IDRC
CRDI
CIID

CANADA

EPIDEMIOLOGY AND STATISTICS IN DIARRHOEA RESEARCH

N.J.D. NAGELKERKE,

F. MANJI, AND

J. MUTTUNGA

The International Development Research Centre is a public corporation created by the Parliament of Canada in 1970 to support research designed to adapt science and technology to the needs of developing countries. The Centre's activity is concentrated in six sectors: agriculture, food and nutrition sciences; health sciences; information sciences; social sciences; earth and engineering sciences; and communications. IDRC is financed solely by the Parliament of Canada; its policies, however, are set by an international Board of Governors. The Centre's headquarters are in Ottawa, Canada. Regional offices are located in Africa, Asia, Latin America, and the Middle East.

Le Centre de recherches pour le développement international, société publique créée en 1970 par une loi du Parlement canadien, a pour mission d'appuyer des recherches visant à adapter la science et la technologie aux besoins des pays en développement; il concentre son activité dans six secteurs : agriculture, alimentation et nutrition; information; santé; sciences sociales; sciences de la terre et du génie et communications. Le CRDI est financé entièrement par le Parlement canadien, mais c'est un Conseil des gouverneurs international qui en détermine l'orientation et les politiques. Établi à Ottawa (Canada), il a des bureaux régionaux en Afrique, en Asie, en Amérique latine et au Moyen-Orient.

El Centro Internacional de Investigaciones para el Desarrollo es una corporación pública creada en 1970 por el Parlamento de Canadá con el objeto de apoyar la investigación destinada a adaptar la ciencia y la tecnología a las necesidades de los países en desarrollo. Su actividad se concentra en seis sectores: ciencias agrícolas, alimentos y nutrición; ciencias de la salud; ciencias de la información; ciencias sociales; ciencias de la tierra e ingeniería; y comunicaciones. El Centro es financiado exclusivamente por el Parlamento de Canadá; sin embargo, sus políticas son trazadas por un Consejo de Gobernadores de carácter internacional. La sede del Centro está en Ottawa, Canadá, y sus oficinas regionales en América Latina, Africa, Asia y el Medio Oriente.

This series includes meeting documents, internal reports, and preliminary technical documents that may later form the basis of a formal publication. A Manuscript Report is given a small distribution to a highly specialized audience.

La présente série est réservée aux documents issus de colloques, aux rapports internes et aux documents techniques susceptibles d'être publiés plus tard dans une série de publications plus soignées. D'un tirage restreint, le rapport manuscrit est destiné à un public très spécialisé.

Esta serie incluye ponencias de reuniones, informes internos y documentos técnicos que pueden posteriormente conformar la base de una publicación formal. El informe recibe distribución limitada entre una audiencia altamente especializada.

IDRC-MR246e
January 1990

EPIDEMIOLOGY AND STATISTICS IN DIARRHOEA RESEARCH

by

N.J.D. Nagelkerke
F. Manji
J. Muttunga

Kenya Medical Research Institute
Medical Research Centre
P.O. Box 20752 Nairobi (Kenya)

Acknowledgement

The authors wish to thank the International Development Research Centre (IDRC) for their generous financial support in the production of this document.

CHAPTER 1

INTRODUCTION

Diarrhoea is a major cause of morbidity and mortality among infants and young children in third world countries where diarrhoea-causing pathogens are endemic. Death as a result of diarrhoea is usually the result of rapid dehydration due to the loss of electrolytes and water. The case fatality rate of diarrhoea is usually low (less than 1%), depending on the type of diarrhoea, the age of the child, and other factors. However, the relatively high frequency of diarrhoea, typically 3-10 episodes per child per year, still makes it a major cause of infant mortality. In addition to its immediate threat to life, diarrhoea is an important cause of malnutrition: not only can it result in malabsorption of essential nutrients, it may also cause loss of appetite (anorexia).

Since the nutritional status of many children in the third world is marginal at most, the extra damage done by diarrhoea constitutes an important threat to their development. It may make such children both more susceptible and more sensitive to the effects of other infections. A large number of public health programmes have been initiated in many parts of the third world in an attempt to reduce infant mortality due to diarrhoea. In many of these programmes, there have been considerable difficulties in either developing intervention strategies whose effects can be meaningfully assessed, or in interpreting the results of long-term studies. In principle, the difficulties have arisen because of the rather complex nature of the etiology of diarrhoea, and because there are usually a considerable number of determinants involved.

Diarrhoea is not a disease. Rather, it is one of the major symptoms of a large variety of infections caused by many different organisms, either bacterial (e.g. cholera vibrio, shigella, salmonella, campilobacter), protozoal (e.g. giardia, entamoeba hystolitica), or viral (e.g. rotavirus). In many cases it is impossible to demonstrate a single causal organism. Often, the etiology of diarrhoea is uncertain because it is found to be associated with infections with a variety of organisms.

Most diarrhoea-causing organisms enter the body through the mouth as a result of the consumption of contaminated food, water and other drinks, or as a result of putting contaminated hands or objects into the mouth.

However, not every infection with one of the many potentially pathogenic organisms results in diarrhoea. This may be due to a number of reasons, the most important of which is probably the general or acquired specific immunity of the host. Much may depend on both the type of infecting organisms as well as the particular strains involved. The capacity of the host to mount an effective immune response seems, in part, to be "dose"

dependent (i.e. dependent upon the number of infecting organisms). The number (or dose) of pathogenic organisms infecting the host may itself be dependent on a number of factors including, for example, the degree of gastric acidity of the individual. This is further complicated by the fact that the infectious dose (i.e. the number of organisms that will be required to cause symptoms) will be higher in those who have already acquired some degree of immunity to the specific pathogen. The effectiveness of both local and humeral 'defence' systems to deal with or cope with a given infection is to a great extent determined by the child's nutritional status and his/her general well-being, and may be considerably weakened by the concomitant presence of other diseases (e.g. measles or malaria).

To reduce its importance as a public health problem and as a cause of death, efforts can be made either to prevent diarrhoea (i.e. reduce its incidence), to cure it when it occurs, or to minimize or alleviate its deleterious consequences (e.g. dehydration, malnutrition). Each of these approaches poses considerable problems when, as is usually the case, no single aetiological factor or determinant can be considered independently or in isolation.

In principle, diarrhoea can be prevented either by avoidance of infection (e.g. through improved hygiene, use of improved food preparation and storage methods and facilities, use of clean water etc.), or by increasing the capacity of the host to defend itself (e.g. through vaccines, by prevent and curing concomitant illnesses, improving nutritional status, etc.). A diarrhoeal episode may sometimes be cured by the use of various drugs and antibiotics. Although popular in some countries, these methods may have deleterious side-effects. For example, the use of drugs that are effectively 'constipants' may result in the retention of pathogens in the gut, whereas antibiotics (especially the broad-spectrum ones) may have the effect, by destroying the normal commensal flora of the gut, of permitting infection by pathogenic organisms. To avert the harmful consequences of diarrhoea, especially death due to acute dehydration, oral rehydration therapy (ORT) can be given by the administration of an oral rehydration solution (ORS) of glucose and electrolytes (salts). If this fails, intravenous rehydration may be necessary.

It is, as yet, not entirely clear which particular mix of strategies is optimal for the control and prevention of diarrhoea. The pathogenic organisms implicated at different times and in different communities vary; modes of transmission may differ; traditions of hygiene are largely socially and economically determined, and various cultural traditions may play important roles in the aetiology; dietary habits and the availability of certain foods are also dependent upon the social, economic and cultural context. Diarrhoea research has, therefore, to be targeted at, and designed for, specific populations at specific periods of time. On each occasion we need to have answers to a series of questions: how serious is the problem of diarrhoea in the given community? What is the ecology of the causative pathogenic organisms? What is the best intervention strategy that is likely to be successful, feasible, and practical in a

given context? And, how to assess the impact of any given strategy?

Clearly a good knowledge of subjects such as microbiology, nutrition, behavioral science, etc., as well as a knowledge of the experiences of other similar programmes conducted elsewhere, is a prerequisite for successful interventions. Of central importance in such studies is a clear understanding of how the principles of epidemiology and statistics can be applied in the particular study. All too often, however, studies are embarked upon without proper knowledge of, or support in, precisely these fields. Often, statistics (except in so far as it may be of relevance for determining sample size or the drawing of a sample) is considered as a necessary evil with which one will inevitably have to deal when one comes to data analysis. What is frequently forgotten is that our capacity to meaningfully interpret data through the use of statistical tools is determined not only by the quality and nature of the data itself (the old adage "garbage in, garbage out" is sufficiently well-known not to require further comment here), but also, and probably more importantly, by the quality of the design of the study or experiment. Glancing at the literature, it is apparent that although a wide scope of statistical techniques are used in the analyses of diarrhoea data, often the capabilities, scope, limitations, and pitfalls of both the study designs and of the subsequent analytical tools used to interpret the data are not always fully appreciated. Furthermore, research workers may not always be aware of the scope of statistical techniques that exist - or which can be developed - to deal with the specific problems posed by their studies.

There are numerous excellent texts on statistics and epidemiology. However most have been written for use in a wide variety of fields and, consequently, research or public health workers are often faced with a difficulty of deciding which of the many types of techniques may be most appropriate for their studies. Research workers in the field of diarrhoea are usually faced with a number of specific problems - for example, both incidence and duration (of diarrhoeal episodes) are variable and may be influenced by many determinants. What are the types of study designs that are appropriate for research on diarrhoea, and what are the advantages or limitations of the statistical tools which can be applied to data derived from such studies? This monograph considers some of these problems in a brief (and, we hope, readable) manner with the emphasis placed on understanding the issues involved, rather than providing merely "recipes". The book is not designed as a substitute for the many essential texts on statistics and epidemiology that are available, and we have assumed throughout that the reader already has some basic knowledge of statistics, computer science and epidemiology. We cannot emphasize too strongly the need that at least one member of a research team possesses such knowledge and is able to fully understand the contents of this book. Too many projects fail because the study is poorly designed, the wrong data is collected and nobody knows how to analyze the data or interpret the results.

At the end of the book we provide a brief suggested bibliography which could (and perhaps should) be used in conjunction with this text.

Implicit throughout the text is the assumption that the kind of diarrhoea with which we are dealing is endemic. This assumption of endemicity permits, to a large extent, the use of epidemiological methods developed for non-infectious diseases. Since the "disease is everywhere" host specific factors will determine who gets diarrhoea. Contact with infected individuals is of lesser importance. Some kinds of diarrhoea, most notably cholera, occur in epidemics, i.e. they display a marked spaciodynamic character. The analysis of such epidemics requires special methods that take this space-time character into account and will not be discussed in this document.

There exists an extensive literature on (acute) diarrhoeal diseases, an overview of which goes far beyond the scope of this monograph. For those interested in such an overview we recommend the bibliographies which are regularly compiled by the World Health Organization.

CHAPTER 2

SOME NOTES ON STUDY DESIGN

Each epidemiological study is meant to answer certain specific questions, e.g. the testing of some well defined hypotheses about the causation of diarrhoea in a given population.

Each study has to be designed in such a way that it can be expected to answer the questions it is supposed to answer. The right individuals have to be sampled, the size of the study has to be sufficiently large, the correct measurements have to be made, etc. Yet, the study has to be feasible with only a limited amount of money and within a limited period of time. The design of the study is therefore of paramount importance and requires careful attention by both subject experts (e.g. experts in microbiology, nutrition) as well as biostatisticians or epidemiologists. Draft protocols have to be written, revised, and rewritten until finally a design is developed which is "optimal". Each study design is in some sense unique. Yet, some choices-common to all investigations - always need to be made. For instance, will the study be observational or experimental? If the study is observational, will the individuals be sampled before they have diarrhoea, or will cases already with diarrhoea (and controls without) be included?

Each of these choices categorizes the study in a specific way. Each of these categories (choices) will have certain possibilities, advantages, pitfalls and inferential implications which should be understood in order to make a rational choice about the design of the study.

Types of Studies

Epidemiological studies can be categorized in several ways. For instance:

i a) Observational studies in which the investigator observes the "natural" course of events and draws conclusions as to the laws and mechanisms governing them. The investigator interferes as little as possible with this "natural" course of events.

and,

i b) Intervention (or experimental) studies in which the investigator applies certain interventions (e.g. a drug, a vaccine, ORS therapy, a borehole in a village) to so-called "experimental units" (e.g. a child given a drug, a household provided with a latrine, a village supplied with a borehole) in order to study the effect of the intervention on the "response variable" (e.g. mortality, morbidity, number of stools).

Alternatively, epidemiological studies may be subdivided into:

ii a) Retrospective studies in which all the events of interest (exposures, diseases etc) have already taken place and the information needs "only" to be collected. This can usually be done on a "cross-sectional" basis, i.e. the study takes place only at one point in time. Inevitably, retrospective studies are always observational. If disease status is the sampling criterion (e.g. we collect information about 100 individuals with and 100 individuals without the disease) then the study is called a "case-control" study.

and,

ii b) Prospective studies in which some of the events (e.g. diseases), about which information is to be collected, have yet to occur after selection by the investigator of the observational units (e.g. individuals).

A third possible subdivision of research designs is:

iii a) Descriptive studies, the purpose of which is the mere gathering of facts. This is the kind of study usually carried out by a census bureau. In the context of diarrhoea research one could think of studies estimating:

1) The prevalence rate of diarrhoea, i.e. the proportion of a given population (e.g. children under 5 years in rural Kenya) suffering from diarrhoea at any given moment.

2) The incidence rate of diarrhoea, i.e. the proportion of a given population getting diarrhoea within a specified unit of time (e.g. a day).

3) The case fatality rate of diarrhoea, i.e. the proportion of diarrhoea cases in a given population with a fatal outcome.

Often such estimates are not, in themselves, very meaningful and a breakdown into subpopulations (regions, age, sex, etc) may be more informative, thereby providing subpopulation-specific incidence, prevalence or case fatality rates.

If diarrhoea follows a seasonal pattern, a breakdown into seasons may be required. Of course, this requires repeated visits to the area of interest.

A breakdown by type of diarrhoea (causative organism, duration, with/without concomitant disease etc.) could also be made. A careful presentation of the results in the form of tables, graphs, histograms, pie charts etc., makes these studies much more accessible for interpretation.

and,

iii b) Analytical studies. Although the distinction between descriptive and analytical (epidemiological) studies is sometimes somewhat blurred (and rightly so), the objective of an analytical study differs from a descriptive study: in the latter one wants only to describe disease patterns, whereas in the former one

wishes to explain them. The key-words in analytical studies are "cause" and "effect".

