<table>
<thead>
<tr>
<th>Serial No</th>
<th>ISBN:978-30958-5-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author 1</td>
<td>AKPALA, C.O.</td>
</tr>
<tr>
<td>Author 2</td>
<td></td>
</tr>
<tr>
<td>Author 3</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Epidemiological Research A Practical Approach for the Medical &amp; Nursing Sciences</td>
</tr>
<tr>
<td>Keywords</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Epidemiological Research A Practical Approach for the Medical &amp; Nursing Sciences</td>
</tr>
<tr>
<td>Category</td>
<td>Medical Sciences</td>
</tr>
<tr>
<td>Publisher</td>
<td>Department of Management, Faculty of Business Administration UNEC</td>
</tr>
<tr>
<td>Publication Date</td>
<td>1994</td>
</tr>
</tbody>
</table>

**Signature**

Digitally signed by Dr Chinwe Nwogo Ezeani
DN: CN = Dr Chinwe Nwogo Ezeani, C = NG, O = University of Nigeria, Nsukka, OU = University Library
Reason: I have reviewed this document
Date: 2008.10.10 14:32:21 +02'00'
EPIDEMIOLOGICAL RESEARCH
A Practical Approach for the Medical and Nursing Sciences
Okey Akpala.

MBBS (Nig); MPH (Glas); FRSH (Lond); FWACP; MFCM (UK)
Senior Lecturer, College of Medicine, University of Nigeria;
Consultant, University of Nigeria Teaching Hospital Enuqu.

Published by the
DEPARTMENT OF MANAGEMENT, FACULTY OF BUSINESS
ADMINISTRATION. UNIVERSITY OF NIGERIA,
ENUGU CAMPUS, ENUGU. 1994.
To My Parents

Professor Agwu Akpala
and
Barrister (Mrs) Sylvia Akpala.
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>ix</td>
</tr>
<tr>
<td>Preface</td>
<td>xi</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>xiii</td>
</tr>
</tbody>
</table>

1. Epidemiological research, health and health care ........................................... 1
2. The research proposal ................................................................. 17
3. Literature review ................................................................. 31
4. Formulating the study objectives .............................................. 39
5. Study population and samples .................................................. 45
6. Sample size determination ...................................................... 57
7. Study Design 1: Descriptive and analytical designs ......................... 66
8. Study Design 2: Experimental or intervention design ......................... 87
9. Data collection ........................................................................ 95
10. Analysis and presentation of data ............................................. 110
11. The research report ................................................................. 130

References 142
Suggested further reading 143
Index 147
FORWARD

This book "Epidemiological Research" is a courageous and highly commendable effort in taking hold of what, to the uninitiated, would seem an unattractive subject matter often cluttered by indigestible definitions and jargon; and converting it into a simple, readable, comprehensible and, indeed, enjoyable little manual. Without sacrificing quality, the bones and sinews have been dissected out of the subject and what is left is a veritable source of quick reference for those interested in the basics and practical aspects of epidemiological research in health.

There is a dearth of indigenous work in this area so this fills a yawning gap which is a welcome development particularly now that the economic dictates of our times have made foreign books financially out of the reach of the average undergraduate or postgraduate student or burgeoning researcher.

To all these people and more the book should find a well deserved place on their bookshelves.

Professor W. O. Chukwudebelu
College of Medicine
University of Nigeria
Enugu Campus, Enugu.
PREFACE

For more than a decade now, there has been a global trend towards the re-orientation of medical and nursing education with greater emphasis on community health and primary health care. Epidemiology, which is a basic science of these disciplines involves the collection, analysis and interpretation of data on the distribution and determinants of health and disease in human populations. This is usually with a view to the application of the resulting information to the solution of health problems facing individuals and communities.

Epidemiological research involves an application of the basic knowledge from three distinct subjects namely: Epidemiology, Biostatistics and Scientific Research methodology. Experience has shown that because these subjects have traditionally been taught as distinct disciplines in isolation to each other, students and even some health professionals have a lot of difficulty in integrating the theory from these three subjects into a practical framework for the design and execution of an epidemiological research project. This practical manual is therefore aimed at bridging this gap by providing an integrated and coordinated approach to epidemiological research.

The book is intended to be read from the beginning to the end. The first chapter introduces the subject by highlighting the role of epidemiological research in the planning and implementation of health care for both individual patients and communities. The reader is subsequently guided through the basic stages involved in the planning and execution of an epidemiological research.
i.e. formulation of the research objectives, review of the relevant literature, study design including sampling methods, data collection techniques, analysis of data and finally the last chapter which deals with the writing of the research report. This last chapter is also expected to act as a guide to the critique of a published biomedical scientific report. Special attention has also been given to the writing of a research proposal and sample size determination.

This book has been written primarily for undergraduate and post graduate students in the medical and nursing sciences. It is however also intended for use as a self-instructional manual by young practitioners in the following professions: medicine, nursing, dentistry, medical laboratory science, public health and primary health care. With the growing trend towards an integration of the biomedical and social science approach to health care, behavioural scientists, social scientists, and demographers will also find the book useful.

C.O. Akpala
ACKNOWLEDGEMENTS

Over the years, I have had the privilege of being consulted by several resident doctors, medical and paramedical students on the best strategy for the successful design, conduct, analysis and reporting of their epidemiological research projects. Initial ideas for this book came from all those students to whom I feel most indebted.

My sincere gratitude goes to several people who have in no small measure contributed to the quality of this book. They include Dr (Mrs) Angela Amah, senior resident doctor in the department of community medicine, University of Nigeria teaching hospital, for her excellent comments on the initial draft of the book. I will also like to thank the following: Tochukwu Nwafor, Edward Nwanegbo, Ifeanyi Nwagbara, Ngozi Osom and Chidi Oranusi. Their crowded academic programme as final year medical students notwithstanding, they were able to read the manuscript and offer suggestions which have immensely contributed to the clarity and simplicity of presentation in the book.

My gratitude also goes to Prof W.O. Chukwudebelu who in spite of his very busy schedule, found time to review and write the forward to the book. I also wish to thank Miss Esther Okonkwo who typed part of the manuscript. Finally, I will not fail to thank my father, Professor Agwu Akpala for his encouragement and guidance in the publication of this book.
INTRODUCTION

Today, even in developing countries, the importance of research in all aspects of individual and national development is no longer in dispute. In Nigeria for instance, this can be attested to by the fact that both undergraduate and postgraduate students are currently required to conduct a research project in part fulfilment of
the requirements for the award of a degree. It is important to note however that this is a far departure from the situation in the not too distant past. Until quite recently, most people in developing countries regarded research as an elitist activity that offered no practical contribution towards the solution of their day-to-day problems. Research was seen as unrewarding and uninteresting while researchers were perceived as eccentric individuals who were most often seen in white laboratory coats and virtually living in isolated offices or laboratories. Although this negative perception of research and researchers was probably attributable to the low level of literacy among the populace, this was further reinforced by the low priority accorded to funding of research by the government. The general apathy and gross neglect of research equally affected the health professionals who considered research activities as an undue and unwelcome intrusion into time meant for patient care. It is gratifying to note that this negative perception of research is gradually disappearing. This is partially attributable to the fact that many more people have acquired western education, especially up to the tertiary level. Most healthcare professionals are today required to undertake a research project either in the course of receiving further training or by virtue of the nature of their jobs. Apart from further training, a major motivating factor for research is the natural joy which follows any achievement in life of which a successful research is one. This motivation may be experienced in so many different ways by the concerned individuals. These include the joy associated with generating new knowledge or applying
existing knowledge towards the solution of the numerous problems afflicting man, the satisfaction of having ones work or name in print in journals, books and magazines or the material reward through promotion which is associated with such achievement. Whatever the motivating factor, an important attribute needed by every researcher to be able to either get a research project off the ground or successfully complete it is enthusiasm. Successful researchers also need to be patient individuals. This is because the research process could sometimes be very slow and unduly protracted either as a result of unforeseen circumstances or because initial ideas need to pass through a number of modifications based on consultations with other researchers and previously published materials. The perceived slow pace of progress of a research project have often resulted in some researchers abandoning the study midway and embarking on a new one which in most cases ends in a similar outcome. These impatient investigators finally abandon research altogether and subsequently end up never completing any successful research project in their lifetime.

DEFINITION

The word research, though in common usage, often means different things to different individuals. It has been
variously described as "A systematic search for new facts; a systematic search for new knowledge, getting additional information and a search for the truth". Although some of these definitions sound more philosophical than scientific, a common recurring theme among them however is the concern that the research process need to be systematic and also should be based on the scientific method. In the scientific method, a problem needing solution is identified, a hypothesis is postulated, observations or data collected and interpreted to prove or disprove the hypothesis. This ultimately leads to a new body of knowledge. In effect, the research process also follows well defined steps or stages.

Research may be defined as the identification of a problem and proffering solutions through a planned and systematic collection, analysis and interpretation of data.

**BENEFITS OF RESEARCH**

The benefits of research to mankind are several. As individuals, research can improve the quality of our lives by broadening the horizon in our jobs through a wider appreciation of the subject area. Research involves a rational and logical interpretation of a mass of information. Through the research process, we are therefore able to improve our thinking process through a critical assessment of issues and facts that we have acquired or received. This leads to better judgements in our day to day activities.
Those involved in research may invariably need to do a lot of reading and handling of figures, an activity that helps to improve one's ability to quickly scan through volumes of documents and immediately deduce the relevant and useful information for decision making.

