Principles of critical appraisal in evidence-based medicine

The term critical appraisal of the literature, as used in the context of evidence-based medicine (EBM), refers to the application of predefined rules of evidence to a study to assess its methodological quality and the clinical usefulness of its results. Critical appraisal represents the most "technical" step in the process of EBM and can be quite demanding for the practitioner. The aim of this paper is to provide the reader with the theoretical skills necessary to understand the principles behind critical appraisal of the literature. These include: (a) the description of the main types of study design used in epidemiological research, (b) the basic statistical procedures used in data analysis, (c) the principles of causal inference and (d) the description of the types of health outcome and measures of effect. These issues are discussed in the present paper and illustrated with several examples from the relevant literature.

ARCHIVES OF HELLENIC MEDICINE 2001, 18(2):192–203
ΑΡΧΕΙΑ ΕΛΛΗΝΙΚΗΣ ΙΑΤΡΙΚΗΣ 2001, 18(2):192–203

P.A. Skapinakis,[1]
N. Stimpson,[1]
H.V. Thomas,[1]
F. Dunstan,[2]
R. Araya,[1]
G. Lewis[1]

[1]Department of Psychological Medicine, University of Wales, College of Medicine, UK
[2]Department of Medical Computing and Statistics, University of Wales, College of Medicine, UK

Αρχές αξιολόγησης της βιβλιογραφίας στα πλαίσια της βασισμένης στις ενδείξεις ιατρικής

Περίληψη στο τέλος του άρθρου

Key words

Causal inference
Critical appraisal
Evidence-based medicine
Study design

1. INTRODUCTION

The term critical appraisal of the literature, as used in the context of evidence-based medicine (EBM), refers to the application of predefined rules of evidence to a study to assess (a) its methodological quality[1] and (b) the clinical usefulness of its results.[2] Critical appraisal represents the most "technical" step in the process of EBM and can be quite demanding for the practitioner.

The aim of this paper is to provide the reader with the theoretical skills necessary to understand the principles behind critical appraisal of the literature. The presentation will follow roughly the order by which the researcher carries out the research. First, the main types of study design used in epidemiological research are discussed. Second, the basic statistical procedures used in data analysis are considered. Third, the ways in which the researcher decides whether any of the associations found have causal implications are analysed. This procedure is called causal inference. Fourth, the process is discussed of how a judgment is made on whether the results are important enough and potentially useful to implement in clinical practice, by assessing the kinds of outcomes studied and the size of the effects observed.

2. THE BASICS OF STUDY DESIGN

The main study designs used in epidemiological research (tabl. 1) can be described as either observational (ecological, cross-sectional, case-control and cohort), experimental (randomised controlled trial), or summary in nature (systematic reviews).[3]

P.A. Skapinakis is supported with a fellowship from "Alexander S. Onassis" Public Benefit Foundation.

Table 1. Types of study in epidemiological research.

Primary research		Secondary research
Observational	**Experimental**	**Summary**
Ecological	Randomised controlled trials (RCTs)	Systematic reviews
Cross-sectional		Meta-analyses
Case-control		
Cohort		

2.1. Ecological or aggregate studies

Ecological studies examine the association between disease and the characteristics of an aggregation of people rather than the characteristics of individuals. The main difficulty with this design is that the association between exposure and disease at an aggregate level may not be reflected in an association at the individual level. In this context, the confounding is often termed the *ecological fallacy*. An example of an ecological study is that conducted by Lewis et al[4] which aimed at examining the association between suicide standardised mortality ratios (SMRs) and the provision of psychiatric services. Both variables are aggregate variables. They found that suicide SMRs were higher in districts with more mental illness consultants and nurses. However this association was reduced after adjustment for the confounding effects of deprivation and whether an area had a teaching hospital. The likely explanation was that more mental illness professionals are employed in deprived areas in response to the greater perceived need. Teaching hospitals in the UK also tend to be situated in inner city areas with high suicide rates.

2.2. Cross-sectional surveys

This type of descriptive study relates to a single point in time and can therefore report on the prevalence of a disease but is adversely affected by the duration of the illness. A cross-sectional survey eliminates the problems of selection bias and has frequently been used for the study of depression and other neurotic conditions. However any association found in a cross-sectional survey could be either with incidence or duration. For example, Skapinakis et al[5] studied the sociodemographic and psychiatric associations of unexplained chronic fatigue in a cross-sectional survey of the general population in Great Britain. They found that chronic fatigue was strongly associated with psychiatric disorder. They also found that other risk factors were independently associated with chronic fatigue (older age, female sex, having children and being in full time employment) after adjustment for psychiatric disorder.

2.3. Case-control studies

In a case-control study individuals with the disease (cases) are compared with a comparison group of controls. If the prevalence of exposure is higher in the cases than in the controls the exposure might be a risk factor for the disease, and if lower the exposure might be protective. Case-control studies are relatively cheap and quick and can be used to study rare diseases. However, great care is needed in the design of the study in order to minimise selection bias. It is important to ensure that the cases and controls come from the same population, because the purpose of the "control" group is to give an unbiased estimate of the frequency of exposure in the population from which the cases are drawn. For example, Kendell et al[6] conducted a case-control study to examine the association between obstetric complications (OCs) and the diagnosis of schizophrenia. They found a highly significant association and concluded that a history of OCs in both pregnancy and delivery is a risk factor for developing schizophrenia in the future. However, in a new paper[7] the same group re-analyzed the data set of the previous study and reported that the previous findings were not valid due to an error in selecting controls. The method used had inadvertently selected controls with lower than normal chances of OCs, thus introducing a serious selection bias. In reality, there was no association between schizophrenia and OCs in this data set.