In the context of diarrhoea studies, one could be interested in the causes or determinants[1] of diarrhoea, or one could be interested in the effects of diarrhoea. Examples of causes or determinants of diarrhoea are:

- The type of water supply (river, borehole, tap etc.)
- The type of food eaten (nutrients, acidity etc.)
- Food preparation and storage facilities
- Vaccination status (e.g. of measles)
- Prevalence of malaria, helminths, Giardia etc.
- Hygienic practices (hand-washing, latrine use etc.)
- Socioeconomic status (income, land ownership etc.)
- Soil type
- Communication with people outside the area
- Altitude and other geographical variables
- Weaning practices (time, kind of food etc.)
- Presence of cattle on a compound or animals in the house
- Knowledge about diarrhoea (causes, prevention etc.)
- Number of people in a house or village (population density)

As the effect of diarrhoea, one could think of mortality (caused by severe dehydration), loss of weight, malnutrition (caused by malabsorption of nutrients and anorexia, i.e. loss of appetite during diarrhoea). Certain anthropometric measures, notably weight for age (indicative of malnutrition in general), height for age (indicative of chronic malnutrition, stunting) and weight for height (indicative of acute malnutrition, wasting) could be used as a measure of malnourishment. Although the use of absolute norms and standards for these measures (e.g. the NCHS standards) is disputable because of genetic variation between peoples, studies of the variation of these measures within a population and their association with other variables (diarrhoea) are valuable. It should be borne in mind that although we can think of such parameters as measures of "effect", they might also be considered as determinants (e.g. those who are judged to be malnourished may themselves be more susceptible or sensitive to developing diarrhoea). For all these anthropometric variables it is crucial to decide when one wants to measure them. Weight measured shortly after a diarrhoea episode is an indication of the immediate effect of an episode, whereas weight taken at an arbitrary point in time (or at a given age) is a measure of accumulated life-time experiences (not necessarily of diarrhoea alone).

A fourth possible subdivision of research designs, closely

---

[1] Determinants are factors which determine an individuals probability of getting or having a disease. They are not themselves necessarily the cause of the disease. In the case of diarrhoea (which is a symptom rather than a disease) the distinction between etiological factors and determinants may be more subtle.

related to the division between retrospective and prospective studies, is the division into:

iv a) Cross-sectional studies. Prevalence and, in principle, incidence can be estimated by means of a (single) cross-sectional survey. To estimate prevalence, selected (sampled) individuals are asked whether they suffered from diarrhoea on the day of (or the day before) questioning.

To estimate incidence in a cross sectional survey, several methods are available.

First, one could enquire whether any current diarrhoea is a new case of diarrhoea, i.e. one that started on the same day or within the previous 24 hours.

Secondly, one could enquire as to the duration of the episode to date. From this information it is possible to reconstruct the distribution of the duration of diarrhoea, and hence the incidence, since

$$\text{average duration of diarrhoea} = \frac{\text{prevalence rate}}{\text{incidence rate}}$$

This holds only in "steady state" situations, an assumption that is reasonable in endemic childhood diarrhoea.

Thirdly, one could identify prevalent cases, ask how long the diarrhoea has lasted (as in the previous method) and follow those cases up till the end of the present episode of diarrhoea. With this method one obtains a distribution of the duration of prevalent cases. It should be noted that this is not identical to the distribution of incident cases since long episodes of diarrhoea will have a disproportionate chance of being included in the sample.

Mathematical expressions for calculating the distribution of incident cases from observations made with the second and third method are given in Appendix 1. It should be stressed, however, that since both of these methods require the recollection of the duration of diarrhoea, they work no better than an individual's capacity to remember when his or her (or the child's) episode of diarrhoea started, which may be very poor.

Analytical studies to establish the "causes" of diarrhoea can be carried out by ascertaining the relevant information (socio-economic status, type of latrine etc.) from the individuals or their household members. Some of these causes may have taken place in the past, that is they are obtained "retrospectively" (when was the child weaned? Did you eat meat the day before yesterday? etc), and memories are notoriously unreliable. Recall of events may also be influenced by having had diarrhoea. This may result in "bias" ,i.e. a systematic distortion of the truth. More information can be obtained if some kind of follow-up of the sampled individuals is carried out.
and,

iv b) Follow-up studies. Two types of follow-up can be distinguished:

1) Cohort follow-up. A cohort is a group of people "identified by a common experience at a specific point of time". For instance the 1985 birth cohort consists of all people born (the common experience) in 1985 (the point of time). Similarly, a cohort can be constructed out of all individuals sampled at a specific point in time. Follow-up of a cohort may be difficult if there is a lot of emigration from the selected population as this could cause serious logistic problems (tracing individuals, transport etc.). The advantage of a cohort is that the selection of certain information (e.g. sex, date of birth) about the individuals in the study needs to be collected only once.

2) Population follow-up. All individuals in a population meeting certain criteria are included in the study. Individuals moving out of the population (village,area, community, cluster) also leave the study, whereas people moving in enter the study. Since it is often not practical to repeat certain parts of the study (KAP, socio-economic questionnaires) this information will not be available for people entering the study later. Certain types of analyses can then only be performed on the data from those individuals who were in the population at the start and who did not move out. Note that this may result in bias.

Follow-up will usually take place by means of surveillance, i.e. field workers visiting the individuals at regular time intervals.

Despite "rigid" definitions and criteria and well designed questionnaires (Appendix 2), certain field-workers may find much more diarrhoea than others for a number of reasons, e.g. because they are more insistent. Rotation of field-workers (if possible), calibration, regular monitoring etc, is therefore advisable in order to avoid, subsequently, spurious relationships being found between diarrhoea and variables operating within a particular field-worker's district.

In diarrhoea research only prospective follow-up studies (i.e. ones in which the follow-up is carried out during the study) are possible. In other contexts, so-called historical follow-up (e.g. historical cohort) studies are sometimes possible. In such instances, the sample (e.g. all victims of some industrial poisoning accident in 1954) is identified and followed-up from records.

Each research design can be categorized according to any of the four (and other) subdivisions of designs we have discussed. Two kinds of design, the intervention study (or the prospective, cohort follow-up, analytical intervention study) and the case-control study (which is retrospective, observational, analytical, cross-sectional) merit special attention because of their wide use. They are therefore discussed in more detail in Chapters 3 and 4, respectively.

CHAPTER 3

INTERVENTION STUDIES

A major purpose of analytical observational studies is to find out which possible interventions could be useful and to target these interventions to groups at risk (which have to be identified by the study). For instance if an association is found between a certain hygienic practice and diarrhoea one could plan an intervention consisting of health education to reduce the incriminated practice. However, even if in an observational study an association has been established beyond statistical doubt (i.e. one which cannot be ascribed to chance alone), this by no means implies that a campaign to reduce this practice would necessarily be successful.

Reasons for this are:

1) The relationship is spurious. The hygienic practice is merely a symptom of the real cause (e.g. hand-washing as a symptom of water availability). Changing the symptom alone would not bring about the desired effect since the real cause (e.g. water availability) exerts its effect not through the hygienic practice. The apparent relationship has been caused by a "confounding" factor.

2) Although the relationship is real and the "efficacy" of the intervention exists, the "effectiveness" of the therapy is disappointing because people find the intervention cumbersome (e.g. getting clean water from a distant source), impractical, or against their beliefs or traditions. Large scale intervention should therefore always (if possible) be preceded by an intervention study to try to assess the effectiveness of the intended intervention. This is not always possible if the required sample size is prohibitive (remember that if the interventions are implemented on a community level then the experimental unit for which the sample size has to be calculated is the community-something that is often overlooked). If a full scale intervention study cannot be carried out, it may still be worthwhile to carry out a feasibility study to see whether the intended intervention can be implemented at all (do people come for vaccinations? Do people use ORS when it is made available?).

To test an intervention one should almost always include a control group for comparison. This is not required if the course of the disease is precisely known (e.g. rabies which invariably results in death), but this is an exception. One should always make sure that the intervention and control units are comparable. One way to achieve this is by means of "randomization", i.e. the experimental units are assigned to experimental and control interventions by some random mechanism (e.g. coin tossing). The very possibility of randomization makes experimental studies superior to non-experimental ones in which the possibility of confounding can never be excluded. Therefore if an experiment

can be designed to answer a question (this is more frequently possible than one expects) it should be done.

One should also be aware of the possibility that the treatment group "infects" the control group. This might happen with health education (people spreading the message) but also with vaccinations (if a certain percentage of a population has been vaccinated, the rest of the population may also be protected because they are less exposed to infective individuals).

Several questions amenable to intervention studies are:

1) Is health education effective in reducing certain hygienic practices? Does it lead to a reduction in incidence of diarrhoea?

There are difficulties associated with the evaluation of health education, namely that it is difficult to individualize. It may be impossible to randomize children within a community into groups receiving different health messages, let alone to randomize them into a group with and a group without health education. An alternative would be to give health education to mothers of children (with diarrhoea) visiting a health centre or an outpatient department of a hospital. In such a context the individualizing of health education is easier. However, follow up may constitute a major problem since mothers may live scattered over a large area.

Since health education cannot be given "blind" the results of the study are sensitive to bias. Investigators aware of the group to which a mother or child belongs may be inclined to observe better hygienic practices in the "treated" group than in the control group even if there are no real differences. Similarly, mothers in the "treated" group may tend to report better hygienic practices, or behave more hygienically (only) when observed because they are more aware of what is expected from them. In addition the effect of the health education message may be obscured by a "placebo" or "Hawthorne" effect: health behaviour has improved (temporarily) not because of the contents of the health message but because of all the attention paid by and interaction with the investigators and field-workers.

2) Which type of ORS is the most effective?

One might want to compare glucose based ORS with cereal based ORS (cereals being more widely available than glucose). Such a study can be carried out in many different ways. The study can be carried out in a community, in an outpatient department or in a hospital. The kind of patients one encounters differs among these institutions. It may well be that cereal based ORS is more effective in outpatients but less effective in inpatients (or vice versa) due to differences in the pathogens involved or in the extent of damage of the intestine. One has to be careful therefore about extrapolating results from one group of patients to another.

An important question is the choice of the response variable to be used for comparison. The most relevant one is mortality. However, mortality is very low (diarrhoea is only a major cause of death because it is so widespread) and the effective comparison of two types of ORS based on mortality would require many thousands of patients. Hardly feasible anywhere. However, it is possible to make a comparison based on response variables such as electrolyte balance, total stool output, water absorption, urine output, nutrient absorption, weight gain and similar variables which are supposed to measure the physiological effects of ORS. With these variables, effective comparison of two types of ORS can be done with only a few dozen patients per group. For the effective measurement of these variables, patients need almost invariably to be hospitalized.

3) Is a specific antibiotic (or other medicine) indicated for the treatment of a certain kind of diarrhoea (e.g. one caused by shigella)?

Here, major response variables seem to be (apart from the difficult case fatality rate) duration of diarrhoea, duration of symptoms (e.g. vomiting, blood excretion, fever, pain, anorexia) and stool frequency.

4) Do certain pills, vaccinations (e.g. rotavirus vaccines) or other treatments (e.g. an anti helminths or anti Giardia treatment) reduce the incidence of diarrhoea? (i.e. prophylaxis studies).

Since the blinding of pills and vaccines is straightforward, neat double blind randomized studies are perfectly possible. The evaluation of the results has to be carefully considered. If the prophylaxis is supposed to prevent only a specific type of diarrhoea (e.g. caused by rotavirus) which constitutes a minority of all diarrhoea cases, the comparison of total diarrhoea incidence or prevalence will be inefficient. However, stool collection and testing for the specific pathogen may be cumbersome and expensive. Sometimes it helps to distinguish between types of diarrhoea by clinical features (duration, blood in stool etc.) if these features are sufficiently sensitive and specific for that type of diarrhoea.

As in all intervention studies, the selection of eligible individuals merits careful consideration. If the future use of the prophylaxis is mainly for international travellers, it should not be tried out in third world infants (or vice versa!). Similarly if certain high risk groups are the prospective beneficiaries of the drug, these groups should, in so far as it is possible, be used for testing the drug.

Anti-diarrhoea drugs for tourists and other international travellers can be tested by soliciting their cooperation in vaccination clinics where most tourists get their "shots" before leaving their country. Special diaries can be given to them to record stool frequency, consistency, and other abdominal complaints.

It is important to decide upon the inclusion and exclusion criteria for an intervention study (clinical trial). Two considerations have to be taken into account in deciding these:

a) What is the group for which the intervention is ultimately targeted? Those patients should be included and others excluded.

b) Patients who are undesirable because they give logistic or measurement problems, e.g. they have complicating concomitant disease or take additional medicines, should be excluded.

These rules are contradictory. Future diarrhoea patients with concomitant disease will be given a new treatment if it has been demonstrated to be effective, so they should be included. On the other hand if the concomitant disease requires medical care which interferes with the smooth running of the study, they should be excluded. The decision should, therefore, depend on the circumstances. Improper inclusion or exclusion criteria, or inclusion and exclusion criteria which are not strictly adhered to, do not invalidate the study in the sense that no bias is introduced (the so-called internal validity) although the generalizability (the so-called external validity) may become somewhat difficult. Improper randomization (e.g. a randomization scheme known to the person who selects patients) however does introduce bias and thereby invalidates the study.

In order to fully adhere to the randomization scheme (methods of randomization are discussed in Appendix 3) one should always decide first whether a patient is to be admitted to the study (i.e. meets all inclusion criteria) and then to randomize.

An essential (although non-medical) inclusion criterion is the giving of "informed consent" by the patient or, in the case of a child, by its parent. This is usually given by the signing of a form ("consent") which indicates the purpose of the study, the fact of randomization, the kind of treatments given (e.g. a "new" one and an old one or a "real" medicine and a placebo) and their potential benefits and risks, the kind of treatment and examinations the patients will undergo etc. In addition it should be made clear that the patient is free to withdraw from the study at any given time and that he will then receive the standard treatment. In some developing countries it may be difficult to obtain written informed consent (e.g. due to illiteracy). In those cases verbal consent will need to be obtained.

Before embarking on an intervention study (or any other study) a detailed study protocol should be made. Appendix 4 gives some guidelines for the design of such a protocol. The WHO has recently published a report giving guidelines for planning clinical trials in diarrhoeal diseases which should also be consulted before starting a clinical trial in this field.

