Appendix A4

Guidance on the epidemiological investigation of outbreaks of infection

Contents

A4.1 Introduction

A4.2 Recognition of an outbreak

A4.3 Initial investigation

3.1 The Epidemiological Study Team
3.2 Obtaining outside support
3.3 Definition of the study population
3.4 Definition of cases
3.5 Case finding

A4.4 Descriptive epidemiology

A4.5 Analytical study

5.1 Setting the objectives
5.2 Hypothesis generation
5.3 Cohort studies

A4.6 Case control studies

6.1 General points
6.2 Bias
6.3 Confounding
6.4 Selection of controls
6.5 Sample size calculations
6.6 Data collection

A4.7 Statistical analyses of case control study data

7.1 Introduction
7.2 Single risk variable analysis
7.3 Multivariable logistic regression analysis

A4.8 Report writing

This Appendix is based upon a report prepared by the Public Health Laboratory Service under contract to the Drinking Water Inspectorate. The principal authors were:

Dr John de Louvois, Head, PHLS Environmental Surveillance Unit, CDSC, Colindale (since retired)
Dr Patricia O’Neill, Consultant Microbiologist/CCDC, Shropshire HA, Shrewsbury Public Health Laboratory
Dr James Stuart, Consultant Epidemiologist, CDSC (South and West), Gloucester Public Health Laboratory
Dr Tony Swan, Head, PHLS Statistics Unit, Colindale
A4.1 Introduction

1.1 High quality epidemiological information is vital in the investigation of possible outbreaks of waterborne infection associated with mains water consumption because microbiological evidence of water contamination by pathogenic organisms is usually difficult to obtain. Epidemiological investigations of such outbreaks are not straightforward; they are relatively uncommon and may be statistically complex, such that individual consultants in communicable disease control (CCDCs) or consultants in public health medicine (CPHMs) may appreciate help with investigations.

1.2 This manual has been produced to give CCDCs/CPHMs and other members of the Outbreak Control Team (OCT) practical advice to assist in the conduct of epidemiological studies. It is aimed at the investigation of outbreaks that may be associated with the consumption of mains water. As *Cryptosporidium* is the most commonly reported organism involved in such outbreaks the emphasis is on investigation of outbreaks caused by this organism. However, the general principles set out in this manual will apply to all outbreaks of potentially waterborne infection and to the investigation of all alternative hypotheses. The manual is not intended to replace local outbreak control plans, but it is hoped that CCDCs/CPHMs will find it a useful supplement and that its use will lead to the establishment of an agreed best practice in the investigation of water associated infection.

1.3 Investigation of suspected waterborne outbreaks involves four complementary activities.

- Epidemiological investigation of persons at risk.
- Microbiological investigation of patients, water, treatment works, animals and the environment.
- Engineering investigations.
- Investigation for evidence of a contamination incident at the water treatment works.
1.4 All four activities are important and need to be carried out in parallel and with close liaison between the investigating agencies. This guidance document will deal only with the epidemiological aspect of the investigation.

A4.2 Recognition of an outbreak

2.1 High quality surveillance is the cornerstone of outbreak recognition and investigation. Surveillance includes:

- recognition by clinicians, especially general practitioners (GPs), of clinical cases of diarrhoeal disease and taking of stool samples;
- appropriate microbiological investigation of all diarrhoeal samples submitted to laboratories;
- reporting of laboratory results from cases;
- maintaining a current health authority (HA) database with regard to water associated events;
- regular review of the incidence of pathogens which might be associated with transmission by the waterborne route;
- regular liaison with water utilities concerning water quality data, for example, faecal indicator organisms, increased turbidity;
- prompt reporting of suspected outbreaks by CCDCs/CPHMs to the regional/national surveillance centres; and
- maintenance of national databases of water associated events.

2.2 Although Cryptosporidium is the most commonly reported organism in outbreaks associated with mains water contamination, other organisms may be responsible for outbreaks of waterborne disease, for example Campylobacter, Giardia, Escherichia coli O157 and enteric viruses. Water should be considered as one of a number of possible vehicles of infection whenever an outbreak due to these organisms is being investigated.