There is no doubt that research has contributed immensely to the improvement of our knowledge of various disciplines and professions such as Law, Business, Engineering and Medicine. With particular reference to medicine, research has had tremendous impact on the improvement and maintenance of the health status of both individuals and whole communities. Medical research have enabled health professionals to acquire new knowledge on the disease process and its causes in addition to new interventions and technology required to promote health, prevent or cure disease. Associated with the rapid growth and developments in medical knowledge is the springing up of new medical specialities of which epidemiology is one example.

EPIDEMIOLOGICAL RESEARCH

Epidemiological research is concerned with the study of the distribution and determinants of health and disease and related states or events in human populations. This is with a view to the application of the findings towards the solution of community health problems. Epidemiological
research essentially deals with the investigation of health and health-related problems of groups or populations rather than individuals. The word "Epidemiology" is derived from the term epidemic, which itself has a Greek origin and means "studies upon the people". (Epi = Upon; Demos = People). The discipline initially was concerned purely with the investigation of outbreaks or epidemics of infectious diseases in the population. Today, the scope of epidemiology has broadened and the same principles are applied not just to communicable diseases but also non-communicable diseases and all health related events such as road traffic accidents, drug abuse, and abortion.

The rapid growth of epidemiology is intricately interwoven with the series of historical developments through which the medical sciences have passed through. First was the era concerned with the improvement of basic environmental factors such as sanitation, housing and working conditions. This approach to the improvement of health originated from the belief during the period that disease occurrence was more as a result of environmental factors. Infact, Hippocrates was the first to highlight this fact in his study on "Airs, waters and places". At that time, the occurrence of disease epidemics was explained through the "Miasma theory", which attributed most illnesses to the breathing of bad quality air. Malaria was then believed to occur as a result of foul air and in fact derived its name from this concept. (Malaria means "Bad air"). The next era was the period associated with the germ theory of disease.
which medical technology was applied to the detection of a specific germ or agent responsible for a disease for subsequent elimination. This so called period of lab-medicine was also accompanied by a rapid growth and development of very expensive high technology equipment for the diagnosis and treatment of disease in individual patients. This trend was most evident in the United States of America (USA) where the Gross National Product spent on health surpassed that of most other developed countries. In spite of the huge health budget, there was no evidence to show that the standard of health in the USA was better than those of other developed countries with much lower health budgets.

The limitations and diminishing returns associated with this system of health care in dealing with current health problems has therefore shifted interest back to the identification of preventable factors as a means of dealing with a wide range of health problems associated with disease occurrence. This concept of health care has further been enhanced by the growing awareness among health care providers that resources are scarce and limited, and therefore need to be equitably distributed to ensure not just the health of a few individuals but that of the highest possible number of people in the entire community, with special attention to those at risk.

One of the consequences of this reorientation of health care and consequently medical research, is that health professionals should be able to understand the
rudiments of epidemiological research. This is necessary to enable them to identify and offer solutions to priority health problems in their local communities. This is particularly important for developing third world countries which cannot afford the huge material and human resources for high technology medical research. In these countries, the potentials for epidemiological research in contributing to the improvement of health and health care cannot be over emphasised. Even if health professionals cannot design and conduct sophisticated epidemiological research, it is still important that they should have an understanding of the subject to be able to critically evaluate the results and conclusions of other researchers. This is particularly relevant to health care managers who may from time to many people and involving a commitment of enormous amount of resources based on new scientific knowledge from an epidemiological research finding. A second implication of the reorientation in health care is that medical education at both undergraduate and post-graduate levels need to be restructured to take into account the immense potentials of epidemiological research in resolving the peculiar health problems in most developing countries.

CONTRIBUTIONS OF EPIDEMIOLOGICAL RESEARCH TO HEALTH & HEALTH CARE

The ultimate aim of epidemiological research is to provide.
information necessary for the prescription of appropriate measures to reduce disease and promote health in the entire population. This information is utilized mainly by the medical specialty of community medicine, a discipline concerned with the identification of health problems of a community, group or defined population and subsequently organizing, providing and evaluating health services aimed at solving these problems. Epidemiology is therefore the basic science of community medicine just as anatomy, physiology and biochemistry are basic sciences of clinical medicine. The contributions of epidemiological research to the health of individuals and communities can be summarized as follows:

1. To describe the health status of a population

The process of describing the health status of a community is sometimes referred to as community diagnosis. Just as in clinical medicine where an accurate diagnosis need to be made before proceeding to treat a patient, a proper diagnosis of the health problems of a community need to be established as a basis for the institution of appropriate measures to deal with the problems.

Community diagnosis is also important for the identification of groups within the population in need of various types of health care. A fundamental problem associated with community diagnosis is a
lack of consensus on the criteria for defining and measuring health, a concept that has received varied interpretations by different groups. Perceptions of health may range from an absence of pathology from a doctor's point of view to absence of illness as perceived by the patient. This problem has to some extent been reduced by the World Health Organisation's definition of health. It defines health as a state of complete physical, mental and social wellbeing, rather than just the absence of disease or infirmity. Despite its general acceptance among health professionals, this definition is not very helpful in the operational measurement of health. It has therefore become customary to measure health status indirectly based on the absence of disease. Indices such as rates of morbidity, mortality or disability are used to describe health status. These information are provided through epidemiological research and is very useful in the planning of health services. In most developing countries, health service resources are allocated based on political or sentimental considerations without reference to prevalent diseases. The result is that such services provided are not utilized by the communities for whom they have been provided, mainly because their priority health problems were not addressed. In more developed countries, it has come to be...
appreciated that resources are scarce and limited while the demand for health care is unlimited. Equitable and rational allocation of resources is therefore based on sound epidemiological information. The provision of hospital beds for preterm babies is for instance not based on the fact that the commissioner for health is a pediatrician. The number of women of childbearing age, number of deliveries per year and incidence of preterm babies by geographical area are some of the epidemiological factors that need to be taken into consideration. This ensures that services reach those in greatest need for them.

2. Describing the Natural History of Disease and Completing the clinical picture

Doctors in hospitals and other health institutions come across relatively small numbers of patients with any particular disease especially the rare ones. In addition, these patients are seen at different stages of the disease. Epidemiological studies enable us to describe the full course of a disease through the follow-up of individuals from the asymptomatic but pathological stage to its outcome which may be recovery, death, or disability. Based on this, health workers can predict the probability of a given outcome which in practice is stated as statements on
prognosis is all about. The prognosis can also be related to other factors such as age and sex.

The natural history of the disease also enables health workers to determine at what stage interventions are likely to be of benefit to the patient. An understanding of the natural history of carcinoma of the cervix for example is the basis for intervention at the early stages of carcinoma in situ.

3. Calculating the Risk of Disease for an Individual

Based on the number of new cases or existing cases of a disease in a given defined population, health workers can estimate the risk or probability that any individual from that community will acquire a particular disease. Statements on disease incidence rate per thousand of the population is in practice a description of the risk to an individual of acquiring a disease or a health related event within a given community.

4. To determine Disease etiology

Epidemiological studies enable us to develop and test hypothesis regarding the causes of a disease or it's associated factors. This is the basis for the formulation of policies and strategies for the
prevention of disease. This is particularly important in diseases where treatment is either non-existent, inadequate or very expensive.

5. Describing Local disease patterns

A description of the local disease pattern also provides useful information which is of assistance to clinicians in making accurate diagnosis. It assists them in assessing the probability that a particular group of symptoms and signs are due to a particular disease. It is therefore very important that doctors who have arrived in a new locality which is quite different from their previous one (e.g. from temperate to tropical climate) first study published reports on the local disease patterns. This will for instance save a doctor who has recently arrived in Nigeria from Japan, the embarrassment of diagnosing malaria as a case of murine typhus. A knowledge of the local disease pattern is also useful for the investigation of disease epidemics since there is a need to first establish the usual baseline level of the disease before ascertaining that an epidemic has occurred.

Describing diseases in terms of distribution enable health workers to formulate a hypothesis regarding possible environmental factors which could account for the disease. This information...
also applied in migratory studies. Individuals who are newly resident in a geographical area will exhibit disease rates similar to their previous area of abode if the disease is genetic or cultural. The opposite will occur if the disease is environmental in origin.

6. Evaluation of health services

There are still a number of clinical services being offered to patients today without any study to support their effectiveness. It is considered ethically wrong to provide a service whether preventive or curative without establishing their effectiveness and safety. This can be achieved through either a carefully designed experimental epidemiological study or by collecting data on changing disease frequency in relation to the service provided.

Some of the individuals who have in the past made notable contributions to the development of epidemiological research include the following:

1. LIND 1747 - Established the relationship between the administration of lime to sea-men and prevention of scurvy. Out of twelve sailors with scurvy, he administered oranges and lemons to some of the sailors while the control
group were given food that did not contain citric acid fruits. Those sailors who received orange and lime showed a marked improvement in their condition and this subsequently formed the basis for an official British government policy then, of using lime to prevent scurvy among sailors in the British navy.

2. Ramazzini
1633
He is otherwise referred to as the father of occupational medicine. He studied and described the association between certain diseases and occupation.