CHAPTER 4

CASE-CONTROL STUDIES

The methodology of case-control studies (also called case-referent studies) is widely used to study the mechanisms (risk factors, determinants) leading to "rare" diseases (Schlesselman, 1982; Breslow and Day, 1980; Lilienfeld and Lilienfeld, 1979).

In case-control studies cases of a disease are identified and controls without the disease are sought for those cases. Cases and controls are then compared for the presence or absence of certain potential risk factors.

For instance, diarrhoea patients can be compared to patients suffering from upper respiratory tract infections for their hygienic habits, or lung cancer patients can be compared to healthy individuals for their smoking habits, dietary habits etc. There are some well known difficulties associated with this type of study. Some of these difficulties are typical for any type of observational study. Some of these are, however, specific to case-control studies.

First, suppose that one examines the effect of sweets consumption (which we know does not cause lung cancer) on lung cancer. Since non-smokers people eat more sweets than smokers it will spuriously turn out that the consumption of sweets has a protective effect for lung cancer. The reason for this is that the smoking of cigarettes is associated with lung cancer and is associated with the consumption of sweets. Note, that both these associations hold even if we keep the third variable constant, i.e. smoking is associated with lung cancer even when the consumption of sweets is kept constant, and sweet consumption is (negatively) associated with smoking within both cases and controls. Here, smoking is a confounder for the relationship between sweets and lung cancer.

As another example consider the relationship between the consumption of a certain food F and diarrhoea. If F tends to be eaten in conjunction with another food G which causes diarrhoea (e.g. it is usually contaminated with pathogenic organisms which cause diarrhoea), then a spurious association between F and diarrhoea is found. The other food, or the presence of pathogens on it is the confounder in this case. The following "equation" of tables can illustrate this.

	overall			G eaten			G not eaten	
	F	not F		F	not F		F	not F
case	80	30	=	40	40	+	2	18
control	20	80		20	20		10	70

It is clear that the association between F and diarrhoea disappears after "controlling" or "adjusting" for the confounding variable "G eaten". Technically, such "adjusting" is usually performed by means of the Mantel-Haenszel method. This method is described in most textbooks on epidemiology.

A confounder is a determinant or cause of both the disease and the "risk factor" under consideration. If potential confounders are known, then one can adjust for them by use of special statistical techniques. However, if a confounder has been overlooked then variables (factors) may appear to be risk factors when in fact they are not.

A different problem occurs when the disease causes the risk factor instead of the other way round. For instance, people who suffer from headaches tend to take more aspirin than people without headaches and children with diarrhoea tend to consume more ORS than children with formed stools. The associations are real and causal but it is not acceptable to conclude that aspirin causes headaches or that ORS causes diarrhoea. This type of reverse causality can easily occur in diarrhoea case-control studies. Spoons, bottles etc. can become contaminated with E.Coli because a child in the household has diarrhoea; behaviour of the mother or child caretaker can be modified by the illness of a child; etc

Case control studies are by their very nature both retrospective and observational, which implies that they are sensitive to all those kinds of bias from which observational and retrospective studies are prone. Some of these are (cf. Sackett, 1979):

1) Recollection and reporting bias. The fact of being a case (or control) influences the recollection of certain risk factors in the past. Mothers of children with diarrhoea may recollect having forgotten to wash their hands much better than mothers of control children even though there may not have been any difference at all. Apart from unintentional recollection bias there may also be a conscious one if the risk factor (or even more the association between risk factor and a certain disease) has strong moral connotations. This kind of bias may occur in diarrhoea studies when mothers of children are asked about their (un)hygienic habits. If mothers are aware of the association between hygiene and diarrhoea then mothers of children with diarrhoea are likely to report differently from mothers of children without diarrhoea. This kind of bias may also occur in follow-up studies. In such cases, however, ascertaining of habits and the diarrhoea surveillance can be done at different times (habits are observed before the diarrhoea surveillance). One can try to "mimic" this in case-control studies by postponing the ascertainment of hygienic behaviour until the episode of diarrhoea has been forgotten as far as behaviour is concerned.

2) Detection bias. A factor may be associated with a disease not because it is causally related to the disease but because the factor increases the probability of the disease being detected, i.e. the probability that a diagnosis of the disease is made. For instance, malnutrition may be a reason to enquire with the mother whether the child also has diarrhoea, but not whether the child has a respiratory infection (if that is the control disease). Conversely, having the disease increases or decreases the probability of the factor being detected. This seldom happens with diseases or risk factors which rarely go

undetected. However, this kind of bias can easily occur if the disease is a reason to look for the risk factor because it is already believed to be associated with it. In this way a case control study is a good way to confirm preconceived ideas. In diarrhoea research detection bias kind of bias can easily occur when looking for pathogenic (i.e. diarrhoea causing) organisms in stools. If diarrhoea stimulates the excretion of certain organisms (e.g. certain protozoa) then a spurious association between those organisms and diarrhoea may be found even though these organisms are perfectly harmless.

3) Selection bias. The way in which cases and controls are selected differs with respect to the probability of including a certain factor, even though this factor has nothing to do with the disease. If cases come from a hospital ward and controls from an outpatient department then factors associated with being hospitalized automatically differ between cases and controls (e.g. distance of home from hospital, ability to pay hospitalization, inability to manage the illness at home etc.).

Note, that selection bias is a design error (although sometimes difficult to avoid) whereas confounding is not. Confounding arises from associations in the population from which cases and controls are taken whereas selection bias is introduced by the investigator through the mechanism of selecting cases and controls. If children coming to a health centre with diarrhoea are compared to a control group of, say, schoolchildren, one is likely to find all kinds of differences. For instance, there may be a disproportionate number of children without siblings at a health centre because mothers may be more inclined to take such a child to a health centre than if the child was one of many. For example, if 80% of children have siblings and children without siblings are twice as likely to be taken to the hospital as children with siblings, we will find that one out of three diarrhoea children is without siblings, whereas one out of five of the control children is without siblings, giving an odds ratio (explained further on) of 0.5 for siblings, leading to the false impression that siblings protect against diarrhoea.

4) Berkson's bias. Berkson's bias occurs when the case-control methodology is used to look for associations between diseases using hospitalized patients (cf. Robin et al, 1979; Boyd, 1979).

Suppose one wants to prove an association between diarrhoea (the cases) and inguinal hernia. Therefore all records of patients with a diagnosis of diarrhoea on their medical (hospital) record are selected and the presence or absence of hernia is established. As a control group, all cancer patients from the same hospital are used, and for them the presence or absence of hernia is recorded as well. Patients with other diagnoses (i.e. other than diarrhoea, cancer and hernia) are excluded. Even without any real association between hernia and diarrhoea in the population at large, one is very likely to find (a positive) one due to Berkson's bias (also called Berkson's fallacy or Berkson's paradox).

To understand this, one first has to note that the hospitalization rate of diarrhoea is (very) low and for cancer (very) high. However, given that one is hospitalized for other reasons, the diagnosis of diarrhoea (which is easy to make) will almost always be recorded on a patient's record because it may relevant for nursing. Patients with both hernia and diarrhoea have, in majority, been hospitalized because they had a hernia, whereas this is not true for patients with both hernia and cancer. As a result there will be few patients with diarrhoea only (without hernia), but there will be quite a few with cancer only. Consequently, a positive association between diarrhoea and hernia will be found.

This example demonstrates that one has to be strongly aware that this kind of bias may occur when associations between diarrhoea cases and other diseases are studied in hospitalized patients (or patients from any other health facility). Since the hospitalization rate of diarrhoea is very low, lower than for most other diseases encountered in a hospital (which could serve as controls), one is likely to find positive associations between diarrhoea and other diseases even if there are in reality none.

Berkson's bias can also occur in the study of "interaction" of pathogens. Suppose stools from diarrhoea patients are collected and tested for the presence of a number of potential pathogens. Presence and absence of pathogens is cross-tabulated (for each pair of pathogens, one 2 x 2 table) and relationships are studied in the usual way (X^2 test, odds ratios etc.). It would be tempting to interpret (positive) associations in a biological way e.g. synergistic action or one organism facilitating infection with another one etc. However, it is clear that getting diarrhoea is comparable (as regards Berkson's bias) to being hospitalized. So, if one organism has a low pathogenicity (relative to the other ones) it is likely to be positively associated with other pathogens in the sample without this being the case in the population at large.

5) Misclassification bias. Sometimes errors are made in determining exposure to risk factors or in classifying an individual as a case or a control. Misclassification of cases and controls can occur when either a child without diarrhoea is classified as having diarrhoea (false positive) or a child with diarrhoea is incorrectly classified as a control (false negative). The probability of this occurring not only depends on the quality of data collection but also on the definitions of case and control. If diarrhoea is defined (most common definition) as passing at least three loose stools within a 24 hour period, then misclassification can occur if, for example, the mother overlooked some stools. However, the definition itself enables one to establish diarrhoea very easily, so that if mothers know what to look for, misclassification need not be frequent. If, however, diarrhoea is defined as loose stools not being a symptom of any non-infectious disease, then it is much more difficult to establish whether an individual is a case or not. The more operational a definition is (that is defined in terms of procedures, test outcomes etc.) the less likely a misclassification is. The same holds true for misclassification of risk factors, of course (for instance in classifying a house as clean or dirty).

If the misclassification is non-differential, i.e. misclassification of cases and controls does not depend on exposure to risk factors and misclassification of risk factor exposure does not depend on being a case or control, then the effect is to weaken the relationship between risk factors and diarrhoea. In particular, odds ratios are biased towards unity and correlation coefficients towards zero. As an example consider a case-control study with 100 cases and 100 controls. Suppose that 80 of the cases and 60 of the controls are infected with a parasite (Giardia, say) which is the risk factor of interest. The true odds ratio is 2.67. Now, the parasite is only detected 80% of the time. We then find the parasite in 64% of the cases and 54% of the controls giving an odds ratio of only 1.51. Non-differential misclassification never introduces a relationship where there is none. If the misclassification is differential then anything can happen and relationships between risk factors and diarrhoea may be complete artefacts. For instance, consider again an (non-pathogenic) organism. In reality 60% of cases and 60% of controls are carriers of that organism. However, the organism is excreted in only 50% of all formed stool but in all loose stools. We then end up finding the organism in 30% of the formed stools, but in 60% of all loose stools. The odds ratio of 3.5 for the organism is of course a complete artefact. Note, that misclassification bias is not restricted to case-control studies but may occur in prospective studies as well. More details and some good examples on misclassification bias are given in a WHO document by Cousens et al (1988).

In addition to misclassification of exposure and outcome variables, misclassification of confounding variables may also cause bias. When a confounder is recognized, the relationship between outcome (case or control) and exposure variable (the risk factor) can be adjusted for by multivariate statistical methods (e.g. multiple regression) or by a Mantel-Haenszel procedure. However, if the confounder cannot be measured exactly, e.g. when it is some behavioral or social (socio-economic status) variable, then adjustment will not be completely successful. The "residual" effect of the confounder can be substantial even when the correlation between "real" and "proxy" confounder is rather high (0.7, say).

It is very important to develop a sensitivity to the presence of all types of bias in a study. Since it is impossible to set up an exhaustive theory about bias (studies can differ in too many respects), this is in part a skill which can only be developed with experience. A skeptical analytical attitude is essential.

-*-

Cases and controls may be collected either in a matched (or more generally, stratified) or unmatched manner. The reason for matching is to eliminate certain known risk factors which are not of interest in the study. Although, sometimes, this elimination can be done by means of statistical techniques, this is not always the case. Classical examples are studies of twins in

which identical twins are compared to identify the effect of the environment on certain conditions. By using identical twins one can completely eliminate genetic differences between individuals which could not have been achieved in any other way. This is because genetic differences cannot be expressed numerically and are therefore not amenable to, say, multivariate statistical methods. Matching or stratifying (matching is a special kind of stratification) is therefore a very powerful methodological tool, but it should always be used with good reason to justify the extra effort involved. It should be remembered that matching and stratification should be taken into account during statistical analysis. More advanced methods are sometimes required for the analysis of stratified data.

The strength of association between a dichotomous (i.e. one that is either present or absent) risk factor and the presence of disease can be expressed in terms of the relative risk or odds ratio.

Relative risk is defined as the probability of getting (if new or incident cases are considered) or of having (if existing or prevalent cases are considered) the disease in the group exposed to the risk factor divided by that probability in the unexposed group. Note, the relative risk for prevalent and incident cases can be different. Factors influencing only the duration of a disease will have relative risks differing from unity for prevalent cases, but not for incident cases. This is especially important in diarrhoea research.

The odds ratio is similarly defined as the ratio of the odds in the exposed group to that of the unexposed group, where the odds means the ratio of the probability of getting (having) the disease to the probability of not getting (having) the disease.

Consider the following table,

	disease +	disease -
risk factor +	a	b
risk factor -	c	d

If the risk factor is sampled and not the disease (e.g. in cohort or intervention studies) the relative risk is,

$$\frac{a/(a+c)}{b/(b+d)}$$

and the odds ratio is

$$\frac{ad}{bc}$$

i.e. the cross product ratio.

As an estimator of the odds ratio, it is perhaps better, especially for tables with small entries, to add 0.5 to each of the cell entries since by doing so infinite or zero estimates are excluded.

In a case control study, however, the relative risk is <u>not</u> estimable, but the odds ratio is.

The reason for this is as follows. Let the relative risk be r and the probability of getting or having the disease in the unexposed group be p. The probability in the exposed group is therefore rp. Let the probability of having the risk factor be f (i.e. the probability that a random individual has risk factor +ve). Then, when an equal number of cases and controls (both equal to n) are sampled, one can easily calculate (try it!) that the expected values of the above table are:

$$Ea = nfrp/(frp + (1-f)p)$$
$$Eb = nf(1-rp)/(f(1-rp) + (1-f)(1-p))$$
$$Ec = n(1-f)p/(frp + (1-f)p)$$
$$Ed = n(1-f)(1-p)/(f(1-rp) + (1-f)(1-p))$$

and from these expressions it can be seen that the cross product ratio of the expected values of the entries gives the correct value of the odds ratio, viz. $r(1-p)/(1-rp)$, whereas the substitution of the expected values into the formula for the relative risk does not.

