No cover image available

Critical Appraisal of Epidemiological Studies and Clinical Trials (3rd edn)

Mark Elwood

https://doi.org/10.1093/acprof:oso/9780198529552.001.0001

Published online: 01 September 2009 **Published in print:** 22 February 2007 **Online ISBN:** 9780191723865 **Print ISBN:** 9780198529552

CHAPTER

9 The diagnosis of causation

J. Mark Elwood

https://doi.org/10.1093/acprof:oso/9780198529552.003.09 Pages 323–358

Published: February 2007

Abstract

This chapter presents a scheme for assessing causal relationships in human health. The approach is that the diagnosis of causation depends on the consideration of both causal and non-causal explanations for the association seen. A reasoned judgement must be reached as to the likelihood of the association seen being produced by causality rather than by any other mechanism. The conclusion as to whether a particular association reflects causation is not a simple yes or no, but requires reasoned and probabilistic judgements.

Keywords: cause and effect, causal relationships, human health, causal reasoning

Subject: Public Health, Epidemiology

Collection: Oxford Scholarship Online

> *What is more unwise than to mistake uncertainty for certainty, falsehood for truth?*
>
> —Cicero (106–43 BC); De senectute, XIX

Part 1: The assessment of causation in one study or in a set of studies

We have now come to the point where we can summarize how to assess whether a particular study or set of studies allows us to decide whether a relationship is causal. As was pointed out in Chapter 1, the question of absolute proof is irrelevant. It can be argued that no amount of data on past experience can ever allow us to predict with absolute certainty the outcome of situations in individuals we have not studied, such as future patients. In health care there are few situations where there is not another factor that could be considered, another study that could be done, or a variation of the hypothesis which could be suggested. To balance this, we have to be able to make judgements in order to make decisions, whether these are decisions about the diagnosis and treatment of an individual patient, or whether they are policy decisions which may affect many people. These decisions must be made by a process of judgement, and that judgement must be based on an objective consideration of the evidence.

In this chapter we present a scheme for assessing causal relationships in human health. The approach is that the diagnosis of causation depends on the consideration of both causal and non-causal explanations for the association seen. A reasoned judgement must be reached as to the likelihood of the association seen being produced by causality rather than by any other mechanism. The conclusion as to whether a particular association reflects causation is not a simple yes or no, but requires reasoned and probabilistic judgements.

An overall scheme to assess causality is shown in full in **Ex.** 9.1, and in a shorter form in **Ex.** 9.2. The questions shown will be dealt within turn.

A SCHEME FOR THE ASSESSMENT OF CAUSATION

A. Description of the evidence
1 What was the exposure or intervention?
2 What was the outcome?
3 What was the study design?
4 What was the study population?
5 What was the main result?

B. Internal validity: consideration of non-causal explanations
6 Are the results likely to be affected by observation bias?
7 Are the results likely to be affected by confounding?
8 Are the results likely to be affected by chance variation?

C. Internal validity: consideration of positive features of causation
9 Is there a correct time relationship?
10 Is the relationship strong?
11 Is there a dose–response relationship?
12 Are the results consistent within the study?
13 Is there any specificity within the study?

D. External validity: generalization of the results
14 Can the study results be applied to the eligible population?
15 Can the study results be applied to the source population?
16 Can the study results be applied to other relevant populations?

E. Comparison of the results with other evidence
17 Are the results consistent with other evidence, particularly evidence from studies of similar or more powerful study design?
18 Does the total evidence suggest any specificity?
19 Are the results plausible in terms of a biological mechanism?
20 If a major effect is shown, is it coherent with the distribution of the exposure and the outcome?

Ex. 9.1. Twenty questions relevant to the assessment of evidence relating to a causal relationship

THE ASSESSMENT OF CAUSATION

A. Description of the evidence
1 Exposure or intervention
2 Outcome
3 Study design
4 Study population
5 Main result

B. Non-causal explanations
6 Observation bias
7 Confounding
8 Chance

C. Positive features
9 Time relationship
10 Strength
11 Dose–response
12 Consistency
13 Specificity

D. Generalizability
14 Eligible population
15 Source population
16 Other populations

E. Comparison with other evidence
17 Consistency
18 Specificity
19 Plausibility
20 Coherence

Ex. 9.2. A scheme for the assessment of causation in note form

p. 324

What evidence do we have?

Consider the practical situation with initially one set of evidence, such as a published study or raw data from our own or others' work. We must critically evaluate the methods used and the results given, and decide whether a causal relationship seems a likely explanation for the results. The questions to be asked are as follows.

Questions 1–5. Description of the evidence

Question 1. What was the exposure or intervention?

Question 2. What was the outcome?

Question 3. What was the study design?

p. 325 Question 4. What was the study population?

Question 5. What was the main result?

The first and often overlooked step is to understand for the particular study exactly what relationship is being evaluated, or, to put it another way, what hypothesis is being tested. We should be able to reduce every study to a consideration of a relationship between an *exposure* or intervention and an *outcome*. It is also necessary to categorize the study in terms of the *design* used; comparative studies of individuals will be intervention studies, randomized or non-randomized, or observational studies of the survey, case–control or cohort design. As was shown in Chapter 3, understanding the design shows what type of analysis is appropriate and indicates which methodological issues will be most important. Then we need to consider the *subjects* studied, in terms of the

source populations, the eligibility criteria, and the participation rates of the different groups compared, as was
p. 326 discussed in Chapter 4. ↳ More than occasionally, describing a published study in this way requires a critical perusal of the methods section rather than simply a glance at the title, because the question which has actually been answered may be somewhat different from the one the investigators would like to have answered.

Two advantages arise from this systematic consideration. First, before getting caught up in the particular details of the study, which may be quite complex, reviewing these questions will give us a clear idea of the overall purpose and relevance of the study. Secondly, it will help us decide whether indeed the study is sufficiently relevant and important for us to review it in more detail. This early consideration of the study design and study population may make it clear that this particular study is not relevant to the question we wish to assess; for example, we may be interested in the use of chemotherapy as part of the primary management of a newly diagnosed cancer patient, but we may find that the study describes the use of the drug in palliation. Thirdly, if there is a great deal of literature available on the topic, this may help us decide which studies are worthy of most attention. Later in this chapter, a system of classifying different types of study in order of their likely relevance will be given; for example, if the question has been addressed in several randomized trials, there is little value in reviewing an uncontrolled descriptive study in any detail.

Having defined the topic of the study, it is very useful to summarize the *main result*: what is the result in terms of the association between exposure and outcome? This step forces us to distinguish the main result from subsidiary issues, which should be considered only after the main result has been dealt with. It should be possible to express the main result in a simple table, and obtain from the paper or calculate ourselves the appropriate measure of association (usually relative risk, odds ratio, or a difference in proportions).

In much current literature, the main result will come from a complex analysis, such as a multivariate analysis. Even so, the raw data are usually available, from which a simple 2 × 2 table of the primary crude result can be generated. Although of course this crude result should not be used in preference to the published more complex analysis, assuming that the analysis is appropriate, it is often a great help in understanding the study and giving us some feeling for the essential content. Moreover, if the simply derived crude result differs dramatically from the published result based on a more complex analysis, it is useful to look at that more complex analysis very carefully.

Internal validity: consideration of non-causal explanations

Having described the study, we assess its *internal validity*, i.e. for the subjects who were studied, does the
p. 327 evidence support a causal relationship between the ↳ exposure and the outcome? This assessment is in two parts; first, we consider the three possible non-causal mechanisms which could produce the result seen. The questions are as follows.

Questions 6 to 8: non-causal explanations

Question 6. Are the results likely to be affected by observation bias?

Question 7. Are the results likely to be affected by confounding?

Question 8. Are the results likely to be affected by chance variation?

These have been dealt with in detail in Chapter 5, 6, and 7. For each, we need to consider how the main result of the study may be influenced. It is useful to consider each separately, making our assessment of the likelihood of the study result being produced by that mechanism compared with a causal effect. Thus, for a study which shows an association between exposure and outcome, the questions can be summarized as follows.

- Could the results seen have arisen by observation bias, if there were no true difference between the groups being compared?
- Do the results show a true difference, but is it due to a confounding factor rather than to the putative causal factor?
- Do the results show a true difference, but one which has occurred through chance, there being no general association between exposure and outcome?

As mentioned previously, while our final assessment will take all these factors into account, and the problems in a particular study may involve all three, considering each in the extreme case of it alone explaining the results seen will often clarify our judgement. The order of these non-causal explanations is relevant. If there is severe observation bias, no manipulation of the data will overcome the problem. If there is confounding, an appropriate data analysis may be able to demonstrate it and control for it; we need to assess if such analysis has been done. The assessment of chance variation should be made on the main result of the study, after considerations of bias and confounding have been dealt with.

Internal validity: consideration of positive features of causation

So far we have considered the recognition of a causal relationship only by the exclusion of non-causal explanations. The new material in this chapter is a consideration of features which when present can be
p. 328 regarded as positive indicators of causality. At this point we will discuss the assessment of these ↳ features within a particular study, and later we will discuss them with regard to all available information relevant to the hypothesis under assessment.

The relevant questions are as follows.

9. Is there a correct time relationship?
10. Is the relationship strong?
11. Is there a dose–response relationship?
12. Are the results consistent within the study?
13. Is there any specificity within the study?

Question 9. Time relationship

For a relationship to be causal, the putative exposure must act before the outcome occurs. In a prospective design where exposed and non-exposed subjects are compared, this is established by ensuring that the subjects do not already have the outcome when the study is commenced. The ability to clarify time relationships is obviously weaker in retrospective studies, and care must be taken to avoid considering as possible causal factors events which took place after the outcome had developed. For this reason, in retrospective studies of disease it is best to enrol incident subjects (those who have just been found to have the outcome), to interview subjects fairly rapidly, and to record only information related to events preceding the outcome.