3. John Snow
1850
Established that cholera was a water-borne disease many years before Robert Koch identified the causative organism as the vibrio cholera. Snow investigated the London cholera epidemic by comparing mortality rates from cholera among the residents of London in relation to their source of drinking water. He was able to demonstrate a much higher rate among those who drank contaminated water supplied by the
Southwark and Vauxhall companies compared to those supplied by the Lambert company.

4. Doll & Hill
1950

Conducted the classical studies on smoking and lung cancer. One of their studies was among British doctors who were followed up for many years. Doll and Hill were able to establish not only an association between smoking and lung cancer but also a dose-response relationship.
INTRODUCTION

It is not unusual for some researchers especially beginners to proceed from a vague idea of a research problem to field data collection with the hope that at the end of the exercise, a suitable research report will materialise. Such investigators realize only too late after investing so much energy, time, money and materials that this is an exercise in futility. The exercise usually ends up in failure simply
because it lacked a sense of direction. There was no plan of action showing what the researcher intended to achieve and the methods to be adopted to reach the expected target.

Like most endeavours in life, success in research depends to a large extent on detailed and careful planning. Although planning a research project can sometimes be more time consuming than its actual implementation, this always turns out to be time well spent. It has been likened to the time spent drawing the plan for a building. This is the time when decisions on the nature, size and type of building being envisaged are taken. No builder will commence construction without going through this planning phase because such crucial decisions cannot be taken on an ad hoc basis as the construction progresses. Similarly, after deciding on a topic for study, the researcher needs to carefully write down the detailed plan of the proposed research project. The resulting document is called the Research Proposal or Research Protocol. Just as in an architectural drawing of a building plan, the research proposal is expected to be strictly adhered to by the researcher as a guide to the successful completion of the project. Another reason for the need for a research proposal is that it is required by funding agencies to enable them assess whether the investigator has asked the relevant research questions and the suitability of the proposed research design. A research proposal also forms the basis for seeking advice or approval for the proposed study from academic supervisors, colleagues, medical statisticians and examining bodies.

IDENTIFICATION AND SELECTION OF A RESEARCH TOPIC

Selecting a suitable topic for research is the first step in the
research process. This first step can sometimes become the most difficult and time consuming especially for beginners. Choosing a topic for an epidemiological research effectively commences with the discovery of a health problem whose solution depends on the answers provided by an appropriate study. Fortunately, a researcher does not always have to go through the rigorous process of problem identification. This task may have been carried out by health managers, health workers, politicians, community leaders or indeed anybody concerned with the health and welfare of any community. Some typical examples include situations in which a community leader has complained to health authorities that many women are dying after delivery at the newly built village maternity centre, a politician raises questions on why the hospital he has worked so hard to attract to the town is not being utilized by the community, or even a schoolteacher who may have discovered a very high rate of absenteeism among the school children due to guineaworm infection and has requested the health ministry for assistance in dealing with the problem. International health agencies may also from time to time express an interest in funding research on specific topics or health problems. In all these situations, all that is left for the researcher to do is to reframe the already identified problems into research topics and then proceed to design an appropriate study as a basis for finding solutions to the problems. This scenario is the exception rather than the rule, especially among academicians and students who will be required to search for and identify a researchable health problem, select a topic and subsequently design and conduct a study whose findings will assist in solving the problem.

Identification of a research problem usually starts as a vague idea which with time is finally nurtured and
refined into a definite research problem and topic for study. Generating ideas for research is not an easy task and requires a high sense of imagination and creativity. Unfortunately some individuals are not as gifted as others in this regard. This does not however necessarily mean that these attributes cannot be cultivated. A keen sense of observation, curiosity or questioning the obvious and conventional beliefs and practices are some of the qualities which can enhance an individual's ability to easily identify and select problems for research. These qualities can be put in practice while reading, observing events at work or in the community and during discussions with patients, colleagues and other members of the public.

A health problem is said to arise when there is a discrepancy between what is observed and what is expected. A clinician may in the course of his medical practice for instance, notice that many more patients than usual are turning up with gastroenteritis. He is likely to ask himself the reason for the high rate, why it is occurring at that particular period, whether anything happened either in the environment or among the people. This could then lead to a research study on "The factors influencing the occurrence of gastroenteritis" in the named community. Similarly, during a visit to a health centre a researcher might discover from the records that very few deliveries had been recorded. On further enquiries, he is told that women prefer to have their deliveries at home instead of at the health centre. He is likely to ask himself the reasons for "non-utilization of health services" in the village.

There are various ways in which an investigator can conceive a research topic while reading through published literature. In the course of reading through a publication on a research study, a researcher may for example, discover that the article identified other aspects of the problem that
will merit further investigation. This is usually found in the concluding part of a research report. The same published report could also provide the researcher with ideas on similar problems in his area of interest. As an illustration, a study on poor compliance with anti-tuberculosis therapy could draw the attention of a researcher involved in the management of diabetic patients to a similar problem among his patients. Finally, a review of different research reports on a health problem may show discrepancies in findings. Such studies are worth replicating to establish their validity.

Another way an investigator can use published literature as a source of ideas for a research topic is by starting from a specific subject area and then proceeding to look up published materials in the same area. The Index Medicus is a good source of published materials arranged according to subject areas and covering all aspects of health.

After identifying a health problem, the next step is to reframe the problem in the form of a research question, the answer of which provides the basis for addressing the problem. As an illustration, if the problem is that many women are dying during childbirth, the research question could be "Why are women dying during childbirth?". This is a pointer to the purpose of the study—To identify the main causes of maternal mortality. This forms the basis for a topic: "The factors influencing high maternal mortality". This title is provisional, and as the planning progresses, it is further refined until the final title for study is obtained.

Many beginners in research are often disappointed when the research topic they had spent so much time in choosing does not receive approval from their supervisors, examination bodies, or funding agencies. This can be minimized if the following criteria are borne in mind when
selecting a topic for study.

a. Feasibility:
The research problem selected for study should be one that is both possible and practicable to accomplish within the given timeframe and in relation to the available manpower, equipment and money. A research study to investigate a given problem could turn out to be too expensive and time consuming for instance either because the population or geographical area to be covered is too large or because the disease under investigation is rare. A study may sometimes not be feasible because the investigator is believed to lack the necessary qualification and experience to mount such a study.

b. Relevance and Importance of the Problem:
A topic selected for study should reflect an important health problem which is in addition relevant to current issues in health care. A health problem may be deemed to be important by virtue of the fact that it has a high morbidity or mortality rate, because it has a marked socio-economic impact on the community or because it constitutes a drain on health care resources. A health problem should be perceived as important by either the community or health care providers, but preferably by both. The anticipated results of the study should help in an overall improvement of the health
status of the community in which it is conducted. It is also important to verify that the study has not previously been carried out in that environment.

WRITING THE RESEARCH PROPOSAL

The research proposal is made up of different sections which include a description of the nature of the problem which is also referred to as a statement of the problem, a justification for the study, its aims and objectives, a brief review of the relevant literature, a description of the population to be investigated, the study design, plan for data collection and analysis, the timetable for the project and an estimate of its total cost. This chapter presents an outline of the component sections of a research proposal. Details of the contents of each of the sections will be discussed in subsequent chapters.

1. Statement of the problem

After the title, the first major section of the research proposal is the statement of the problem. In this section the researcher thinks the problem merits investigation and finally the anticipated benefits of the results. Defining a research problem involves a description of situation in which there is a departure from normal or the expected. This can easily be worked out from the title of the research project. As an example, in a project titled "Factors influencing infant mortality in Enugu", it is obvious that the research problem is that children are dying before the
age of one year when they should not or that the number of deaths among infants is beyond what is expected. Defining the problem in this case may take the form of highlighting the fact that children in Enugu rarely reach the age of one year when compared with those in another town or country.

In some research proposals there is a separate section on the rationale or justification for the study. One of the main reasons why a researcher will want to investigate a problem is because it is believed to be important. This importance may be from the point of view that either many people or a wide geographical area is affected or that it is associated with a very high mortality rate. A problem may also be important because it has serious socio-economic consequences on either the community or the health care system. Another justification for embarking on a study could also be the lack or inadequacy of information to tackle the problem. This can be done by critically analysing what is already known about the problem and the area that calls for further study. This section is concluded by pointing out the benefits to health and health care of the anticipated results from the study. This is particularly important to funding agencies considering the proposal because they are interested in seeing how the results from the proposed study will improve the work of their organization.

2. The literature review

This section gives a brief description of some published studies related to the one under consideration. The location, methodology and findings in these studies are highlighted. Important relationships between these previous studies and
the proposed one are discussed especially in relation to deficiencies or gaps in knowledge which the proposed study will address.

3. Aims & objectives of the study

This section sets out the research questions which the study will answer. It describes the purpose of the study. The aims are stated first in general terms and then as specific measurable objectives.

4. Methodology and plan for data analysis

The study design for the proposed research is described in this section. This also includes a description of the geographical area and population where the study will be conducted, the sampling method and sample size. The logistics and method of data collection is described and finally the type and method of data analysis. Very detailed proposals will also contain dummy tables which show details of the type of analysis envisaged to provide answers to the research question.