This is perhaps best understood with an example. First consider population 1. Let (in population 1) the risk of people with factor F to have disease D be 40%. Without factor F the risk is 20%. The relative risk is therefore 2. If factor F occurs in 50% of all people, then in the population 20% has both F and D, 30% has F but not D, 40% has not F and not D, and 10% has D but not F. In a case-control study 3/7 of the controls and 2/3 of the cases will have F.

Now consider population 2. With F the risk of having D is 2/3000 and without F the risk is 1/4000. The relative risk is 8/3. Factor F occurs in 3 out of 7 individuals in population 2. In a case-control study we find again that 2/3 of the cases have F and 3/7 of the controls. Yet, the relative risks differ.

If the disease is "rare" then there is little difference between the odds ratio and the relative risk, and the estimator for the odds ratio can be used as an estimator of the relative risk. This assumption that the disease is rare is almost universally made in the epidemiological literature.

What is the scope for case-control studies in diarrhoea research? It is evident that many factors play a role in the causation of diarrhoea and similarly that many other factors can cause the same effects as diarrhoea (e.g. malnutrition). The danger of confounding and similar problems is therefore enormous. One could think however of a few instances where case control

studies could be valuable especially where other types of study design are impossible.

An example is the study of risk factors which operate on village or area level (e.g. water supply, presence of health facilities etc.). Such a study could be set up in a hospital or health centre which services many villages or areas which differ with respect to the factor under study. As controls, one could use children suffering from other diseases not related to the potential risk factor. For instance, if the risk factor is the kind of water supply (Briscoe et al, 1985; 1986) then diseases like fractures, measles, (upper) respiratory diseases, otitis, mumps, malaria and many others are suitable candidates for consideration. There may however be a relationship between the distance from a health centre and water supply, e.g. the health centre was set up in a village with good water supply. Since the further one lives from a health service facility the less likely one is to make use of it, distance is clearly related to the probability of being included as either a control or a case. The dependence on distance of coming to the facility may however depend on the kind of disease (case or control), more serious (or believed to be more serious) diseases being more likely to be reported from greater distances than are minor afflictions. Thus:

i) The risk factor (e.g. water supply) and distance are related.
ii) The disease (or its probability of being seen) and the distance are related.

Consequently, distance may seem to act as a confounding variable. However, the phenomenon is more accurately described as selection bias since it is not the probability of becoming a case that is associated with distance, but rather the probability of being selected as a case.

If the quality of water deteriorates with distance, the quality of water may be found to be negatively associated with diarrhoea (i.e. the expected result) if for controls less serious diseases are used, whereas the quality of water may be found to be positively associated with diarrhoea (which is rather bizarre) if more serious diseases are taken for controls. It is therefore essential to make sure that control diseases are "as serious as" diarrhoea, i.e. have the same "selection dependence" on the distance.

The case-control methodology and terminology is also sometimes used for studies which use a (arbitrary) subdivision of a continuous response scale (e.g. number of days of diarrhoea in a 3-months period, or duration of episode of diarrhoea) to define cases and controls. This, in itself, is not invalid but may be very inefficient since differences within the case and control groups are ignored. Such a dichotomization (i.e. division into two parts) is only justified when certain quantitative differences are believed to constitute qualitative differences (for example diarrhoea lasting longer than 14 days is believed to belong to the qualitatively different category of persistent or chronic diarrhoea).

CHAPTER 5

SAMPLING

To estimate a parameter (e.g. prevalence or incidence rate or the regression coefficient of regression weight on diarrhoea experience) one could examine the whole population for which the parameter is meant to apply. This however is rarely feasible and almost always unnecessary. First of all it is not feasible if the parameter is supposed to be representative of its value in the (near) future. Obviously, it is impossible to obtain observations from the future. Note, sampling does not overcome this problem. Therefore, some kind of assumption about "constancy" or "stationarity" of the disease has to be made. Secondly, it is unnecessary because sampling from a population is almost always as effective as a full enumeration.

There are two fundamentally different ways in which a sample from a population can be drawn, viz. unstratified and stratified. In unstratified sampling, the sampling units (e.g. individuals) are drawn from only one "homogeneous" population, e.g. newborns in Kenya. In stratified sampling several subpopulations (e.g. individuals with and without measles vaccination, or children with and without current diarrhoea) are distinguished: from these, separate samples are drawn.

Stratified sampling precludes the answering of certain questions. If healthy and diseased individuals are sampled separately, then it is impossible to estimate the prevalence of the disease. The reason for using stratified sampling is that questions pertaining to differences between the subpopulations can be answered much more efficiently with stratified sampling, especially if one of the subpopulations is small, i.e. comprising only a small proportion of the population at large. If the long term effects (e.g. in terms of risk of getting diarrhoea) of a stay in an incubator (due to premature birth) has to be studied, it is extremely inefficient to do a follow-up study of a sample of all children born within a certain period. Rather, one should take separate samples from babies with and babies without a stay in an incubator. Similarly, if the vaccination status of children with a rare type of diarrhoea (e.g. very persistent) is to be compared to that of healthy individuals, a random sample of the population at large is not the correct method (separate samples should be taken from diarrhoea patients and healthy individuals, i.e. a case-control design).

In addition to answering questions pertaining to differences between subpopulations, stratified sampling (or matching) can also be used to control for differences, i.e. to eliminate differences. For instance, if we want to study the effect of measles vaccination on the risk of developing (chronic) diarrhoea, it is not a good method to follow up children with and without measles vaccination without taking into account other risk factors for diarrhoea which may be associated with measles vaccination (e.g. age, area, socio-economic status). To adjust

for these other risk factors we can stratify the population with respect to the levels of these risk factors and take separate samples from the (strata of) vaccinated and unvaccinated children in each of the strata. It is noteworthy, however, that adjustment for additional risk factors can also be made by means of statistical methods after the study has been carried out. This tends to be slightly less efficient than stratification. This is because then the number of vaccinated and unvaccinated children is often unbalanced. The gains in efficiency of stratification, however, do not always offset the extra effort that goes into it. By contrast, stratification with respect to subpopulations to be compared (incubator or not) is often enormously efficient. If a certain characteristic (e.g. incubator history) occurs in a (very small) fraction f of a population, then a stratified sample of n individuals with and n without the characteristic (i.e. a sample size of 2n) is approximately as efficient as a simple (unstratified) random sample of size n/(2f). So if 1% of all children are ever incubated , then an unstratified sample needs to be 25 times as large as a stratified sample to get the same efficiency.

Unstratified sampling is commonly used in descriptive studies and in analytical studies with many different objectives (questions). The more specific the question(s) addressed by an analytical study, the more likely it is to benefit from some form of stratified sampling.

Unstratified sampling

The classical sampling method, also called simple random sampling, works as follows. A list of all units (individuals) in the population (the so-called "frame") is drawn up and the required sample size is drawn randomly from this list. This is rarely feasible in diarrhoea research in developing countries. First of all, a list of individuals (e.g. under 5's in Kenya) is hard to come by; secondly, the logistic difficulties (e.g. obtaining data from a thousand children scattered all over Kenya) are insurmountable. Therefore, the method of cluster sampling (Cochran,1963) is more common. Certain clusters (e.g. villages) are sampled and all individuals in the clusters enter the study (if only a sample of the cluster enters the study, then the method is called "two-staged" or "multi-staged" sampling).

The efficiency of cluster sampling depends on the heterogeneity of the population (or rather the heterogeneity of the variables in which we are interested) within a cluster. If, for instance, certain (explanatory) variables operate at cluster (village) level, e.g. water supply, the presence of a health centre etc., then the effects of these variables (on diarrhoea) is hard to assess since clusters tend to differ in many other respects. Similarly if the diarrhoea experience (i.e. the response variable) of individuals within clusters is roughly the same, whereas there is substantial between cluster variation in diarrhoea incidence or prevalence, analysis of the data (for an analytical study) may be very difficult indeed.

The effect of variables which operate on cluster level on a disease are sometimes better estimated from a case-control study: cases (of diarrhoea) are selected from a wide range of clusters (e.g. serious diarrhoea cases from a district hospital) and

compared with a comparable control group. The construction of a comparable control group (which should not be matched on the variables of interest, i.e. the cluster variables) is a gigantic problem. It has only rarely been applied to diarrhoea studies.

True cluster sampling (i.e. the random selection of clusters) is not always possible and clusters are selected because their selection is logistically feasible (near the centre, a road, not too distant from each other etc.). Similarly, patients hospitalized for dehydration are usually not taken at random but from a few cooperating hospitals.

The generalization of results obtained from such a "sample" (i.e. its external validity) is not a purely statistical matter, but has to be answered by subject experts (are these village typical enough? Are the cooperating hospitals the ones getting the most serious cases? etc.).

Certain questions are certainly more sensitive to non-random sampling than others. Prevalence and incidence of diarrhoea (and other "tropical" diseases) fluctuate widely and estimates obtained from a few selected villages in a country cannot be expected to accurately reflect the situation in the country as a whole. Similarly, case-fatality rates of diarrhoea obtained from an academic hospital in a Capital city is a poor indicator of these rates in other hospitals, let alone of all diarrhoea cases. Answers to questions such as whether ORS prevents malabsorption of nutrients, or whether cholera vaccination increases the infectious dose of cholera bacteria, can be expected to depend less on the specific population studied and can therefore be more confidently answered from a selected "sample". Note, that in intervention studies (e.g. clinical trials) the use of true random samples is exceptional.

Stratified sampling

The difficulty in unstratified sampling, viz. that of obtaining a truly random sample, besets the investigator who wants to take a stratified sample even more than the one taking an unstratified sample. Not just one, but several (sub)populations have to be sampled. Besides, these subpopulations are often defined by attributes of individuals (hand-washing or not, immune for measles or not, having diarrhoea or not) for which no (reliable) register is available. Selection in sampling is therefore unavoidable. However, since the objective of deciding for a stratified sampling was the comparison between the strata, this selection effect is less serious if samples from the subpopulations involved are affected by selection in a similar manner.

For example, if in a cohort study, the effects of incubation (as a result of prematurity) on future diarrhoea incidence is to be studied, one has to make sure that the selection, induced by the case-finding method - here, the decision of the parents to take the premature newborn to hospital - affects the controls in a similar manner. The specific way to do this depends on the local circumstances and the inventiveness of the investigator(s).

Since it can never be proven that samples from subpopulations are affected by selection in the same way, it is recommendable to take several samples from (one of the) subpopulations taken in totally different ways (therefore affected by different selection mechanisms). In the above example, on the effects of incubators, one could consider several "samples" of healthy children as "control" groups, one taken from those born in the same hospital, one taken from those born in the same village etc. If the difference between incubated children and all control groups is consistent (i.e. pointing in the same direction), then the evidence for the effect to be demonstrated is much stronger than that obtained from one control group alone. Note, that one can take several different control groups for a totally different reason as well, viz. each one taken for the answering of a different (etiological) question.

For case-control studies (for which we have already discussed selection bias in chapter 4), where the sub-populations consist of diseased and healthy individuals, the same principles apply: try to take controls affected by the same selection mechanism, and, if possible, take several different control groups. When cases are selected from a hospital, these principles are often best met by taking as controls, not healthy individuals from the general population, but patients from the same hospital, suffering from diseases with comparable hospitalization rates and unrelated to the determinants of interest. If several diseases qualify as control diseases then one can take several control groups.

Stratified sampling can also be used to eliminate the effect of certain (confounding) variables. If, for instance, one wants to study, in a case-control study, the effect of hygiene on diarrhoea incidence or prevalence independent of the socio-economic status of the family (a potential confounder), one could decide to form socio-economic "strata" (e.g. low,middle and high) and take controls from the same stratum as the cases. The alternative, forming such strata afterwards (during data analysis) may prove very inefficient since it may well turn out that all cases are from the low stratum whereas controls are from the high stratum, making comparison difficult. In the above example double stratification has been used: sampling has been stratified with respect to the variable case/control and to the variable socio-economic status. If a too high degree of stratification is used it may be very difficult to fill some of the (sub) strata. This effectively limits the degree and complexity of stratification one can use.

CHAPTER 6

DATA ANALYSIS

Each field study on diarrhoea is bound to yield a huge amount of data which needs to be properly analyzed after collection. However, the precise way in which the data is to be analyzed should be thought out before the study has started, i.e. when preparing the study protocol.

Each type of study design entails its own characteristic complexity of data structure. Cross-sectional (e.g. case-control studies) are usually quite simple: individuals (e.g. cases and controls) are representable as single records containing all the information about the individual. Such records are directly processable by most statistical packages.

Longitudinal studies produce data of a far more complex nature. The data structure obtained from a population follow-up diarrhoea surveillance study is usually even more complex than the one obtained from an age cohort follow-up study, due to the irregularity of measurement points caused by the differences in ages (of individuals in the study) at the start of the study. For instance, in a one-year follow-up study of under 5 children, some of the children will be born during the study period, others we be included from the age of 15 months to the age of 27 months etc. For each individual (child) a huge amount of data will have been collected: diarrhoea surveillance, anthropometric (weight, height) surveillance etc. In addition, there is usually also information which is not collected on the individual (child) level, but on e.g. household or village level (e.g. wealth of household, number of children in a household, water-supply of village). In short, the data structure is far removed from the "classical rectangular data structure" of statistical textbooks and statistical computer packages, with "cases" as rows and variables as columns.

Somehow, the data has to be brought into this rectangular format before analysis can take place. To be able to do this it is effectively we would strongly recommend the use of relational database techniques for data storage. This is discussed in more detail in Appendix 5.

For instance, suppose one wants to analyze the effect of the wealth (e.g. measured as monthly income) of the household of a child on its diarrhoea experience between birth and the age of 1 year. First, one has to select only observations made within the first year of life. From these observations one should calculate a diarrhoea score for each child (e.g. the total number of days of diarrhoea divided by the total number of days exposed), link this observation to the wealth of the households and then (note that the data is now rectangular) calculate correlation coefficients, regression analyses etc.

However, if the whole study (period of surveillance) took one year only, only a few children will have been included from birth to the age of one year. Most children will have been included for several months only. Diarrhoea experience during the first six months is very different from diarrhoea experience during the second six months of life. This makes analysis far more difficult. Inclusion of the age or average age of the child as an extra covariate (covariable, explanatory variable) in a multivariate regression model is usually indicated.

Even more complex is the situation where two surveillance data sets (e.g. diarrhoea and anthropometry) have to be related to each other. Within each child there is a continuous mutual influencing of both "processes": diarrhoea (negatively) influences nutritional status (weight, height, weight for height etc), and a poor nutritional status increases (or is likely to do so) the probability of getting diarrhoea. Cause and effect are thus hard to disentangle.

