

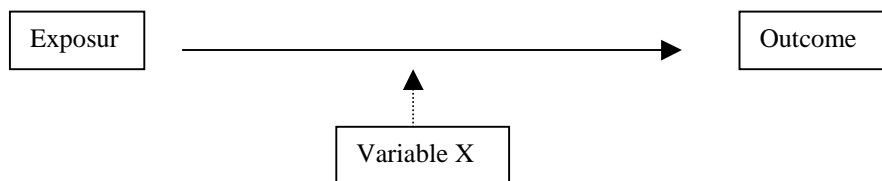
Chapter 11

Establishing the Cause

Usually many factors operate together to cause disease. There are predisposing causes, and precipitating causes. Koch's postulates are useful only when there is a single causative factor, as is the case with infective illnesses. In the majority of non-infective illnesses the occurrence of disease may be determined by a variety of factors like genetic predisposition, environmental agents, or behavioral causes. These are often referred to as risk factors. Knowledge of risk factors may help to develop preventive approaches even before the full pathogenic mechanisms are known e.g. reduction in mortality from coronary heart disease in many industrial countries by 30% in recent years.

Before making a causal inference about whether a given exposure is the cause of an outcome, one must ask, "compared to what?" For example, smoking one pack of cigarettes per day is a cause of lung cancer compared with not smoking, but not compared with two packets per day.

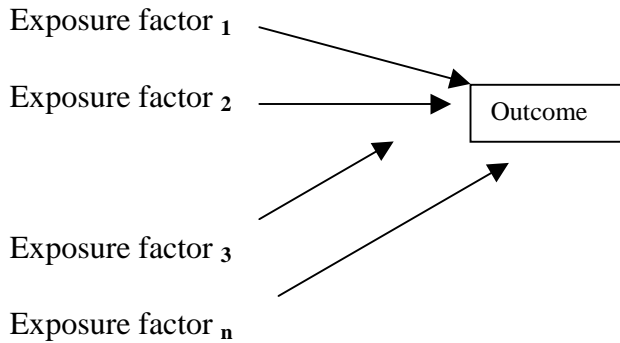
A given exposure is considered a **necessary cause** of an outcome if the outcome does not occur in its absence. All changes in the outcome are preceded by changes in the necessary cause. Exposure is a **sufficient cause** if it inevitably leads to the outcome without requiring the presence or absence of any other factors. In some cases a cause may need the presence of one or more other factors. These are called effect modifiers. In the illustration below X is an **effect modifier**.



Exposure to tubercle bacillus does not necessarily cause tuberculosis in all individuals. Age, immune status, nutritional status, living conditions all play a role in determining whether an exposed person develops the disease or not. Variable X in the above figure represents these conditions.

In real life, for many health problems causality is multifactorial. For example young mother, poor maternal nutrition, infections like rubella in pregnancy, smoking in pregnancy are all factors which alone or in combination can cause low birth weight. Each factor may independently contribute or augment the risk, and so each can be considered a true cause.

The multifactorial model can be written as:

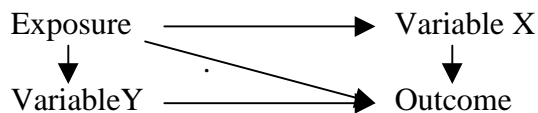


There may be interactions among the exposure factors as well as with other extraneous variables.

Conversely, we know that certain factors can cause more than one pathological outcome. Smoking is a good example to illustrate this point, since it can cause coronary artery disease, chronic lung disease, cancer and so on.

Causal Paths

Exposure may cause an outcome through an intermediate factor. These are called mediating variables. A series of causally linked variables is called a causal path or causal chain. For example, low socio-economic status causes disease in a variety of ways by affecting material resource, life-style, and general attitude. These are the more proximate determinants. Sometimes they are described as direct causes, and the more distal ones as indirect. Direct and indirect causes often interact in a variety of causal paths and networks to bring about disease. This is depicted in the model below.