2.3 Particular attention should be paid to monitoring and investigating cases of cryptosporidiosis (Furtado et al 1998). Sporadic cases of cryptosporidiosis should be interviewed to identify common risk factors using a short questionnaire. An example is given in Sub-appendix A4.1. Interviews by EHOs are already standard practice in many districts.

2.4 Thresholds for investigation according to level of Cryptosporidium incidence above baseline rates of disease should be agreed (Table 1).

<table>
<thead>
<tr>
<th>Table 1: Suggested thresholds for investigation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed number of cases in one week</strong></td>
</tr>
<tr>
<td>Usual weekly rate*</td>
</tr>
<tr>
<td>&gt;=1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

* Allowing for seasonal variation
These numbers have been calculated to have only a 1 in 20 ‘check’ or a 1 in 100 ‘alert’ probability of occurring by chance assuming that the baseline numbers of cases per week follow a Poisson distribution (and could be used for any organism). Reaching the check or alert level in any one week is suggested as an indication to review recent cases, and to take further action if appropriate. Reaching the alert level for two consecutive weeks (which should only occur by chance about 1/10,000 weeks without an outbreak) is a suggested indicator for calling an OCT meeting. The CCDC/CPHM may decide that responses are appropriate at other levels of incidence.

A4.3 Initial investigation

3.1 The Epidemiological Study Team

3.1.1 One of the major responsibilities of the OCT is to carry out appropriate epidemiological studies of the outbreak. These will normally be dealt with by a sub-group of the OCT. Typically, this will be led by the CCDC/CPHM, with input from EHOs and other members of the team. The regional epidemiologist (RE) or Scottish Centre for Infection and Environmental Health (SCIEH) representative will usually be invited to join the OCT for large or multi-district outbreaks.

3.1.2 Specific resources will be specifically required to undertake epidemiological studies effectively. These include access to relevant information e.g. water supply zones, population densities, HA databases for choice of controls; computer software, including maps and statistical packages to store and analyse the data; and staff trained to use them; access to expert statistical advice; data entry clerks and trained staff to carry out interviews for the questionnaire. Interviewers may be the most difficult resource to obtain, particularly if it is a large outbreak. Interviews may be conducted by EHOs and staff of the Department of Public Health Medicine. In England and Wales the RE may be able to recruit public health staff from elsewhere within or outwith the region to assist. However, the number of interviewers should be kept as small as possible to ensure that interviewing style is as uniform as possible throughout the whole data collecting exercise. The Information that is kept by water utilities on their public records is detailed in Sub-appendix A4.2.

3.1.3 Outbreaks can dramatically increase in size during the progress of an investigation. The burden of interviewing all new cases can be great, and the possible need for additional interviewers should be assessed periodically during an outbreak. It may be necessary to recruit interviewers from outside the groups recommended above. Other potential sources include hospital infection control nurses, health visitors and school nurses. However, as these groups are likely to be less familiar with the use of the relevant questionnaires, it is vital that they are appropriately briefed. Some of the factors to be considered in briefing interviewers are summarised in Sub-appendix A4.3.

3.2 Obtaining outside support

3.2.1 The RE, Public Health Laboratory Service (PHLS) Communicable Disease Surveillance Centre (CDSC), or SCIEH will be able to provide support. This support could include:
■ providing regional/national incidence data on the relevant pathogen;
■ advising or assisting with epidemiological studies;
■ seeking or providing specialist registrar support;
■ seeking or providing statistical support;
■ co-ordinating epidemiological investigations if several districts are involved.

3.2.2 Issues in the design and analysis of such epidemiological investigations are complex. Expert statistical support is recommended, for example, from the PHLS Statistics Unit, as well as advice from experts in waterborne disease at national centres such as CDSC, SCIEH, the PHLS Cryptosporidium Reference Unit or the Scottish Parasite Diagnostic Reference Laboratory.