The easiest way to analyze such data is by means of "pseudo cases". For instance, the observation period is divided into 3-month-periods and variables such as the nutritional status at the end of each of these periods, the number of diarrhoea days in the preceding 3 months, (possibly) the nutritional status 3 months earlier, and also other (co)variables, are treated as a "case". In this instance, we would probably wish to use nutritional status as the dependent variable. It should be apparent that this will result in one individual yielding several such cases. If the "cases" of one individual are (strongly positively) correlated even after adjustment for the covariates, then the p-values resulting from e.g. regression analysis should be considered with suspicion since the "cases" are not independent. Although this is not very likely to happen, if it does, then there are two ways to "remedy" the situation:

i) Modify the (e.g. regression) model so as to allow for correlated residuals (i.e. the observations minus their predicted values).

ii) Use standard procedures to estimate parameters (correlations etc.) but use jackknife techniques to compute the standard errors of these estimates. The jackknife is based on leaving out of the sample each individual (in this context, this would involve leaving out several "cases") at a time and re-estimating the parameters, correlations etc. from the reduced data set. (For a more detailed discussion, see Efron, 1982).

Many software tools (packages, programs) are available for the whole process of data entry, record/file linkage and statistical analysis (descriptive statistics, significance tests, regression analysis, discriminant analysis, factor analysis, cluster analysis, survival analysis, correspondence analysis etc.). For instance, in DBASE III or in the SPSS Data Entry module, it is possible to design screen forms quite similar to those used in the field to facilitate data entry. Programs can be written in Data Entry (easy) or DBASE III (more difficult) to check for "impossible" (year-of-birth 91 for a child) or "unlikely" (birth height 60 cm) data, or for digit preference (e.g. in

one study diarrhoea of 7 days duration were several times as likely to occur as diarrhoea lasting 6 or 8 days).

Double coding and double entry to reduce errors is also possible but more expensive and does not remove field-worker errors. It cannot be emphasized strongly enough that there is no substitute for ensuring that the data actually collected in the field is accurate. Strictly, any data collected in the field should be validated in the sense that repeated "measurements" of a sub-sample should be undertaken so as to ensure reproducibility.

Subsequently for each type of analysis record linkage has to be performed. This can be done within DBASE III (although this package has only limited possibilities to do this) or another database manager (e.g. Paradox), or in the statistical package SPSS. Once this is done the required analysis can be carried out.

Some useful statistical techniques

Regression Analysis

A very useful statistical technique is (multiple) regression analysis (Rao, 1973), which is used to explore the relationship between a response variable y (e.g. weight) and one or several explanatory ("causal") variables $x_1,...,x_p$ (also called covariables, covariates) e.g. diarrhoea experience, weight 1 year earlier, age, sex, breast-feeding, season etc.

The regression model assumes the following linear relationship between y and $x_1,...,x_p$

$$y = E(y|x_1,..,x_p) + e = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + e$$

where e is the "error" term, i.e. that part of y which cannot be explained by the x variables. $E(y|x_1,..,x_p)$ is the expected value of y conditional on (i.e. given) the x variables.

The regression equation is easy to understand, although not always easy to interpret. The equation simply indicates that the increase of x_j with one unit is associated with an expected increase of y with β_j units. The regression model is additive, the contributions from the different x variables are added together to form the expected value of y.

Statistical packages can be used to estimate $\beta_0,....,\beta_p$ and test whether these differ from zero (in which case the corresponding x did not need to be included in the regression equation).

It may sometimes be necessary to apply a transformation to either the dependent y variable or any of the x variables before using regression analysis. For instance, if the effect of the x (co)variables is assumed to be multiplicative rather than additive (e.g. in relative risk models), a logarithmic transfor-

mation is required of both y and the x variables before the linear (additive) regression model is applicable. The reason for this is that the logarithm of a product of numbers equals the sum of the logarithm of those numbers.

In addition to transformations it is sometimes necessary to include so called "interaction terms" into the regression model, e.g. x_1x_2 which reflects an interaction or synergism between the variables x_1 and x_2. A positive β associated with this interaction term implies a positive synergism between these two variables. However, an interaction that is detected in a regression analysis on an additive scale may well disappear after a (e.g. logarithmic) transformation of the y and/or x variables. Then the transformed regression model is preferable to the untransformed one because it requires less parameters (viz. there will be no coefficients for interaction terms).

A problem which sometimes occurs in regression analysis (and other multivariate techniques) is that of collinearity, two or more of the x variables are so highly correlated (e.g. duration of diarrhoea and total number of stools) that it is difficult to separate their effects. One then has to decide on theoretical a priori grounds which of the two variables is the most likely to be a real causal factor. If two variables are highly correlated and one of them, which is not a real causal variable, is used in the regression model and the other one, which is the real causal variable, is not, then the variable in the regression will turn out to be important, although in reality it is not. The real causal variable acted as a confounder.

If regression analysis is used for predictive purposes (for example to predict from household characteristics the probability of getting diarrhoea) then the question whether the predictive variable(s) are causally related to the outcome variable (diarrhoea) is irrelevant. Only the quality of the prediction - as for instance measured by R^2 (the multiple correlation coefficient) - is of importance. If, however, the purpose of the investigation is to identify (potential) etiological factors, then interpretation of regression analyses (and other multivariate techniques) can be problematic in non-experimental studies.

In diarrhoea research many variables are not directly measurable, for example socio-economic status of a household or personal hygiene. Instead so called "proxy" variables are used. As a "proxy" for socio-economic status land ownership or the number of cows owned is used and, as a proxy for personal hygiene, one uses the consumption of soap or the number of times a child caretaker washes hands. The effect of using a proxy variable in lieu of the real variable is to bias the regression coefficient associated with it to zero. That is, the relationship becomes less strong. Essentially this kind of bias is the same as misclassification bias discussed earlier.

Discriminant Analysis

Another useful technique, closely related to regression analysis) is discriminant analysis (Kendall and Stuart,1968),

which is used to explore whether two (or more) identifiable groups (e.g. healthy versus ill) can be distinguished by a set of observable variables $x_1,...x_p$ (e.g. age, sex, WBC, blood-pressure etc.).

In discriminant analysis a score consisting of a linear combination of variables

$$\beta_1 x_1 + ... + \beta_p x_p$$

is constructed which best discriminates between the two groups. Observations with a low score are assigned to one group whereas observations with a high score are assigned to the other group. One always has to check how well the method performs by inspecting the number of misclassifications.

It may sometimes be useful to use sample splitting to get a good idea how well the method really works. The sample is split into two halves, one half is used to estimate the discriminant score function which is then used to classify the other half and vice versa. Of course, sample splitting is applicable to other techniques (e.g. regression analysis) as well. It should be borne in mind that interpretation of the results of discriminant analysis poses problems similar to those of regression analysis.

A popular technique related to discriminant analysis is logistic regression also called logistic discriminant analysis (Cox,1970). In logistic regression the probability that an observation belongs to group 1 (versus group 0) is modelled as

$$\ln \left\{ \frac{Pr(1)}{1-Pr(1)} \right\} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

If the groups to be distinguished are people suffering from a specific illness (cases) and healthy people (controls) and the observable variables are "risk factors" then discriminant analysis/logistic regression provide estimates of odds-ratios (in an epidemiological context, relative risks) associated with these risk factors.

The odds ratio of a individual with covariable values x_{i1} (i=1,2,...,p) and one with covariable values x_{i2} is,

$$\frac{\exp(\beta_1 x_{11} + ... + \beta_p x_{p1})}{\exp(\beta_1 x_{12} + ... + \beta_p x_{p2})}$$

If $x_{i1} = x_{i2}$, for i = 1, 2, ... , p-1, i.e. all risk factors except the p-th one have the same value, then the odds ratio is

$$\frac{\exp(\beta_p x_{p1})}{\exp(\beta_p x_{p2})}$$

Cluster analysis

Cluster analysis (Everitt,1980) works in the opposite direction to discriminant analysis. Here, groups are not yet known/defined, but the observed variables are used to see whether certain configurations of these variables occur in "clusters" thereby defining "groups".

Cluster analysis is an exploratory technique aiming at pattern recognition. It is used to explore whether meaningful groups exist. The statistical properties of cluster analysis methods are not fully understood and it is therefore difficult to establish how "probable" the existence of a certain cluster/group is. This can only be done in a second, confirmatory stage of research.

There are many types of cluster analysis differing in the distance measure used or in the computational method (algorithm). Results from these different types of clustering may differ substantially thereby posing interpretational difficulties. Cluster analysis could be used to group stools based on organisms found in them, or to group individuals based on their (hygienic) habits.

Survival analysis

In survival analysis (Kalbfleisch and Prentice,1974) one studies the time to the occurrence of a specific event (e.g. death, end of diarrhoea, occurrence of a second episode of diarrhoea, dismissal from hospital etc.) and the factors which influence this time.

A distinguishing feature of survival studies is that some of the events are not yet observed to have occurred. Children may have been followed up for some time during which some, but not all, children have been observed to have had a second episode of diarrhoea. Such observations are called censored observations. These censored observations should not be confused with missing observations. It may be very informative to know that a child has been free of diarrhoea for at least a year.

Survival curves can be calculated and plotted by means of the Kaplan-Meier method. Survival curves can be compared by means of the logrank test and the effect of covariables $x_1,..,x_p$ on survival can be estimated by means of Cox's proportional hazards regression model which enjoys an ever increasing popularity. In Cox's proportional hazards model, the incident risk or hazard of getting the event of interest (e.g. death or a second diarrhoea episode) of an individual is proportional to

$$\exp(\beta_1 x_1 + ... + \beta_p x_p)$$

As in regression analysis the coefficients β_j can be estimated from the data and one can test whether they differ from zero.

Factor analysis

Factor analysis (and the related principal component analysis) is useful in a situation with many variables which are all highly correlated. Factor analysis (Kendall and Stuart,1968) explores whether these many variables can be thought of as being "built" of a few unobservable common variables (also called factors or dimensions) which account for the observed correlations. Those factors can (with pluck and luck!) be interpreted and given names.

Solutions of factor analysis are not unique. By "rotations" one can obtain different solutions, some of which may be easier to interpret than others.

Factor analysis originates from psychometrics, where researchers were confronted with the problem of how to extract fundamental capacities (e.g. verbal, spacial, IQ etc.) from a large number of mutually correlated test items. In diarrhoea research one could apply this technique in the analysis of hygiene behaviour studies or knowledge questionnaires to explore whether hygiene behaviour or knowledge can be explained in terms of a few (understandable) dimensions. For instance, it may turn out that hygiene has two dimensions, one relating to personal hygiene and one related to hygiene of the house, latrine etc.

Some investigators use factor analysis as a "solution" to the multicollinearity problem in regression analysis, i.e. when a large number of mutually highly correlated covariables thwart the identification of "significant" variables. Factor analysis or principal components analysis is then used to extract a few mutually uncorrelated "factors" which are interpreted (given names) and subsequently used as "input" (as covariables) for a regression or discriminant analysis.

Such a procedure is not recommended. First of all, the results are usually difficult to communicate. It is easy to explain that blood-pressure predicts something. To explain that "0.7*bloodpressure - 1.2*income + 0.2*race - 2.9*cigarettes" predicts something is quite a feat. At times factor analysis gives clear cut results with easily interpretable factors. But this is an exception. Second, if in a regression analysis a factor appears to be significantly related to the dependent variable, one still does not know which of the variables in that factor is related to the dependent variable and which one is not. The method tends to obscure rather than clarify cause effect relationships.

A technique related to factor analysis is multidimensional scaling (Schiffman et al, 1981). In this technique pairs of variables are assigned similarity measures (e.g. correlation coefficients) and multidimensional scaling then maps all the points on a two (usually, since two dimensions can be plotted) dimensional space such that similar point are close together, dissimilar points far from each other etc.

A program for calculating multidimensional scaling is available in the SYSTAT statistical package.

As in factor analysis this technique is very useful to explore and depict relationships among many variables (e.g. from questionnaires).

Correspondence analysis

Correspondence analysis (Greenacre,1984) can be used to explore the relationships in (large) contingency tables (e.g. a table of diagnoses by village) with a high X^2 value, i.e. the table is very inhomogeneous (one diagnosis being relatively more frequent in one area than in another). Plots (so-called biplots) can be drawn which evince the correspondence between row variables (e.g. diagnoses) and column variables (e.g. areas) and also the correspondence among row and column variables themselves.

Despite its extreme usefulness this technique has only recently been incorporated in the BMDP statistical package and is not yet available in other packages. Separate programs can also be used. A very user friendly program is the program CORAN which is available from the Central Bureau of Statistics in Voorburg, The Netherlands.

-*-

All these multivariate techniques (multivariate meaning dealing with more than two variables simultaneously) can only be performed by means of a (micro)computer. Hand calculations would take ages. If one has no access to a computer, then it is absolutely necessary to acquire one (before the protocol is even written include a PC in your budget!). The PC can also be used for patient management, budget management and text processing. Typical prices of IBM compatible PC's are 1000 to 2500 US$ (tax free prices). Software is also rather expensive unless one works (as unfortunately many people do) with illegal copies which are widely available.

Studying causal pathways

In large studies involving many individuals and many variables one is likely to find several factors being associated with diarrhoea (even when these relationships are not very strong). Usually these factors also influence each other (e.g. socio economic status and type of food eaten) and for possible interventions it is worthwhile to examine which factors are associated due to confounding, i.e. are correlated with causes (the confounders) but are not a cause themselves, which factors constitute a direct cause and which factors constitute an indirect cause.

Although statistical methods may offer some help in this field (there are even formal methods like LISREL (Joreskog,1981)

to do this), knowledge of the subject field (e.g. microbiology) and intuition are at least as expedient in solving this problem. If some parts of the causal pathway are well known, e.g. drinking from a contaminated source of water causes diarrhoea, then instead of studying factors explaining diarrhoea, one can also look into the factors which "cause" the drinking of contaminated water.

It is may be illuminating to try to draw the causal pathways among the variables in a study in the form of a graph, i.e. to represent variables as dots and to draw arrows between variables which are causally linked, with the arrow pointing in the direction of the causation.