5. Schedule or time-table

Every research project should have a definite time schedule for completion to enable a timely utilization of the results. Health managers may for instance require answers to their questions within a reasonable time frame for effective decision making. A student preparing a research proposal for approval by an examination body has a target date in
mind and will need to convince himself and the examiners that his proposed research can be completed and a report submitted within this timeframe. Finally, proposals for submission to funding agencies need to specify a schedule within which the project will be completed. This schedule provides the agencies with a framework for the monitoring and evaluation of the progress and success of the project especially in relation to the release of funds.

The schedule for the research project consists of all the tasks previously described in the proposal which are now listed in the order they will be carried out with the dates and duration of each task. Some funding agencies may require specific details of manpower responsible for each task. A work schedule is better illustrated in a tabular or graphical form. In the graphical presentation which is also known as a GANTT CHART, a bar covering both the commencement and termination dates of each task is used to depict the expected duration of the task. A table of workschedule is similar to the gantt chart but instead of bars, the actual dates for the tasks are specified. An illustration of a work schedule is shown below.

<table>
<thead>
<tr>
<th>TASK</th>
<th>JAN.</th>
<th>FEB.</th>
<th>MAR.</th>
<th>APR.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Obtaining Permission</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Training of Interviewers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Pilot test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Field work</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Report Writing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Every research project needs to be tailored to match the available resources. Most funding agencies calling for proposals will usually state the limits of their funding. The budget is an important guide to the scope of the proposed study. With the identification of all the tasks involved in the completion of the project in terms of the necessary manpower, materials and the duration of each task, a detailed budget can be prepared. Like the work schedule, the budget is also presented in a tabular form with the resources grouped into four main categories: Manpower; Travels; Consumables and Equipment. Each budgeted item needs to be justified under a sub-heading titled budget justification. The researcher needs to make a good case on why money should be spent on each of the listed items. This is very important as a means of avoiding a drastic cut in the proposed budget especially in situations where the researcher will not be present to defend the budget before the appropriate committee. Funding agencies may for instance not appreciate the importance of an electricity generator because they are not accustomed to power failures in their country. Irregular public power supply can therefore be used by the researcher as a justification for the need for an electricity generator. An illustration of the budget for a proposed research is shown below.
## Budget

<table>
<thead>
<tr>
<th>BUDGET CATEGORY</th>
<th>UNIT COST</th>
<th>TOTAL COST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. PERSONNEL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Five Interviewers</td>
<td>N20,000</td>
<td></td>
</tr>
<tr>
<td>Ten Contact persons</td>
<td>N10,000</td>
<td></td>
</tr>
<tr>
<td><strong>SUB TOTAL</strong></td>
<td>N30,000</td>
<td></td>
</tr>
<tr>
<td><strong>2. EQUIPMENTS</strong></td>
<td>N50,000</td>
<td>N100,000</td>
</tr>
<tr>
<td>Two Microcomputers</td>
<td>N2,000</td>
<td>N10,000</td>
</tr>
<tr>
<td>Six Bicycles</td>
<td>N100,000</td>
<td>N300,000</td>
</tr>
<tr>
<td>One Car</td>
<td>N100,000</td>
<td>N300,000</td>
</tr>
<tr>
<td><strong>SUB TOTAL</strong></td>
<td>N160,000</td>
<td>N300,000</td>
</tr>
<tr>
<td><strong>3. SUPPLIES</strong></td>
<td>N15,000</td>
<td>N100,000</td>
</tr>
<tr>
<td>Fifty Reams paper</td>
<td>N100</td>
<td>N100,000</td>
</tr>
<tr>
<td>Three hundred syringes</td>
<td>N100</td>
<td>N100,000</td>
</tr>
<tr>
<td><strong>SUB TOTAL</strong></td>
<td>N180,000</td>
<td>N100,000</td>
</tr>
<tr>
<td><strong>4. TRANSPORT/ TRAVELS</strong></td>
<td>N20,000</td>
<td>N40,000</td>
</tr>
<tr>
<td>Accomodation x 20 Nights</td>
<td>N200,000</td>
<td>N40,000</td>
</tr>
<tr>
<td>Air Ticket (Enugu-Lag x 20)</td>
<td>N200,000</td>
<td>N40,000</td>
</tr>
<tr>
<td><strong>SUB TOTAL</strong></td>
<td>N260,000</td>
<td>N50,000</td>
</tr>
<tr>
<td><strong>GRAND TOTAL</strong></td>
<td>N289,000</td>
<td></td>
</tr>
</tbody>
</table>

---

7. **References**

Published literature which have been referred to by an investigator in the course of the research are usually cited in both the proposal and the final report. This is aimed at providing interested readers with sufficient information to enable them locate and retrieve the publications for more detailed study. The two major systems of references commonly used in biomedical journals are the Vancouver and Harvard systems. In the Harvard system, references are cited in the text using the name of author(s) and date of publication, for example: "Tuberculosis is one of the major..."
causes of mortality in Nigeria (Smith and Adewole, 1993). At the end of the text, all the references cited are listed in alphabetical order. Each reference should contain the following information: Name of author(s) with initials, year of publication, Title, Journal, Volume, Issue and Pages. The article by Smith cited above will appear as: Smith, C and Adewole, K. (1993) "Factors associated with occurrence of tuberculosis in Oyo state, Nigeria". Annals of Medicine, 62, 1, 11-18. If the reference is a book instead of a journal article, the following information are included: Authors name, Date of publication, Title, Publishers name, Place of publication and pages. For example, Smith (1993) An Introduction to Occupational Medicine. Butterworths, London, 6-9.

The Vancouver system of references differs from the Havard system by virtue of the fact that references are cited in the text using sequential numbering in the order in which the references were first mentioned. At the end of the text, the references are listed chronologically as they have appeared in the text. If the article by Smith which was previously used as an illustration were to be the first article cited in a text in the Vancouver system, it will appear as: Tuberculosis is one of the major causes of mortality in Nigeria (1). At the end of the text, the reference will be listed with the following information: Name of author(s), Title, Journal, Year of publication, Volume, Issue and Pages. The article by smith will therefore be listed as:


The names of all the authors are listed when six or less. If there are more than six, only three are listed and the
prefix "et al." added. The format for other types of publications other than journal articles are shown below.

a. Author of a Book.

Name of author(s), Title, Edition, Place of publication, Name of publisher, Year of publication and Page.

b. Editor or Compiler of a Book.

Name of editor(s) followed by "eds.", title, place of publication, name of publisher, year of publication and page(s).


Name of author(s) of the chapter, title of chapter, Name of author(s) of the book (preceded by the word "In:"), title of book, place of publication, name of publisher, year of publication and page(s).
The literature review is one of the most important aspects in the planning of a research project. As the term implies, it involves a critical re-examination of what has been
published in the past in relation to the research project an investigator is proposing to embark on. This review serves many useful functions. Firstly it is very common for researchers, especially beginners to assume that their research idea has never been previously considered by other investigators. After investing so much time and other resources in the study, they discover from the medical literature that the topic had actually been overflogged by other researchers. The literature review therefore helps the investigator to avoid a duplication of studies previously conducted by other researchers in the same environment. Related to this is the fact that at the end of the literature review, an investigator should be in a position to identify gaps in knowledge on the topic. This could serve as a rationale for embarking on the study. Research should as a matter of principle, contribute new knowledge rather than duplicate on what is already known.

At the initial conception of a research project, some of the ideas on what is to be investigated and how to go about it remain vague in the researchers mind until the stage of literature review. This review throws more light on the possible methodologies for investigating the problem including the advantages and disadvantages of different methods based on the past experiences of other researchers. The feasibility of the proposed research in terms of time, manpower and money could also be ascertained from the review of published reports of similar studies. The researcher is therefore able to fine tune the scope of the problem to be investigated, adopting an appropriate methodology and striking out aspects of the problem that may not be easily amenable to investigation.

A review of previously published reports will show that some articles end with a recommendation on some aspects of the problem that merit further investigation.
This can serve as either a source of other research topics or ideas on how to redirect the proposed research towards producing better results. The literature review on its own is an educational material for health professionals by keeping them abreast of new developments in their areas of interest.

ASSEMBLING MATERIALS FOR LITERATURE REVIEW

An important requirement necessary for an investigator to be able to successfully collect materials for literature review is a knowledge of the various sources of published materials and where they can be located. In terms of location, most of the materials needed for a literature review will usually be found in libraries, which could be departmental, university or national libraries. Libraries of international health agencies may also be found useful. In addition to libraries, consulting senior health professionals or researchers may yield very useful material.

From the foregoing, it is obvious that every researcher needs to be conversant with how to use the library to access information needed for a literature review. Speaking to a librarian can be quite helpful but accessing the library materials personally saves an investigator a lot of valuable time which would have been spent on the examination of piles of materials which at the end might even turn out to be irrelevant to the proposed study. The librarian can however assist an investigator with the system of classification and cataloguing of the materials in the library to enable the researcher retrieve them quickly. Published materials stocked in standard medical libraries are mostly in the form of textbooks, journals, research
reports, conference proceedings, thesis and official publications of various health organizations. Among all these, journal articles contain the most recent and current information especially if the journal has a high frequency of publication such as weekly. Published articles usually have a list of references on related studies at the end. This also provides an invaluable source of other related literature especially those contained in local journals, that may not be indexed or abstracted. The more recent the publication and especially if it is well researched, the greater the chances of accumulating most of the relevant related literature.