A *nested case-control study* is one based within a cohort study or sometimes a cross sectional survey. The cases are those that arise as the cohort is followed prospectively and the controls are a random sample of the non-diseased members of the cohort.[3]

In a *matched case-control study* one or more controls are selected for each case to be similar for characteristics which are thought to be important confounders.

The analysis of case-control studies results in the reporting of odds ratios, case-control studies cannot directly estimate disease incidence rates. If the study is matched, a more complex matched analysis needs to be performed (conditional logistic regression).

2.4. Cohort or longitudinal studies

A cohort (or longitudinal, or follow-up) study is an observational study in which a group of "healthy" subjects who are exposed to a potential cause of disease, together with a "healthy" group who are unexposed, are followed up over a period of time. The incidence of the disease of interest is compared in the two groups. Ideally, the exposed and unexposed groups should be cho-

sen to be virtually identical with the exception of the exposure. The ability of a cohort study to rule out reverse causality as a reason for an observed association is of great benefit. Schizophrenia is more common in cities, but it had been widely accepted that this was due to "geographical drift" of people with schizophrenia. Lewis et al[8] studied a cohort of 50,000 Swedish male conscripts who had been asked before the onset of schizophrenia about where they had been brought up. The incidence of schizophrenia was 1.65 times higher in those brought up in cities compared with those brought up in rural areas. These results could not have occurred because of the "geographical drift" hypothesis. Since this cohort study had asked about upbringing before the onset of disease, the authors concluded that environmental factors found in cities increase the risk of the disorder, though drift after onset might also occur.

Cohort studies always "look forward" from the exposure to disease development, and therefore can be time-consuming and expensive. To minimise costs historical data on exposure i.e. information already collected, can be used. The Lewis et al[8] paper above is an example of this. The disadvantage of these studies is that exposure measurement is dependent on the historical record that is available.

The completeness of follow-up is particularly important in cohort studies. It is essential that as high a proportion of people in the cohort as possible are followed up and those who migrate, die or leave the cohort for any reason should be recorded. The reasons for leaving the cohort may be influenced by the exposure and/or outcome and incomplete follow-up can therefore introduce bias.

The analysis of cohort studies involves calculation of either the incidence rate or the risk of disease in the exposed cohort compared to that in the unexposed cohort. Relative and absolute measures of effect can then be calculated.

2.5. Randomised controlled trials

Randomised controlled trials (RCTs, fig. 1) are most frequently used (when possible) to investigate the effectiveness of medical interventions.[9] They constitute the strongest design to investigate causality between an intervention and outcome, because randomly allocating sufficient patients to two or more treatments should eliminate both selection bias and confounding when comparing outcomes.[10] Selection bias and confounding are explained later but the principle of the RCT is that the subjects in the randomised groups should be as similar as possible. The main argument for randomisation is that it is impossible to measure the many confounding variables that affect outcome. If outcome could be predicted very accurately then a longitudinal study would be a satisfactory design.

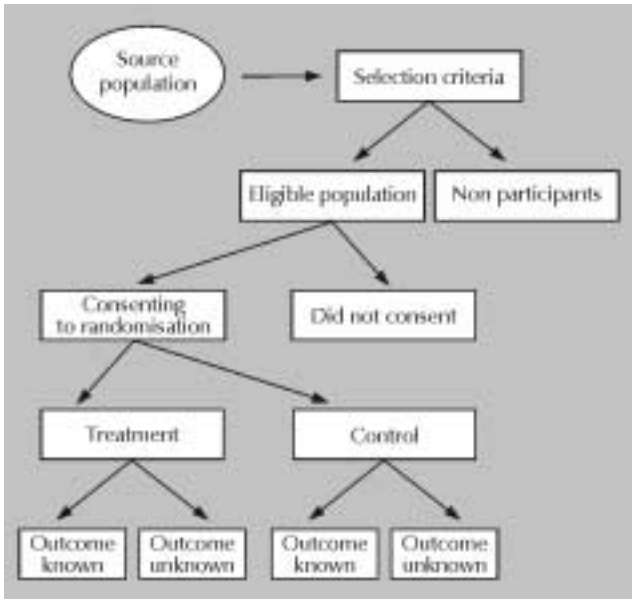

Figure 1. Design of randomised controlled trials.

For RCT to influence clinical practice it must address an area of clinical uncertainty. If there is a consensus that a treatment is effective then there is little point in conducting a trial without some other good reasons. The more common the dilemma the more important and relevant becomes an RCT. It is important that areas of clinical uncertainty are recognised in order to design future RCTs. Clinical uncertainty is also related to the ethical justification for randomisation. If a clinician is uncertain about the most effective treatment then randomisation becomes an ethical option or even an ethical requirement. It is therefore important that RCTs address important clinical dilemmas.

Subjects must be allocated to the treatments in an unbiased way. This is done by concealing the process of randomisation so that the person who has assessed the patient cannot interfere with the randomisation. The concealment of randomisation is an important aspect of RCT methodology and has been used as a proxy for the quality of an RCT.[11]

The validity of the comparison between the randomised groups in RCT depends critically on ensuring that the measurement of outcome is not affected by the allocation of treatment. This is usually done by disguising the random allocation from the person making the

assessment; or "blinding" the person as to the allocation. A double blind study refers to one in which both the patient and assessor are blind. A triple blind study refers to those in which the person analysing the data is also unaware of the treatment allocation.

One of the main difficulties in interpreting the results of an RCT concerns the influence of subjects withdrawing from treatment or from follow-up. As subjects drop out of an RCT the treatment groups depart from the balanced groups created at randomization. If the drop-outs are substantial in number then there is a possibility that confounding is reintroduced. Even more importantly, since non-compliers usually tend to be those subjects at a higher risk of adverse health outcomes, there is a risk of bias creeping in especially if there is differential drop-out between the groups. Therefore it is important to minimise the non-compliance rate and loss to follow-up rate.