Some general statistical principles may be useful though.
1) If an association between a factor and diarrhoea is caused by confounding (this can be depicted by arrows pointing from the confounder to both diarrhoea and the factor) then the association disappears after adjusting for the confounder. Such adjustment is ideally done by conditioning on the confounder, i.e. by considering the relationship for constant levels of the confounder. This is done e.g. in the Mantel-Haenszel method. Alternatively, adjustment can be made by including the confounder in a multivariate statistical model, e.g. regression analysis.
The relationship between a confounder and diarrhoea (or any other outcome variable) does not disappear after adjustment for the study factor.
2) If a factor is antecedent to another intervening variable, e.g. income is antecedent to the intervening variable protein consumption (this can be depicted by an arrows pointing from income to protein consumption and from protein consumption to diarrhoea), then the relationship between the antecedent factor and diarrhoea after adjusting for the intervening variable. The converse, however, is not true.

It is clear from the foregoing that it is not possible to distinguish between a confounder and an intervening variable on purely statistical grounds, i.e. by studying the association between a study factor and diarrhoea before and after adjustment of the third variable (confounder or intervening variable). Whether an arrow points from the study factor to the third variable (an intervening one) or from the third factor (a confounder) to the study factor cannot be determined from mere associations. Additional (e.g. biological) arguments, temporal relationships (causes preceding effects) or experiments are needed to decide this.

Poisson variation

Statistical methods like regression are very useful in attributing variation in observed diarrhoea to observed causative factors (e.g. age, sex, water-supply etc.). However, a large part of the observed variation is bound to remain unexplained by the observed causative factors, because it is totally random (with, for the incidence of diarrhoea, a Poisson distribution). This is easily understood: two identical children under the same circumstances (as measured by the observed determinants) will not

get diarrhoea at exactly the same time and for the same duration. Circumstances, household environment, food ingested etc. vary constantly and unpredictably. All one can try to "capture" in a statistical model is the risk of a child of getting or having diarrhoea. The actual pattern of diarrhoea, given a specific risk, remains unexplained.

For ease of elaboration let us consider only diarrhoea incidence (i.e. we ignore the length of diarrhoea). We can simplify matters further by treating such incidence cases as point events (in practice of course they are not, during a bout of diarrhoea it is impossible to observe a new incident case). If the risk of getting diarrhoea is constant (e.g. an average of μ attacks during the observation period) and is the same for all children, and the incidence of attacks of diarrhoea are mutually independent (likely to be a reasonable approximation in the case of diarrhoea if there are many causative organisms), then the probability of getting k episodes of diarrhoea is

$$Pr(\underline{k}=k) = \frac{(e^{-\mu})\mu^k}{k!}$$

For instance, for $\mu=1$ and $\mu=2$, one gets the following probabilities:

$\mu=1$	$\mu=2$
$Pr(\underline{k}=0)=0.37$	$Pr(\underline{k}=0)=0.14$
$Pr(\underline{k}=1)=0.37$	$Pr(\underline{k}=1)=0.27$
$Pr(\underline{k}=2)=0.18$	$Pr(\underline{k}=2)=0.27$
$Pr(\underline{k}=3)=0.06$	$Pr(\underline{k}=3)=0.18$
$Pr(\underline{k}=4)=0.02$	$Pr(\underline{k}=4)=0.09$
	$Pr(\underline{k}=5)=0.04$

This variation will be observed even without there being any factor or variable explaining it (i.e. all children have the same risk).

It is illuminating to play a little with the formula for Poisson variates (using different values of μ) to get a "feel" for the kind of variation that is to be expected as "background noise".

Inspection of the distribution of diarrhoea incident cases is very valuable for locating the sources of variations of incidence. If the incidence distribution within a village (or other type of cluster) roughly follows the Poisson distribution, but there exists a substantial between village variation in average incidence, then the causative factors (in an epidemiological sense) operate on village level and a search for within village between individual factors e.g. hygienic practices is not promising (with large numbers of individuals subtle differences can of course be proven, but these differences are not likely to be important predictors or determinants and therefore not useful in intervention). Note, however, that this only holds true if the observation period is such that the average number of diarrhoea episodes per individuals is larger than unity. For

short observation periods "Poisson" distributions may be observed even when there exists a large variation (in terms of relative risks) in underlying risk.

Since age is an important determinant (but one we are usually not interested in) in diarrhoea incidence, it may be worthwhile looking at the incidence distribution by age group as departures from the Poisson distribution may be attributable to an age effect. Formal statistical tests to test for a Poisson distribution are derived in appendix 6.

If the distribution of the diarrhoea incidence is not Poisson due to overdispersion (more variation in incidence than expected under the Poisson distribution), then there must be a real variation in underlying risk which may be attributed to (a variation in) certain risk factors. Often, a good model for the incidence distribution then is the negative binomial distribution,

$$Pr(k=k) = \frac{\Gamma(k+r)}{\Gamma(r)\,\Gamma(k+1)}\, p^r(1-p)^k$$

where $\Gamma(.)$ denotes the Gamma function. (If r is an integer, then $\Gamma(r)=(r-1)(r-2)(r-3)....321.$) This distribution has 2 parameters r and p. The expected value (mean) of k is $r(1-p)/p$, whereas the variance is $r(1-p)/p^2$, so that the variance is always larger than the mean. The ratio of "excess" variance to variance, 1-p, indicates approximately how much variation is due to variation in risk and how much is random, i.e. Poisson variation. For r very large and p close to one the negative binomial "resembles" the Poisson distribution.

As an example consider the following (yearly) incidence data from a community in Western Kenya. The number of children (0-5 years old) having 0,1,2,...,11 episodes of diarrhoea was 3878,2033,976,563,272,162,101,58,24,12,7,5. The best fitting (using maximum likelihood) negative binomial distribution was the one with r=.93 and p=.46, so that roughly half the variation was attributable to variation in "real" risk. A large part of this variation in real risk was however due to differences between age groups, so that the within age group variance attributable to potential risk factors is less than 50%. Under these circumstances correlations between risk factors and diarrhoea incidence can only be low.

It is important to note that the longer the observation period, the more an existing underlying heterogeneity in risk can be detected in the data. On the other hand, if the observation period is so short that most individuals had only none, one or two episodes of diarrhoea (mean number less than unity), the true incidence distribution becomes indistinguishable from a Poisson distribution. This is called "the law of small numbers" (Quine and Seneta, 1987). The reason for this is that even when there are sizable risk ratios, the risk differences, and hence the variance in underlying risk, can only be small. Under these circumstances an apparent Poisson distribution does not vitiate a search for risk factors.

The negative binomial and the Poisson distribution can be used to estimate the number of individuals with zero frequencies if these are not observed directly. Suppose that a health centre is interested in the size of the (child) population it "covers". The centre has records of individuals who visited the health centre at least once, only. Fitting a negative binomial to the observed number of visits and extrapolating to the number of zero visits gives an estimate of the latter. Applying this method to the Kenya data (leaving out the zero cases) gives an estimate of 4437 individuals without diarrhoea, which is reasonably accurate.

CHAPTER 7

SAMPLE SIZE CALCULATIONS

The purpose of an epidemiological study is to gather information. This information can serve two purposes, a purely scientific one, or an applied scientific one (that is, one in which the information is gathered to improve decision making). Epidemiological diarrhoea research is mostly of the latter kind. The kind of decisions such research could help make are, for instance, whether a health education campaign is more cost effective than the installation of piped-water installations, or what the optimal time for measles (often accompanied by diarrhoea) vaccination should be.

In deciding what sample size or study size is required one must realize that the bigger the study the more information it yields and therefore the better the decisions based on it.

Research costs money and more research costs more money. In addition, research costs time. The decisions which the study is supposed to help one to make has to be delayed until the completion of the study. Inevitably, this will result in unnecessary deaths, potentially inadequate treatments etc. The optimal study size is therefore one which takes into account all these aspects and maximizes the total "gain" or minimizes the total "loss".

This approach to sample size determination is called a "decision theoretic approach" and should, in our opinion, be used whenever possible. Unfortunately, in medical research (unlike e.g. oil exploration or military submarine surveillance) it is rarely used. Even if all aspects of the decision process can be quantified (and this is rarely the case) then one still has the problem of balancing "losses" measured in different units (e.g. money and deaths: what, for example, is the monetary value of the life of a child of 1 year old?).

An additional problem is that although a research project is designed to help decision making, the precise decision process is not known in detail because it is not the researcher who decides but instead a committee in a ministry of health or a politician. All these considerations lead to the conclusion that although a decision theoretic approach is logically sound and consistent it is of limited value.

It may however still be useful as a general way of looking at things. For instance, if all possible decisions the study can help make can be made without that study (e.g. because the required information is available from other studies) then it should not be carried out.

A more practical approach is the "classical" approach to sample size determination which we shall illustrate with a few examples.

Example 1

Suppose one wants to estimate a "parameter" in a population e.g. average (population) bloodpressure, average number of diarrhoea days per child per year etc. It is clear that the required sample size depends on:

i) The precision one wants to know the parameter.
ii) The variability (both biological and measurement error) or standard deviation σ of the bloodpressure, number of days etc. in the population.

It is natural to estimate the population average μ by the sample mean

$$x = (x_1 + \ldots\ldots + x_n)/n.$$

The standard deviation $\sigma(x)$ of x (also called standard error) is:

$$\sigma(x) = \frac{\sigma}{\sqrt{n}}$$

If the required precision (the width of the confidence interval), with $1-\alpha$ confidence, of x is d, then the required sample size is approximately

$$n = 4\,\frac{\{k(\alpha/2)\sigma\}^2}{d^2}$$

where e.g. $k(0.025) = 1.96$.

If one wants to estimate a fraction (e.g. fraction of children with Giardia) p in a population which is supposed to be of the order 0.5, then the standard deviation is $\sqrt{[p(1-p)]}=0.5$ (of course p is still unknown but a rough guess is sufficient) within an interval of 0.04 (4%) with 0.95 confidence, then the required sample size is

$$n = 4\,\frac{\{1.96 \times 0.5\}^2}{0.04^2} = 2400$$

Note that the required sample size increases with the square of the precision, so that an interval of 0.08 width requires only 600 observations. If no reasonable guess for the standard deviation σ can be made, then it may be necessary to carry out a pilot study to obtain one.

Example 2

Suppose the study has the objective of discriminating between two hypotheses, a so-called null hypothesis H_0 (which is usually the "uninteresting" one of no difference, no effect etc.) which we would like to reject (in favour of the alternative hypothesis) and the alternative hypothesis H_1 that there exists a (specific)

difference, an effect etc. In the case of a diarrhoea study one could think of the null hypothesis e.g. to be that there exists no difference in diarrhoea mortality between ORS treated children and of the alternative hypothesis that ORS therapy reduces mortality (this is a one sided alternative).

To see whether we can reject H_0 in favour of H_1 we apply a statistical test (t-test , X^2 test etc.) to the data. Assume that H_0 is true. Then, if the probability of finding a result as extreme or more extreme than the one obtained (more extreme in the direction of H_1) is less than a certain value α (the so called significance level e.g. 0.05 or 0.01), we will reject H_0 in favour of H_1. We will then say that there exists a statistically significant difference between the groups.

The probability of incorrectly rejecting H_0 (error of the first kind) is therefore α.

However, what is the probability of correctly rejecting H_0 when H_1 is true (i.e. the "power of the test" or 1 minus the probability of incorrectly not rejecting H_0, i.e. $1-\beta$, where β is the probability of an error of the second kind)? Of course, when H_1 is close to H_0 (very small effect, very small difference between treatments etc.) this will not be much bigger than α unless the sample size is very large. If H_1 is far away from H_0 then a smaller sample size will suffice.

The specific choice of H_1 can be based on either the difference one can reasonably expect, or on the minimal difference which is worth demonstrating (e.g. if the difference is less the intervention will not be used in practice). The power (for a given value of α) therefore depends on both the kind of alternative hypothesis one wants to demonstrate (small difference or a large one) and on the sample size.

Conversely, if H_1 and the power are given one can calculate the required sample size.

As an example we present here the formula for the required sample size for the comparison of two means (e.g. the mean number of diarrhoea days in a treated group versus this number in an untreated group). If the standard deviation (the variability of the response in the populations) in both populations is approximately the same and equal to σ, we find for one-sided tests:

$$n = \frac{2\{k(\alpha)+k(\beta)\}^2\sigma^2}{d^2}$$

as the required number per group, where d is the mean difference under H_1 and $k(\alpha)$ and $k(\beta)$ are numbers which depend on the required significance level α and power $1-\beta$ of the test. For two-sided tests with significance level α, we replace $k(\alpha)$ in the above formula by $k(\alpha/2)$.

For instance, $k(0.20)= 0.85$, $k(0.10)= 1.28$, $k(0.05)= 1.65$, $k(0.025)= 1.96$, $k(0.01)= 2.33$.

The derivation of this formula can be found in Appendix 7.

It is immediately clear from the sample size formula that, for given α and β, the demonstration of a difference of 0.5d requires 4 times (and not two times) as many observations as of a difference of d.

As an example, let us consider two groups of children, one treated with X and the other group treated with a placebo. Each child is observed for 6 months. The alternative hypothesis is that the average incidence (per 6 months) goes down from 1.1 to 0.9. For α we want a value of 0.05 and for the power a value of 0.8. Assuming a Poisson distribution for the incidence, we find that the standard deviation in each group is $\sqrt{1.1}$ and $\sqrt{0.9}$ respectively. Taking 1 as a reasonable average of σ, we find that the required sample size per group is 313.

In all these calculations it is assumed that the population is infinite (in practice more than 10 times the sample). If the population one samples from is finite then less observations are required. One must bear in mind however that the "population" here refers to the reference population to which generalization is required (this may include an unknown number of future patients) which in science is almost always infinite. It is not always practical or possible to make both groups of equal size. One of the treatments (in a clinical trial) may be more costly than the other, or the availability of one of the treatments is limited. In that case one could use unequal sample sizes n_1 and n_2. The test will have the same power as long as,

$$1/n_1 + 1/n_2 = 2/n$$

In case control studies in which cases and controls are sampled in a fixed predetermined proportion (unlike in clinical trials where the treatment or risk factor distribution is predetermined) we have to reverse the role of response (i.e. disease) and risk factor to calculate the required sample size. Mathematically there is no difference between:

a) two groups with different levels of a risk factor. Between these groups a difference between p_1 and p_2 in probability of getting a disease (or other response) has to be demonstrated;

and

b) two groups - one with and one without a disease. Between these groups a difference between p_1 and p_2 in probability of having a certain risk factor has to be demonstrated.