So much literature is generated everyday within the medical profession that no library can boast of stocking all the books, journals and other publications that may be needed by every researcher. To ease access to the vast accumulation of medical literature, these publications have been compiled as indexes or abstracts which are located in good medical libraries. The most popular of the indexes is the INDEX MEDICUS which is published by the national library of medicine in the United States of America. It contains details of about 300 journal articles every month. A list of journal articles is provided in two separate sections - subject and author, each of which is listed alphabetically. This in effect means that if a researcher has either the authors' name or the title of an article, using the appropriate volumes of the index medicus, all related articles needed to write a literature review can be traced. Having identified the relevant articles the next step will be to search within the library and retrieve the appropriate volume of the journal containing the articles.

Some typical examples of publications of abstracts of articles appearing in the medical literature are the EXERPTA MEDICA and the QUARTERLY.
BIBLIOGRAPHY OF TROPICAL DISEASES: The first contains all types of medical literature and appears in separate subject areas for ease of reference. The second is a specialized publication limited to some tropical infectious diseases and is divided into two sections - subject and author. Most of the abstracts or summaries contained and these publications are detailed enough to enable a researcher write a literature review without further reference to the full article. Computerization of these abstracts have markedly reduced the time and energy spent by researchers on literature search. In the recent past, most standard medical libraries have acquired facilities for computerized literature search. An investigator can therefore in a short time have a computer print out of summaries or abstracts of publications which are relevant to his proposed study. A sample of a page in the author section of the quarterly bibliography of tropical diseases is shown below.


Malaria is a parasitic disease about which there is much biomedical knowledge on causation, prevention, treatment and control. Attempts in eradication, as well as control in the past have been mainly a technical and biomedical endeavour. With the policy shift from worldwide eradication of malaria to control up to primary health care, there is increasing interest in dealing with possible determinants of the problem both at local as well as regional levels as part of the search for an effective intervention. This paper presents the results of a study into community perceptions and practices relating to causation, treatment and prevention of malaria in a rural Adangbe farming community in Southern Ghana. Malaria is common in this community. Crude parasite rates among adolescent girls (10-19 years old) in the community were 49% towards the end of the major rainy season, and 47% in the early dry season. The symptoms and signs of malaria are rudely described by lay people as well as traditional healers. Diagnosis and treatment of uncomplicated episodes of malaria at home, according to biomedical perceptions, is the predominant behaviour in this community. Very few of uncomplicated malaria are brought to health facilities. Ethnomedical perceptions and practices relating to causation and treatment of malaria in this community are complex and multifactorial. The following morphological lesions were observed: (1) 62.1%: signs -1 (21.71%); (2) 50.4%: signs -2 (14.54%); (3) 47.7%: signs -3 (14.29%); (4) 76.2%: signs -4 (14.29%); (5) 62.1%: signs -5 (14.29%); (6) 50.4%: signs -6 (14.29%).
Many students have found themselves stuck and confused after amassing volumes of published materials. They are at a loss as to how to proceed with the literature review even after reading most of the collected materials. To make a success of the literature review, there are a number of basic steps which may be followed either sequentially or simultaneously.

Just as the *Index Medicus* and publications of abstracts have been designed to assist the investigator in the writing of the literature review, a researcher will first need to organize the assembled materials as abstracts or summaries on an index card. One sheet of quarto sized paper could be used to summarise each article. It is however preferable to use thick cardboard paper cut to a similar size. This can better withstand repeated handling since the summaries may need to be consulted more than once and even for future projects. The index cards may be arranged alphabetically for ease of reference. Each index card should contain a minimum of the following information: Author(s) name with the surname first, Title of article, Name of journal, year of publication, volume, issue number and pages. Eg Rink, B; Swan, S; and Swansfield S. "Communicating coverage data with non-literate communities: beans, sticks or pie charts?" Health Policy and Planning; 1993, 8(1): 57-60. The remaining space on the card is left for a summary of the journal article as illustrated in the next page.
It is important to bear in mind that since the summary is purposely to assist the researcher in writing a literature review, only aspects of the article which are relevant to the proposed study should be included on the index card. It may also be useful to summarise the articles using appropriate headings for rapid and easy identification of its contents. Examples of headings that could be used are:
objectives of the study, methodology and conclusion. After assembling all the materials on index cards, the researcher can then proceed to write the literature review.
The word objective, is one of the research terminologies which have origins in the military where it is used to refer to a point or target which an operation is expected to reach.
or achieve. The military will not set out on any exercise without clearly defining their objective and the operation will be assumed to have come to a successful end only if this objective has been achieved.

Setting an objective and aiming to accomplish it is not however peculiar only to the military. In our day to day living, we have come to appreciate the need for a clarification of the purpose or aim of any activity we wish to embark upon. Although sometimes these objectives may not be explicitly stated, they still remain in our subconscious as a guide toward the successful accomplishment of a programme. In fact, success is actually the attainment of one's set objectives.

**DEFINITION**

An objective can be described as a statement of what we aim to accomplish or the target we aim to reach at the end of a given activity or programme. The success of a research project depends to a very large extent on a clear statement of its objective(s). It describes the purpose of the study or the specific questions which the study is expected to answer. A health problem usually poses some questions whose answers are provided by a research project or study. The objectives of a study are statements of the questions which the study is expected to answer. In fact, some researchers prefer to state their objectives in the form of study questions. As an example, a study on the health status of a community could have its objective as determining the number of infant deaths and the factors influencing it. Another research might pose this as a study question such as: "How many infants die in this community every year and what are the major causes of death?" The study objectives describe the knowledge or
information which the researcher wishes to acquire at the end of the study. It can be easily appreciated that the type of data to be collected and how we collect them is guided by the study objectives. A statement of the objectives at the beginning of the study also saves a researcher the misfortune of dissipating energy in collecting and analysing data that will eventually not assist him in providing answers to a particular health problem.

GENERAL AND SPECIFIC OBJECTIVES

The objectives of a study are in two main parts. The first part is called the general objective or sometimes referred to as the aim or purpose of the study. It describes in very broad terms what the study is set out to achieve. In the example of the study on the health status of a given community, the general objective might be stated as: "An assessment of the patterns of mortality and factors influencing it".

The second part of the statement of the objective describes the specific objectives of the study. These are numbered serially and outline the component sub-sections of the general objective which are now stated in operational, measurable or quantifiable terms. This serves as a guide to the type of data to be collected to answer specific questions in the study. As an example, the study aimed at assessing the health needs of a community may be broken down into the following specific objectives.

1. To determine the total number of deaths in the community.
2. To determine the crude mortality rate.
3. To establish the demographic factors associated with infant mortality.
4. To establish the major causes of death in the community.

In another study on "The smoking habits of secondary school children in Enugu", the specific objectives may be stated as follows:

1. To establish the proportion of school children who are smokers.
2. To establish the average age of commencement of smoking.
3. To determine the average number of cigarettes smoked per day.
4. To establish the main reasons for smoking.

It is clear from the specific objectives above, that the initial general objective of the study has now been broken down into quantifiable or operational forms. As a guide, the specific objectives could be derived from the general objective by breaking it down into its main components including key factors that influence it. A study whose general objective is to assess the health status of a community can be broken down into the following components of health status:

1. Number of sick people.
2. Number of deaths in the community.
3. Number of disabled people in the community.
4. Characteristics of the sick people, eg. age, sex, occupation.
5. Proportion of the population utilizing health facilities.
6. Levels of environmental hazards in the community.
HYPOTHESIS

Objectives can also sometimes be framed in the form of a hypothesis. This is common in analytical studies which are aimed at assessing the relationship between cause and effect. A hypothesis is a supposition which is tested and ultimately either accepted or rejected. The supposition is tested based on the collection and analysis of data which will eventually either support or disprove the hypothesis. In the study of the health status of a community, one of the objectives might be to assess the influence of the use of health services on their health status. When this is restated in the form of a hypothesis, it will appear as follows: Those who use preventive health services are more healthy than those who do not. Similarly, if it is postulated that there is an association between infant mortality and religion, the hypothesis may be stated as follows: There is a difference in infant mortality rates among christians and atheists. The opposite of the research hypothesis is called the Null Hypothesis, which in this case will appear as "There is no difference in the infant mortality rates among christians and atheists".

A hypothesis should be very specific and clear if a statistical test of significance is to be successfully applied. In other words, the nature of the statistical test can be influenced by the degree of specificity of the hypothesis. This can best be illustrated by the fact that a hypothesis is also sometimes stated in relation to the magnitude of the variable under consideration. In the above example for instance, the hypothesis did not state which of the two religions has a higher infant mortality rate. To take this into consideration, the hypothesis will appear as follows: The infant mortality rate is higher among christians than among atheists. This in effect changes the nature of the
statistical test of significance to be applied to the hypothesis. Data is subsequently collected and analysed to prove or disprove the hypothesis. One disadvantage of stating objectives in the form of hypothesis is that the scope of the study is greatly narrowed down by making a specific assumption at the beginning and therefore restructure the scope of the research to this assumption or supposition.