3.3 Definition of the study population

3.3.1 It is necessary to identify and define as precisely as possible the population considered to be at risk. This may be restricted to a specific age range, or by other factors, as well as to an identifiable administrative area. Definitions involving a potential risk factor, such as ‘those living in given “water supply” areas’ should be avoided because that could pre-empt the result. The population defined as ‘those on the lists of health or local authorities in which cases have occurred’ has proved useful in the past and is probably the safest definition to use unless there are particular local reasons against it. Less rigorous, but sometimes useful is the population defined by ‘the combined practice lists of the GPs of the cases’. The cases and controls to be included in the epidemiological study must be recruited in such a way as to ensure that they are members of this population. For outbreaks in rural or seaside areas during holiday periods the population at risk is likely to include visitors who are temporary residents. Inclusion in the case definition should be considered but follow up may be difficult especially when urgent investigation is required.

3.4 Definition of cases

3.4.1 The initial case definition should be designed to include all those who could reasonably be part of the outbreak. It needs to define geographical, clinical and temporal parameters and whether temporary residents are included.

3.4.2 An example of an appropriate case definition is:

■ Probable case: Any person resident in the area ______________ who became ill with symptoms of ______________ with date of onset on or after ______________.

■ Confirmed case: As probable case with stool specimen positive for ______________.

3.5 Case finding

3.5.1 Cases should be actively sought through enhanced surveillance as the cases detected initially usually represent only a small proportion of the total cases occurring. Alerting local microbiology laboratories, GPs and
hospital doctors, and other public health departments, should help to identify additional cases. Regional and national case finding may be helpful especially if temporary residents are included in the case definition.

A4.4 Descriptive epidemiology

4.1 The cases should be described by age, occupation, sex, time, place of residence and work/school, and by relevant known exposures obtained from initial interviews, for example, farm visits, recreational water contact, drinking water source and supply zones, food consumed and milk supply (see Sub-appendix A4.1).

4.2 If local microbiological screening for Cryptosporidium includes samples from adults, positives during waterborne outbreaks will usually be distributed throughout the age range. The median age for primary cases is often increased in comparison to the median for sporadic cases. A predominance of cases in school-aged children may indicate that infection has arisen from farm visits or swimming pools.

4.3 An epidemic curve should be created, and kept up-to-date, by plotting a bar chart indicating the numbers of cases, as the height of the bars on the horizontal scale according to the date on which symptoms first appeared. The pattern of the epidemic curve may suggest whether the outbreak originated from a point source, continuing source or as a result of person to person spread. The attack rates by area, eg local authority, water supply zone should be calculated. The denominator is the population at risk in the relevant area. The proportion of cases exposed to each of the suspected risk factors should be examined.

4.4 Water supply zone. Cases should be plotted on a map of water supply zones to determine whether distribution of cases matches the distribution of water supplies. Such maps should be provided by the water utility. Water from different treatment works is often blended so supply maps should ideally include information on the percentage of water from each treatment centre supplied to each zone. Statistical tests can be used to estimate the associations between infection rates and the percentage from each centre and the probability that differences in rates of illness in the different water supply zones occurred by chance.

4.5 Proceeding to analytical study. The OCT should review the evidence for water contamination and other suspected risk factors. Finding a statistically significant association between being ill and residence in a particular water supply zone does not necessarily mean that illness was caused by the water. However, in conjunction with evidence of mains water contamination, this can provide strong support for a hypothesis that mains water is the vehicle of infection and for the immediate implementation of control measures. In the absence of clear evidence to identify a vehicle of infection, an individual based analytical study is recommended. This may need to be done rapidly. In the presence of clear evidence of the vehicle of infection, if a boil water notice has been issued and/or population immunity is high, these factors should be borne in mind when deciding whether to proceed with an analytical study (see paragraph 5.2.6 of this Appendix).
5.1 Setting the objectives

5.1.1 The objectives of the epidemiological study should be clearly defined. These are usually to determine the size and extent of the outbreak and to identify the source of infection and its mode of transmission so that appropriate measures can be implemented to control the current outbreak and prevent a recurrence.

