The statistical evidence here is supposed to produce a result of 100%, to be in effect completely deterministic in accordance with the requirements of Claude Bernard. Now, when he compared the cholera situation of Hamburg with that of Altona in 1892, Koch did use more genuinely statistical evidence. The difference between the two places was very striking, but fell short of being 100%. Koch relied on this statistical evidence to establish his case, and so, in his practice, recognised the value of statistical evidence. However, he never modified his postulates to allow them to include statistical evidence of this kind. Perhaps this was partly because statistical evidence was still at that time held to be a bit suspect, as a result of Claude Bernard's attack.

3. Contribution of Bradford Hill

Although I am not an enthusiast for Bradford Hill's classification of evidence for causal claims in medicine, I am a great admirer of the contributions he made to the development of statistical evidence in medicine. He organised the first randomized controlled trial in medicine which was the trial of streptomycin as a possible cure for tuberculosis which took place between January and September 1947. He was also a pioneer of the prospective cohort epidemiological survey. He was here one of the first, though not the first, to conduct such a survey. This was designed to investigate whether smoking was a cause of lung cancer and other diseases. It was begun in 1951. He and Doll wrote at the end of October that year to all the doctors on the British Medical Register who were believed to be resident in the United Kingdom to ask them if they would participate in a survey concerning smoking. 34,440 agreed to take part and they were then followed for the next forty years. Their smoking habits were monitored from time to time, and when they died the cause of death was noted. Reports on the results were published occasionally as the study progressed. The great interest of this investigation undoubtedly helped to promote epidemiological surveys of this kind, which became subsequently increasingly common. Indeed statistical evidence became so widely used in medicine that the Evidence-based Medicine (EBM) movement, which began in the 1990s, suggested at least in its hard form that evidence of mechanism be completely replaced by statistical evidence, thus exactly reversing the views of Claude Bernard. I will consider EBM in the next section.

4. Evidence-based Medicine (EBM)

My account of evidence-based medicine is based on Jeremy Howick excellent book on the subject (Howick, 2011). Howick broadly supports the evidence-based medicine position. According to Howick, the proponents of EBM arrange evidence for medical claims in a hierarchy (see Howick, 2011, p. 5). At the top are randomized control trials, then come other types of statistical evidence concerning human groups. However, evidence about mechanisms is considered to be much less important. As Howick puts it (2011, p. 23):

" … when it comes to appraising evidence, randomized trials remain at the pinnacle of a hierarchy … , while expertise and mechanistic reasoning are omitted altogether … or are at the very bottom … ."

As an instance of EBM proponents, who omit evidence about mechanisms altogether, Howick cites Guyatt, Oxman, Vist *et al* (2008), and of those who place it at the very bottom of the hierarchy, he cites Phillips, Ball, Sackett *et al* (2001). The latter is a publication of the Oxford Centre for Evidence-Based Medicine.

There are two EBM positions. Hard EBM is the view that the evidence used to assess medical claims should be entirely statistical in character, and not include any evidence of mechanism. This point of view is to be found in Guyatt, Oxman, Vist *et al* (2008). This expounds GRADE, which the authors claim represents an emerging consensus on rating quality of evidence. In fact they write (2008, pp. 925-6):

"The GRADE system is used widely: the World Health Organization, the American College of Physicians, the American Thoracic Society, UpToDate (an electronic resource widely used in North America, www.uptodate.com), and the Cochrane Collaboration are among the more than 25 organisations that have adopted GRADE."

In their exposition of GRADE, Guyatt, Oxman, Vist *et al* (2008) mention observational studies and randomized trials. They say (2008, p. 926): "observational studies (for example, cohort and case-control studies) start with a 'low-quality' rating." However, they add that in some circumstances this can be up-graded. By contrast (2008, p. 995): "In the GRADE approach to quality of evidence, randomised trials without important limitations, constitute high quality evidence." They also say (2008, p. 925):

"Expert reports of their clinical experience should be explicitly labelled as very low quality evidence, along with case reports and other uncontrolled clinical observations."

These are the only types of evidence considered in the GRADE system. Evidence of mechanism is not mentioned at all. The position is the exact opposite of Claude Bernard's.