It is important to note that at the initial conceptualization of a study, the formulated objectives may initially appear vague. It is usually after the literature review that they are then fine tuned in line with what has been learned from the experiences of other researchers within the limits of available time and resources. Finally a study or research project without stated objectives can rarely be assessed in terms of its quality. The objectives or hypothesis tell a reader what the study set out to achieve and therefore enables the audience to determine whether the conclusions reached at the end of the study are in line with its stated objectives.
INTRODUCTION

The section on methodology describes how the study will be conducted in a specified population. It provides details on the study population, sampling methods, the study design and methods of data collection. This chapter will be
restricted to a discussion of the concepts of study population and samples. Other aspects of research methodology will be treated in subsequent chapters.

THE STUDY POPULATION

In epidemiological research, there is a need for a proper definition of the study population. This forms the basis for defining those to whom the findings of the study may be applicable. A researcher whose study population is for instance made up of survivors of a second cardiovascular accident cannot generalize his findings to all patients that had a cardiovascular accident. It is therefore very important that the characteristics of all those who are eligible for inclusion (or exclusion) in an epidemiological study need to be defined at the onset. A definition of the study population also enables anybody reading the research report to assess the appropriateness of the study population towards realizing the objectives of the study. This helps to explain the extent to which variations in the findings of similar studies are as a result of differences in populations selected for study. Finally, a definition of the study population guides the field workers as to those who are eligible for interview or physical examination.

The study population is a totality of all the units available for inclusion in a study. Although in most cases, these units are persons, in epidemiological studies, other examples of units of study are houses, toilets, wells, families, medical records and death certificates. The type of population chosen for study will depend on the objectives of the research. As an illustration, a study on maternal mortality among rural women will be restricted to a population of women residing in a defined rural area. The study population may also consist of medical records of
women who utilize maternity services and have come from the specified rural area.

THE SAMPLE

Though epidemiological studies aim at providing information on disease distributions and its associated factors within entire populations, a researcher is rarely required or in fact is able to study every individual in the whole population. Including all individuals within the population in the study will involve an enormous commitment of time and resources in terms of money, materials and manpower. Studying an entire population also increases the chances of errors in the final result. To avoid these constraints, only a subsection of the population called a SAMPLE is customarily selected for study. The process of selecting this subsection of the population with the aim of extrapolating the results from the subgroup to the entire population is referred to as SAMPLING. Selecting a sample rather than studying the entire population therefore enables the researcher to save time and money and also to study individuals in the sample in greater details. Because fewer study units are involved, expensive and more accurate tests can be used. As an example, in a study on the prevalence of anaemia in a community, it will be too expensive and time consuming to examine the blood specimens of all individuals in the entire population. If a small sample is chosen, the researcher can go into great details to document not just the prevalence of anaemia but also the most common types prevalent in the community. It is also worthy of note that in examining the blood of the individuals for anaemia, the researcher did not collect the entire blood for study but a sample specimen as a basis for a statement concerning the whole blood.
There are two main conditions that need to be satisfied for the results from a sample of the population to be generalizable to the parent population. Firstly, the sample has to be chosen in a way that it is representative of the parent population. Secondly, the sample size needs to be adequate.

**SAMPLING METHODS**

Sampling can be classified as either probability or non-probability. In probability sampling, every unit of the study population has an equal or known probability of being selected for inclusion in the study. The sample that is eventually selected is therefore representative of the study population from which it is chosen. This is in contrast to non-probability sample where the likelihood of the selection of any individual unit in the study population depends on the whims and caprices of the investigator. The individual units therefore do not have an equal or known chance of selection and such samples are unrepresentative of the study population. Some groups will be either grossly over or under-represented. The importance of choosing a representative sample is illustrated in a situation in which a study is aimed at investigating the prevalence of helminthic infections among children in a community. The researcher goes to a nearby market and selects the first one hundred children that came for treatment. Majority of the children may have been at school at the time and therefore will be under-represented. In addition, children in school are likely to belong to a higher socio-economic class and therefore a distribution of those selected will be skewed towards children from very low socio-economic groups. Results from such a study will be highly distorted and
biased and not represent the true helminthic situation in that community.

Depending on the nature of a study, there are various methods of choosing a sample to ensure that it is representative of the parent population. In addition, the existence of a sampling frame will also determine the choice of an appropriate sampling method. A sampling frame is a list of all the study units from which a sample is to be drawn. Examples include school registers, census or voters registers, age-sex registers of general practitioners and a register of medical doctors in a state.

TYPES OF PROBABILITY SAMPLING.

In epidemiological research, there are five probability sampling methods that in most cases will satisfy the objectives and design of the study. These include the following:

1. Simple random sampling,
2. Stratified sampling
3. Systematic sampling
4. Cluster sampling
5. Multi-stage sampling.

Simple Random Sampling.

This is a sampling method in which all individuals in the study population have an equal chance of being included in the sample. In selecting a simple random sample, there is first the need for preparation of a sampling frame, if none is in existence. Depending on the previously agreed sample
size, the required number of units are selected from the the sampling frame. This can be done using the lottery method in which each number from the sampling frame is written on a piece of paper and all are put into a box or basket and properly shuffled. The sample is selected by picking the appropriate number of folded pieces of paper from the box or basket and recording their numbers. Another technique for choosing a simple random sample is by the use of a simple random number tables which can be found either in most statistical textbooks or as published tables. A specimen of the simple random number table is shown below.

95 19 70 36 92 85 05 39 25 78 54 34 14 28 76
95 61 59 19 81 87 14 58 61 75 53 44 19 12 00
97 55 66 76 74 68 47 68 66 86 49 47 63 51 43
90 51 34 31 18 74 55 41 42 81 70 15 36 55 16
99 56 78 99 98 77 87 25 77 90 34 18 82 02 11
27 24 80 09 77 14 13 96 19 16 22 46 88 26 25
34 63 66 89 97 29 99 91 27 17 14 56 41 05 32
26 86 45 23 35 60 69 82 66 37 43 44 27 92 07
06 86 34 21 67 66 12 56 74 35 91 64 66 15 01
19 62 94 14 94 83 15 22 30 16 82 69 70 27 07
44 36 92 62 39 56 96 89 04 43 89 96 58 17
78 96 88 88 98 78 41 35 91 77 68 88 50 31 77
31 61 87 98 86 30 43 65 84 51 94 88 95 85 33
42 96 12 76 32 33 23 70 68 71 76 83 28 45 92
95 47 30 03 82 83 30 79 07 95 87 66 83 53 36
The table is usually arranged in groups of five digits, either horizontally or vertically. In effect, the numbers can be read horizontally across or vertically up or down the column. The numbers are selected as individual digits or as groups of two, three, four or five digits depending on the maximum sample size required for a particular study. If for instance the required sample size is 300, the numbers are selected as groups of three digits such as 003, 014, 134, 256, etc. Similarly, for a sample size of 3000, the numbers are selected as groups of four digits such as 0002, 1178, 2356, 0034, etc. This procedure is carried out either horizontally or vertically along the column until the total sample size is obtained.

The simple random sample is easy to select but has the disadvantage that it can only be carried out in situations where a sample frame exists or can be constructed. Compiling a list of sampling units in situations where no sampling frame already exists can be very time consuming and expensive. In such circumstances other sampling methods are preferable. Researchers need to exercise caution in using existing sampling frames by making sure that they are complete and up-to-date.

**Systematic Sampling**

This is a sampling method in which units are selected systematically to include all those that appear at regular intervals on specific locations in a sampling frame, list or a queue. As an illustration, an investigator may decide to select for study every fifth house in a street or every fifth woman attending an ante-natal clinic. The sampling interval chosen by an investigator depends on both the required sample size and the total study population. If for instance, a sample size of 150 is required from a primary school
population of 3000, the sampling interval is given as 3000/150 or 20. The investigator then proceeds to choose every 20th pupil on the school register. The inverse of the sampling interval, which in this case is 1/20 is also referred to as the sampling fraction.

Selecting a systematic sample from a queue follows a similar procedure. Supposing that an investigator requires a sample size of 100 to be selected from women attending an ante-natal clinic, he will first have to ascertain the population of women attending the clinic over a given period of time. Assuming 500 women attend the clinic every month, the sampling interval for a study in which data collection have been planned to last one month, will be 500/100 or 5. Every fifth woman who arrives at the antenatal clinic is then selected for inclusion in the study. Systematic sampling is easier to perform than a simple random sample and does not necessarily require a sampling frame.

Stratified Sampling

This sampling method derives its name from the term strata which in epidemiology refers to units of a study population sharing similar characteristics such as age, sex occupation or living conditions. To obtain a stratified sample, the investigator therefore has to first divide the study population into different strata. A simple random sample is then selected from each strata. The sample size from each strata may either be the same or proportional to the total number of units in the strata.

A stratified sample is very useful and appropriate when various subgroups of the population are not uniformly distributed. This type of sample will ensure a representation of important subgroups of the population.
with respect to specified characteristics. Stratified sample is also very useful if the study involves a comparative analysis of various subgroups of the same population. As an illustration, to select a stratified sample for a study on Drug abuse among secondary school children, the researcher will start by dividing the children into various strata based on age or sex. In terms of age, the various classes could be regarded as already existing strata. Simple random samples are then selected from each of the classes one to six. Alternatively a simple random sample is selected each among the boys and girls.