Assessment of causality inevitably involves a statement of probability. The probability need not be 1 (that is, 100 per cent) to justify action. In assessing the contribution of an exposure to the outcome, one first arrives at a probability statement, and the probability is then calculated at between 0 and 1.

Three **general conditions** must be fulfilled before a given agent can be considered to be a risk factor for a disease. These are:

1. The cause must co-vary with the disease. It is not enough to just demonstrate association. The frequency of the disease must vary with the prevalence of the factor thought to be a cause. As a corollary, the severity of the disease should be shown to vary with the intensity of the factor.
2. The proposed causative agent must be shown to precede the outcome.

3 The observed relationship must not be due to any error in the study design, or due to chance, or some extraneous other factor, or errors in data analysis and interpretation of findings.

Causality can be assessed in terms of 3 different questions relating exposure to outcome viz.

1. Can it cause the outcome?

What is the probability that exposure can, at least in some persons under certain circumstances, cause the outcome?

2. Will it cause the outcome?

What is the probability that the next person exposed will develop the outcome because of the exposure? Is the exposure a quantitatively important cause of the outcome?

3. Did it cause the outcome?

What is the probability that a given person who has already developed the outcome did so because of the exposure?

Let us next examine each of the three questions in detail.

Can exposure cause the outcome?

This question is usually posed in relation to groups rather than individuals, and then too for diverse groups in order to find out whether there are any effect modifiers. If reduction or elimination of exposure leads to a lower risk or lesser degree of risk of the outcome in a given population the exposure can be said to cause the outcome.

Epidemiological studies are used to answer the "Can it?" type of question. The stronger the epidemiological evidence in favour of causality, the higher the probability for "Can it?" Assessing whether the findings of such studies reflect a true cause – effect relationship is a matter of ruling out the likelihood of alternative explanations like chance, bias and confounding.

The different elements to take into account in weighing up the evidence are the following:

Chance – It is necessary to quantify the degree to which chance variability accounts for the difference observed between the subjects in the study. This is done by performing an appropriate test of statistical significance like the *t* test or the χ^2 test. The *P* value so obtained gives the probability of the observed result occurring by chance alone. A more informative measure is the confidence interval, which provides the range within which the true magnitude of effect lies with a degree (90% or 95%) of assurance. The narrower the confidence interval the better, because it means less variability was present in the estimate of the effect. Statistical significance, however strong, should never be viewed as a definite yes or no statement, but rather as a guide in arriving at a final decision about causality.

Bias – A second and equally important explanation for any observed association is the possibility that some aspect of the design or conduct of the study has introduced a systematic error (or bias) in the results. The evidence in favour of causality is strengthened if all possible sources of bias in selection of subjects, gathering of information or in measurement have been eliminated. Different sources of bias and ways of avoiding them are discussed below.

Information bias:

The measurements of exposure and outcome should be precise and reproducible, and neither measurement should be influenced by the other. Imprecise measurements obscure true relationships.

A systematic difference between the research question and the actual question answered by the study can lead to wrong conclusions being drawn. It always helps to write down the research question and the study plan side by side, and then to consider the following:

- 1). Do study subjects accurately represent the target population?
- 2). Does measurement of the predictor accurately represent the predictor variable of interest?
- 3). Does measurement of the outcome accurately represent the outcome variable of interest?

Whenever the answer is "No" or "May be" to any of these questions one needs to consider whether the bias would be large enough for a wrong conclusion to be made. Bias is best dealt with by vigilance at the design stage of a study. However, if a potential source of bias is discovered after the data has been gathered there are two possible ways of dealing with it:

- Check the information which has already been gathered and/or consider obtaining additional information.
- Check the results of other studies for consistency of conclusions. When several studies have come up with similar results, the association is less likely to be due to bias.