5.1.2 In the protocol for the analytical study it is necessary to specify aims and objectives which identify the hypotheses to be tested, the effects to be estimated and the information needed for both.

5.2 Hypothesis generation

5.2.1 In epidemiological studies it is usual to begin with the premise that there is no association between the exposure most under suspicion and the illness (null hypothesis) and then assess the extent to which the evidence is in conflict with this by carrying out an analytical study. The hypothesis will generally take the form of specifying that ‘There was no association between exposure to a specified risk factor and an increased risk of infection’.

5.2.2 To generate hypotheses the value of good descriptive epidemiology cannot be overemphasised. The descriptive epidemiology from early cases should be reviewed in conjunction with the results of microbiological investigations, information on water quality and treatment and on any possible contamination of water sources.

5.2.3 More sensitive hypotheses for the investigation of waterborne infections use dose-response relationships. These take the form of specifying that ‘There was no increase in the infection rate associated with increasing exposure to a specific risk factor’. Testing these hypotheses requires that data on habitual levels of exposure are collected for those factors which could have a dose-response relationship with the rate of infection.

5.2.4 Ideally, cases used to identify a risk factor and for generating a hypothesis should not be used in a subsequent analytical study to test that hypothesis. However, if there are few cases, the power of the study may be compromised by excluding these early cases. They can be included if full data can be collected from them. The final statistical analysis should be performed with and without these particular cases.

5.2.5 As soon as an outbreak is identified in which mains drinking water may be the vehicle of infection, investigators should use a full outbreak protocol and questionnaire. Examples derived from documents which have been used successfully to investigate previous waterborne outbreaks subsequently identified as due to Cryptosporidium, are contained in Subappendices A4.4 and A4.5. If the same questionnaire is used for cases and controls throughout the investigation, then cases from the initial interviews can be included in any subsequent case control study without it being compromised. Controls should be recruited as early in the investigation as possible.

5.2.6 In carrying out analytical studies, it may be helpful to recognise that:
the power of epidemiological studies may be reduced if the population at risk has a higher level of immunity due to previous contamination incidents; and

introduction of an ‘advice to boil-water’ notice may influence responses of cases and controls.

Dose-response relationships (and water supply zone attack rates) should be less affected by the above factors.

5.3 Cohort studies

5.3.1 A cohort study is one which includes all of a complete cohort of individuals who may have been at risk whether or not they have been ill and will generally have more statistical power than a case control study. As outbreaks associated with mains water mostly have large at risk populations, a cohort study is usually not feasible. Although a sample of the at risk population could be taken, attack rates are unlikely to be high enough for adequate statistical power. As a general rule, a cohort study should only be undertaken if:

(a) most of the cases have arisen from a readily identifiable cohort;

(b) it is feasible to identify and interview all members of the cohort, and

(c) it is plausible that the exposure responsible for the illness was limited to the members of the cohort.

5.3.2 An example of this situation would be an outbreak arising as a result of contamination of a private water supply. If a cohort study is undertaken, analysis follows much the same lines as that for case-control studies, which is described fully below.

A4.6 Case control studies

6.1 General points

6.1.1 Where an analytical study is done and where the population at risk is large, that is in most outbreaks where mains water is a possible vehicle of infection, a case-control study is the most appropriate method of analytical investigation. The numbers of cases and controls should be sufficient to ensure that associations between potential risk factors and infection, which are large enough to explain the outbreak, can be shown to be statistically significant (see paragraph 6.5 below). The assumption that the vehicle of infection is water must always be tested.

6.1.2 Enquiries about the period of exposures normally refer to the incubation period before illness for the cases and to an equivalent period before interview for controls. Occasionally it may be known that exposures occurred during a specific calendar period. In that situation if the investigation is sufficiently timely, information should be requested on exposures during that period for both cases and controls. It is important that epidemiological studies should start as early in an outbreak as possible.

6.1.3 The case definition may be revised, if necessary, to exclude those individuals who were ill for reasons not related to the suspected source.
This would include travel abroad in the incubation period, or a household contact being ill during the incubation period prior to onset of symptoms in the case.