Cluster Sampling

A cluster refers to a sub-unit of the study population assembled together in the same geographical area. Examples are all individuals in a village, school, workplace or institutions. Cluster sampling involves the selection of clusters and studying every individual within the selected clusters. The researcher therefore needs a listing of all the clusters in the study area such as villages, households or wards.

In field epidemiological studies it is always useful to obtain a map showing the clusters of interest and their respective boundaries. Using the map a simple random sample of clusters are selected and every individual residing within the boundary limit of the chosen clusters are selected for study. It is always better to select many clusters with few individuals than to select few clusters with large number of study units.

Cluster sampling is very useful when a sampling frame of individual units of the study population do not exist or when the units are scattered over a very wide area.
geographical area that will involve a lot of time, money and manpower to cover.

Multistage Sampling

In multistage sampling, the various sampling methods previously discussed are adopted in a number of stages to finally select a sample for study. It is the most appropriate method of selecting a sample from a very large population such as that of an entire town, district or local government area. As an illustration, supposing a local government authority intends to select a representative sample of 5000 women to ascertain the proportion of women utilizing various family planning methods. As a first step, a sampling frame of all the wards making up the local government area is prepared. From this a simple random sample of the wards is selected. In the second stage, for each ward, using a list of the streets, a simple random sample is selected. In the third and last stage, a systematic sample of houses in each street is selected and women living within the selected houses are interviewed on their use of various family planning methods. Any mixture of the different types of sampling methods, including cluster sampling, can be used to select a sample for study.

TYPES OF NON-PROBABILITY SAMPLING

An unrepresentative sample can invalidate the results of a well conceived epidemiological research study. Non-probability samples are therefore being discussed here just for the sake of completeness and also to enable readers to be able to critically evaluate the suitability of the sampling method adopted in a published epidemiological report.
1. Accidental or Convenience Sample

This form of sample is selected on the basis of convenience, availability or accessibility of the sampling units. A researcher will in this case therefore select only those who happen to be available or who volunteer to participate in the study. All individuals passing through a street corner may for instance be interviewed by the researcher until the required sample size is achieved. Individuals who are selected in this may have special features which may differ from the general population. As an example, those who decide to volunteer to participate in a study may come from the very high social classes or may be unduly health conscious people both of whom may differ remarkably from the general population.

2. Quota Sampling

Quota sampling is used in situations where the researcher feels certain characteristics are related to the the problem under investigation and therefore wants to ensure their representation in his sample. The researcher may for instance want to ensure that individuals within each of the ethnic groups in Nigeria are represented in his sample. Although this sounds similar to a stratified random sample, the difference here is that the investigator selects a predetermined number of individuals by methods not governed by chance. From each strata of ethnic group, the researcher merely selects the first available set of individuals who meet his criteria until the required number or quota for each stratum is satisfied.

Investigators sometimes resort to this type of sampling when the characteristic under investigation are difficult to locate. Typical examples of such health related characteristics are stigmatizing diseases like leprosy, sexually transmitted diseases, tuberculosis or even teenage pregnancy. The investigator starts by deciding on the characteristic he wishes to study. The first available set of individuals who meet these criteria are selected. They are then instructed to assist in identifying other individuals with the same characteristics who are also selected. The process is repeated with each new group until the required sample size is met.
SAMPLE SIZE DETERMINATION

INTRODUCTION.

One of the most common criticisms of published epidemiological reports, whether as journal articles or
thesis submitted for examinations, is the inadequacy of the sample size. Some researchers never give any serious consideration to the sample size during the planning stage while others are simply unaware of the criteria for the determination of the minimum sample size for a study. Every study requires a certain sample size in order that generalizations can be confidently made from the sample to the entire study population. It is therefore important that researchers should be acquainted with the theoretical basis for the determination of the minimum sample size for a study. This will not only improve the standard of their work, but also will enable them to be able to critically evaluate published reports of other researchers.

A representative sample is chosen for study, usually with a view to making generalizations on the entire population. As an example, if the mean height of women in a sample is 170 cm, this is used as a basis for the estimation of the mean height of women in the entire population. The precision and degree of confidence with which this estimate of the population parameter can be made is related to the sample size. Provided that the right sampling method has been adopted, the bigger the sample, the more precise the estimate of the population parameter will be and vice versa. Statistical formulae are available for the calculation of the desirable sample sizes for various types of epidemiological studies. This calculation has been further simplified by the availability of published graphs and tables which can assist researchers in choosing the minimum sample size required for their study. There are now also computer programs which can easily perform sample size calculations. It will be recalled however, that the bigger the sample size, the greater the commitment of resources to the study in terms of time, money and manpower. The degree of accuracy of measurements also
suffers when there are too many study units. The crucial question that therefore remains to be answered is: at what point should the sample size be regarded as adequate but not too large to result in a waste of resources? The answer is that the sample size that is eventually used for a study is a balance between what is feasible in terms of available resources and what a statistically acceptable minimum size ought to be.

In this chapter, the theoretical basis for sample size determination will be considered. An application of the formulae for the calculation of the sample size for various types of epidemiological studies will also be illustrated. The assumption has been made that the reader is conversant with the statistical concepts of probability and the normal distribution curve, which is briefly reviewed here. To be able to fully appreciate issues raised in this chapter, some readers may therefore first need to refer to statistical textbooks for a more detailed discussion of the normal distribution curve.

Computation of a sample size depends on whether the study is aimed at estimating a population parameter such as mean, proportion (e.g., percentage or prevalence), and rate (e.g., incidence), or whether the purpose of the study is to measure differences between two population groups. The discussion in this chapter will be limited to sample size determination for the first type of study. Computation of sample sizes for the second type of study is more complex and is beyond the scope of a book of this nature. Interested readers are referred to the statistical textbooks which have been suggested for further reading.

THE NORMAL DISTRIBUTION

An empirical frequency distribution resulting from the tabulation of values in a sample can be represented in a
The graphical form called histogram. If the midpoints of each of the bars in the histogram are connected to each other, the result is a frequency polygon. As the number of values increase, the polygon is transformed into a smooth curve. A frequency distribution of the total or infinite number of values in a population is a theoretical distribution as opposed to an empirical distribution from a sample. A typical theoretical distribution is the NORMAL OR GAUSSIAN DISTRIBUTION CURVE.

It has numerous applications in epidemiological studies due to the fact that it has been found to approximate to the empirical distributions of many naturally occurring biological characteristics such as height and weight. Because of this relationship, the properties of the normal distribution are usefully applied in making generalizations from a sample to the entire population if the sample data follows a normal distribution. Some of the important properties of the normal distribution curve (Fig. 6.1) are as follows:

1. It is bell-shaped
2. The mean, median and mode are located at the same point and therefore identical.
3. It is symmetrical about its mean which therefore is in the middle of a distribution of all the values.
4. The tails of the curve approach but never touch the base.
5. Approximately 68% of all the observations lie within one standard deviation from the mean.
6. Approximately 95% of all the observations or values lie within two standard deviations from the mean. This in effect means that the probability that a given value lies within a range of two standard deviations from the mean is 95%.
In choosing a representative sample for a study aimed at estimating a population parameter such as a mean or prevalence, it is assumed that the sample is similar in every way to the parent population. In actual practice it is known that this assumption is not exactly true. If for instance a study is aimed at assessing the mean height of pregnant women, and even if all the pregnant women in the study population had an equal chance of being selected, a sample of six pregnant women is likely to include slightly more or less tall or short women compared to the parent population. The mean height of the pregnant women in the sample will therefore not be exactly the same as that of the parent population. This difference arises due to sampling error which simply means the tendency for sample values to differ from population values due to chance. Supposing that the sample size is increased to sixty pregnant women, their mean height is more likely to be closer to the true mean height.
height of all pregnant women in the population from which the sample was drawn. If different samples of the same size are repeatedly chosen from the parent population of pregnant women, their mean heights will be different from each other but cluster around the population mean to form a normal distribution called **sampling distribution of the means**. Since the sampling distribution is a normal distribution, it has all the properties of a normal distribution curve. It therefore has a mean which is equal to the true population mean and its standard deviation is called standard error of the mean. This also means that 95% of all the values will lie within two standard deviations of the mean (in this case two standard errors of the true population mean). In effect, if a single sample is selected and the mean computed, the probability of this mean lying within two standard errors from the true population mean is 95%. In other words, the sample mean will lie within two standard errors of the true population mean at the 95% level of confidence.

If a population proportion were being estimated, the same principle applies and the distribution is called **sampling distribution of proportions** and the standard deviation referred to as the standard error of proportions.

Before the appropriate formula can be used to calculate the minimum sample size for a study aimed at estimating either a population proportion such as prevalence or a population mean, the researcher needs to specify the following:

1. An estimate of the value of the condition in the population.

Since this is usually unknown, the prevalence or mean as estimated in a previously...
published study is substituted for the estimate of the prevalence or mean of the condition in the population. For this to be valid however, the sample size in the published report should be greater than 50 for a population proportion and greater than 30 for a population mean.

2. The degree of precision or margin of error required for an estimate.

This is measured by the sampling error (SE) of the estimate (standard error of the mean or sampling error of proportion), and represents the margin of error an investigator is prepared to allow in the estimation of the population value.