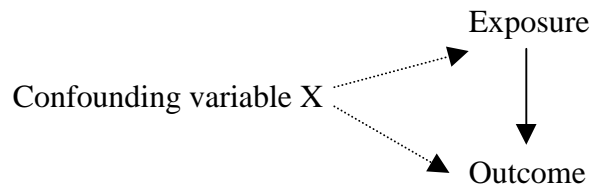
Sample distortion bias:

It commonly arises during assembly of the study sample or because of unbalanced dropouts during follow-up. Vigilance and care are also needed at the data gathering stage of a study.

Confounding bias:

The third alternative explanation to exclude is that an observed association could in fact be due to a mixing of effects between the cause, the disease and a third factor that is associated with the cause and also independently affects the outcome. For example, anaemia in pregnancy is said to be associated with low birth weight in the infant. But a poor maternal diet can cause both anaemia in the mother and low birth weight in the infant. Anaemia is thus only a marker and not a cause. Here the confounding variable makes the cause and effect appear connected even though their association may be spurious. If

the confounding variable is controlled for (e.g. for different levels of blood haemoglobin) the spurious association disappears.



Confounding can also occur when the exposure factor is closely linked to a third factor e.g. an adverse drug reaction may be caused by an adjuvant but gets attributed to the drug.

Randomized controlled trials are a good protection against confounding. When they are double blind, and standardized measuring techniques have been used together with careful follow-up, RCTs provide the most convincing evidence of the Can it? type. This is so because in intervention studies one variable at a time can be altered and observed, so that any effects that are observed can be attributed to that variable thereby removing uncertainty and also establishing temporality.

In observational studies the first step is to make a list of variables which may be associated with the predictor variable and may also be a cause of the outcome. The next step is to work out strategies for the design and the analysis phases to control the influence of these potential confounders.

Design stage strategies for the avoidance of bias comprise specification and matching. Both involve changes in sampling. Whether sampling is done by exposure (e.g. cohort studies) or by outcome (e.g. case-control studies) the subjects are selected in such a way that they have the same value for the confounding variable.

Specification involves describing inclusion criteria which specify a value of the confounding variable and exclude subjects with different values. The problem is that the sample size gets restricted.

Matching has the advantage that subjects at all levels of the confounder can be studied, and so the sample size does not suffer. Also generalizability is unaffected. The other advantages with matching are as follows:

- It is an effective way of preventing confounding by factors like age, sex, social class etc.
- Confounders who cannot be measured or dealt with in any other way can be controlled by matching.
- Precision of comparison between groups is increased by balancing the number of cases and controls at each level of confounding.
- Matching is often used as a sampling convenience to narrow down the number of controls.

Matching has been further discussed in Chapter 13. There are also disadvantages with matching. Additional time is spent and costs may be incurred in identifying matched subjects. Cases for whom no satisfactory match can be found will have to be discarded, and the sample size gets reduced. There is also the problem of overmatching.

Matching is a sampling strategy and must be made at the beginning of the study. Once made the decision is irreversible. This is a serious disadvantage because it can constrain the researcher's ability to observe real associations. Another major difficulty arises when a predictor variable is also a cause of the potential confounder. A less serious problem occurs when the matched variable is not fixed (e.g. age, sex, race and so on) but modifiable, and may be influenced by the predictor.

As the name suggests design phase strategies require deciding at the outset of the study as to which variables are to be treated as predictors and which as confounders. But the question of confounders can also arise as the study progresses, or even towards the end of data collection. In such cases analysis phase strategies for the avoidance of bias like stratification and adjustment are the answer.

- Stratification ensures that only cases and controls with similar levels of a potential confounding variable are compared. Several stratified analyses are performed and the results are compared with those obtained without stratification. If there are no substantial differences, it is possible to identify variables, which are true predictors and not confounders.
- Adjustment requires multivariate statistical modelling in which the influence of several confounding variables is controlled simultaneously.

Reverse Causality bias.