6.1.4 If possible, dose-response relationships should be assessed for all appropriate parameters, e.g. consumption of water and milk, frequency of immersion in swimming pool etc. This means collecting quantitative information on these exposures. Nearly all the controls and cases will have been exposed to some tap water. It is therefore important to collect detailed information on the amounts of unboiled water consumed by cases and controls, as well as its source. Questions on water consumption should be related to usual consumption (Sub-appendix A4.5, Q26).

6.2 Bias

6.2.1 Bias is defined in epidemiology as the ‘deviation of results or inferences from the truth, or processes leading to such deviation’. There are many ways in which the design, execution, analysis and interpretation of epidemiological studies can introduce bias (Table 2). A major task in epidemiology is the recognition and avoidance of such biases. Incompleteness or changing completeness of case ascertainment is unlikely to introduce bias in the results of case control studies unless a substantial proportion of cases is misclassified as controls.

<table>
<thead>
<tr>
<th>Type</th>
<th>Explanation</th>
<th>How to reduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td>Controls may not be representative of the population at risk</td>
<td>Selection of controls (See section 6.4)</td>
</tr>
<tr>
<td>Information bias</td>
<td>Cases may be more motivated to answer carefully than controls. If the suspected risk factor is known to cases and controls, this may influence response e.g. if an advice to boil water notice has been issued. (NB. This should not affect information on other risk factors and would be less likely to affect a dose response relationship)</td>
<td>Undertake interviews of both cases and controls within the same time frame as early as possible in the investigation. Use prompts and memory aids such as big local or national events</td>
</tr>
<tr>
<td>(a) Recall bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) Interviewer bias</td>
<td>The interviewer may question cases and controls in a different way thereby influencing the response.</td>
<td>Keep the number of interviewers to a minimum and ensure that they receive adequate instruction (Sub-appendix A4.3). Use the same interviewers and same interview method for cases and controls</td>
</tr>
</tbody>
</table>

6.3 Confounding

6.3.1 Confounding occurs when the relationship of an exposure to a disease is distorted by another (confounding) exposure. This occurs when the confounding exposure is not only associated with the exposure under investigation but is also an independent risk factor for the disease.

6.3.2 Confounding can be minimised by careful study design. The effect of confounding factors can be managed in the statistical analysis of the study, but only if information on the confounding variable is collected. Therefore if any confounding factors are identified, data on them must be collected in the questionnaire.
6.3.3. For example if some cases had acquired their infection while travelling abroad their exposure patterns to other risk factors might well be the same as the controls, but not the other cases. This means that combining them with non-travellers in the analysis will make the exposures to tap water (say) of the cases more similar to the exposures of the controls and hence tend to hide the effect of this exposure. Such confounding with travel is usually dealt with by excluding travellers from the study.

6.3.4 It is important to remember that any one outbreak may have two or more vehicles of infection. For example, a higher incidence of cryptosporidiosis from farm animal contact may coincide with an episode of mains water contamination.

6.4 Selection of controls

6.4.1 Controls must be selected in a way that ensures they are an unbiased sample of the unaffected members of the population at risk. Controls should be free of the disease being studied, and should be within the same overall age range and area of residence as the majority of cases. They do not need to be the same sex unless the cases are predominantly of one sex. Inclusion or exclusion criteria, which apply to the cases (except their disease status), should apply equally to the controls.

6.4.2 Matching. This is a traditional way of dealing with confounding devised before statistical software for stratified multivariable analyses was widely available. Other than to ensure a roughly even distribution of age and sex between cases and controls (frequency matching), matching in studies of water associated outbreaks of infection where controls are relatively easy to obtain is rarely of much benefit. Matching can lead to wasted data, and confounding variables are best handled in the analysis of the data. Formal matching hardly ever increases the power and reliability of this kind of study and should only be used when there is clear evidence that it will be of benefit.