The main way in which this problem is circumvented is by use of an intention-to-treat strategy in the analysis in which all the randomised subjects are included irrespective of whether they continued with the treatment or not. If there is missing follow-up data, data from a previous time-point can be used to assume a poor outcome for drop-outs. There are also more complex ways of substituting values for missing data that rely upon multivariate methods. An intention-to-treat strategy ensures that all the randomised individuals are used in the analysis. In this way, the benefits of randomisation are maintained and the maximum number of subjects can be included in the analysis. Using an intention-to-treat analysis is one of the characteristics of pragmatic trials.[9] They aim to study the long-term consequences of one specific clinical decision e.g. to prescribe the treatment or not, and to follow best clinical practice after that. The treatment effect may be less (i.e. the effect is diluted) than in the ideal case of 100% compliance, but it gives a far more realistic estimate of the treatment effect.

There is an ongoing debate between those who argue that randomisation is the only safe, unbiased means of assessing new interventions, and those who view randomisation as a narrow methodology of limited usefulness except for assessing drug treatments.[12] There are three sets of arguments:

a. *External validity.* RCTs might lead to findings that overestimate treatment effects or do not have relevance to the settings which most interest clinicians.

b. *Feasibility.* Sometimes it is impossible to mount RCTs for practical reasons. For example, an RCT of suicide prevention would need to randomise tens of thousands of people.

c. *Rarity.* The number of well conducted RCTs of sufficient size to draw conclusions will always be limited. There are going to be many clinically relevant issues that will not be addressed by using RCTs.

Perhaps the main criticism is the limited external validity or generalisability.[12,13] RCTs are strong on internal validity i.e. drawing conclusions about the effectiveness of the treatment used in that particular setting, on those patients. However, clinicians are also, if not primarily, interested in the external validity of a trial. The key question is "Do the results apply to the circumstances in which the clinician works?".

There are probably 3 main reasons why this can be a problem:

a. *The professionals.* The doctors and other professionals involved in trials are atypical, often with a special interest and expertise in the problem under investigation.

b. *The patients.* It is often difficult to recruit subjects to RCTs and the group of patients included is often very unrepresentative of the group eligible for treatment. This difficulty is often exacerbated by the investigators choosing a large number of "exclusion criteria".

c. *The intervention.* Many studies are carried out in prominent services, perhaps with dedicated research funds providing additional services. It is often difficult to know about the effectiveness of a similar intervention applied to other services either in the country of the study or elsewhere in the world.

Pragmatic RCTs are designed to answer clinically relevant questions in relevant settings and on representative groups of patients.[9] One of the priorities of pragmatic trials is to ensure external validity as well as internal validity. Choosing clinically relevant comparisons is also essential and pragmatic trials are designed to reduce clinical uncertainty. Assessment of a pragmatic trial should consider the representativeness and relevance of: (a) the patients in relation to the intended clinical setting, (b) the clinical setting, (c) the intervention(s), and (d) the comparisons. Economic assessment is often an important aspect of pragmatic trials. Clinicians, patients and commissioners need to know how much an intervention costs as well as whether it works. There will always be limitations on the resources available for health care and economic assessment should help to make judgments on the best place to invest. This has to be done in conjunction with knowledge about the size of treatment effect.

In addition to clinical outcomes, trials also need to examine outcomes concerned with the "quality of life" of the subjects. Measures of quality of life should assess whether subjects are working, pursuing their leisure activities or require additional support.

2.6. Systematic reviews and meta-analyses

Secondary research aims to summarise and draw conclusions from all the known primary studies on a particular topic (i.e. those which report results at first hand).[14] Systematic reviews apply the same scientific principles used in primary research to reviewing the literature. In contrast, the more traditional or narrative review relies upon the ability of an expert to remember the relevant literature and to extract and summarise the data he or she thinks important. Systematic reviews ensure that all the studies are identified using a comprehensive method and that data are extracted from the studies in a standardised way. Meta-analysis provides a summary estimate of the results of the studies identified using a systematic review. It enables the results of similar studies to be summarised as a single overall effect, with confidence intervals, using formal statistical techniques.

The main advantage of these studies is the resulting increase in the combined sample size (tabl. 2).

A problem of secondary research is the presence of publication bias, i.e. small negative results are less likely to be published. Therefore, ideally a comprehensive search strategy should be attempted which includes not only published results but also those reported in abstracts, personal communications and the like. Systematic reviews have mostly been used to summarise the results from randomised controlled trials (see Cochrane Collaboration below) but the same arguments apply to reviewing observational studies.

A central issue in secondary research is heterogeneity.[15] This term is used to describe the variability or differences between studies in terms of clinical characteristics (clinical heterogeneity), methods and techniques (methodological heterogeneity) and effects (heterogeneity of results). Statistical tests of heterogeneity may be used to assess whether the observed variability in study results (effect sizes) is greater than that expected to occur by chance. Heterogeneity may arise when the populations in the various studies have different characteristics, when the delivery of the interventions is variable, or when studies of different designs and quality are included in the review. Interpreting heterogeneity can be complex, but clinicians are often interested in heterogeneity in order to practice informed clinical decision making.[16] For example, clinicians want to know if a particular group of patients respond well to a particular treatment. Meta-analysis has also been criticised for attempting to summarise studies with diverse characteristics. Investigating heterogeneity can also be used to address such concerns.

Table 2. Advantages and disadvantages of secondary research.