A difficulty in all study designs, but particularly in retrospective studies, is that the occurrence in biological terms of the outcome of interest may precede the recognition and documentation of that outcome by a long and variable time; often some arbitrary assumption about this time is used. For example, in the retrospective study described in Chapter 14, drug histories of case and control subjects were taken from medical records, but

excluding the time period of 1 year prior to clinical diagnosis in the cases, and an equivalent time in the controls. In the cohort study described in Chapter 12, it is suggested that differences between studies may be due to the key association changing with the time since exposure.

A similar issue may arise in the definition of exposure. For example, in assessing an association between an occupational exposure and disease, it may be reasonable to define exposure as a minimum of, say, 5 years in a particular occupation. In that event, the follow-up period begins immediately the 5-year period is completed.

A study may show no association because the time scale is inadequate; a treatment comparison may give irrelevant results if based on a short follow-up, and long-term effects of an exposure factor such as radiation or oral contraceptive use will be missed by studies with a short time scale.

p. 329

Question 10. Strength of the association

A stronger association, i.e. a larger relative risk, is more likely to reflect a causal relationship. One reason is that as the measured factor approaches the biological event in the causal pathway more closely, the relative risks will become larger. The deterministic ideal is that the factor is the necessary and sufficient cause, which gives a risk of zero in the unexposed and 100 per cent in the exposed, and a relative risk of infinity. However, this is a very rare situation in health issues. Suppose that a rare disease is in fact caused totally by exposure to a specific chemical used in the manufacture of photographic film. In sequential studies, we might detect a weak association with employment in a photographic plant, a stronger one with working in the film manufacture process, and a very strong association with heavy exposure to the particular chemical.

However, a true causal factor may be related to a small increase of risk, as the factor may be one of a number of such factors operating. Consider the role of air pollution in the causation of chronic bronchitis. Where there are few other factors operating to cause the disease, for example in non-smoking subjects who are not exposed to occupational hazards, the role of air pollution may be major, producing a high relative risk which is relatively easy to demonstrate. However, in a heavy smoker, the smoking factor is of such overwhelming importance that the extra risk contributed by atmospheric air pollution will be relatively small; if the attributable risk of air pollution is similar to that in a non-smoker, the relative risk will be small because of the very high baseline risk produced by smoking. This does not alter the causal nature of the relationship, but it does make the strength of the relationship less and makes it more difficult to demonstrate.

The fact that a relationship is strong does not protect against certain non-causal relationships. Severe observational bias may produce very strong relationships. Suppose we identify mothers who have recently been delivered of babies with abnormalities, ask them about exposure to drugs in early pregnancy, and compare their responses with those of mothers of healthy babies. We should anticipate that bias in selective recall might be very considerable. If it operated at all, there is little reason to assume that the bias it could produce would be small; it could quite easily be very large. Similarly, strength does not protect against confounding caused by closely associated factors. An example of this is where a disease risk may be related to a previous drug exposure, or to the reason for that exposure. There may be a close relationship between the indication and the drug, and therefore if one of them is a true causal factor the association of disease with the other factor will be strong despite the fact that it is due only to confounding.

p. 330

However, if a strong relationship is due to bias, the bias must be large and so should be relatively easy to identify. If a strong relationship is due to confounding, either the association of the exposure with the confounder must be very close, or the association of the confounder with the outcome must be very strong. For example, the relative risk of lung cancer in heavy smokers compared with non-smokers is of the order of 30. It has been suggested that this relationship is due to confounding by a genetic predisposition to lung cancer, linked to a genetically determined personality trait leading to smoking. If so, that genetic predisposition factor

must have a relative risk of about 30 for lung cancer. Even given the difficulties of assessing personality, it should be possible to demonstrate the existence of such a relationship.

Question 11. Dose–response relationship

The consideration of a dose–response relationship is similar to that of strength. The major issue which it does not protect against is the relationship being due to a confounding factor closely related to the exposure, such as in the drug versus indication for drug situation. In some circumstances the demonstration of a smooth dose–response relationship may be a strong argument against the relationship being due to bias. For example, it could be argued that women who use oral contraceptives might be more likely to have certain symptoms recorded, simply because they visit their general practitioners more frequently than women who do not use oral contraceptives. However, it is less likely that there is a close relationship between the oestrogen dose of the oral contraceptive and the frequency of visits. Therefore, if the outcome under study shows a regular dose–response relationship with the oestrogen dosage, this bias is unlikely to be the explanation.

We usually expect unidirectional dose–effect relationships. Obviously other types of associations, showing a threshold or all-or-none effect, or a complex relationship may be in fact the true situation. However, the general assumption that if a causal relationship holds, the frequency of the outcome should show a unidirectional increase with increasing exposure, even though the relationship may not be linear, seems very reasonable—so reasonable, that evidence that that is not the case should be considered carefully. For example, the age distribution of Hodgkin's disease does not show the common unidirectional increase of incidence with age as is seen in many other cancers, but instead shows a complex pattern with a peak at younger ages followed by a decrease followed by a further increase. On the general assumption that complex relationships are unlikely, this pattern suggested that there are two distinct diseases, of which one shows the steady increase of incidence
p. 331 with age characteristic of many other cancers, and the other shows a peak incidence at young ages. ↳ This suggestion, made in the 1950s simply from descriptive data [1] was later confirmed by the demonstration of differences in clinical and pathological features between the previously unseparated types of disease [2]. The demonstration of a dose–response relationship is an important component of the study presented in Chapter 14, and is one of the prime objectives of the study described in Chapter 15.

Question 12. Consistency of the association

A causal relationship will be expected to apply across a wide range of subjects. If a new painkiller is effective, it is likely to be effective in patients of both sexes and different ages, for a wide variety of causes of pain. In other circumstances, specificity (see below) rather than consistency might be predicted; for example, a hormonal treatment for breast cancer might be expected to work best for cancers which are positive for hormone receptors. If an association within one study is seen to be consistent in different groups of subjects, that may well be regarded as support for causality, particularly if the likely sources of bias and confounding are different in those subgroups. Similarly in reverse: when a new study showing a positive association between the consumption of artificial sweeteners (mainly saccharine) and bladder cancer was published, the association was seen only in males, and in the absence of a biological explanation for that lack of consistency, it weakened the case for causality [3]. The international trial described in Chapter 11 demonstrates consistency of the main effect with regard to different geographical areas and the main subtypes of the outcome.

The difficulty with consistency is that large data sets are required to assess the similarity or otherwise of associations in different subgroups of subjects; the effective sample size is the number of observations in each subset. Even with adequate numbers, the subgroups to be compared need to be defined on a priori grounds, and not merely generated from the analysis. In a large analysis where many subgroups are defined, it is to be expected that some will show different results by chance alone. This has been a major problem in clinical trials; even where no overall benefit of a new treatment is shown, a benefit may be apparent in one subgroup of

patients. Such post hoc analysis is misleading and best avoided; such findings should be regarded as new hypotheses which require testing.

Question 13. Specificity of association

It has been argued that a specific association between one causal factor and one outcome is good evidence for causality. This may be misleading; some took the view that the fact that smoking was shown to be associated
p. 332 with the ↳ occurrence of a number of cancers and other serious diseases, and therefore demonstrated non-specificity of action, made the hypothesis of a causal link with lung cancer less likely. In the medical area specificity is often contrived by definition. If we define tuberculosis as a clinical disease comprising various signs and symptoms, which is produced by infection by the tubercle bacillus, we end up with a specific association between that disease so defined and the infectious agent. Without that definitional convenience, the associations between infection with tubercle bacillus and chronic meningitis, swollen joints, and lung disease do not appear to be specific.

However, in many situations demonstration of specificity may be valuable, as it may show that bias or confounding is unlikely to explain the results. For example, consider a retrospective study in which recently delivered mothers are interviewed, which shows that use of a certain drug is much more frequently reported by mothers of infants with cardiac malformations than by mothers of healthy babies. We would have to question whether recall bias is the explanation of that association. However, if mothers of babies with a range of other defects were questioned, and their reported histories of drug use were similar to the mothers of the healthy babies, this would be a strong argument against recall bias being the explanation of the association seen with cardiac disease. In a study of birth defects in relation to vitamin A intake, an excess risk was found with high intakes only before pregnancy and in early pregnancy, but there was no association with high intake after 6 weeks of pregnancy; this specificity of exposure is consistent with causality and provides some protection against both confounding and observation bias [4]. The trial described in Chapter 11 was set up to distinguish the effects of two different exposures, and showed a result specific to one of the exposures. Previous observational studies could not separate the effects of these two related exposures. The study described in Chapter 15 also illustrates specificity, as an association between breast cancer and alcohol intake was found for only certain types of alcohol intake, but further work on this result has produced conflicting evidence.

A hospital based study showed that women who had developed endometrial cancer had a higher frequency of past use of oestrogenic drugs than did patients who had cervical cancer [5]. Endometrial cancer is more common in high socio-economic groups, while cervical cancer is less common. The use of this drug is likely to be greater in the higher socio-economic groups, and so the association seen may be due to confounding. However, patients with ovarian cancer, which has a similar socio-economic distribution to endometrial cancer, were also assessed, and their usage of oestrogen drugs was also much lower than that of the endometrial cancer patients. This makes confounding by socio-economic status a less likely explanation for the association seen.

p. 333 Therefore specificity may be useful if we do not make it an absolute criterion, as one causal agent may in truth produce various outcomes, and one outcome may result from various agents. The concept is often useful in study design; as a check on response bias we may deliberately collect information on factors which we expect to be the same in the groups compared, as similar results will indicate a lack of observation bias. We may choose control groups to capitalize on similar effects, as noted above.