6.4.3 In investigating outbreaks of waterborne illness health authorities (HA) controls are recommended where feasible and especially if rapid recruitment is not compromised. These will be associated with particular cases and may be kept to a similar age range, but should not be considered as formally matched. However, although matching should generally be avoided ‘case nominated’ controls are a useful alternative when rapid recruitment is necessary. The analysis must then take full account of the matching between the cases and their ‘nominated’ controls.

6.4.4 Health Authority/GP controls. Controls are selected from the same health authority or general practice and from the same age band, that is <6 years, 6-15 years and over 15 years, as the case. This method is more feasible now that HAs have computerised population registers and can produce a more random sample of the population at risk. The main disadvantage in an urgent investigation by telephone interviews is that this method of control selection involves ‘cold calling’ and may reduce the response rate of controls. Random selection from the HA population is preferable to controls from the same GP list who would have to be regarded as matched controls. Delays can also arise from the need to gain approval of the GP.
6.4.5 **Case nominated community controls.** Each case is asked to nominate friends of similar age who live in the same community and who would be willing to act as controls. This is a quick and convenient method of obtaining controls when they might otherwise be difficult to obtain, and may control for some confounders. This method avoids ‘cold calling’ and overcomes the difficulties of contacting those with numbers not listed in the telephone directory. The main disadvantage is that such controls may be so similar to cases in relation to the exposure of interest that a true association with disease is not found. This is known as overmatching. This is more likely to occur in relation to social activity such as swimming pool exposure or farm visits than in relation to drinking water consumption but there is also a risk of overmatching on water supply.

6.4.6 Reluctance on the part of cases to give names of controls can often be overcome by suggesting that the case contacts potential controls for permission and that the interviewer phones the case again to confirm this before contacting controls.

6.4.7 Ideally, cases and controls should be interviewed in a standardised way, using the same interview method by a limited number of experienced interviewers as soon as possible after the identification of the outbreak.

6.5 **Sample size calculations.**

6.5.1 The main aim of the study will be to test for an association between the risk of becoming infected and a range of risk factors such as ‘household pets’, ‘farm visits’ or ‘tap-water consumption’. These associations are generally measured by odds ratios (ORs). The size of the study will determine how much power it has for detecting odds ratios of different magnitudes.

6.5.2 Since the size of the outbreak will not be easy to predict the number of cases available for the study will be uncertain. However, it is wise to calculate the likely power of the study for differing numbers of cases covering the numbers likely to be available.

6.5.3 Small outbreaks where the number of cases are in the range 10 to 50 will only have sufficient power to detect large ORs, but studies of this size will still give estimates of the effects of the various risk factors and can help identify which risk factors are and are not likely to be associated with the infection.

6.5.4 Unless there are many cases they should all, as far as is possible, be included in the study, but the number of controls is a matter of choice. Where the number of cases is small recruiting up to three controls for each case will increase the power of the study. Conversely if there are many cases sufficient power may be achieved with more cases than controls up to a ratio of three to one. Increasing either ratio to more than three to one is generally not very useful because the power of a study is largely limited by the size of the smaller group.

6.5.5 In general it is sensible to start recruiting two or three controls per case and review this strategy as the outbreak progresses.

6.5.6 Finally for a given study size, the power to detect a dose-response relationship between the risk of infection and increasing tap-water consumption is somewhat greater than that for detecting an equivalent
association with an all or nothing exposure. Thus quantitative data on tap-water consumption (and on other exposures where dose response can be readily measured) should be collected in the questionnaire to ensure that this type of analysis is possible. The volume estimation may be inaccurate and some guidance in the questionnaire is recommended (for example, one glass is equivalent to a third of a pint or about 200 millilitres). See Sub-appendix A4.5, Q26.

6.6 Data collection

6.6.1 Questionnaire design. Questions should be short, clear and designed to obtain unambiguous answers that are free from bias. Details of illness in the case and household members, of recent travel abroad, and of all likely known risk factors should be included. Postcodes are important for mapping to water supply zones.