A given exposure can cause an outcome only when it precedes it. In cross-sectional studies it may not be easy to determine which occurred first, unless the exposure factor was known to be present since birth. So in general, the evidence from such studies is relatively weak, and usually open to reverse causality bias.

Case-control studies can avoid the reverse causality bias by using incident (newly occurring) outcomes and by specifically inquiring about prior exposure. This will depend on adequate records or valid recall. Cohort studies can avoid this problem if the study sample was known to be free of the outcome at the time the exposure began.

We have considered the question of bias at length. There are also other points to consider in answering the question “Can exposure cause the outcome?” These are now described below:

Strength of association - This is a measure of the effect on the outcome produced by a given amount of exposure. For a full assessment of a causal connection one needs to know not only of the existence but also of the degree of association. When the exposure and outcome are both dichotomous (of the yes/no variety) the relative risk provides the measure. When the exposure is dichotomous but the outcome continuous, the mean difference in outcome between exposed and not exposed is the indicator.

All the elements of the epidemiological evidence being equal, the larger the effect size the greater the likelihood of exposure causing the outcome. Small relative risks or mean differences always raise the question as to whether some unknown factor or bias is responsible for the results. Large effects are naturally more convincing.

Graded effect - When exposure is graded (e.g. mild/moderate/severe) or continuous, the "Can it?" question is better answered by a size of dose to size of response relationship. With higher categories of exposure the relative risk should increase in dichotomous outcomes. When both exposure and outcome are continuous the regression coefficient indicates the amount of change in outcome for a given increase or decrease in exposure.

Consistency - The more the number of studies coming up with similar answer, the stronger is the evidence. If different investigators in different settings and using different study designs find a significant association then the probability of exposure causing the outcome is increased. In experimental studies replication is particularly important in removing the element of chance. But repeated failure to control for sources of bias can also give consistent findings, which are invalid; e.g. anaemia and low birth weight not controlling for maternal nutrition.

Biological plausibility - The exposure-outcome association is plausible when supported by what is known about the mechanism of action and the underlying disease process. If the results contradict established biological factors, the plausibility is reduced until further studies confirm the alleged association.

Similarity to known cause - If a study shows an association between a new drug and marrow depression, then an older drug with a similar chemical structure and known to cause marrow depression helps to strengthen the evidence.

Will exposure to a known risk factor result in the outcome?

In other words, how important is the contribution of the exposure to the outcome?

If in the "Can it?" question the probability of an exposure being a cause turned out to be low, then it follows that the probability is also low for the exposure to be an important cause of the outcome. If the "Can it?" probability is high then the strength of the cause-effect relationship merits further study.

The "Will it?" probability is the difference in the probability of the outcome occurring in an exposed person, and the probability of the same outcome occurring without exposure. It is estimated by the attributable risk i.e. the difference in the incidence of the outcome in exposed and unexposed persons who are similar in all other aspects. Attributable risk is the measure of the additional number in the outcome owing to a given exposure.

Did exposure to a known factor result in the outcome?

The question "Did exposure cause the outcome?" is normally asked in the clinical setting. The focus is on the individual. A rough rule of thumb for etiological inferences is that a sharp relative risk gradient usually indicates a causal relation even in the absence of an adequate biological explanation. On the other hand, a weak empirical association may indicate a causal relation only when the supporting biological evidence is overwhelming.

Exposure to a known factor causing an outcome is measured by estimating the proportion of subjects developing the outcome who do so because of the exposure. This is called the Etiologic Fraction (EF) or Population Attributable Risk.

EF is measured by

$$EF = P_e(RR-1) \div P_e(RR-1) + 1$$

where RR is relative risk and P_e is prevalence of exposure in the community.