Summary of internal validity

By this point, we should be able to decide whether the internal validity of the study is adequate. A positive decision means that we accept the results to be a valid measure of the true association in the subjects studied, and if an association between exposure and outcome is present, we regard it as likely to be due to a causal relationship.

A negative decision means that we decide that one or more of the non-causal explanations is likely to hold; the association seen is due to observation bias, to confounding, or to chance, and we should be able to specify the likely biases or confounding factors. Often we will be able to eliminate some but not all of the options, and decide for example that the result is likely to be due to either causation or to confounding; such a conclusion is very valuable as it makes clear what further information is necessary.

Observation bias is discussed in all the studies presented in the later chapters, but is of particular interest in the studies described in Chapter 11, 12, and 14; all these studies demonstrate efforts to minimize observation bias. In the studies described in Chapter 12 and 14, the major question of interpretation is whether the association is causal or is due to confounding; for the studies discussed in Chapter 10 and 13, the main issue is whether the result is causal or due to chance variation.

External validity: generalization of the results

If the internal validity of the study is very poor, there is no point in proceeding further, for if the study result is not valid even for the subjects studied, its application to other groups of subjects is irrelevant. However, if we conclude that it is a reasonably valid result and that a causal relationship is a reasonably likely explanation, we need to go on to consider the external validity of the result. The relevant questions are as follows.

Question 14. Can the study results be applied to the eligible population?

Question 15. Can the study results be applied to the source population?

Question 16. Can the study results be applied to other relevant populations?

p. 334 The relationship between the study participants and the population of eligible subjects should be well documented. Losses due to non-participation have to be considered carefully as they are likely to be non-random, and the reasons for the losses may be related to the exposure or to the outcome. These issues were discussed in Chapter 4.

Beyond this, it is unlikely that the study participants will be a ‘representative sample’ of a definable source population, and even if they were, we would want to extrapolate the results further, for example to our own community, future patients, and so on. The issue is not whether the subjects studied are ‘typical’ or ‘representative’, but whether the *association* between outcome and exposure given by the study participants is likely to apply to other groups. In assessing the applicability of results, we need to be specific about the factors which are likely to affect the association. Most clinical trials are done on patients in teaching hospitals. If a new therapy for breast cancer is shown to be effective in such a trial, we would readily apply the results to patients in a district hospital who had a similar stage and type of tumour and were of similar age, even though the trial patients cannot be said to be ‘representative’ of district hospital patients in a general or statistical sense. Similarly, women in the USA and Japan have very different incidence rates of breast cancer, and very different diets; but if a causal relationship exists between saturated fat intake and breast cancer incidence, we should expect to see it in both populations, even though its strength might be modified by the relative importance of other factors. However, other considerations may apply. If we read of a clinical trial of a new drug therapy used for severe depression in a well-known teaching centre, we should not apply the results to patients in general

practice uncritically; the general practice patients, even with the same diagnosis, are likely to be different (e.g. in the severity and duration of disease) from those in the teaching centre, and the effects of the therapy may well differ in inpatients and in ambulant patients. In general, the difficulties of applying results from one group of subjects to another will be minimal for issues of basic physiology and genetics, and maximal for effects in which cultural and psychosocial aspects are dominant. The generalizability of the results is important for all the studies discussed in subsequent chapters, but is particularly interesting for the trial discussed in Chapter 11, where generalization from a high-risk population to a general population is a major issue. The applicability of the results of the trial described in Chapter 10, and of the cohort study described in Chapter 12, are important issues in the interpretation.

Comparison of the results with other evidence

We have now made a critical assessment of the evidence presented by one study. We have assessed the internal
p. 335 validity, and come to a reasoned judgement ↳ as to whether the results of the study are consistent with a cause and effect relationship. We have explored the external validity of the study, and come to a decision concerning how far we can generalize the result beyond the subjects who participated in the study.

We can now move to the issue of comparing the result of this particular study with the evidence from other studies and other types of experience. As we did with the evidence from within the study, we shall make these comparisons with specific questions in mind. In comparing the results of a particular study with those of other studies, we will ask the following questions.

Question 17. Are the results consistent with other evidence, particularly evidence from studies of a similar or more powerful study design?

Question 18. Does the total evidence suggest any specificity?

Question 19. Are the results plausible, in terms of a biological mechanism?

Question 20. If a major effect is shown, is it coherent with the distribution of the exposure and the outcome?

Question 17. Consistency with other studies

This is the most important characteristic used in the judgement that an association is causal. To say that the result is consistent requires that the association has been observed in a number of different studies, each of which individually can be interpreted as showing a causal explanation, and which have enough variation in their methodology and study populations to make it unlikely that the same biases or confounding factors apply in all the studies. For example, when a British case–control study demonstrated a new relationship between hormone replacement therapy and venous thromboembolism [6], the interpretation of this was aided by the simultaneous publication, because of cooperation between investigators, of a case–control study and a cohort-based study from the USA [7,8] with consistent results.

Consistency of results between studies, each of which is individually unsatisfactory, is of little value, as is consistency between studies which all suffer from the same design defect. For example, it is difficult to assess the effect of breast self examination in producing early diagnosis of breast cancer from observational studies, because women who practise self-examination are likely to have many other characteristics which may lead them to an earlier diagnosis of a tumour. A meta-analyses of such studies has been performed [9], but it is still open to the alternative explanation of confounding due to other factors associated with earlier diagnosis. Lack of consistency argues against causality, but care must be taken in its assessment also. The failure to find an
p. 336 association in a study which is limited in its methodology and size so that it has very little ↳ power to detect

an association, if one were present, is of no value. When a new and controversial result is published, weak, badly designed, and small studies which show no association are often presented to refute it; these studies have to be examined with the same critical approach as is applied to the original.

Consistency is assessed statistically within meta-analyses by the assessment of heterogeneity, testing by methods such as those described in Chapter 8 whether the results for each study are consistent with the overall pooled result. The difficulty arises when there is significant heterogeneity. Often, the studies with outlying results are simply noted and then removed from the overall analysis, and a summarized result based on the rest is published. It is more satisfactory if the reasons for these discrepancies can be explored, as was discussed in Chapter 8. Consideration of the quality of the studies and their specific characteristics, using meta-regression methods if appropriate, may be valuable.

Question 18. Specificity

Specificity relates closely to consistency. Whether a difference in results between two studies is interpreted as inconsistency or specificity depends on whether the difference is anticipated by a hypothesis set up before the comparison is made. If not, but a plausible mechanism can be found or if the difference is itself found consistently, then the hypothesis may be modified to take into account the specificity which has been shown. This creates a new hypothesis which should be assessed by a further independent study.

For example, the two congenital defects of spina bifida and anencephalus have similar embryological and epidemiological features, and both are prevented by folic acid, as discussed in Chapter 11. However the antiepileptic drug valproic acid shows a strong and apparently specific association with spina bifida, while no excess of anencephalus has been recorded. If this is not due to some observation bias, it suggests a specificity of effect, showing one substantial difference in the aetiology of the two conditions [10]. A highly specific association is easier to recognize; an example quoted by Bradford Hill in a celebrated paper [11] was that of nickel miners in South Wales, where there were 16 deaths from lung cancer and 11 from cancer of the nasal sinuses, compared with expected numbers of about 1 and much less than 1, respectively; deaths from all other causes numbered 67, similar to the expected number of 72, pointing to a risk specific to these two cancer types.

Question 19. Plausibility

Plausibility refers to the observed association being biologically understandable on the basis of current knowledge concerning its likely mechanisms. The consideration of plausibility is useful, particularly as it may
p. 337 indicate biases ↳ or confounding factors which should be considered. The interpretation of the positive association between ice-cream sales and drowning at summer holiday resorts is an example, as a consideration of its plausibility will suggest a confounding factor.

However, any dramatically new observation may be in advance of current biological thinking and its lack of plausibility may reflect deficiencies in biological knowledge rather than an error in the observation. John Snow effectively prevented cholera in a district of London 25 years before the isolation of the cholera bacillus and acceptance of the idea that the disease could be spread by water. Percival Pott demonstrated the causal relationship between exposure to soot and scrotal cancer some 150 years before the relevant carcinogen was isolated and the mechanism further understood. The greatest value of the concept of plausibility is to emphasize that where an association does not match a known biological mechanism, further studies are indicated to clarify this, but these need not necessarily delay appropriate action if the evidence for causality is strong enough.

There has been considerable debate on the possible relationship between exposure to electromagnetic fields at low frequencies from electric power sources and the occurrence of childhood leukaemia. One of the major

reservations about accepting such an association is the lack of plausibility. Even after extensive research, there is little evidence from cellular or animal studies that indicates a mechanism for a carcinogenic effect. On the other hand, meta-analyses of several epidemiological studies shows an empirical association with the highest level of exposure found in the home environment [12,13]. At present, it is unclear whether these epidemiological investigations are the first indication of an important effect, with the elucidation of a biological mechanism to follow, or whether these observations are explicable by non-causal relationships such as confounding and selection biases.

The dramatic reduction in the frequency of a major developmental defect by a relatively low dose of a common vitamin is described in the study in Chapter 11. Although plausible in general terms, this empirical demonstration of effect preceded specific knowledge of a mechanism, but has stimulated extensive work to identify the key biochemical process and its genetic aspects [14].

Lists of the expected features in causality often include the concept of *analogy*, meaning that a relationship is regarded as more acceptable if it is analogous to some other well-established relationship, but clearly this concept comes within the overall concept of plausibility.