6.6.2 For investigation of outbreaks of cryptosporidiosis suspected of being related to drinking water contamination the same questionnaire (for example, see Sub-appendix A4.5) should be used to interview all cases and controls. Additional questions generated by local circumstances or the preliminary interviews should be included. Local points of relevance e.g. names of swimming pool, farms open to public should be added where appropriate. Questions should be worded so that they are completely unambiguous. Optional responses must be exhaustive and mutually exclusive. ‘Not applicable’, negative, and ‘Don’t know’ responses should be recorded in a distinguishable way. Apart from ‘skip sections’, no answer space should be left blank. It may be helpful to consult with regional/national epidemiologists and a statistician before the questionnaire is finalised.

6.6.3 Questionnaire administration. Questionnaires can be administered in face to face interviews, by telephone or by post. Face to face interview is the best way of administering a questionnaire and is recommended in preliminary interviews but is frequently not possible for use in the main case control study because of resource constraints. Telephone administered questionnaires are likely to be the best available option unless telephone ownership in the study population is unusually low. The interviewer can ensure the questions are understood and answered correctly, but does not have to travel to find the person. Evening or weekend calls are more likely to produce an unbiased group of controls than daytime calls. Considerable efforts should be made to ensure that all questions are answered so that the interaction between the effects of several exposures can be thoroughly investigated. Clear interviewing instructions must be given for both face to face and telephone interviewers (Sub-appendix A4.3) to increase the reliability of the data collected.

6.6.4 Questionnaires sent by post are less likely to be completed correctly and fully than those completed by interview and have longer return times. Postal questionnaires should be considered if the other suggested methods are not practicable, or if large numbers of people have to be contacted and there is no urgency.

6.6.5 Data entry. Data should be entered into a computerised database. Direct input of interview data via a data entry screen representing the questionnaire is the best way to avoid data transcription errors. Accurate data entry should where possible be facilitated by double entry comparison together with range and consistency checks at entry.
A4.7 Statistical analysis of case-control study data.

7.1 Introduction

7.1.1 The analysis must produce and the results be presented to give:

(i) a clear summary of the data available in terms of the age and sex distributions of cases and controls and the symptomatology, microbiology and outcomes of the cases;

(ii) a set of simple estimations and tests of the associations with infection for all the risk factors considered individually as single risk variables; and

(iii) a multivariable analysis assessing the affects of various risk factors considered together to allow for the possibility of confounding.

7.1.2 The magnitudes of the associations found should be indicated by ORs or 95% confidence limits. The actual \( p \) values from tests of ORs against 1.0 should also be given to indicate the strength of the evidence for a genuine association. Formal testing of primary hypotheses stipulated in the protocol (for example, there is no dose-response relationship between the rate of infection and the consumption of unboiled tap water) is usually made at the 5% significance level. If multiple variables are being tested, \( p \) values should be treated with caution.

7.1.3 If a matched case-control study design has been used the analyses differ slightly from those described below, in that the case-control pairs must be kept as a unit throughout and the important information is whether and how the members of each pair differ in their exposures. Strong positive associations will lead to a large proportion of pairs with the cases exposed and the controls not exposed. A full discussion of the analysis of matched and unmatched case-control data is given in Schlesselman (1982), and Tillet (1986) gives a good overall summary. The analyses described below assume that the controls were not formally matched with specific cases, for example in HA selected controls.

7.2 Single risk variable analysis

7.2.1. This should consist of a sequence of analyses estimating the associations between illness and exposure for each exposure considered individually, including investigations of dose-response relationships and simple stratified analyses to assess and adjust for confounding between two exposure variables.