Not all followers of EBM are quite so hard, however. There is also soft EBM, an exposition of which is to be found in Sackett, Rosenberg, Gray *et al* (1996). Soft EBM does allow the inclusion of some evidence of mechanism. For example, Sackett, Rosenberg, Gray *at al* (1996) do say (p. 72):

"And sometimes the evidence we need will come from the basic sciences such as genetics and immunology."

This does seem to be a concession to evidence of mechanism, but the authors then go on to stress that for them, and they think, most people, randomized trials have become the gold standard (1996, p. 72):

"Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to

mislead us, it has become the 'gold standard' for judging whether a treatment does more good than harm."

Either version of EBM must surely lead to the conclusion that causal claims in medicine can be established using only statistical evidence, and without taking account of evidence of mechanism. In particular, the EBM position is that causal claims such as 'drug M, taken in such and such quantities, cures disease D without harming the patient' could be established by randomized controlled trials (RCTs) without using any evidence of mechanism.

5. EBM+

I now come to the position which I and my colleagues (Brendan Clarke, Phyllis Illari, Federica Russo, and Jon Williamson) have been developing for the last decade or so. It is I think a position which arises naturally from the preceding development. Our argument is that EBM, especially hard EBM, is wrong to limit evidence in medicine to statistical evidence. Statistical evidence and evidence of mechanism should both be considered and treated on a par when evaluating a causal claim in medicine. As the proposal is to add evidence of mechanism to EBM's statistical evidence, we refer to our position as EBM+. Naturally as we are living in the internet age, we have a website to promote our ideas: ebmplus.org.

On this approach the first basic distinction in evidence for medicine is between statistical evidence and evidence of mechanism. I will now make a few remarks about each type of evidence. First of all statistical evidence here means statistical evidence in human populations, if we are dealing with human diseases. If we were dealing with diseases of say sheep, statistical evidence would be in sheep populations. In either case we are concerned with whole humans, or whole sheep. By contrast evidence of mechanism is concerned with parts of humans or parts of sheep. I am here confining myself to physical, as opposed to mental illnesses. As the disease is physical, the mechanism associated with it will be a bodily mechanism, concerned with organs, tissues, cells, chemicals etc.

For our classification of evidence in medicine, however, we need another distinction, which arises out of the analysis of causation. One group of theories of causation could be called AIM theories of causation, where AIM = Action, Intervention, Manipulation. The idea behind these theories of causation is that causation is closely connected with actions, interventions, and manipulations. In my book, I have argued for a particular theory of this type, which I call an action-related theory of causality. If we adopt an AIM theory of causality, then it becomes important to distinguish between observational and interventional evidence. Observational evidence arises from observing some phenomena without intervening in it in any way. Interventional evidence arises as the result of an intervention, typically an experiment of some kind. AIM theories of causality suggest what could be called the *Principle of Interventional Evidence*, which states:

A causal law cannot be taken as established unless it has been confirmed by some interventional evidence.

In my book, I argue that this should be adopted in medicine.

So now we have a 2x2 classification of evidence in medicine. One dimension is statistical evidence v evidence of mechanism, and the other dimension is observational v interventional evidence. This can be shown in the following 2x2 table.

Table 1 2 x 2 Classification of Types of Evidence in Medicine

	Observational	*Interventional*
Statistical Evidence	Epidemiological Surveys	Clinical Trials
Evidence of Mechanism	Autopsies	Laboratory Experiments on Animals, Tissues, Cells, etc.

As can be seen, the various types of evidence used in medicine fit neatly into this table. This then is the classification of evidence in medicine which I propose to use as the basis of an attempt at formalisation. This formalisation will further be based on an important principle which I will expound in the next section.

6. Strengths and Weaknesses of Different Types of Evidence in Medicine

In the evidence-based medicine approach, types of evidence are arranged in a hierarchy, some being considered better than others. For example, in the quotations I gave earlier, epidemiological surveys were considered to be of lower quality than randomized controlled trials. To obtain high confirmation on this approach, we should try to obtain as much evidence of the best type as we can, largely ignoring lower quality evidence. This point of view I consider to be mistaken. Each type of evidence has both strengths and weaknesses. So there is no type of evidence which is best in some absolute sense. The strategy for obtaining high confirmation is to combine different types of evidence in such a way that the strengths of one type cancel out the weaknesses of another and vice versa. This is what I call the *Principle of Strength through Combining.* It leads to the strategy of looking for evidence of different types and fitting these different kinds of evidence together. This is quite different from the EBM approach of focussing as much as possible on a single type of evidence, considered to be the best, while largely ignoring other types of evidence.