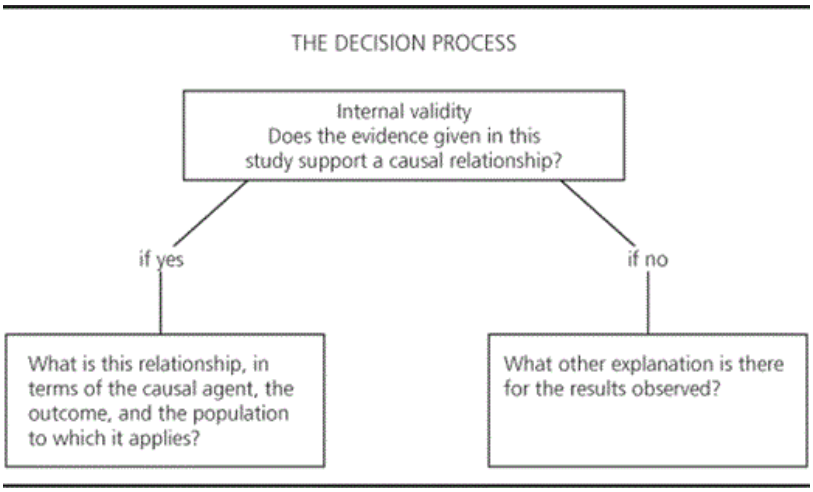
Question 20. Coherence

An association is regarded as coherent if it fits the general features of the distribution of both the exposure and
p. 338 the outcome under assessment; thus if ↳ lung cancer is due to smoking, the frequency of lung cancer in different populations and in different time periods should relate to the frequency of smoking in those populations at relevant earlier time periods. The concept of coherence has several limitations: it is assumed that the exposure and the outcome are the same in different populations, and it holds only if a high proportion of the outcome is caused by the exposure, and if the frequency of the outcome is fairly high in those exposed. If the factor causes only a small proportion of the total disease, the overwhelming influence of other factors may make the overall pattern inconsistent. A comparison of an exposure such as smoking in different countries, or a general category of disease such as lung cancer, may not take sufficient note of differences in types of smoking and in types of lung cancer.

As an example of an argument based on coherence, it was suggested some years ago that neural tube defects were caused by a teratogen in damaged potatoes. In support, the high frequency of the defects in Ireland and the strong association with low social class were quoted, on the argument that these populations had a high consumption of potatoes. However, it was also noted that the condition is fairly common in Taiwan, and is also more common in the poor, although potato consumption is probably higher in the upper social classes [15]. The case–control study described in Chapter 14 shows an association between a drug and a disease outcome, and in this instance the question of coherence is of great importance in separating a causal relationship from the alternative explanation of confounding. However, in the other studies described in subsequent chapters, coherence is not particularly helpful, either because the association suggested is not strong enough, or the information on the distribution of the relevant exposure and outcome is inadequate.

Assessing causal relationships and making decisions

A general method has now been set out which should assist the reader to assess written evidence, his or her own experience, and the experience of colleagues. The system is obviously only a framework, and issues specific to each subject will influence the relative importance of different aspects of the process; a minor issue in one subject may be a major issue in others. The entire process is summed up in **Ex.** 9.3. Presented with new results from a study of a putative causal relationship, the question we must ask is: 'Does the evidence given in this study support a causal relationship?' This involves the assessment of the internal and external validity of the study, and its relationship to other evidence, as expressed in the scheme given in this chapter.



Ex. 9.3. Assessment of causal relationships: the decision process

If our judgement is that the evidence does support a causal relationship, we should be able to reinforce this by answering the question: 'What is this ↳ relationship, in terms of the causal agent, the outcome, and the population to which it applies?' p. 339

If the answer to our question is in the negative, we need to be able to answer the question: 'What other explanation is there for the results observed?' The results we have been presented with do not go away; they are the facts from which we are arguing. If we reject a causal explanation, we must be able to propose an alternative hypothesis. The specification of the alternative hypothesis, or hypotheses, will help us to see the weaknesses in the evidence we have and guide us in how to search for better information.

Some further applications of the assessment of causality

In this book, we have concentrated on the assessment of cause and effect relationships, and almost exclusively on evidence provided by studies of groups of individuals. There are many important relationships which are not necessarily of a cause and effect type, and there are other types of data which can be considered. While there is no intention to go into these fully, a few comments on the relevance of the concepts expressed in this approach to these other questions may be helpful.

Associations can be useful even if not causal

If we can establish that an association exists, even if it is not causal, this information may be very valuable. In
p. 340 particular, *diagnosis* is dependent on the reliable ↳ demonstration of associations which are not causal associations. Diagnosis, the separation of subjects with a pathological condition from subjects who do not have that condition, is based on the features associated with that condition; whether they are part of the causation of the condition is irrelevant. Diagnostic symptoms, signs, and laboratory-measured abnormalities are usually features which are produced by the same causal mechanism as produces the disease which we are trying to diagnose, or are produced by the disease process. Despite that, many of the principles outlined in this volume are applicable to the study of diagnostic concepts, particularly the issues of observation bias and external validity.

Similarly, in the social sciences the emphasis has generally been on exploring associations between factors, for example whether attitudes to health care are different in different social groups, without of necessity considering cause and effect relationships. Often, however, assumptions and judgements about cause and effect relationships lie not far below the surface, and again the principles set forth in this volume should be of assistance in assessing such questions.

Application to studies of populations, and of physiological or biochemical parameters

This book is primarily about studies of individuals, where each contributes one exposure and one outcome event. Causal relationships may often be suggested, and sometimes be assessed, by data comparing different population groups; e.g. routine statistics of disease incidence, morbidity, or case fatality. Such studies yield comparisons between rates of exposure and rates of outcome in populations, without having data showing which individuals within those populations are involved. While such studies are inherently weaker than studies of individuals, they can be analysed in the same way. Most can be considered as cross-sectional surveys or cohort studies. For example, in a correlation study it was shown that populations (Canadian provinces) with more intense programmes of cervical cytology (the exposure) showed a larger decrease in mortality from uterine cancer (the outcome) [16]. The comparisons of firearm-related outcomes in Seattle and Vancouver [17], noted in Ex. 3.11, page 70, is similar. The analysis of these associations is conceptually the same as for a study of individuals: Is the association likely to be due to bias, confounding, or chance? Does it show the positive features of causation? The same logical approach can be applied.

Similarly, we have not dealt directly with the other main group of human studies, those where the results are based on series of physiological or biochemical measurements within an individual. Thus, to show the causal
p. 341 relationship between external temperature and peripheral blood flow, measurements could ↳ be made in different subjects at different temperatures, but it is much more efficient to compare the same individuals, making a series of measurements of peripheral blood flow while varying the external temperature. The logic of causal inference presented here is totally applicable: bias, confounding, and chance must be considered, and the positive indicators of causality sought. Such studies can be considered as matched cohort studies, with each group of subjects under one set of conditions being one study cohort.

While the logic can be applied to both descriptive epidemiological and physiological studies, often the outcome and/or exposure factors will be continuous rather than discrete variables, such as temperature, peripheral blood flow, and morbidity rate. The statistical methods applicable are those for continuous variables, including Student's *t*-test, correlation and regression methods, analysis of variance, multiple linear regression, and non-parametric methods such as rank correlation techniques. Such methods are well described in most statistical textbooks.

Applications in designing a study

This scheme can also be used in the design of studies. A useful general approach to the design of a study is that the investigator should attempt a *forward projection* to the point where the study has been completed and the results have been compiled. It is useful for the investigator to consider all possible results, i.e. a positive association or one in the direction expected, no association or a negative one, or an association in the opposite direction from that expected. The investigator can then ask the question, given that those results arise, 'How will I assess this study in terms of its internal and external validity?' or in simpler terms, 'Will I believe the results that I have obtained?' A good test of a well-designed study is that the investigator will be willing to accept and act on the results, even if they are different from the anticipated result. This is also a useful question to ask colleagues who wish you to be involved in their research. By projecting yourself to the point of assessing the possible results of the study, you as investigator can consider the major issues of bias, confounding, and chance variation, and consider methods by which these issues can be recognized and dealt with. You can also consider the desirability of being able to demonstrate strong associations, dose–response effects, specific relationships, and so on. By doing this, you give yourself the opportunity to incorporate into the design of the study the potential for demonstrating such features. The discussion of the study design with colleagues who are prepared to adopt a similar critical approach and who bring to bear specialist knowledge of the issues involved will be a major safeguard against embarking on a study which is inadequately designed.

p. 342

Part 2. The application of causal reasoning to clinical care and health policy: Hierarchies of evidence

In appraising the science relevant to an issue, each piece of evidence has three relevant dimensions: What type of study is it? Is it a high-quality study? Is it relevant to the question? So far in this text we have concentrated on assessing the quality of studies within the major study types, and in assessing their external validity, which will determine whether they are relevant to a particular question.

For most clinical and policy questions, a large amount of evidence from different types of study is available, and so it is useful to consider a hierarchy of evidence. Given that the studies are adequately performed within the limitations of the design used, the reliability of the information from them can be ranked in the manner shown in **Ex.** 9.4. Most generally accepted classifications follow this format.

A SIMPLE HIERARCHY OF EVIDENCE

Level 1	Randomized trials	1m: Results from a meta-analysis of trials 1s: One or more individual trials
Level 2:	Cohort and case-control studies	2m: Results from a meta-analysis of such studies 2s: One or more individual studies
Level 3:	Other comparative studies	
Level 4:	Case series, descriptive studies, professional experience, etc.	