Advantages	Disadvantages
All evidence is used to asses an intervention	Publication and citation bias
Increased statistical power	Limited by the quality of the primary studies
Can investigate heterogeneity and test generalisability	Pooling disparate studies may be invalid (but such heterogeneity can be investigated)

The use of systematic reviews for the assessment of the effectiveness of health care interventions has been promoted largely by *Cochrane Collaboration*. Archie Cochrane, a British epidemiologist who was based in Cardiff for much of his working life, recognised that people who want to make better informed decisions about health care do not have ready access to reliable reviews of the available evidence. Cochrane emphasised that reviews of research evidence must be prepared systematically and they must be kept up-to-date to take account of new evidence. In 1993, 77 people from eleven countries co-founded "The Cochrane Collaboration". The Cochrane Collaboration aims to review systematically all the RCTs carried out in medicine since 1948 and is committed to update the reviews as new evidence emerges. The mission statement of the Cochrane Collaboration is "Preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions". The Cochrane Collaboration's web site is www.cochrane.org. This has links to the Cochrane library which contains the Cochrane database of systematic reviews and the Cochrane controlled trials register.

3. THE BASICS OF STATISTICAL ANALYSIS

3.1. The study population

The *study population* is the set of subjects about whom it is wished to learn. It is usually impossible to learn about the whole population, so instead a subset or *sample* of the population is looked out in detail. Ideally a sample is chosen at random so that it is representative of the whole study population. The findings from the sample can then be extrapolated to the whole study population.

Suppose that two random samples are selected from a large study population. They will almost certainly con-

tain different subjects with different characteristics. For example if two samples of 100 are chosen at random from a population with equal numbers of males and females, one may contain 55 females and the other 44 females. This does not mean that either sample is "wrong". The randomness involved in sample selection has introduced an inaccuracy in measurement of the study population characteristic. This is called *sampling variation*. The aim is to extrapolate and draw conclusions about the *study population* using findings from the *sample*. Most of the statistical tests are therefore trying to infer something about the *study population* by taking account and estimating the sampling variation.

3.2. Hypothesis testing

Studies are usually designed to try to answer a clinical question, such as:

"Is there any difference between two methods for treating depression?".

In hypothesis testing this question formulated as a choice between two statistical hypotheses, the null and alternative hypotheses.

The *null hypothesis* H_0 represents a situation of no difference, no change, equality, while the *alternative hypothesis* H_1 specifies that there is a difference or change:

H_0: There is no difference between two treatments for depression

H_1: There is a difference between the methods.

A decision must be made as to which is thought to be true. This is usually based on the *p-value*, which is essentially the probability of obtaining results at least as extreme as those obtained if the null hypothesis is true. A small p-value, such as 0.05, means it is unlikely that such a result would be obtained by chance and this offers evidence against H_0 while a large p-value, such as 0.5, tends broadly to support H_0. How small should the p-value be to reject H_0?

Traditionally the critical level has been set at 0.05 or 5%. If $p<0.05$ is the criterion for rejecting H_0 then we say the result is *significant* at 5%. Other levels can be taken but this is the most common.

If the p-value exceeds 0.05, the decision is that H_0 is not rejected, rather than H_0 is accepted. It is difficult to prove that there is absolutely no difference. It is concluded that it cannot be shown that there is a difference.

3.3. Type I and type II errors

There are two types of wrong decision that can be made when a hypothesis test is performed.

A *type I error* occurs when the null hypothesis H_0 is true but is rejected. Five per cent of all tests that are significant at the 5% level are type I errors. Carrying out repeated tests increases the chance of a type I error.

A *type II* error occurs when the null hypothesis H_0 is false but is not rejected. For example, in a small study it is possible to record a non-significant p value despite large true differences in the study population. Type II errors need to be considered in all "negative" studies. Confidence intervals will help to interpret negative findings (see below).

3.4. Statistical power

The statistical power of a study is the probability of finding a statistically significant result assuming that the study population has a difference of a specified magnitude. It is the probability of not having a type II error. The power depends upon:

– The level of statistical significance, usually 5%
– The size of effect assumed in the study population
– The sample size.

Calculating the power of a study is useful at the planning stage. The power calculation depends critically on the size of effect one wishes to find. When designing studies the power is often set to 80% in order to determine the sample size. 80% is an arbitrary value, similar in that way to the 5% significance value.

3.5. Interpreting "statistically significant" results

Five percent (one out of every 20) of statistical tests will be statistically significant at the 5% level by chance. The 5% significance level is fairly arbitrary. There is no real difference in interpreting a 4% and 6% significance. If a study reports twenty p-values, one would be expected to be "significant" by chance. Repeated tests increase the chance of type I errors.

3.6. Confidence intervals

It is known that the sample estimate from a study (e.g. a proportion, a mean value, or an odds ratio) is subject to sampling error. The primary interest is in the size of effect, so it needs to be known how accurately the proportion is being estimated. Confidence intervals are based on an estimate of the size of the effect together

with a measure of the uncertainty associated with the estimate of the size.

The standard error (SE) shows how precisely the sample value estimates the true population value. If a lot of similar sized samples are taken from the same population, then the SE can be thought of as the standard deviation of the sample means. If the sample size is increased, the standard error decreased as the study population value is being estimated with more accuracy.

A 95% confidence interval is constructed so that in 95% of cases, it will contain the true value of the effect size. It is calculated as:

$$95\%\ CI = \text{estimated value} \pm (1.96 * SE)$$

Different levels of confidence can be used if desired. Based on the same information a 99% confidence interval will be wider than a 95% one, since a stronger statement is made without any more data; similarly a 90% confidence interval will be narrower. In recent years it has been generally agreed that results should be summarised by confidence intervals rather than p-values, although ideally both will be used. The p-value does not give an indication of the likely range of values of the effect size whereas the confidence interval does.

3.7. Interpreting "negative" results

When a trial gives a "negative" result, in other words no statistically significant difference is demonstrated, it is important to consider the confidence intervals around the result. It must be remembered that a study estimates the result inaccurately. The confidence intervals give the range within which one can be 95% confident that the "true" value lies. Small negative trials will usually have confidence intervals that include differences that would correspond to potentially important treatment effects. One way of thinking about results is that they are excluding unlikely values. The confidence interval gives the range of likely values and an effect size outside the confidence interval is unlikely.