A considerable reduction in the required sample size can sometimes be achieved by improving the design of the study. If two pills (e.g. anti-asthmatics, sleeping pills, headache pills) are to be compared and the between patients variability is large but the within patient variability is small, it may be more efficient (instead of using two groups) to use a "cross-over" design in which each patient is his own control (patients get

both pills in randomized order). Typical reductions in the order of a factor 5-10 can thus be achieved.

Always consult your statistician!

In practice it is not always easy to specify what kind of alternative hypothesis one wants to demonstrate. Since the required sample size is very sensitive to exactly this specification, sample sizes differing by a factor 4 to 9 can often be reasonably defended.

Another practical problem is that one study has often many objectives each of which requires its own sample size calculation. Taking the maximum of all these sample sizes often leads to gigantic samples to be taken. This method of sample size calculation clearly also has its limitations. In addition, sample sizes are often to a large extent dictated by logistics (budget, time available, laboratory capacity, children in a village etc.). If this is the case, then a practical way of sample size determination is as follows. Take the maximum number logistically possible (for some questions addressed in the study this is inevitable very little) and see for each of the hypotheses formulated which difference can be demonstrated with which power. If the result is disappointing, then it is not wise to carry out the study at all (a reformulation of the objectives might sometimes help as well, less ambitious objectives can be very valuable). If for some objectives there is reasonable hope to get a satisfactory answer, then one can go ahead. If not, one should seriously think of not carrying out the study at all.

APPENDIX 1: Calculating duration distributions of incident cases

In a steady state situation the following relationship holds between incidence rate I, prevalence rate P and duration (i.e. duration of incident cases) D,

$$P = IE(D) \qquad (1)$$

where E(D) denotes the expected or mean duration. E(D) is calculated from the distribution of the durations using,

$$E(D) = \Sigma\ D.Pr(D) \qquad (2)$$

When prevalence rate (its estimation is comparatively simple) and the average duration of incident cases are known, then so is the incidence rate.

To estimate the distribution of durations of incident cases from duration-to-date data, consider the following:

Let n_i (i=1,2,3,...) be the number of cases of diarrhoea who report that their diarrhoea started i days ago. Those cases have diarrhoea that lasts at least i days. The fraction of those cases is $f_i = n_1/N$ ($N=n_1+n_2+.....$). To get the fraction of cases p_i with diarrhoea that lasts exactly i days we have to take the difference between the fraction lasting at least i+1 days and the fraction lasting at least i days, or

$$p_i = f_i - f_{i+1} \qquad (3)$$

which is the required probability distribution.

The length-specific incidence, i.e. the incidence of cases lasting exactly i days is of course, $I.p_i$. For small samples it is quite likely that one or more of the p_i will turn out to be negative. Since probabilities are necessarily positive such p_i's are unacceptable as estimates. Proper smoothing of the estimates may then be a solution. Alternatively a suitable curve, e.g. an exponential curve or, for discrete probabilities, a negative binomial curve could be fitted to the observed fractions. If only the average duration is of interest, e.g. to relate prevalence rates to incidence rates, negative estimates of probabilities are merely an aesthetic evil.

The average duration of incident cases can be estimated using

$$E(D) = \Sigma i.p_i = \Sigma f_i \qquad (4)$$

i.e. the sum over all i of the fractions of duration-to-date of i days.

When prevalent cases are followed-up until the end of their episode of diarrhoea, so that (in combination with the reported duration-to-date) the distribution of the durations of prevalent cases are known, the method of computing the distribution of the durations of incident cases is slightly different. The probability of being included in the cross-sectional sample is proporti-

onal to the duration of the episode of diarrhoea. Long episodes are more likely to be included than short episodes. This phenomenon is called "length biased sampling".

Let r_i (i=1,2,..) be the prevalence probability of an episode of i days and p_i the incidence probability of an episode of i days. Due to sampling proportional to the length of an interval, and since the sum of all r_i must be unity, we find

$$r_i = i.p_i/E(D) \tag{5}$$

where E(D) again denotes the expected (mean, average) duration of incident cases.

Conversely, p_i can be expressed as a function of r_i by solving the p_i from the above set of equations, yielding

$$p_i = \frac{r_i/i}{E_p(1/D)} \tag{6}$$

where $E_p(1/D)$ denotes the prevalence expected value of the inverse of the duration, or

$$E_p(1/D) = \Sigma r_i/i \tag{7}$$

Using these formulae, the relationship between incidence expected duration E(D) and prevalence expected duration $E_p(D)$ is easily found to be

$$E_p(D) = \frac{E(D^2)}{E(D)} \tag{8}$$

or,

$$E_p(D) = \frac{E(D)^2 + var(D)}{E(D)} \tag{9}$$

and

$$E(D) = \frac{1}{E_p(1/D)} \tag{10}$$

which can be approximated by

$$E_(D) = \frac{E_p(D)^2 + var_p(D)}{E_p(D)^3} \tag{11}$$

where $var_p(D)$ denotes the variance of the prevalence distribution of durations, and var(D) denotes the variance of the duration under the incidence distribution.

Note, that the problem of negative estimates of probabilities which could occur in the "duration-to-date" method does not occur here.

A derivation of the expressions in this Appendix can be found in Freeman and Hutchison (1980).

APPENDIX 2: Some notes on questionnaire design

In an epidemiological study of complex design, including an important economic or behavioral part (as most diarrhoea studies are likely to be, many forms and questionnaires will be used. These forms and questionnaires will have to be filled by field-workers, heads of households, caretakers of children etc.
To be sure that these are filled correctly and unequivocally, one has to design these with utmost care.

Some problems that may occur are:

i) The language in which the questions are put is not (fully) understood by the person who has to fill the questionnaire. This may happen in Africa when the study involves different tribes. Even if there exists some lingua franca or "national language" (like Kiswahili in Kenya) which everybody knows to some extent, this knowledge may be insufficient to understand subtle nuances in a questionnaire. Even if the language is fully understood, the wording of the questions may still be too difficult because sentences are too long or difficult words are used. If the questionnaire is originally designed in English (or Spanish or French) and is thereafter translated into local languages, then one should have these translated back into English to check whether the questions are still asking the same thing. The results may not only be revealing but also very amusing.

ii) Ambiguous or vague questions. For instance "did your child have diarrhoea recently?". In this question it is not specified what is meant by diarrhoea. Is one loose stool enough? Or two? Do people still call a disease diarrhoea if there are other symptoms (e.g. with fever they may call it malaria)? Similarly, a question like "don't you think that diarrhoea is a health problem?" may well be answered with yes if people don't think that diarrhoea is a health problem. Double negatives (e.g. don't you think that contaminated water is not a serious problem in this area?) are to be avoided at all costs.

Another common form of ambiguity is caused by questions offering several choices which can be chosen by ticking a box, but without making it clear whether only one box is to be ticked (as in multiple choice questions) or as many as the respondent thinks are correct.

These examples are trivial of course. Real ambiguity may often be discovered long after the study has started.

iii) The questionnaire is too lengthy. Too often questionnaires are overloaded with all sorts of questions many of which the investigator does not (yet) know what to do with but have been included because they were seen on another questionnaire.
This causes several problems. Interviewees will get bored answering too many questions and answers will be less reliable. Data checking, coding, entering etc. will also be more difficult. So one should ask only those questions which are really necessary and of which one knows beforehand how the answers are going to be analyzed. Questions should be directly related to the objectives set out in the protocol.

iv) Unnecessary open ended questions. Open ended questions may sometimes be useful to explore certain fields. However, whenever open ended questions are useful questionnaires are probably not the best tool of research. Interviews, informal observations by trained anthropologists (possibly from video recordings) are likely to be better tools for such exploratory work. Besides, open ended questions are much resistant to formal statistical analysis. A question like: "what in your opinion are the main causes of diarrhoea?" may yield such a huge variety of answers that any definite conclusion about differences in opinion (e.g. between mothers and grandmothers) is impossible to draw.

v) Poor lay-out of questionnaire. Frequently a questionnaires consists of main questions which everybody has to answer and sub-questions which only have to be answered if the main question was answered by yes or by (fe)males. It is very important to design the questionnaire in such a way that the hierarchy of questions is made clear.

vi) Suggestive questions. Questions like: "you don't think you can wean a child as early as 6 weeks after birth, do you? suggests that the opinion that you can will not be the "correct" answer. If people are inclined to give "desirable" answers (as many are) then you will not get a correct impression of what people really believe or feel or do.

vii) Unnecessary grouping of values in a question. Very frequently researchers include questions into their questionnaires like:

"You were born between 1950 and 1960?
between 1961 and 1970?"
etc

instead of just asking the precise year an individual was born. The reason for this is that researchers think of the way they intend to present data in tabular form when they ask such questions. This however is completely unnecessary since the desired grouping of values can be done during the analysis of the data. A major drawback of asking "grouped" questions is that it is impossible to retrieve the ungrouped value. Neither is it possible to decide on a different grouping if, after all, the original grouping appears less suitable.

For statistical analysis (rather than presentation) grouping is anathema since it throws away all the information about differences within groups. However, grouping may be appropriate when it increases the response rate. In some countries income is a sensitive issue, people are very reluctant to give a precise answer about it even when they are able to do so. By using grouping questions about income, the issue can be made less sensitive. Saying that one earns between 5000 and 10000 units may be easier than reporting an income of 6900 units.

viii) Un-piloted questionnaire. The above given guidelines do not provide a foolproof guarantee that the questionnaire will be understood or that the given answers will have anything to do

with reality. The best way to make sure that the questionnaire is understood and correctly answered, it has to be piloted. Piloting can be done in stages. First a few colleagues, cleaners, field-workers are shown the questionnaire and they are requested to fill the questionnaire and explain the meaning of the questions. If an anthropologist with recent experience in the project area is available then his/her opinion may prove very valuable. In this way some obvious flaws can be removed. Then the questionnaire should be piloted in the field. Apart from requesting local people to fill and explain the questionnaire one should also try (but this is not always possible) to verify the answers given by means of direct observation.

APPENDIX 3: Randomization

Before a clinical trial starts the randomization code has to be prepared for the envisaged number of patients (sample size). Randomization can effectively be carried out in several ways:

1) Use of sealed envelopes. Individual patient assignments are placed in a series of ordered (numbered) sealed envelopes. Each time a patient is admitted to the trial the next envelope is opened and the patient is assigned to the treatment therein indicated. Unless the medication itself is blinded (e.g. simply marked as pill "A" and pill "B") the study will not be double blind (i.e. both patient and investigator are unaware of the treatment received) but at most single blind (only patient is uninformed as to which of the two treatments he receives).

2) Computer randomization. Patients are allocated to a therapy (preferably kept blind, so coded "A" and "B") by the computer after the data of the patient has been entered. The computer can then check whether the patient indeed meets all inclusion criteria and will refuse to randomize a patient if the patient is ineligible.

3) Use of coded medications. Patients receive the next numbered container of medicine. The medication in the box corresponds to the treatment indicated on the randomization list. If both medications look (and taste!) identically, double blindness is easy to maintain.

It is recommendable to randomize in blocks. The randomization list is constructed of consecutive blocks (of e.g. 10 patients) each of which constitute of the random assignment of half the block to one treatment and half to the other treatment. This is done both to ensure equal numbers in the two groups (although it certainly has this effect) and to avoid the (rare but not impossible) possibility that most patients in one group are allocated in the early part of the study and the other group in the late part. A shift in the kind of patients admitted or in the quality of care may then introduce spurious differences between the treatment outcomes. Although both events are unlikely and not very serious (e.g. a slight imbalance hardly affects the power of statistical tests) this method of block-randomization should still be used since it costs nothing to use it.

Some people also recommend stratification i.e. patients with differing characteristics that could affect the outcome of the treatment are randomized separately to ensure an equal distribution of those characteristics over the two groups. Although this may be a good idea in exceptional cases when a factor of an overriding importance exists, this is not to be recommended as a rule. It is difficult to maintain and practically impossible if more than one such factor exists. In addition, imbalances in risk factors can be corrected by taking these risk factor into account during the statistical analysis. One of the exceptions to this rule is when the trial is carried out in several centres simultaneously. Not only is randomization by centre easier than centralized randomization, but treatment

quality and (even with rather rigid inclusion and exclusion criteria) patient characteristics may also differ.

APPENDIX 4: Some notes on protocol design

A full research protocol should contain the following items:

1) Project title. The individuals and institutions participating in the study, including names, addresses, qualifications. For some funding agencies curriculum vitae of the principal investigator(s) are also required. If field-workers or other staff has to be recruited, explain the minimum qualifications required and the kind of training (if any) they are going to receive.

2) A brief summary of the proposal. This is of particular importance when initial approaches are being made to potential funding agencies.

3) A background of the problem to be addressed including some literature review. An overview of the state of the art.

4) The general and specific objectives of the study and how these relate to policy and practice of diarrhoea management. Formulate the specific objective(s) in terms of research questions.
If possible make clear what main and what subsidiary research questions are.

Spin-offs in terms of developing research capacity, institutional development etc. could also be mentioned.

5) Ethical considerations and implications. What benefits (or harm) accrue to the participants? In the case of a clinical trial ethical review procedures and the way informed consent is obtained should be explained. A consent form should be included.

6) All hypotheses to be tested or explored in the study. If possible formulate these in terms of "alternative hypotheses" i.e. the hypotheses to be demonstrated.

7) The design of the study, whether it is prospective, retrospective, experimental etc.. Reasons for choosing a specific design if several were possible, should be given.
Sample size calculations to justify the proposed size of the study.

8) The data to be collected and its relevance to the objectives and hypotheses. Materials (instruments) and methods to be used for data collection. Operational definitions of concepts (e.g. how is diarrhoea or a low income household defined). Whenever possible (i.e. when they can be made operational and are also relevant to the study at hand), "standard" definitions of concepts (e.g. diarrhoea is the passage of at least 3 loose stools within 24 hours; persistent diarrhoea is a diarrhoea episode of at least 14 days) should be used to facilitate comparison with other studies. Inclusion and exclusion criteria for patients or subjects should be stated and their relevance to the study explained. The way data is to be stored. Examples of questionnaires should be included.

9) Data analysis. How the data is going to be analyzed. Statistical methods to be used. Which software will be used.

Sometimes dummy output (dummy tables) are helpful. Explain (if necessary) why these analyses answer the questions addressed in the study.

10) A time-table or time chart of the study. What is done when and by whom. Explain that the specified time is sufficient. State which operations and actions have been piloted.