Ex. 9.4. A simple hierarchy of types of evidence relevant to human health studies: as well as the level of evidence, the consistency or otherwise of the results, the quality and detail of the studies, and the relevance of the studies to the situation and question of interest have to be assessed

At the top are randomized intervention trials, if properly performed on adequate numbers of subjects and, of course, in the human situation. Evidence from such studies should be given the greatest weight because of the unique advantages of these studies in overcoming the problems of bias and confounding.

Second come observational studies of appropriately selected groups of subjects, i.e. the cohort and case–control designs. There is logic in placing cohort studies somewhat ahead of case–control studies because, if well performed, cohort studies should have less observation bias, give clearer evidence of the time relationships of the association, and have a comparison group whose results are more easily interpreted. However, both these observational designs can have severe problems, and a well-performed case–control study may be of more value than a poorly performed cohort study.

p. 343 As we have seen in Chapter 8, the data from many randomized trials can be summarized in a meta-analysis, and meta-analyses can also be done for cohort and case–control studies, although they are generally more difficult as there are usually greater methodological differences between the studies. Thus in categories 1 and 2 in Ex. 9.4 we can give higher ranking to an appropriately performed meta-analysis which combines the results of several relevant trials. The best possible evidence relating to an intervention would come from a clear result from a well-performed meta-analysis of all available randomized trials.

The third level of evidence comes from studies comparing groups of subjects not chosen specifically for the purpose of the study but representing different population or subject groups. This includes correlation studies of populations in which data on each individual are not assessed separately, and also informal comparisons between patients in different hospitals, patients treated at different time periods, and so on.

In the fourth category there is evidence which is largely anecdotal, based on the unsystematic recollection of personal or group experience (sometimes referred to as 'clinical judgement' or 'experience'), conclusions based on traditional practice, and information derived from other species, *in vitro* testing, physiological principles, and other indirect assessments. Meta-analyses may also be applied to evidence in the third and fourth categories; this is less frequently done, but undoubtedly more efforts will be made in the future.

This hierarchy is useful in assessing the very large amounts of information that may be available on a particular topic. It is sensible to concentrate on the best possible evidence. If there are randomized trials available on the question, they should be evaluated first, and if they provide strong evidence for or against causality, the results of the other less rigorous types of study may add little. For many topics, randomized trial evidence will not be available, and therefore we must look particularly at the results of well-performed cohort and case–control studies.

This general hierarchy of evidence is followed in the now numerous specific versions of hierarchies, some of which are becoming quite complicated. It has become quite customary for people to refer to 'level 1 evidence', generally meaning randomized trials, 'level 2 evidence', generally meaning comparative observational studies, and so on. This quite useful shorthand may disappear as more complex systems develop. For example, because the best source of evidence for studies of the effects of hazardous exposures comes from cohort and case–control studies, some feel that such studies should be referred to as level 1 evidence in this context; but this would lose the attractive simplicity of the overall framework described here. Of course, many systematic
p. 344 reviews use an ↳ even simpler ranking system, concentrating solely on randomized trials and ignoring other evidence.

The value of such ranking systems is now widely accepted. One of the first uses of such a system was in the reports, starting in 1979, of a Canadian multidisciplinary group on the effectiveness of procedures for inclusion in regular medical examinations [18,19]. The group concerned itself only with clinical efficacy, not with economic value or social acceptability. They assessed 88 suggested procedures including testing newborn babies for phenylketonuria, routine tests for urinary tract infection in pregnancy, and screening of preschool children for hearing impairments and adolescents for spinal deformity. Of the 88 situations, evidence from randomized trials was available in only 20 (of these 12 were questions of immunization), for 17 conditions evidence from cohort or case–control studies was available, for 22 only descriptive or uncontrolled observations were available, and for 29 there was nothing other than subjective opinion and 'experience'. The

relationship between the type of evidence available and the final decision of the multidisciplinary committee (**Ex.** 9.5) shows, as we would expect, that where the quality of evidence was high, a firm recommendation on whether to include or to exclude the procedure could be made. The situation where a firm recommendation was based on only category 3 evidence was one with a long clinical tradition—the use of silver nitrate drops in a neonate's eyes to prevent ophthalmia neonatorum.

	EVIDENCE AND DECISION						
Best evidence	Recommendation					Total	With firm recommendation
	include			exclude			
	firm	weak	none	weak	firm		
1. Randomized trials	11	0	1	2	6	20	17
2. Cohort, case-control	3	5	1	8	0	17	3
3. Other comparative	1	11	2	8	0	22	1
4. Other types	0	0	29	0	0	29	0
						88	

Ex. 9.5. Evidence and decision: relationship between decisions of a multidisciplinary group regarding regular health examinations and the best type of evidence available for 88 procedures for the general population. Derived from Canadian Task Force on the Periodic Health Examination [18]

This report was unusual at that time because it documented both the committee's recommendations and the type of evidence available for each procedure. This was done in terms of both the hierarchical system of
p. 345 classification of evidence, which is a gross measure of internal validity, and its applicability ↳ to the relevant population, patients in primary care in Canada, giving a measure of external validity. Since then many groups have used such ranking systems, including the US Preventive Services Task Force [20].

From evidence to recommendations

The classification of the type of evidence needs to be taken together with an assessment of the quality and relevance of the study in coming to an overall decision as to whether the evidence supports, rejects, or is neutral concerning a specific intervention or recommendation. Clinical and health policy groups have come up with various schemes to summarize this process. For example, the US Preventive Services Taskforce presents its recommendations in five categories [21] (**Ex.** 9.6). Many other groups have schemes which are similar in their objectives, while differing in the details.

GRADES OF CLINICAL RECOMMENDATIONS

A. The USPSTF strongly recommends that clinicians provide [the service] to eligible patients. *The USPSTF found good evidence that [the service] improves important health outcomes and concludes that benefits substantially outweigh harms.*

B. The USPSTF recommends that clinicians provide [the service] to eligible patients. *The USPSTF found at least fair evidence that [the service] improves important health outcomes and concludes that benefits outweigh harms.*

C. The USPSTF makes no recommendation for or against routine provision of [the service]. *The USPSTF found at least fair evidence that [the service] can improve health outcomes but concludes that the balance of benefits and harms is too close to justify a general recommendation.*

D. The USPSTF recommends against routinely providing [the service] to asymptomatic patients. *The USPSTF found at least fair evidence that [the service] is ineffective or that harms outweigh benefits.*

I. The USPSTF concludes that the evidence is insufficient to recommend for or against routinely providing [the service]. *Evidence that [the service] is effective is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.*

Ex. 9.6. The U.S. Preventive Services Task Force (USPSTF) grades of recommendations: these are based on the strength of evidence and magnitude of net benefit (benefits minus harms). From US Preventive Services Task Force [21] (www.ahrg.gov/clinic/pocketgd.pdf)

The recommendations result from a detailed assessment of individual studies and the use of a hierarchy of evidence as has been described, to assess the overall quality of evidence on a five category scale, with an approximate estimate of the relative size of the expected net benefits (**Ex.** 9.7). Thus to give an 'A' p. 346 recommendation, the US Preventive Services Taskforce scheme requires ↳ high-quality evidence which shows a substantial net benefit for the intervention in relevant populations [20]. Such recommendations include chemoprevention by aspirin for adults at increased risk of coronary heart disease, and routine screening for blood lipids in adults. A 'B' recommendation may have good evidence of a moderate benefit, or fair evidence of a moderate or substantial benefit; an example is counselling and behavioural interventions for weight reduction in obese adults, where the main caution is around the evidence for effectiveness of the interventions, which is mixed. In many situations the evidence is judged insufficient to yield clear recommendations, for example in screening for prostate cancer, and in others, the evidence is firm enough to recommend that such interventions have a zero or negative net benefit and should not be used, for example routine screening of adolescents for scoliosis.

DERIVATION OF CLINICAL RECOMMENDATIONS

1. *Rating of quality of the overall evidence in three categories:*

   Good: Evidence includes consistent results from well-designed well-conducted studies in representative populations that directly assess effects on health outcomes.

   Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalizability to routine practice, or indirect nature of the evidence on health outcomes.

   Poor: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.

2. *Assessment of size of net benefit (benefit–harm) in four categories: comparison with rating of quality of evidence determines the final recommendation as A, B, C, D or I.*

	Estimate of Net Benefit (Benefit Minus Harms)			
Rating	Substantial	Moderate	Small	Zero/Negative
Good	A	B	C	D
Fair	B	B	C	D
Poor	I – Insufficient Evidence			

Ex. 9.7. Derivation of the overall grades of recommendations shown in Ex. 9.6. From: US Preventive Services Task Force [21] (www.ahrg.gov/clinic/pocketgd.pdf)

For clinical interventions, many specialist and professional groups produce clinical guidelines, which vary considerably in their content and presentation, but inherently they all assess the available evidence according
p. 347 to a process involving critical appraisal and using a hierarchy of evidence similar to that in Ex. 9.4. ↳ The type of advice can range from specific advice on particular issues representing the consensus of the group producing the guidelines, to the more sophisticated approach of encouraging practitioners to assess and use the evidence efficiently themselves. The latter approach is shown by the manuals for evidence-based clinical practice produced by the Evidence-based Medicine Working Group of the American Medical Association and the Centre for Health Evidence, which has a valuable website (www.userguides.org) [22].