I will now go through the types of evidence given in the 2x2 table, and in each case consider the characteristic strengths and weaknesses of that type of evidence.

Let us start with epidemiological surveys, such as prospective cohort studies. The strength of this type of evidence is that typically it involves a large sample (the cohort or cohorts), and the participants are followed for a long time, maybe 30 or 40 years. The weakness is that the results of the investigation consist of correlations, and, as everyone knows, correlation is not necessarily causation, because of confounding.

Next let us consider clinical trials which nowadays are nearly always randomized controlled trials (RCTs). The strength of such trials is that randomization is a good method for dealing with unknown confounders, and so overcomes the problem of confounding in epidemiological surveys. On the other hand, RCTs have their weaknesses as well. The two principal ones are time limitations and sample limitations. The very first RCT designed to test the effectiveness of streptomycin as a

treatment for tuberculosis is a good example of time limitations, because the treatment gave good results over a period of a few months, but many of those who had initially improved as a result of streptomycin relapsed later. Now those conducting the test realised that there was a problem here, because they considered evidence of the mechanism of the treatment as well as carrying out a RCT. Streptomycin took a long time to kill the tubercle bacilli, and in this period many of the bacilli developed immunity to the drug. The extent of the immunity could be measured. This problem was brought to light by combining the RCT with evidence of mechanism, and the latter also suggested a solution. This was to give as a treatment a cocktail of different anti-tubercle drugs.

Note also that the weaknesses of RCTs (time limitations and sample limitations) can often be overcome by the strengths of epidemiological surveys, which usually involve large representative samples, and are carried out for long periods of time. This shows once again the advantage of combining different types of evidence.

As regards evidence of mechanism, it is usually based on laboratory studies of animals, tissues, cells etc. The strength of the evidence is the strength of controlled experiments, but the weakness concerns whether these results can be carried over to living humans. Results *in vitro* (in the Petri dish) may not apply *in vivo* (in the living animal or human). Results in an experimental animal may not carry over to the human case. These difficulties can be partly overcome by combining the experimental evidence with the observational evidence of autopsies. This evidence has the weakness of all observational evidence, but it has the strength that we know it applies to humans.

Many more examples could be given, but I hope I have said enough to show that it is very advantageous to combine different types of evidence in such a way that the weakness of one type is cancelled out by the strength of another. In my next seminar, I want to look at whether and how we might formalise the principle in the context of confirmation theory.

References

Bernard, Claude (1865) *An Introduction to the Study of Experimental Medicine*. English Translation, Dover, 1957.

Bovens, Luc and Hartmann, Stefan (2003) *Bayesian Epistemology*, Oxford University Press.

Gillies, Donald (1998) Confirmation Theory. In D.M.Gabbay and P.Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, Volume 1, Kluwer, pp. 135-167.

Gillies, Donald (2016) Establishing Causality in Medicine and Koch's Postulates, *International Journal of History and Philosophy of Medicine*, **6**, pp. 1-13. (Open Access Journal)

Guyatt, G.H.; Oxman, A.D.; Vist, G.E.; *et al*. (2008) GRADE: an emerging consensus rating quality of evidence and strength of recommendations, *British Medical Journal*, **336**, pp. 924-926.

Hill, A.B. (1965) The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, **58**, pp. 295-300.

Howick, Jeremy (2011) *The Philosophy of Evidence-Based Medicine*. Wiley-Blackwell.

Landes, Jürgen, Osimani, Barbara, and Poellinger, Roland (2018) Epistemology of causal inference in pharmacology, *European Journal of Philosophy of Science*, **8**(1), pp. 3-49.

Phillips, B.; Ball, C.; Sackett D.; *et al*. (2001) *Oxford Centre for Evidence-Based Medicine Levels of Evidence*, Oxford: CEBM. Available from www.cebm.net/?o=1021.

Sackett, David L.; Rosenberg, William M.C.; Gray, J.A. Muir; Haynes, R.Brian; Richardson, W.Scott (1996). Evidence Based Medicine: What It Is And What It Isn't: It's About Integrating Individual Clinical Expertise And the Best External Evidence, *British Medical Journal*, **312**(7023), pp. 71-72.