3.8. Correlation and regression

Linear regression allows the relationship between 2 continuous variables to be studied. The regression line is the straight line that fits the data best. The correlation coefficient varies between –1 and 1. The more the correlation coefficient departs from 0 the more the variation is explained by the regression line. A negative correlation coefficient arises when the value of one variable goes down as the other goes up.

Each observation can be thought of as a "predicted" value i.e. that which would lie on the regression line, and a "residual" that is the difference between the predicted value and the observed value (fig. 2). The total variance is therefore the predicted variance added to the residual variance. The correlation coefficient is the predicted variance divided by the total variance. If all the points lie on the line then the correlation coefficient is 1. The slope of the line is sometimes called the regression coefficient. It gives the increase in the mean value of y for an increase in x of one unit.

4. CAUSAL INFERENCE

The main task of critical appraisal is to decide upon the presence of a causal association between a treatment, a possible causal agent or prognostic factor and a disease outcome. An association between an exposure and a disease can be explained by *chance*, *bias*, *confounding*, *reverse causality or causation* (fig. 3).[17] It is important to emphasise that all study designs, including the RCT, are concerned with causal inference. In an RCT, the interest lies in whether a treatment "causes" an increased rate of recovery.

a. *Chance.* Significance testing assesses the probability that chance alone can explain the findings. Calculating confidence intervals gives an estimate of the precision with which an association is measured. A type I error occurs when a statistically significant result occurs by chance (see section 3.3). It is a particular problem when many statistical tests are conducted within a single study in the absence of a clearly stated prior hypoth-

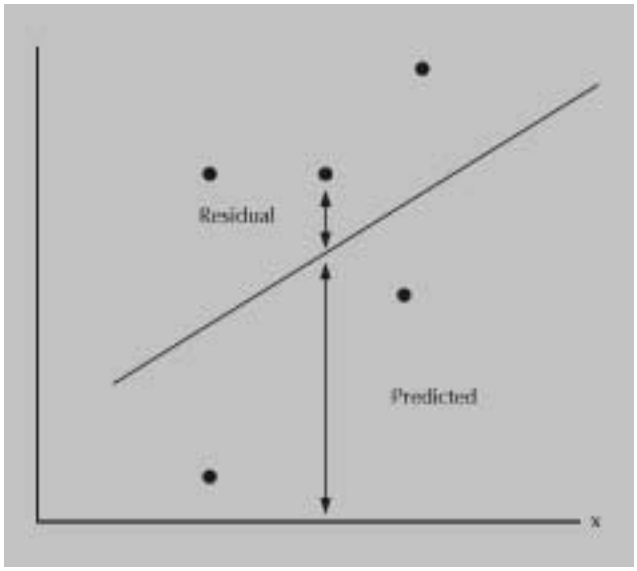

Figure 2. The regression line.

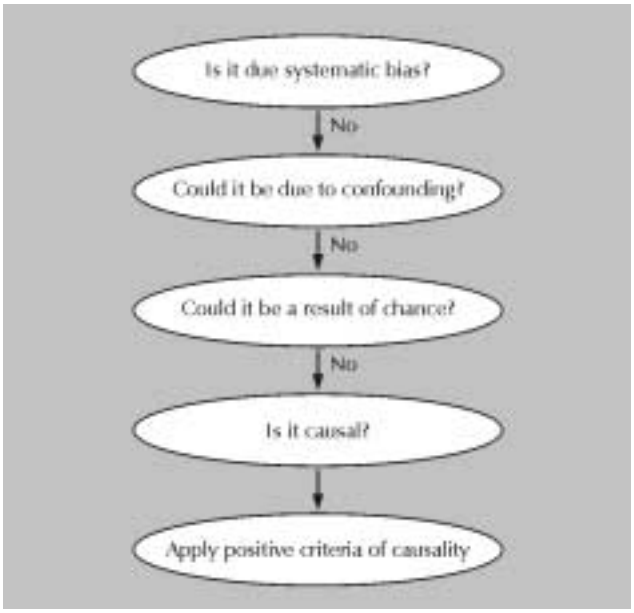

Figure 3. Interpreting an association.

esis. A type II error occurs when a clinically important result is obscured by chance or random error, often made more likely by inadequate sample size. For "negative" findings, the confidence interval gives a range of plausible values for the association.

b. *Bias. Systematic error* or bias can distort an association in any direction, either increasing or decreasing the association. No study is entirely free of bias, but attention to the design and execution of a study should minimise sources of bias. There are two main types of bias in epidemiological studies: selection bias and information bias.[3,18]

Selection bias can occur in any investigation but is a particular problem in case-control studies. It arises if the sampling method used to identify cases and controls results in a poor representation of the diseased and non-diseased individuals in the same population. An example is seen in the study of Brown and Harris[19] who identified community cases of depression using a cross-sectional survey and patient cases of depression by contact with psychiatric services. When the community cases were compared with community controls there was an association between depression and having a young family (odds ratio 3.77), but this association was absent when comparing the patient cases and community controls. Selection bias is acting here because, for someone who is depressed, having young children might reduce the likelihood of receiving treatment from psychiatrists. Since the majority of cases of depression never get referred to a psychiatrist, a community sample would not be an appropriate control group for cases seen in psychiatric services.