Conflicting recommendations

In many controversial issues, different conclusions are reached not because of fundamental disagreements about the evidence available, but because of different methods of ranking different types of evidence. For example, there has been much controversy about breast cancer screening at younger ages, with some expert groups recommending it as a beneficial procedure, while others have concluded that the evidence for benefit is unclear and do not recommend it. In general, those less willing to support such screening argue primarily from the results of randomized trials, which have not shown clear benefits. Some of the groups reaching a different conclusion, such as the American Cancer Society, make it clear that they put considerable emphasis on the results of other types of study, such as a large uncontrolled demonstration project carried out in the USA [23]. The inherently different rankings given to different types of evidence influence the conclusion reached, and cultural differences in different societies are important [24,25]; for example, some argue that if the evidence is unclear, an established intervention like screening should be continued unless there is clear evidence of lack of benefit, and some argue that it should not proceed until there is clear evidence of benefit. This issue came to a head in 1993, when the National Cancer Institute in the USA assessed the available randomized trial evidence on breast cancer screening in women under age 50, concluded that the benefits were uncertain, and changed its previous recommendation which supported screening in this age group. This decision resulted in a report from a US Senate subcommittee, which criticized the dependence on randomized trial results, pointing to the uncontrolled demonstration projects in the USA, and argued that screening should continue unless the absence of benefit could be proven. A similar scene was repeated when a consensus conference was held by the National Institutes of Health in 1997; again, this group, putting most emphasis on randomized trial results, concluded that there was no clear evidence of a mortality reduction and did not recommend screening [26]. This decision was rejected by the director of the National Cancer Institute and again by the US Senate, which voted unanimously in favour of the value of mammograms for younger ↳ women, using what one journalist described as ‘some mysteriously acquired medical insight’ [27]. The result seemed to be that in this context a scientific conclusion that mammography in younger women was not effective was not politically acceptable, whereas in some other countries it was more accepted that it was necessary to prove the benefits of screening, rather than to prove the lack of benefit [24].

p. 348

The application of critical appraisal to clinical medicine and health care policy

In this book we have reviewed a system for the critical appraisal of evidence relating to cause and effect relationships in health. The application of critical appraisal methods to health care issues is often described as *evidence-based medicine*, with a primarily clinical orientation, or as *knowledge-based health care* or *evidence-based policy* with a wider community perspective. Of course, these are both huge topics about which only a few words can be said here. What follows is a brief comment on the development of evidence-based medicine and health policy, rather than any attempt at a comprehensive approach to the area.

The development of evidence-based medicine

Clinical decision-making has had three main influences: authority and tradition, laboratory-based science, and empirical assessment. The predominant influence through most of the twentieth century was authority and tradition: the viewpoint of accumulated experience and consensus. Many such 'consensus' views are not supported by empirical evidence. In 1991, a *British Medical Journal* editorial concluded that 'only about 15 per cent of medical interventions are supported by solid scientific evidence' [28], noting several previous similar estimates. More eloquently, David Naylor, head of the medical school and subsequently President of the University of Toronto, said 'clinical medicine seems to consist of a few things we know, a few things we think we know (but probably don't), and lots of things we don't know at all' [29].

Many health interventions arise from a consideration of mechanisms studied in the laboratory, but this can be dangerous if empirical testing in a whole person and community situation is not done. One of the worst examples of this was the use of drugs to suppress cardiac arrhythmias that were observed in patients who had had a myocardial infarction. The logic was that patients with such arrhythmias had a higher risk of death, the drugs could effectively suppress the arrhythmias, and therefore the drugs should reduce mortality. The drugs
p. 349 were developed, marketed, approved, and widely used on this basis, ↳ before large randomized trials demonstrated that although reducing the ectopic heartbeats, the drugs produced an increase in mortality [30]. Similarly, the evidence from basic science and observational epidemiological studies that beta-carotene should protect against cancer was strong, and led to its widespread use as a preventive factor outside mainstream medicine. However, randomized trials of beta-carotene have shown no benefit or even higher risks of cancer, so that supplementation in these trials has had to be stopped, as was discussed in Chapter 6 [31].

The tensions between different sources of wisdom are not new. In the nineteenth century pioneers such as the French physician Louis promoted empirical methods of assessing therapies using the 'numerical' method, i.e. observing and counting the results in patients, which demonstrated amongst other things that the popular use of leeches in therapy was unsupported by evidence of efficacy [32]. The contrast between assessments based on clinical experience and on numerical observations was shown in the debates about thyroid immunization in the early part of the twentieth century. Supporting immunization was Sir Almroth Wright, a prominent pathologist (and later mentor of Alexander Fleming, of penicillin fame), who put his faith in what he called the 'experiential' method. He described this as 'unconscious automatic induction by an expert', or more fully: 'we let the two streams of experience which correspond to the two series of substantive and controlled experiments filter through our minds and then compare the impressions which have been imprinted' [33,34]. Opposing him was Karl Pearson, one of the pioneers of statistical methods, arguing from empirical data that the efficacy of vaccination was unproven; Pearson used correlation techniques, having written that 'this new conception of correlation brought psychology, anthropology, medicine and sociology in large parts into the field of mathematical treatment' [35]. In this debate, described eloquently by Susser [36], Wright was correct, because his clinical experience showed him that some of the data used by Pearson was poorly collected and unreliable. In our terminology, he was aware of the problems of observation bias and error, whereas Pearson's statistical technique gave equal weight to all the observations he had available, irrespective of their quality. However, subsequently, Wright's methodological weakness showed in his development of, and fervent support for, an autoimmunization process for treating infections, which gained immense popularity but was not subjected to empirical trial and led to a great deal of inappropriate and probably dangerous therapy [37]. His process, and the issues it raised in terms of the use of resources and access to care, were described by George Bernard Shaw, a close friend, in his play *The Doctor's Dilemma*, first performed in 1906.

p. 350 In 1937, introducing a series of articles on statistics by Bradford Hill, which later emerged as an important textbook [38], the editor of *The Lancet* wrote: 'in clinical medicine today there is a growing demand for adequate proof of the efficacy of this or that form of treatment'. Bradford Hill's papers, which embodied most of the principles of assessment of causality covered in this book, went beyond arithmetical methods to the assessment of observation bias, confounding, and the evidence for causality [39].

Evidence-based medicine was developed prominently at the McMaster Medical School, in Canada, led by David Sackett, and was expressed in *Clinical Epidemiology*, first published in 1985 [40]. It was firmly rooted in individual patient care, with four steps described in 1995 [41]: to formulate a clear clinical question of a patient-based problem, search the literature for relevant clinical articles, critically appraise the evidence for its validity and usefulness, and implement useful findings in clinical practice. The focus of this development was on the efficacy of clinical interventions. Sackett and his colleagues reported that, in contrast with the estimates given earlier, the great majority of interventions in a specialist medical unit at Oxford under Sackett's direction were demonstrated to be based on empirical evidence [42]. However, this experience in a pioneering academically based unit cannot be generalized to clinical medicine as a whole.

Critical appraisal threatens the traditional approach. The *Lancet* editorial in 1937 commented on the threat which doctors saw from statisticians: 'It is exasperating, when we have studied a problem by methods which we have spent laborious years in mastering, to find our conclusions questioned, and perhaps refuted by, someone who could not have made the observations himself' [39]. The development of meta-analyses, the insistence on systematic standards even for review articles, and the extension of evidence-based medicine to policy and management may be viewed by many as further threats. Some may feel vulnerable if individual professional experience counts for little in comparison with a well-planned study, and if an individual cannot even be trusted to collate experience, that also requiring an objective and systematic process. Even in 2005, a Canadian contributor to *The Lancet* wrote: 'Yet if everything has to be double-blinded, randomized, and evidence-based, where does that leave new ideas?' [43], and more scholarly papers have asked if evidence-based practice does more good than harm [44]. The authoritarian tradition is still dominant in many cultures, such as that of Japan [45]. However, there is no doubt that the greater attention paid to the justification of policies from empirical evidence has been beneficial, and the requirements for such objectivity will continue to increase.

From clinical efficacy to health policy

The individual patient-based approach has been criticized as paying insufficient attention to issues of
p. 351 efficiency, cost, and other social considerations; ↳ it has been labelled 'narrow scienticism'. The prominent British health economist Alan Maynard wrote: 'If evidence- based medicine and the individual ethic are allowed to determine treatment choices, resources will be used inefficiently and unethically' [46], and a leading social science commentator borrowed the traditional clinical criticism of empirical scientific evidence, arguing that the heterogeneity of individual patient problems requires the mysterious process of 'clinical judgement' [47].

However, a wider concept of evidence-based medicine allows for the consideration of cost and utility in addition to clinical efficacy [48]. The emphasis in this book is on the assessment of a causal relationship between an exposure and an outcome. Where we are dealing with an intervention, this is the question of *efficacy*: does the intervention produce the outcome desired? In clinical and policy decision-making, efficacy is the most important but not the only component. *Efficiency*, i.e. efficacy in comparison with the resources necessary, is also relevant, and so is *acceptability*, which brings in many other considerations from cultural approaches to ethical questions. Efficacy is central, and arguments about efficiency or acceptability are rather pointless where efficacy is dubious. Efficiency and acceptability will be of overriding importance where a choice has to be made between different efficacious approaches.