When exposure and outcome are both known to have already occurred in the individual, the Etiologic Fraction among the exposed (EF_e) can be derived from the following formula:

$$EF_e = (RR - 1) \div RR$$

In the absence of any other information except that an individual was exposed EF_e provides an estimate of the probability that the exposure caused the outcome. If for example, RR of lung cancer in a population is 10 in smokers as compared to non-smokers, the probability that a smoker developed lung cancer because of smoking should be

$$\frac{10 - 1}{10} = 0.90$$

Often one knows much more about the patient e.g. dose of exposure, its duration, social and other background factors, and so on. This information may be used to refine the "Did it?" estimate.

When more than one factor act together synergism can occur and the effect is more than the sum of individual factors. In such a situation it is often difficult to isolate the effect of each individual factor. But even then a substantial impact on health can be made by removing the effect of one of them. Effect modification occurs when the strength of the cause and effect relationship between two variables is different according to the level of some third variable, called effect modifier. For example, the effect of contraceptive use and myocardial infarction in women is modified (in the direction of enhancement) by cigarette smoking.

Establishing the Cause

Having considered the evidence that needs to be mustered for establishing a causal association, let us now turn to study designs for the purpose.

Studies in Individual Subjects

In clinical work, cause and effect relationship can be determined only by empirical evidence, which can then be built up to the extent that all possible doubt is removed. Similarly, evidence against

a cause can be mounted to the extent that a cause and effect relationship looks impossible. The same logic also applies in experimental studies, which are on firmer ground, because evidence is gathered under carefully controlled conditions. There is thus less scepticism about the proof in experimental studies.

Association and cause

The suspected cause and effect must be shown to be associated before the investigator can begin to consider one a cause and the other an effect.

In order to do so, a decision is first to be made whether the association is real or merely an artifact created by bias or random variation. Checking for selection and measurement biases and the role of chance as well as methods of dealing with each was discussed earlier. If these factors are excluded, a true association exists. If confounding also is excluded, a causal relationship now appears more likely. Even then at a future date some another hitherto unknown factor may come to light which is more directly causal. Factors which are considered causal at one time are sometimes found to be indirectly related to disease later, when more evidence becomes available.

Association and Cause

EXPLANATION	ASSOCIATION	FINDING
Bias in selection or measurement	Yes	No
Chance	Likely	Unlikely
Confounding	Yes	No
Cause		Cause

Hierarchy of research designs

The most important evidence for a cause and effect relationship is the strength of the research design used to establish the relationship.

Well conducted randomized controlled trials, with adequate number of patients; blinding of therapists, patients and researchers; carefully standardized methods of measurements and analysis are the best evidence for cause and effect relationship. Such studies are best suited for studying the effects of a single factor. They guard against differences in the groups being studied.

As we have seen randomized controlled trials are commonly used in studies about treatment and prevention. But practically speaking they cannot be used to show that a particular agent is a cause of a disease. The reason is that a potentially harmful agent cannot be assigned to individuals for ethical reasons. There are also the problems related to long latent periods, and the large number of patients needed. Because of these and other reasons, randomized controlled trials are rarely feasible. Instead, observational studies must be used.

The more one departs from randomized controlled trials the less does the research design protect against possible biases, and the weaker gets the evidence for a cause and effect relationship.

Well-controlled cohort studies are the next best designs. They can be conducted to minimize the effects of selection and measurement biases, as well as known confounding biases. Cross sectional studies are vulnerable because they provide no direct evidence of the sequence of events, as to what came first.

Prevalence surveys, cross sectional studies of a defined population, guard against selection bias, but are subject to measurement and confounding biases. Case control studies also are vulnerable to selection bias. Weakest of all are case series because they have no defined population, and no comparison group.

Summarizing the evidence for or against cause.

The following criteria may be used to guide decisions whether a given environmental agent is a cause of a disease:

1). Relationship in time (Temporal Relationship).

Cause should precede effects. This fundamental principle may get overlooked when interpreting cross-sectional studies and some case control studies in which both the proposed cause and the effects are measured at the same time. In both these types of studies it is *assumed* that one variable precedes another without establishing that this is actually so.