Information bias in an analytical study occurs when subjects are misclassified according to their exposure status, disease status or both. If this misclassification of disease status is dependent on exposure status or vice versa, the estimate of association will be biased. Examples of such differential misclassification include recall bias, reporting bias and observer bias. *Recall bias* occurs especially in case-control studies and cross-sectional surveys when individuals with the disease may be asked retrospectively about the exposure. In cohort studies exposure is determined before onset of the disease and so it is less likely to be biased by the presence of disease. Case-control studies may also be able to use exposure information collected before the onset of disease. *Observer bias* occurs when the observer is aware of the hypothesis and the measurement of exposure or disease is biased by the knowledge. Keeping the assessments blind to disease or exposure will reduce the chance of this happening.

c. *Confounding.* Confounding occurs when an estimate of the association between an exposure and disease is an artefact because a third confounding variable is associated with both exposure and disease (fig. 4).[3] If an association results from confounding it does not mean this association is wrong, but that there is an alternative explanation for it. An epidemiologist should identify potential confounders at the design stage of the study and collect information on them, otherwise at the time of analysis it will be difficult to reject alternative explanations for any associations found in the data. A variable is said to be a confounder if the analysis that controls for this variable produces results that are markedly different from the crude or uncontrolled analysis.

All observational studies are susceptible to confounding. Potential confounders must always be thought of when interpreting studies. Even in an RCT there can be an imbalance (by chance) in important confounders between the allocated interventions. It must be asked if the authors have measured the potentially important confounders and if they have adjusted their findings for them.

d. *Reverse causality.* This is the possibility that the exposure is the result rather than the cause of the disease. This is more likely to occur in case-control studies and cross-sectional surveys that assess exposure after the onset of disease. Cohort studies usually eliminate this possibility by selecting people without the disease at the be-

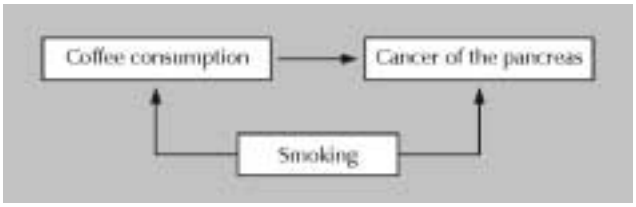

Figure 4. Smoking is a confounder for the association between coffee and cancer of the pancreas.

ginning of the study. RTCs also select people at the beginning of the trial who are ill in order to examine outcome. However, it can remain a problem for some conditions, such as psychosis, where the timing of the onset of the disease remains a matter of debate.

e. *Causation.* An association may indicate that the exposure causes the disease. Trying to infer causation is a difficult task. It is usually helpful to review the epidemiological literature to help decide whether there is a consistent finding, irrespective of the population or study design. When there is a strong association then the likelihood that the relationship is causal is increased. For example, for relative risks over 3 or 4, confounding and bias have to be quite marked to explain the findings. However, there is generally little confidence in findings when the relative risk is 1.5 or below. A dose-response relationship can also provide additional evidence for causality, depending upon the hypothesised mechanism of action. For example, one would expect that more severe obstetric complications would lead to higher rates of schizophrenia than milder forms if obstetric complications were a causal agent. Finally, the scientific plausibility of the findings has to be considered.

A number of criteria have been suggested for deducing that exposures have a causal role in disease (tabl. 3).[18] These usually require evidence from a variety of sources and it would be expected that a number of different studies using different approaches would all produce consistent results before a conclusion could be made about causality.

The most important criteria usually are:

a. Strength of relationship measured by relative risk. Large relative risks are more likely to be causal. A relative risk below about 1.5 should be treated with great caution.
b. Specificity of effect. Does the possible risk factor also cause other diseases?
c. Consistency of findings across studies. A variety of different studies in different populations and with different strengths and weaknesses in the design should all produce the same results.
d. Biological plausibility and dose-response. Is the relationship biologically plausible and does it show a dose-response relation i.e. the greater the exposure to a risk factor, the more likely the disease.

Table 3. Causality criteria.

The Bradford Hill criteria
Temporality (the exposure occurs before the outcome)
Strength of the association (strong associations more likely causal)
Consistency (same results with different methods)
Dose-response relationship
Specificity of the association
Biological plausibility
Coherence (no conflicts with current knowledge)
Experimental evidence
Analogy (similar factors cause similar diseases)

5. CLINICAL IMPORTANCE OF THE RESULTS: TYPES OF HEALTH OUTCOMES AND MEASURES OF THE EFFECT

5.1. Health outcomes

Here the interest lies in the changes (referred to as outcomes) amongst the research subjects which are associated with exposure to risk factors or therapeutic or preventive interventions. There are two main types of outcomes (tabl. 4):[20] (a) biological or psychosocial parameters not directly related to disease (for example cholesterol values or scores on a scale measuring social support) and (b) clinical outcomes directly related to disease.

Non-clinical outcomes can only be viewed as surrogates for the clinical outcomes of interest and cannot be used directly as a reason to change clinical practice unless there is a clear causal association between this and a clinical outcome. Clinicians are thus more interested in research papers which have used relevant clinical outcomes. Outcomes in the course of a disease include the following: death, disease status, discomfort from symptoms, disability, dissatisfaction with the process of care. These can easily memorised as the five Ds of health outcomes. In establishing the clinical importance of a study it should always be checked that the outcome is relevant.

Table 4. Outcomes in the course of disease. Adapted from Muir Gray.[20]

Death
Disability
Disease status
Dissatisfaction with process of care
Discomfort about the effects of disease

5.2. Clinical importance

A study may be methodologically valid, with an outcome of interest to clinicians but still not be clinically relevant because, for example, the effect of treatment is negligible. A new antihypertensive drug which lowers systolic blood pressure by 5% compared to routine treatment is probably not clinically significant in the sense that it has no implications for patient care. There are 2 broad categories of measures of effect, relative measures (e.g. relative risk, odds ratio) and absolute measures (e.g. attributable risk) (tabl. 5).