The evidence-based approach can now be applied to the appraisal of all health interventions, and terms such as *knowledge-based health care* have been used. While John Swales, a national director of research and development for the British National Health Service, noted [49] 'in the 1980's, it was inconceivable that purchasers of health services would enquire in any but the most rudimentary way about scientific evidence

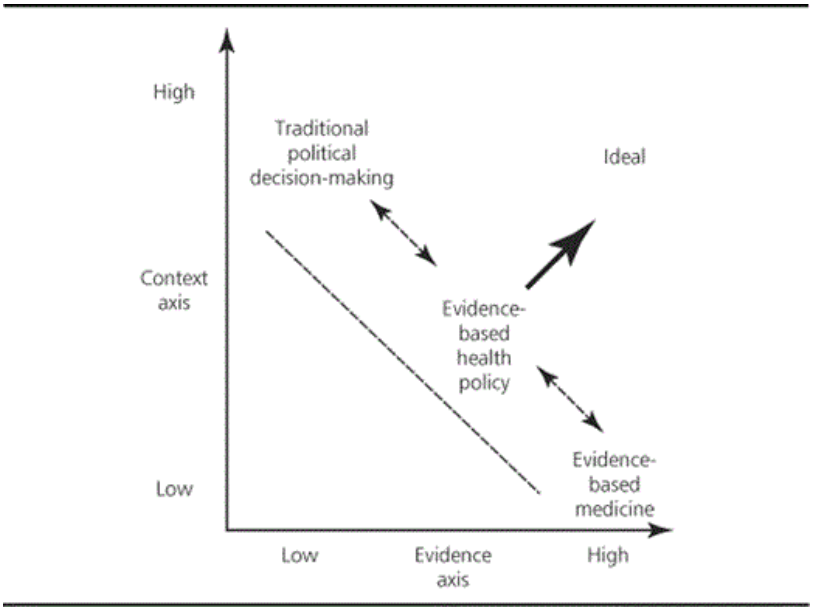
before funding new developments', such assessments have since become an essential part of NHS development, and other countries have followed to various extents.

Methods have been developed which combine the critical appraisal of scientific evidence with cost–benefit or cost–utility analysis, and can incorporate considerations of equity, acceptability, and feasibility. For example, one such method is programme budgeting marginal analysis, which has been used to compare options in health care development in several countries [50–52]. These methods take into account more of the various influences which impact on decision-making in our societies. However, the relationships between objective appraisal of evidence and the realities of policy-making are far from simple. Just as proponents of evidence-based clinical medicine may assume that decisions will be driven only by the quality of scientific evidence, most economic models assume that decision-makers are pure and rational; they are 'benevolent and unbiased, seeking to maximise efficiency and equity subject to budgetary constraints'. In a more cynical (or realistic)
p. 352 view, it has been said that 'just as companies seek to ↳ maximise profits and consumers seek to maximise utility, clinicians seek to maximise their autonomy, bureaucrats seek to maximise control, and politicians seek to maximise support' [53].

Decision-makers operate in a different environment from clinicians, researchers, or indeed economists. Politicians and bureaucrats need to make decisions under time pressure and with incomplete information, must take account of the perceived urgency and importance of the issue, and must operate with short time horizons. Indeed, these characteristics are often used to criticize research evidence or researchers and to dismiss them as irrelevant [54]. Some regard the gaps between decision-makers and researchers as insurmountable. Comments include 'Researchers are from Mars; policymakers are from Venus' and 'research is actually a dirty word to many policymakers' [55]. On the other hand, health care professionals may feel that policy-makers set the policy first and use scientific evidence, if at all, only to support already agreed policies; they use policy-based evidence rather than evidence-based policy.

Such gaps may be attributed to the poor communication skills of researchers, or to the lack of consultation by policy-makers, but such blame attribution is unhelpful. Researchers need to be aware of the constraints policy-makers work under, and anticipate the counter-arguments that will arise. Approaches that will be more likely to bridge the gaps include an emphasis on presenting solutions rather than problems, and paying attention to wide consultation and to gaining broad professional and consumer support. Policy-makers generally have neither the time nor the expertise to assess the value of a proposal, and so they often assess who is promoting an idea and who would be opposed to it. In this context, the consumers' voice has become very important in health planning in most countries. Decision-makers will accept decisions that appear good enough, ('satisficing' decision-making), rather than seeking the optimum, and they like incremental decisions, in which options are kept open and irrevocable commitments are not made, to keep an escape route open. Risk assessment is an integral part of any policy development, including the assessment of financial and political hazards. Doing something new has more risks than continuing current policy: starting friction is greater than sliding friction. The greatest political risks are in taking something away; witness the almost inevitable outcry over the closure of a facility, even if a good argument has been made on grounds of clinical efficacy or efficiency. It has been found that systematic attempts based on good scientific and economic logic to reduce ineffective care in several countries have not been successful [53].

An elegant synthesis is given by Dobrow *et al*. [56], who consider decision-making in terms of the importance of scientific evidence and the importance of context, which includes economic, social, and political
p. 353 considerations (**Ex.** 9.8). The evidence-based clinical model rates high on the evidence axis but low on ↳ the context axis, while the traditional political process rates high on the context axis but low on the evidence axis. They argue for an ideal situation where these are balanced. The challenge is how to achieve this. Even recently, in a major review of several massive social interventions in the UK, the King's Fund group concluded that these multi-million pound programmes were developed 'on the basis of informed guesswork and expert hunches, enriched by some evidence and driven by political and other imperatives' [57,58].



Ex. 9.8. The context and evidence axes of evidence-based decision-making as given by Dobrow *et al*. [56] Reprinted from *Social Science in Medicine*, **58**, Dobrow *et al*. Evidence based health policy: context and utilization, pp. 207–17, 2004, with permission from Elsevier.

This was in part because the necessary research had not been done. The choice of what research is done is haphazard, and if we base programmes only on the existing research our scope is very limited. The further development of evidence-based policy requires identifying key questions that require new research. A good example of this is screening for neuroblastoma, a cancer affecting young children. It can be detected by a urine test in infancy, and screening has been advocated, especially in Japan. Before routine screening was accepted in Canada or the USA, a trial was done in Canada. This was not a randomized trial, but a comparison of a screening programme applied to births in Quebec, with control groups unscreened in Ontario. The results
p. 354 showed that screening produced a great increase in the incidence rate of ↳ diagnosed neuroblastoma, but the mortality rate was not reduced. The interpretation was that screening produced a substantial number of false-positive diagnoses, and also ‘silent tumours’, which are tumours which were diagnosed as neuroblastoma and treated, but which if left alone would not have produced any clinical problems. The trial cost $US 8.8 million. If a screening programme had been set up, the costs for the USAand Canada would have been $570 million over a 12-year period [59]. Thus as well as avoiding substantial morbidity, the trial investment showed a rate of return of 6400 per cent, which even the most cynical policy-maker would regard as impressive.

Finale

The appraisal of evidence is an essential component of any wider decision-making process. Like any other process, it can sometimes be misused. It is important that the central issue in evidence-based health care, the critical appraisal of the available evidence, receives adequate attention. This indeed is the purpose of this book.

The remaining six chapters of this book present the application of this method of critical appraisal to six studies, illustrating a range of study designs and topics. The reader is encouraged to work through these and apply the system which has been presented.

References

1. MacMahon B. Epidemiological evidence on the nature of Hodgkin's disease. *Cancer* 1957; **10**: 1045–1054. 10.1002/1097-0142(195709/10)10:5(1045::AID-CNCR2820100527)3.0.CO;2-0
WorldCat Crossref

2. Glaser SL, Jarrett RF. The epidemiology of Hodgkin's disease. *Baillieres Clin Haematol* 1996; **9**: 401–416. 10.1016/S0950-3536(96)80018-7
WorldCat Crossref

3. Howe GR, Burch JD, Miller AB, *et al*. Artificial sweeteners and human bladder cancer. *Lancet* 1977; **ii**: 578–581. 10.1016/S0140-6736(77)91428-3
WorldCat Crossref

4. Rothman KJ, Moore LL, Singer MR, Nguyen US, Mannino S, Milunsky A. Teratogenicity of high vitamin A intake. *N Engl J Med* 1995; **333**: 1369–1373. 10.1056/NEJM199511233332101
WorldCat Crossref

5. Smith DC, Prentice R, Thompson DJ, Herrmann WL. Association of exogenous estrogen and endometrial carcinoma. *N Engl J Med* 1975; **293**: 1164–1167. 10.1056/NEJM197512042932302
WorldCat Crossref

6. Daly E, Vessey MP, Hawkins MM, Carson JL, Gough P, Marsh S. Risk of venous thromboembolism in users of hormone replacement therapy. *Lancet* 1996; **348**: 977–980. 10.1016/S0140-6736(96)07113-9
WorldCat Crossref

7. Jick H, Derby LE, Myers MW, Vasilakis C, Newton KM. Risk of hospital admission for idiopathic venous thromboembolism among users of postmenopausal oestrogens. *Lancet* 1996; **348**: 981–983. 10.1016/S0140-6736(96)07114-0
WorldCat Crossref

8. Grodstein F, Stampfer MJ, Goldhaber SZ, *et al*. Prospective study of exogenous hormones and risk of pulmonary embolism in women. *Lancet* 1996; **348**: 983–987. 10.1016/S0140-6736(96)07308-4
WorldCat Crossref

p. 355 9. Hill D, White V, Jolley D, Mapperson K. Self examination of the breast: is it beneficial? Meta-analysis of studies investigating breast self-examination and extent of disease in patients with breast cancer. *BMJ* 1988; **297**: 271–275. 10.1136/bmj.297.6643.271
WorldCat Crossref

10. Elwood JM, Little J, Elwood JH. *Epidemiology and Control of Neural Tube Defects*. Oxford: Oxford University Press, 1992.
Google Scholar Google Preview WorldCat COPAC

11. Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965; **58**: 295–300.
WorldCat

12. Crumpton MJ. The Bernal Lecture 2004. Are low-frequency electromagnetic fields a health hazard? *Philos Trans R Soc Lond B Biol Sci* 2005; **360**: 1223–1230. 10.1098/rstb.2005.1663
WorldCat Crossref

13. Ahlbom IC, Cardis E, Green A, Linet M, Savitz D, Swerdlow A. Review of the epidemiologic literature on EMF and Health. *Environ Health Perspect* 2001; **109** (Suppl 6): 911–933. 10.2307/3454653
WorldCat Crossref