Although it is absolutely necessary for a cause to precede the effect temporal sequence alone is a weak evidence.

2). Strength of the association

A strong association is indicated by a large relative risk or odds ratio. For example, the relative risk of 4 to 16 in different studies for lung cancer amongst smokers, and of 300 for hepatocellular cancer and hepatitis B. Sometimes bias can result in large relative risks, but usually unrecognized bias is less likely to produce very large values of relative risks.

3). Dose-response relationship.

Such a relationship is present when varying amount of the cause produces varying amount of the effect. For example, the number of cigarettes smoked (dose) and the varying likelihood of lung cancer. If a dose-response relationship can be demonstrated it strengthens the argument for cause and effect.

Although a dose-response relationship is a good evidence, especially when coupled with a large relative risk, it does not rule out confounding factors.

4). Reversible association.

A factor is most likely to be a cause if its removal results in a decreased risk of disease. In other words the association between suspected cause and effect is reversible.

Confounding can still explain a reversible association.

5). Consistency

When several studies, conducted at different times, in different settings, and with different kinds of patients come up with similar results the evidence is much strengthened. But they can all make same mistakes. So evidence is strengthened when studies with different research designs come up with similar results.

6). Biologic plausibility.

Biologic plausibility depends on whether the cause and effect relationship is consistent with our existing knowledge about the mechanism of disease. However, such plausibility depends on the state of medical knowledge. For example acupuncture for anaesthesia.

7). Specificity.

One cause one effect is mostly found in infectious diseases, or for inborn errors of metabolism. Usually there are many causes for the same disease, or many effects from the same cause. (Cigarette smoking causes bronchitis, coronary artery disease, lung cancer, peptic ulcer and so on). Absence of specificity is not a strong point against cause-effect relationship.

8). Similarity.

The cause and effect relationship is strengthened if there are examples of similar effects. For example, slow virus disease like kuru and bovine spongiform encephalitis (the mad cow disease).

Studies in Population Groups


In some studies exposure is known for a population group and not for individuals, e.g. radiation following the Chernobyl disaster.

Such studies are known as *aggregate risk studies* or *ecological studies*. In these studies people are classified by the general level of exposure in their environment. The problem here is that people in a generally exposed group may not themselves be exposed. For example exposure to lead of children living in homes near motorways. There can be many confounding factors. Such studies are helpful in raising hypotheses, which must be tested by more rigorous research.

Time series studies are one way forward. In these studies the effect is measured at different times before and after the purported cause has been removed. If there is an appreciable fall in the effect after the purported cause has been removed the evidence is strengthened. For example, reduction in the incidence of the toxic shock syndrome after the removal from sale of a particular type of tampon.

Weighing the evidence

When the evidence for a cause-effect relationship is conflicting, as is usual the case, clinicians must have to rely on their own judgment. Such judgment has to be based on the strength of the available evidence, the type of study design used to obtain the evidence, the rigor of the study, sample size, sampling methods used, and so on. The different types of studies, and the findings from them are summarized in the table below:

STRENGTH	DESIGN	FINDINGS
Strong  Weak	Clinical trial	Temporality
	Cohort study	Strength
	Case control study	Reversibility
	Cross-sectional	Dose-response
	Aggregate risk	Consistency
	Case series	Biologic plausibility
	Case report	Specificity Analogy

A study purporting to describe a cause-effect relationship can be assessed as follows:

Is there a clear statement of the hypothesis?

Research that arises from vague speculation is likely to lead to vague conclusions. It is necessary to have a well-defined statement as to which relationships are proposed as cause and effect.

Is the study design appropriate to the hypothesis?

See the figure above. Choice of study is determined partially by the current state of knowledge about the relationship being investigated. Cross-sectional and case-control studies are normally undertaken in the early stage of inquiry. Cohort studies and clinical trials are typically reserved for subsequent investigation.