In the clinical context, absolute measures are of greater interest because the relative measures cannot discriminate between large and small treatment effects. For example, in a clinical trial if 90% of the placebo group developed the disease compared to 30% of the treatment group the relative risk reduction would be (90–30%)/90%=66% and the absolute risk reduction 90–30%=60%, a clinically important result. In a trial however with 9% for placebo compared to 6% for treatment, the relative risk reduction is the same but the absolute risk reduction is 3%, a figure not important from the clinical perspective. In the following paragraphs the basic measures of effect used in clinical research are presented.

Attributable risk (ARR) (risk difference, rate difference) is the absolute difference in risk between the experimental and control groups. A risk difference of zero indicates no difference between the two groups. For undesirable outcomes a risk difference that is less than zero indicates that the intervention was effective in reducing the risk of that outcome useful for interpreting the results of intervention studies.

The *number needed to treat (NNT)* is an alternative way of expressing the attributable risk between two groups of subjects. It has been promoted as the most intuitively appealing way of presenting the results of RCTs and its use should be encouraged in interpreting the results of trials.[2]

The NNT is the number of patients that need to be treated with the experimental therapy in order to prevent one of them from developing the undesirable outcome. It is calculated as the reciprocal of the absolute difference in risk (probability of recovery) between the groups. An NNT of 5 indicates that 5 patients need to be treated with treatment A rather than treatment B if one person is to recover on treatment A who would not have recovered on treatment B.

The following example can help to better understand these measures: An RCT of depression finds a 60% recovery with an antidepressant and a 50% recovery on placebo after 6 weeks treatment. The absolute risk difference is 10% (or p=0.1). The NNT is 1/0.1=10. Therefore, if 10 patients were treated with the antidepressant, one would get better who would not have got better if treated with placebo. Another way of thinking of this is if there were 10 patients in each group, 6 would get better on the antidepressant and 5 on the placebo.

Relative risk is a general and rather ambiguous term to describe the family of estimates that rely upon ratio

Table 5. Measures of effects.

	Absolute	Relative
Effect measures for binary data	Absolute risk reduction (ARR): The absolute difference in risk between the experimental and control groups Number needed to treat (NNT): The number of patients that need to be treated with the experimental therapy in order to prevent one of them from developing the undesirable outcome. It is the inverse of ARR	Odds: The number of events divided by the number of non-events in the sample Odds ratio (OR): The ratio of the odds of an event in the experimental group to the odds of an event in the control group Risk: The proportion of participants in a group who are observed to have an event Relative risk (RR): The ratio of risk in the experimental group to the risk in the control group
Effect measures for continuous data	Mean difference: The difference between the means of two groups Weighted mean difference: Where studies have measured the outcome on the same scale, the weight given to the mean difference in each study is usually equal to the inverse of the variance Standardised mean difference: Where studies have measured an outcome using different scales (e.g. pain may be measured in a variety of ways) the mean difference may be divided by an estimate of the within-group standard deviation to produce a standardised value without any units	
Effect measures for survival data	Hazard ratio: A summary of the difference between two survival curves. It represents the overall reduction in the risk of death on treatment compared to control over the period of follow-up of the patients	

of the measures of effect for the two groups. They are not the best way of summarizing treatment trials. This is because it is the absolute change in risk rather than the relative risk which is of interest. Ratio measures are more useful in interpreting possible causal associations. Ratio measures estimate the strength of the association between exposure and disease.

Incidence rate ratio is a further "relative risk" measure, derived when incidence rates are compared.

Epidemiologists often prefer to use odds rather than probability in assessing the risk of disease.

a. The mathematics of manipulating odds ratios is easier and can be performed using a handheld calculator.
b. The results of logistic regression can be expressed as odds ratios. Therefore it is possible to present results before and after multivariate adjustment in terms of odds ratios.
c. Finally, odds ratios are the only valid method of analysing case-control studies when the data is categorical. The odds ratio from the case-control study corresponds to the odds ratio in the population in which the case-control study was performed.[21]

For rare outcomes the odds ratio, risk ratio and incidence rate ratio have the same value. To illustrate calculating odds and odds ratios, the following table can be thought of as the results of either a cross-sectional survey, cohort study or case-control study.

	Cases	Controls
Exposed	a	b
Unexposed	c	d

$$\text{The odds of an event} = \frac{\text{number of events}}{\text{number of non}-\text{events}}$$

$$\text{An odds ratio} = \frac{\text{odds in the treated or exposed group}}{\text{odds in the unexposed group}}$$

The odds of being a case in the exposed group is a/b. Similarly in the unexposed group the odds of being a case is c/d. The odds ratio (OR) is therefore (a/b)/(c/d), and after manipulating algebraically, (ad)/(bc). The odds ratio is therefore a "relative odds" and gives an estimate of the "etiological force" of an exposure.

If the OR is greater than 1, the exposure is a risk factor for the disease.

If the OR is less than 1, the exposure (often a treatment) protects against the disease.

If the OR is exactly equal to 1, there is no association between exposure and disease.

ΠΕΡΙΛΗΨΗ

Αρχές αξιολόγησης της βιβλιογραφίας στα πλαίσια της βασισμένης στις ενδείξεις Ιατρικής

Π.Α. ΣΚΑΠΙΝΑΚΗΣ,[1] N. STIMPSON,[1] H.V. THOMAS,[1] F. DUNSTAN,[2] R. ARAYA,[1] G. LEWIS[1]

[1]Department of Psychological Medicine, University of Wales, College of Medicine, UK

[2]Department of Medical Computing and Statistics, University of Wales, College of Medicine, UK