14. Mitchell LE, Adzick NS, Melchionne J, Pasquariello PS, Sutton LN, Whitehead AS. Spina bifida. *Lancet* 2004; **364**: 1885–

1895. 10.1016/S0140-6736(04)17445-X
WorldCat Crossref

15. Emanuel I, Sever LE. Questions concerning the possible association of potatoes and neural-tube defects, and an alternative hypothesis relating to maternal growth and development. *Teratology* 1973; **8**: 325–332. 10.1002/tera.1420080315
WorldCat Crossref

16. Miller AB, Lindsay J, Hill GB. Mortality from cancer of the uterus in Canada and its relationship to screening for cancer of the cervix. *Int J Cancer* 1976; **17**: 602–612. 10.1002/ijc.2910170508
WorldCat Crossref

17. Sloan JH, Kellermann AL, Reay DT, *et al*. Handgun regulations, crime, assaults, and homocide. *N Engl J Med* 1988; **319**: 1256–1262. 10.1056/NEJM198811103191905
WorldCat Crossref

18. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J* 1979; **121**: 1193–1203.
WorldCat

19. Woolf SH, Battista RN, Anderson GM, Logan AG, Wang E. Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. A report by the Canadian Task Force on the Periodic Health Examination. *J Clin Epidemiol* 1990; **43**: 891–905. 10.1016/0895-4356(90)90073-X
WorldCat Crossref

20. Sibbald B, Addington-Hall J, Brenneman D, Freeling P. Telephone versus postal surveys of general practitioners: methodological considerations. *Br J Gen Pract* 1994; **44**: 297–300.
WorldCat

21. U.S Preventive Services Task Force. *The Guide to Clinical Preventive Services 2005*. Washington, DC: Agency for Healthcare Research and Quality, 2005.
Google Scholar Google Preview WorldCat COPAC

22. Guyatt GH, Rennie D (eds). *Users' Guide to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago, IL: JAMA & Archives Journals, American Medical Association, 2002.
Google Scholar Google Preview WorldCat COPAC

23. Mettlin C, Smart CR. Breast cancer detection guidelines for women aged 40 to 49 years: rationale for the American Cancer Society reaffirmation of recommendations. *CA Cancer J Clin* 1994; **44**: 248–255. 10.3322/canjclin.44.4.248
WorldCat Crossref

24. Elwood JM. Breast cancer screening in younger women: evidence and decision making. *J Eval Clin Pract* 1997; **3**: 179–186. 10.1046/j.1365-2753.1997.00002.x
WorldCat Crossref

25. Jatoi I, Baum M. American and European recommendations for screening mammography in younger women: a cultural divide? *BMJ* 1993; **307**: 1481–1483. 10.1136/bmj.307.6917.1481
WorldCat Crossref

26. National Institutes of Health Consensus Development Conference. *National Institutes of Health Consensus Development Statement: Breast Cancer Screening for Women ages 40–49: January 21–23, 1997*. Washington: National Institutes of Health, 1997.
Google Scholar Google Preview WorldCat COPAC

27. Begley S. The mammogram war. *Newsweek* [Feb 24], 55–59. 1997.
WorldCat

28. Smith R. Where is the wisdom…? *BMJ* 1991; **303**: 798–799. 10.1136/bmj.303.6806.798
WorldCat Crossref

29. Naylor CD. Grey zones of clinical practice: some limits to evidence-based medicine. *Lancet* 1995; **345**: 840–842. 10.1016/S0140-6736(95)92969-X
WorldCat Crossref

30. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992; **327**: 227–233. 10.1056/NEJM199207233270403
WorldCat Crossref

31. Omenn GS, Goodman GE, Thornquist MD, *et al*. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *N Engl J Med* 1996; **334**: 1150–1155. 10.1056/NEJM199605023341802
WorldCat Crossref

32. Armitage P. Trials and errors: the emergence of clinical statistics. *J R Statist Soc Ser A* 1983; **146**: 321–334. 10.2307/2981451
WorldCat Crossref

33. Wright AE, Morgan WP, Colebrook L, Dodgson RW. Observations on the pharmaco-therapy of pneumococcus infections. *Lancet* 1912; **ii**: 1701–1705. 10.1016/S0140-6736(01)41480-2
WorldCat Crossref

34. Wright AE. *The Unexpurgated Case Against Women's Suffrage*. London: Constable, 1913.
Google Scholar Google Preview WorldCat COPAC

35. Pearson ES. *Karl Pearson: An Appreciation of Some Aspects of His Life and Work*. Cambridge: Cambridge University Press, 1938.
Google Scholar Google Preview WorldCat COPAC

36. Susser M. Judgment and causal inference: criteria in epidemiologic studies. *Am J Epidemiol* 1977; **105**: 1–15.
WorldCat

37. Nachbar J, (Ed.). *Vaccine Therapy: Its Administration, Value and Limitations. A Discussion Opened by Sir Almroth E. Wright, MD, FRS*. London: Longmans, Green, 1910.
Google Scholar Google Preview WorldCat COPAC

38. Hill AB. *Principles of Medical Statistics*. London: The Lancet Ltd, 1937.
Google Scholar Google Preview WorldCat COPAC

39. Anonymous. Mathematics and medicine. *Lancet* 1937; **i**: 31.
WorldCat

40. Sackett DL, Haynes RB, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical medicine*. Boston, MA: Little, Brown, 1985.
Google Scholar Google Preview WorldCat COPAC

41. Rosenberg W, Donald A. Evidence-based medicine: an approach to clinical problem-solving. *BMJ* 1995; **310**: 1122–1126. 10.1136/bmj.310.6987.1122
WorldCat Crossref

42. Ellis J, Mulligan I, Rowe J, Sackett DL. Inpatient general medicine is evidence based. *Lancet* 1995; **346**: 407–410. 10.1016/S0140-6736(95)92781-6
WorldCat Crossref

43. Wu J. Could evidence-based medicine be a danger to progress? *Lancet* 2005; **366**: 122. 10.1016/S0140-6736(05)66867-5
WorldCat Crossref

44. Hammersley.M. Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice. *Evid Policy* 2005; **1**: 85–100. 10.1332/1744264052703203
WorldCat Crossref

45. Yokota T, Kojima S, Yamauchi H, Hatori M. Evidence-based medicine in Japan. *Lancet* 2005; **366**: 122. 10.1016/S0140-6736(05)66866-3
WorldCat Crossref

46. Maynard A. Evidence-based medicine: an incomplete method for informing treatment choices. *Lancet* 1997; **349**: 126–128. 10.1016/S0140-6736(96)05153-7
WorldCat Crossref

47. Klein R. The NHS and the new scientism: solution or delusion? *Q J Med* 1996; **89**: 85–87.
WorldCat

48. Sackett DL, Rosenberg WMC. The need for evidence-based medicine. *J R Soc Med* 1995; **88**: 620–624.
WorldCat

49. Swales J. Scientific basis of health services. Science and medical practice: the turning tide. *J Health Serv Res Policy* 1996; **1**: 61–62.
WorldCat

50. Mitton C, Donaldson C. *Priority Setting Toolkit: A Guide to the Use of Economics in Healthcare Decision Making*. London: BMJ Publishing, 2004.
Google Scholar Google Preview WorldCat COPAC

51. Peacock S, Ruta D, Mitton C, Donaldson C, Bate A, Murtagh M. Using economics to set pragmatic and ethical priorities. *BMJ* 2006; **332**: 482–485. 10.1136/bmj.332.7539.482
WorldCat Crossref

52. Mitton C, Donaldson C. Twenty-five years of programme budgeting and marginal analysis in the health sector, 1974–1999. *J Health Serv Res Policy* 2001; **6**: 239–248. 10.1258/1355819011927558
WorldCat Crossref

53. Hauck K, Smith PC, Goddard M. *The Economics of Priority Setting for Health Care: A Literature Review*. Washington, DC: Health, Nutrition, and Population Family (HNP) of the World Bank's Human Development Network, 2003.
Google Scholar Google Preview WorldCat COPAC

54. Innvaer S, Vist G, Trommald M, Oxman A. Health policy-makers' perceptions of their use of evidence: a systematic review. *J Health Serv Res Policy* 2002; **7**: 239–244. 10.1258/135581902320432778
WorldCat Crossref

55. Greenlick MR, Goldberg B, Lopes P, Tallon J. Health policy roundtable—view from the state legislature: translating research into policy. *Health Serv Res* 2005; **40**: 337–346. 10.1111/j.1475-6773.2005.0b360.x
WorldCat Crossref

56. Dobrow MJ, Goel V, Upshur RE. Evidence-based health policy: context and utilisation. *Soc Sci Med* 2004; **58**: 207–217. 10.1016/S0277-9536(03)00166-7
WorldCat Crossref

57. Coote A, Allen J, Woodhead D. *Finding Out What Works: Building Knowledge About Understanding Complex, Community-Based Initiatives*. London: King's Fund, 2004.

Google Scholar Google Preview WorldCat COPAC

58. Bowen S, Zwi AB. Pathways to 'evidence-informed' policy and practice: a framework for action. *PLoS Med* 2005; **2**: e166. 10.1371/journal.pmed.0020166
WorldCat Crossref

59. Soderstrom L, Woods WG, Bernstein M, Robison LL, Tuchman M, Lemieux B. Health and economic benefits of well-designed evaluations: some lessons from evaluating neuroblastoma screening. *J Natl Cancer Inst* 2005; **97**: 1118–1124. ↳ 10.1093/jnci/dji203
WorldCat Crossref