11) The budget of the study. This should include:

i) Personnel costs (salaries, fringe benefits, per diem etc.).
ii) Costs for training.
iii) Overhead costs charged by institutions for use of office space, utilities etc.
iv) Insurance premiums.
v) Local and international travel expenses (scientific meetings).
vi) Computer costs (hardware and software).
vii) All instruments, laboratory equipment etc. bought or hired.
viii) Chemicals and other disposables.
ix) Fees of consultants.
x) Stationary.
xi) Publication costs (including interim reports).
xii) Unexpected (e.g. 10% of budget).

All items should be as specific as possible. Quotations should be given. Inflation should be accounted for. It may sometimes be very difficult to make an accurate budget if costs will be made in one currency while the budget (because of the funding agency) has to be in another currency. Consult the funding agency about this problem.

Budget development can be facilitated by so-called "spreadsheet" programs like LOTUS 1-2-3. Proper use of such a program is also very expedient in keeping track of the expenses during the study.

APPENDIX 5: Relational databases

To store effectively data of a complex nature it is highly recommendable to use "relational database" techniques (Date, 1983).

Within a relational database, data are stored in several rectangular files, consisting of records of the same length, i.e. the same number of fields or variables, at least one of which is a key field or variable. The key is unique for that record.
A rectangular data structure however is not "natural" in many situations. For instance if we want to register time, duration, etc. of diarrhoea episodes per child, we could make a record per child which in addition to date of birth, sex, name and other personal data of the child, has a fixed number of sets of fields each describing one diarrhoea episode. However, if this number of sets is large enough for the child with most episodes of diarrhoea to be recorded, then a lot of computer storage is wasted since most children have far fewer episodes of diarrhoea. Alternatively, one could make a record for each diarrhoea episode and enter the personal data on each of these records. This too causes a lot of wasted computer memory space. In addition, if a personal data variable (e.g. date of birth) was originally mis-coded and needs to be changed, then the changing of this variable is quite complex since it has to be carried out on a variable number of records. Sometimes even more important is that it is difficult to check whether the data is consistent i.e. whether the same name has been entered in all records of the same individual.

To avoid this waste of space and other problems and to allow for a variable number of diarrhoea episodes per child, one should bring the data into what is called the first normal form.

First normal form.
In the first normal form the data is stored in two files. One file contains all the personal data of the child (date of birth, sex, name etc.) and a key (this could be the name) which is usually some unique identification number. A second file is made in which each record represents one diarrhoea episode. The key of such a record consists of both the identification number of the child and some variable which is unique for that period of diarrhoea, which could be some special number but could also be the starting date of the diarrhoea episode (no two diarrhoea episodes of one child can start on the same day). Note that this structure allows for any number of diarrhoea episodes per child and that no needless storage space is occupied.

However, it could well be that some of the data in the personal file does not depend on the whole key. For instance, let the key of the child consist of an identification number of the area it lives in, followed by a household identification number followed by a within-household identification number (this is a very common situation). If some of the data in the personal file depends only on the household the child lives in (e.g. number of rooms, cooking facilities) and some of it depends on the area it lives in (e.g. altitude) then again storage space is

wasted, inconsistency may arise etc. To avoid this, data has to be brought into the second normal form.

Second normal form.
Data is in the second normal form if each variable in a record (sometimes rather confusingly called a relation in data base jargon) depends only on the whole key.

In our example we could achieve this by making separate files for areas and households. Since part of the key of a child consists of an area identification number and a household identification number, this information can always be linked to the child's data. Of course, if there is only one child per household, the distinction between child data and household data is unnecessary and may look pedantic.

Closely related to the second normal form of data storage is the third normal form.

Third normal form
Data is in the third normal form if all variables within a record only depend on the key and not on other variables in the record (relation). For instance if households obtain their data from several sources, e.g. a river or a borehole, and the variable "bacterial contamination" depends only on the source of water, then the inclusion of two variables, viz. source of water and bacterial contamination of water, in the household file records violates the third normal form. To bring data into the third normal form, only source of water should appear on the household records and a separate file should be made which relates source of water (the key variable) to its bacterial and other properties.

These three normal forms together guarantee:

1) Efficiency in storage. No disk space is wasted.
2) Modifications and updates of data have to be carried out only once.
3) Data consistency. All children from the same area live at the same altitude. All children drinking from the same borehole are exposed to the same bacterial contamination.

APPENDIX 6: Tests for Poisson distribution

Suppose we want to test whether the observations on diarrhoea incidence come from a Poisson process, i.e. whether all children have the same risk of getting diarrhoea and variations in observed incidence are purely random.

Consider N individuals each observed during a period T (e.g. a year). Let m_i (i=0,1,2,....) denote the number of individuals with i periods of diarrhoea. Clearly $N = m_0 + m_1 + ...$

The maximum likelihood estimator of μ i.e. the Poisson distribution parameter is:

$$\mu = \Sigma i m_i / N \qquad (1)$$

that is the mean number of episode per individual.

The expected number of individuals with i episodes of diarrhoea is

$$E(m_i) = \frac{N\, e^{-\mu}\, \mu^m}{m_i!} \qquad (2)$$

The hypothesis of a poisson distribution can now be tested using,

$$\Sigma\, (m_i - E(m_i))^2 / E(m_i) \qquad (3)$$

which has (approximately) a X^2 distribution with r-2 degrees of freedom, where r denotes the total number of cells (i.e. the number of different values of m_i) used for comparison. If this is large then the test has little power, so either one should leave out cells with a small expected value of m_i, or one should group cells with small expected values.

The above test is very general. It tests against all different kinds of departure from a Poisson distribution. In reality most of these alternatives are not of interest because they do not occur. The only type of alternative we are interested in is the alternative of overdispersion due to the presence of several Poisson distributions with different parameters (a mixture of Poisson distributions or a compound Poisson distribution). Overdispersion is defined by the variance of the distribution being larger than the mean.

This can be tested as follows. Take of each Poisson variate K_i its square root $\sqrt{K_i}$, or rather better $\sqrt{(K_i+3/8)}$. This is a so-called variance stabilizing transformation since the variance (if the K_i are Poisson distributed) of these variables is approximately 1/4.

Then calculate

$$4 \Sigma \{\sqrt{(K_i+3/8)} - \sqrt{(\mu +3/8)}\}^2 \qquad (4)$$

which under the assumption of a Poisson distribution has a X^2 distribution with N degrees of freedom. Large values of this statistic are evidence for overdispersion and hence against a Poisson distribution.

Both the above tests generalize easily to the situation with unequal durations of observation per individual. Let T_i be the duration of observation of the i-th individual. Then using

$$\mu = \Sigma\ i\ m_i / \Sigma\ T_i \qquad (5)$$

and

$$E(m_i) = \Sigma \frac{e^{-\mu T} (\mu T_i)^m}{m_i!} \qquad (6)$$

in (3) instead of (1) and (2) yields the required generalization of the first test for a Poisson distribution.

Generalization of the overdispersion test is even simpler. It is obtained by using

$$4 \Sigma \{\sqrt{(K_i+3/8)} - \sqrt{(\mu T_i+3/8)}\}^2 \qquad (7)$$

as a test statistic.

If a Poisson variable is used as a dependent variable in a regression model, a square root transformation is also indicated to stabilize the variance as unequal variances vitiates the assumptions of the standard linear regression model. However, all kinds of "regression" models for Poisson variates are easily fitted by means of the maximum likelihood method. This goes as follows:

Let K_i be Poisson with parameter μ_i.

Let $\mu_i = g(x_{i1},\ldots,x_{ip};\ \alpha)$

where α is some parameter vector to be estimated and g( ) is some known function of covariates $x_1,\ldots,x_p$ and α. Straightforward differentiation of the loglikelihood function yields as orthogonality relations

$$\Sigma \frac{(g_i-K_i)}{g_i} g_i' = 0 \qquad (8)$$

where g_i' denotes (the vector of) derivatives of g_i with respect to (the elements of) α.

The NONLIN module of SYSTAT can be used for such likelihood maximization.

APPENDIX 7: Derivation of sample size formula

We have seen that the formula for the comparison of two means with the t-test is:

$$n = \frac{2\{k(\alpha)+k(\beta)\}^2\sigma^2}{d^2} \qquad (1)$$

where n is the required sample size per group, σ is the standard deviation of each individual observation, α is the required significance level (probability of incorrectly rejecting the null hypothesis of no difference, i.e. a type I error), β is the probability of incorrectly not rejecting the null hypothesis (i.e. a type II error, note that $1-\beta$ is the power of the test) when the difference between the two means in reality equals d. $k(\alpha)$ and $k(\beta)$ are numbers such that if X is a variate with a standard normal distribution (i.e. one with zero mean and unit standard deviation) then:

$$Pr(X>k(\alpha)) = \alpha \qquad (2)$$

To derive this sample size formula we assume that the number of observations per group n, is large enough to make the mean values per group observations approximately normally distributed with standard deviation (i.e. the standard error of the mean) equal to $\sigma/\sqrt{n}$. The difference between the two means then has a normal distribution with standard deviation $\sigma\sqrt{2}/\sqrt{n}$ and a mean value of zero under the null hypothesis and d under the alternative hypothesis.

To test with a significance level of α, we must check whether

$$\frac{\sqrt{n}\ (x_1 - x_2)}{\sigma\sqrt{2}} > k(\alpha) \qquad (3)$$

or,

$$x_1 - x_2 > \sigma\sqrt{2}\ k(\alpha)/\sqrt{n} \qquad (4)$$

The probability $1-\beta$ under the alternative hypothesis (i.e. the hypothesis that the mean of group 1 is the mean of group 2 + d) that the null hypothesis is rejected is (i.e. the power of the test), is the same as the probability under the null hypothesis that

$$\frac{\sqrt{n}\ (x_1 - x_2 - d)}{\sigma\sqrt{2}} > k(\alpha) - \sqrt{n}\ d/\sigma\sqrt{2} = k(1-\beta) = -k(\beta) \qquad (5)$$

where the latter identity holds in virtue of the symmetry of the normal distribution. Hence,

$$\{k(\alpha) + k(\beta)\}^2 = n\ d^2/2\sigma^2 \qquad (6)$$

from which the formula for the sample size follows immediately.

REFERENCES

Boyd AV (1979): Testing for association of diseases. J Chron Dis 32, 667-672.

Breslow NE,Day NE (1980): Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies. Lyon: International Agency for Research on Cancer.

Briscoe J, Feachem RG, Rahaman MM (1986): Evaluating Health Impact, Water Supply, Sanitation and Hygiene Education. Ottawa: International Development Research Centre.

Briscoe J, Feachem RG, Rahaman MM (1985): Measuring the Impact of Water Supply and Sanitation Facilities on Diarrhoea Morbidity: Prospects for Case-Control Methods. Geneva: World Health Organization (WHO/CWS/85.3/CDD/OPR/85.1).

Cochran WG (1963): Sampling Techniques. New York: Wiley.

Cousens SN, Feachem RG, Kirkwood B, Mertens TE, Smith PG (1988) Case-Control Studies of Childhood Diarrhoea:I. Minimizing Bias. WHO/CDD/EDP/88.2

Cox DR (1970): The Analysis of Binary Data. London: Methuen.

Date CJ (1983): An Introduction to Database Systems. Reading (Massachusetts): Addison-Wesley.

Efron B (1982): The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: SIAM.

Everitt BS (1980): Cluster Analysis. London: Heineman.

Freeman J, Hutchison GB (1980): Prevalence, incidence and duration. Am J Epidemiol 112: 707-723.

Greenacre MJ (1984): Theory and Applications of Correspondence Analysis. New York: Academic Press.

Joreskog KG (1981): Analysis of covariance structures. Scand J Statist 8: 65-92.

Kalbfleisch JD, Prentice RL (1973): The Analysis of Failure Time Data. New York: Wiley.

Kendall MG, Stuart A (1968): The Advanced Theory of Statistics. (Vol 3). London: Griffin.

Lilienfeld AM, Lilienfeld DE (1979): A century of case-control studies: progress? J Chron Dis 32: 5-13.

Quine MP, Seneta,E (1987) Bortkiewicz's Data and the Law of Small Numbers. International Statistical Review 55,2,173-181.

Rao CR (1973): Linear Statistical Inference and Its Applications. New York: Wiley.

Robin S, Spitzer WO, Delmore T, Sackett DL (1978): An empirical demonstration of Berkson's bias. J Chron Dis 31: 119-128.

Sackett DL (1979): Bias in analytic research. J Chron Dis 32: 51-63.

Schlesselman JJ (1982): Case-Control Studies: Design, Conduct, Analysis. New York: Oxford University Press.

Schiffman SS, Reynolds ML, Young FW (1981): Introduction to Multidimensional Scaling: Theory, Methods and Applications. New York: Academic Press.

World Health Organization (1987): Guidelines for Planning Clinical Trials in Diarrhoeal Diseases. Geneva: World Health Organization (WHO/CDD/CMT/87.2).

World Health Organization (1987): Bibliography of Acute Diarrhoeal Diseases. (Vol 7: Nos 1 & 2). Geneva: World Health Organization (WHO/CDD/BIB/87.12 and 87.13).

Some recommended further reading

A classical, and still very interesting for its good coverage of the principles, text on medical statistics is:

Bradford Hill A (1966): Principles of Medical Statistics. 8th ed. London: Lancet.

Another, rather comprehensive book in this field with good coverage of formal statistical methods, is:

Armitage P (1971): Statistical Methods in Medical Research. Oxford: Blackwell.

More up to date than the previous two books, also covering topics like survival analysis and logistic regression is:

Matthews DE, Farewell VT (1985): Using and Understanding Medical Statistics. Basel:Karger.

On epidemiology a good general text is:

MacMahon B, Pugh TF (1970): Epidemiology. Boston: Little Brown and Co.

Two texts on epidemiology more specificly tailored to developing countries are:

McCusker J (1978): Epidemiology in Community Health. Nairobi: AMREF.

Barker DJP (1976): Practical Epidemiology. Edinburgh: Churchill Livingstone.

A good introductory text on the use of microcomputers in research is:

Madron TWM et al (1985): Using Microcomputers in Research. Sage Publications.

The manuals supplied with the SPSS/PC+ statistical package provide some excellent information about problems of coding of data for analysis by computer, as well as useful descriptions of various statistical techniques.

Survey methods, including the design of questionnaires is discussed in:

Moser CA, Kalton G (1979): Survey Methods in Social Investigations. (2nd ed.) London: Heinemann.