Is the information about exposure and disease appropriate for the hypothesis and the study design?

Objective measures are to be preferred over subjective ones e.g. interview.

Are the methods of analysis appropriate to the research question, study design and data collection?

Significance level is influenced both by the magnitude of the association and sample size. A large sample size will produce a significant result that has little or no clinical significance.

Two Illustrative Examples

The above discussion on establishing the cause would have indicated to the reader that the progress from hypothesis to final identification of a cause is along a slippery and uncharted path. Flaws in study designs, bias, confounders, errors in data analysis, statistical methods employed as well as in interpretation of results dog every footstep. Pathways from cause to effect are at best conjectural, and subject to a variety of pitfalls. The usual requirements of consistency, strength of association, temporal relationship, and so on, are only guides. But once the cause is established beyond all reasonable doubts major benefits can follow. Two examples are described below of the kind of effort needed in proving a cause-effect relationship. These are about the role of folate deficiency in the etiology of Neural Tube Defects; and of copper in that of the Indian Childhood Cirrhosis.

The Role of Folate Deficiency in Neural Tube Defects

The possibility that folic acid may have a part to play in Neural Tube Defects was first raised in 1964. Several small-scale studies followed with inconclusive results. Then in 1980 an intervention study was reported in women who had previously given birth to affected children, and were therefore at greater risk than the general population for future pregnancies. Those planning another pregnancy but not yet pregnant were given a mixture of eight vitamins including folic acid to take daily; those already pregnant or refusing to participate served as controls. The risk of recurrence in the supplemented group turned out to be one-seventh of the unsupplemented women. This study was followed by a randomized trial in 1981 with folic acid alone. But the results were inconclusive because of small numbers and non-compliance. Clearly a large randomized controlled trial was needed, but a lively debate ensued about the ethics of withholding folic acid in case it was truly protective. Finally, a multi-centre double blind randomized trial involving 33 centres was planned. A factorial study design was employed in order to assess the effect of folic acid and a selection of other vitamins. Subjects known to be at high risk as judged by a previous affected pregnancy were recruited into the study. The results of the trial were published in 1991. Folic acid was shown to have a 72% protective effect. The other vitamins had no demonstrable protective action. Folic acid is now recommended for supplementation during pregnancy.

Indian Childhood Cirrhosis.

The first description of this condition was in 1887. To start with there was the problem of case definition. Despite many publications and several reviews there was confusion until 1960 when the matter was finally resolved, and a case definition agreed based on histological criteria. Several clinical trials with a variety of drugs gave inconclusive results. Then in 1970 a group of investigators seeking to demonstrate hepatitis B surface antigen in biopsy specimens of the liver used orcein staining. An unexpectedly large number of hepatocyte turned out to be orcein positive, believed to be due to copper associated protein. Copper was demonstrated histologically in liver biopsies obtained from children with the illness in 1978. This was followed by the demonstration of high hepatic concentration of copper by means of atomic absorption spectrophotometry. A high concentration of copper in milk stored in copper vessels was demonstrated by case control studies, and the large amounts of copper leeching out of brass vessels into milk stored for six hours was demonstrated experimentally.

There may still remain a few more mysteries to solve, but clear warning about the dangers of storing milk in untreated brass vessels is now widely recommended in India.

Appendix 11.1

CHECKLIST FOR STUDIES ON ETIOLOGY

1). How **STRONG** was the method used?

STRONGEST

WEAKEST

Randomized clinical
trial

Case series

Cohort study

Case - control study.

2). Is there **EXPERIMENTAL EVIDENCE** in humans?

3). Is the **ASSOCIATION** strong?

4). Is the association **CONSISTENT** from study to study?

5). Is there a **DOSE-RESPONSE** gradient?

6). Does the association make **EPIDEMIOLOGIC** and **BIOLOGIC** sense?

7). Is the association **SPECIFIC**?

8). Is the association in line with previously proven causal association?