7.2.2. These analyses should be performed by generating two-by-two tables summarising the numbers of cases and controls exposed and not exposed to each of the risk factors. The tables should be used to obtain ORs (or relative risks for cohort studies) to estimate the magnitude of associations with 95% confidence limits and \( p \) values to indicate the strength of evidence for such associations. Exposure variables with \( k \) categories (\( k \) being the number of categories) where \( k \geq 2 \), e.g. types of occupation or categories representing increasing exposure or dosage, will require \( k \) by 2 tables and odds ratios will need to be obtained for each category against one chosen as a reference category. This is usually the first in an ordered sequence of categories.
7.2.3. The analysis should include $\chi^2$ tests of dose-response trends in such k x 2 tables, where the categories represent a steadily increasing exposure such as increasing consumption of tap water. These can readily be performed with statistical software using a method originally due to Armitage (1955) for testing linear trends in a sequence of proportions. In this case the proportions are, for each dosage category, (no. of cases)/(no. of cases + no. of controls). Note that this must not be considered or presented as independent information to that obtained from a test of association with the same risk factor dichotomised, for example, tap water consumption into ‘no tap-water’ and ‘any tap-water’ unless the zero consumption category is omitted from the trend analysis. In the single variable analysis the latter approach will generally be the most appropriate.

7.2.4. It may be possible to demonstrate an association between illness and another factor such as age or area of residence. In that case it will be useful to investigate the association between illness and tap water consumption separately in two or more strata representing different age groups or areas. The associations if found would then be adjusted for any confounding due to age group or area effects.

7.2.5. Such a stratified analysis can be particularly useful to assess how dose-response trends with tap water consumption differ between individuals living in areas with water supplies which differ in the proportions of water they receive from a suspect source.

7.3 Multivariable logistic regression analysis

7.3.1 This method of analysis provides a comprehensive way of assessing associations with a variety of exposures and of estimating and testing dose-response relationships allowing for the possibility of the effects of one factor being confounded with another. To ensure confounding of any sort does not bias the results this type of analysis is essential. However, since it requires complete questionnaire responses for all risk factors included in each regression it may not always be possible to use all the relevant data in estimating the effect of an individual risk factor. In addition there are rarely more than one or two risk factors in acute outbreaks of infectious disease. For these reasons it is generally best, as long as their direction and magnitude are confirmed by the regression results and not influenced by confounding, to consider the single variable results, which use the maximum information on that variables association with disease, as the most reliable. Logistic regression can be performed using the most professional statistical software including the public domain software EPI-Info which was designed for outbreak investigations with an additional logistic regression module. Both are available from a World Health Organization Internet site. See References.

7.3.2. The logistic regression analysis should endeavour to include all the risk factors where the single variable result was insufficient to exclude them as associated with illness. Since studies may include many potential risk factors and questionnaire responses are often incomplete, some strategy is needed for restricting which risk factors are selected. One such strategy which has proved effective is to omit potential risk variables with odds ratios from the single variable analysis with $p > 0.2$ excepting, at this stage, any risk factors known to have been associated with this infection in the past if a substantial proportion of the cases have been exposed. The results of the logistic regression including the selected factors can then be used to restrict the selection further by identifying, for omission, those
factors with no evidence of an association once other factors were taken into account and the process repeated. In this sequence factors with any substantial proportion of responses missing should be omitted as soon as possible unless they show strong evidence of an association. If this is not done the analysis will suggest conclusions based only on part of the data. If this is not a substantial part of the data set, ie if it is not at least 70% of both the cases and controls the results cannot be considered reliable. The resulting model can then be used to identify those associations that are reduced to insignificance when other factors are taken into account and those that persist.

7.3.3. As stated above, because the logistic regression is generally unable to use all of the data, the final conclusions will be better based on the single variable odds ratios, given that they are supported by the logistic regression.

7.3.4. In general the conclusions will convey whether or not there was a dose-response relationship between tap water consumption and infection rates, what the evidence was for associations between infection rates and other factors and what, based on the analytic study results, was the most likely vehicle by which the infection was spread.

A4.8 Report writing

8.1 A preliminary report should be written once an outbreak is recognised, ideally within 24 hours. Follow-up reports should be written as required. The final report should include the study protocol, questionnaire and examples of press releases and letters used in the study for the benefit of other investigators. Report writing is the responsibility of the OCT. However, the epidemiology study team should prepare the report of the epidemiological study. A copy of the report should be sent to regional/national epidemiological centres to ensure that information on outbreaks is collated and used to inform future practice.

Useful references


CDC Atlanta Epi Info software download site: http://www.cdc.gov/epo/epi/software.htm