Αρχεία Ελληνικής Ιατρικής 2001, 18(2):192–203

Ο όρος αξιολόγηση της βιβλιογραφίας, όπως αυτός χρησιμοποιείται στα πλαίσια της βασισμένης στις ενδείξεις Ιατρικής, αναφέρεται στην εφαρμογή προκαθορισμένων κανόνων ιεράρχησης των ενδείξεων με στόχο την εκτίμηση της μεθοδολογικής ποιότητας μιας μελέτης και της κλινικής της χρησιμότητας. Η αξιολόγηση της βιβλιογραφίας αποτελεί ένα από τα πιο καίρια στάδια της βασισμένης στις ενδείξεις Ιατρικής και, όντας το πιο τεχνικό, απαιτεί από τον κλινικό μια επαρκή θεωρητική γνώση των αρχών και μεθόδων της ερευνητικής μεθοδολογίας. Στόχος του παρόντος άρθρου είναι ακριβώς η παροχή στον αναγνώστη των βασικών θεωρητικών δεξιοτήτων, που είναι απαραίτητες για την επιτυχή ολοκλήρωση του σκοπού της αξιολόγησης. Η παρουσίαση ακολουθεί σε αδρές γραμμές τη σειρά με την οποία ο ερευνητής σχεδιάζει και εκτελεί τη μελέτη του. Αρχικά, περιγράφονται τα βασικά σχέδια που χρησιμοποιούνται στην επιδημιολογική και κλινική έρευνα. Στη συνέχεια, αναφέρονται πολύ συνοπτικά οι στατιστικές αρχές που διέπουν την ανάλυση των δεδομένων και με τις οποίες μπορούμε να απαντήσουμε στο ερώτημα της «στατιστικής σημαντικότητας». Μετά, αναλύεται η έννοια της αιτιολογικής συμπερασματολογίας, δηλαδή πώς μπορούμε να συμπεράνουμε εάν η συσχέτιση δύο

παραγόντων σημαίνει και αιτιολογική τους σχέση. Τέλος, περιγράφεται ο τρόπος με τον οποίο μπορούμε να αξιολογήσουμε την κλινική σημαντικότητα των αποτελεσμάτων μιας έρευνας, εξετάζοντας τόσο το είδος της έκβασης που μελετήθηκε (είναι η έκβαση κλινικά σχετική;), όσο και το μέγεθος του αποτελέσματος που εκτιμήθηκε (είναι κλινικά σημαντικό, ακόμη και αν είναι στατιστικά σημαντικό;).

Λέξεις ευρετηρίου: Αιτιατική συμπερασματολογία, Αξιολόγηση βιβλιογραφίας, Ιατρική βασισμένη στις ενδείξεις, Σχεδιασμός έρευνας

References

1. LAST JM. *A dictionary of Epidemiology*. 3rd ed. Oxford University Press, New York, 1995
2. SACKETT B, STRAUS S, RICHARDSON WS, ROSENBERG W, HAYNES RB. *Evidence Based Medicine. How to practice and teach EBM*. 2nd ed. Churchill Livingstone, Edinburgh, 2000
3. MACMAHON B, TRICHOPOULOS D. *Epidemiology. Principles and methods*. Little, Brown and Co, Boston, 1996
4. LEWIS G, APPLEBY L, JARMAN B. Suicide and psychiatric services. *Lancet* 1994, 344:822
5. SKAPINAKIS P, LEWIS G, MELTZER H. Clarifying the relationship between unexplained chronic fatigue and psychiatric morbidity: results from a community survey in Great Britain. *Am J Psychiatry* 2000, 157:1492–1498
6. KENDELL RE, JUSZCZAK E, COLE SK. Obstetric complications and schizophrenia: A case-control study based on standardised obstetric records. *Br J Psychiatry* 1996, 168:556–561
7. KENDELL RE, McINNERY K, JUSZCZAK E, BAIN M. Obstetric complications and schizophrenia: Two case-control studies based on structured obstetric records. *Br J Psychiatry* 2000, 176:516–522
8. LEWIS G, DAVID A, ANDREASSON S, ALLEBECK P. Schizophrenia and city life. *Lancet* 1992, 340:137–140
9. HOTOPF M, CHURCHILL R, LEWIS G. The pragmatic randomised controlled trial in psychiatry. *Br J Psychiatry* 1999, 175:217–223
10. POCOCK SJ. *Clinical trials. A practical approach*. John Wiley & Sons, Chichester, 1983
11. KUNZ R, OXMAN AD. The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *Br Med J* 1998, 317:1185–1190
12. BLACK N. Why we need observational studies to evaluate the effectiveness of health care. *Br Med J* 1996, 312:1215–1218
13. McKEE M, BRITTON A, BLACK N, McPHERSON K, SANDERSON C, BAIN C. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. *Br Med J* 1999, 319:312–315
14. LEWIS G, CHURCHILL R, HOTOPF M. Systematic reviews and meta-analysis. *Psychol Med* 1997, 27:3–7
15. THOMSON SG. Why sources of heterogeneity in meta-analysis should be investigated. *Br Med J* 1994, 309:1351–1355
16. LAU J, IOANNIDIS JPA, SCHMID CH. Summing up evidence: one answer is not always enough. *Lancet* 1998, 351:123–127
17. LEWIS G, THOMAS H, CANNON M, JONES P. Epidemiological methods. In: Thornicraft G, Szmukler G (eds) *Textbook of Community Psychiatry*. Oxford University Press, Oxford, 2001
18. ROTHMAN KJ, GREENLAND S, ROTTHMAN KJ, GREENLAND S. *Modern Epidemiology*. 2nd ed. Lippincott, Williams and Wilkins, Philadelphia, 1998
19. BROWN GW, HARRIS T. *Social Origins of Depression*. Tavistock, London, 1978
20. MUIR GRAY J. *Evidence based health care. How to make health policy and management decisions*. Churchill Livingstone, New York, 1997
21. SCHLESSELMAN JJ. *Case-control studies: design, conduct, analysis*. Oxford University Press, New York, 1982

Corresponding author:

P.A. Skapinakis, Heath Park, Cardiff CF 14 4XN, UK
e-mail: skapinakisp@cardiff.ac.uk.