3. The level of confidence required for the estimate.

This is a measure of probability that the true population parameter lies within a specified interval. In most cases the 95% confidence level is used. At this level the true population value is expected to lie within the interval mean ± 2SE. This confidence interval which is expected to contain the true population value is therefore dependent on the size of the sampling error. In other words, the smaller the sampling error, the smaller confidence interval.

4. The degree of variability of the variable under study.

This is represented by the standard deviation and is necessary for the calculation of the standard error which is also related to the degree of variability of
the parameter in the population. Usually, the higher the standard deviation, the larger the sample size required to attain the same level of precision.

FORMULÆ

The formula for the calculation of sample size for qualitative data expressed as a proportion is given by:

\[ N = \frac{p \times q}{SE^2} \]

where \( p \) = percentage of those affected, \( q = 100 - p \), \( SE = \text{Sampling error tolerated (\%)} \).

As an example, an investigator intends to determine the minimum sample size for a study aimed at assessing the prevalence of tuberculosis in a given community. From previous studies, the investigator is able to estimate the prevalence of tuberculosis at 20%. The investigator is prepared to allow a margin of error of 5\% in his estimate. In other words, at the 95\% confidence level, he prepared to accept that the prevalence lies between 10\% and 30\%. The sample size is therefore:

\[ N = \frac{p \times (100 - p)}{SE^2} = \frac{20 \times 80}{5^2} = 64 \]

If the investigator wants his estimate to be more precise by accepting a smaller error,
margin of only 2.5%, which at the 95% confidence level translates to a confidence interval of $20 \pm 2 \times 2.5$ or $15\% - 20\%$, the required minimum sample size will be:

$$20 \times 80 = 256$$

In effect, a much bigger sample size is required to achieve a higher degree of precision in the estimated population value. For quantitative data expressed as a mean the formula for the calculation of sample size is given by:

$$N = \frac{SD^2}{SE^2}$$

where $SD = $ Std. deviation of the value in the sample.

$SE = $ std. error of the mean.

As an example, supposing an investigator intends to estimate the sample size required for a study aimed at determining the mean height of pregnant women attending for ante-natal care. The estimated height from a published article is 170 cm with a standard deviation of 10 cm. The investigator is prepared to accept a margin of error of 2.5 cm. At the 95% confidence level, the margin of error of the estimate will be $mean \pm 2(SE)$, and the interval will therefore be $170 \pm 2.5$.

The sample size $N = \frac{SD^2}{SE^2} = 16$. 
The need for a researcher to choose the appropriate study design if the objectives of his study are to be attained cannot be over emphasized. An inappropriate design may lead to the wrong conclusion or the collection of data that...
is irrelevant to the objectives of the study. Choosing the most appropriate epidemiological study design will depend on the answers to a number of important considerations which can be summarized as follows.

1. **What is the purpose of the study?**

   One of the major aims of most epidemiological studies is to provide information necessary for the planning of health services. Such studies will therefore be directed at the description of disease patterns in a community. In other words, information is needed on the size and type of health problems mainly for planning purposes. As an example, an international health agency involved in the control of tuberculosis in a community might want to know how many people are currently suffering from the disease, its geographical, age and sex distribution.

   The purpose of other studies may be to provide information needed for the institution of an intervention program. Institution of such programs will depend on a knowledge of the factors or influences which determine the disease occurrence. Why is it for instance that in the same community, some individuals suffer from tuberculosis while others do not. Finally, information may be required on the effectiveness of a new therapeutic procedure or the better of alternative methods of treatment.

2. **How much resources are available?**

   What resources are available to the researcher in terms of money and manpower to conduct his
study. A study being funded by a health agency will be more elaborate than one personally financed by a researcher from his earnings.

3. What is the time frame for the study?
   How much time is available to conduct the study. A hospital management board may for instance require that information be made available to it in the next one week on the number and the age-sex distribution of all cases of tuberculosis seen in the hospital every year.

4. What is the incidence rate of the disease under study?
   How common is the disease under study and what is the incubation period? A study designed to review the occurrence of cholera will be inappropriate for the cancer.

The answers to the issues raised above will guide the researcher in choosing from any of the following epidemiological study designs: Descriptive, Cross-Sectional, Case-control, Prospective or Cohort and Experimental or Intervention study designs.

Cross sectional, case-control and prospective studies are collectively referred to as analytical studies. Discussions in this chapter will be limited to descriptive and analytical study designs while experimental design will be considered in the next chapter.

DESCRIPTIVE EPIDEMIOLOGICAL STUDIES

This type of study is the first step in understanding the
health or disease problems of a given community. The size and nature of the health problems are described without any attempt at establishment of the relationship between cause and outcome. Data is collected and analysed to describe the health problems of a community in terms of the frequency of occurrence, WHO (Persons) is affected WHERE (Place) it occurs, and WHEN (Time) it occurs. The characteristics considered in descriptive epidemiology are age, sex, educational status, occupation, ethnic group or race, religion, marital status and family size. Disease occurrence and death rates show a marked variation in relation to these variables. Taking age for instance, the major causes of morbidity and mortality vary with age. Respiratory infections and malaria are quite common in pre-school children in contrast to the elderly whose major health problems are mainly degenerative disorders such as hypertension and cardiovascular accident. The severity of diseases are also known to vary with age. A typical example is pneumonia, which is known to be more severe and with higher case fatality rates at the extremes of age i.e. infants and the elderly. Analysis of health problems based on characteristics of persons is important in highlighting population groups that are particularly at risk of acquiring specific diseases. This is useful for the planning of services in relation to the needs of the various segments of the population. Description of disease occurrence in relation to these variables may lead to the formulation of a hypothesis on the influence of any of them on the aetiology of a
particular disease. This hypothesis can then be tested using analytical epidemiological methods. As an example, the influence of age on disease occurrence could be related to a number of factors such as immune status, occupational exposure to toxic substances, and lifestyles. A number of diseases are known to be specific for various sexes and this is invariably related to the reproductive systems. It is however well known that for some other diseases, one sex may be more susceptible than the other. In general, mortality rates are higher among males compared to females. This could possibly be attributed to a number of factors such as hormonal imbalance, life styles or differential use of health services. The various hypothesis on the influence of age, sex, and the other variable on health and disease can further be investigated using analytical studies.

Place or Where

The place of residence of individuals and populations may partly determine the diseases they suffer from. This may be related to exposure to toxic environmental factors or cultural habits. Disease occurrence in relation to geographical location of individuals correlates more with natural than political boundaries. Variables that may be considered under place include climate, urbanization, vegetation, geographical location, cultural factors, water supply, presence or absence of vectors and other environmental hazards. Description of disease occurrence by place is quite useful for the equitable allocation of health care resources to different regions in a country. This information is also useful in the generation of hypothesis on the influence of the environment on disease aetiology. The
localization of onchocerciasis infection in a given geographical area may for instance be associated with the presence of fast flowing rivers. Representation of disease distribution on a map showing the physical and social characteristics of a geographical area could at a glance highlight environmental factors which are likely to have contributed to the occurrence of a particular disease.

The occurrence of a disease in a given geographical area is likely to be due to environmental factors if a similar frequency of occurrence is observed in all ethnic groups residing in the area, or if different frequencies of occurrence are observed in people of the same ethnic group living in another geographical area. A disease may also be likely to be due to environmental factors if healthy persons entering a new geographical area acquire a disease with a frequency similar to those already residing in that geographical area, for example caucasians visiting Nigeria, with time will develop malaria attack rates similar to that of Nigerians.

Time

This refers to the hour, day, week month or year of onset or occurrence of a disease. Disease is described in relation to any of the above measures of time depending on whether the disease is acute or chronic in nature. Most diseases will change in their frequency of occurrence with time. The frequency of occurrence may be highest or lowest at certain times in a week, month or year. Clinic attendance for sexually transmitted diseases are for instance known to be highest within the first three days of the week which in related to exposure during the weekend, Malaria, on the other hand is seasonal in occurrence. Changes in disease
frequency over short periods of time, such as weeks or months is described as cyclic, periodic or seasonal variation. This may act as a pointer to the presence of an agent at specific time periods or throw light on the mode of transmission of the disease. Some seasonal activities such as farming are also likely to bring members of a community in contact with various disease vectors at specific times in the year. An example is the peak exposure of rural agricultural communities to the bites of the simuliun fly with the resultant seasonal variation in onchocerciasis infection.

Variations in disease frequency over long periods of time such as years, is described as secular trends. It may point to specific health measures that have affected the disease frequency over such periods. An example is the influence of the introduction of vaccines on the trends of some childhood diseases such as pertussis. It can also be used to predict the future patterns in relation to the disease trends.

Descriptive studies are very useful when results are needed urgently, resources for conducting the study are very limited and when very little is known about the problem. Descriptive studies are also very useful for the provision of the type of information required by managers for the planning of health services. The main disadvantage of descriptive studies is that they are often based on routinely available health data and therefore subject to all the limitations of such statistics.

Analytical studies are concerned with the assessment of the
determinants or causes of disease. They are expected to provide answers as to why or how the disease occurs. This is done by testing a hypothesis which has linked a particular disease with a specific causative agent or factor. There are three types of analytical studies:

a. Cross-sectional study.
b. Case-control or Retrospective study.
c. Prospective or Cohort study.

CROSS-SECTIONAL STUDIES

Like other analytical studies, cross-sectional studies aim at testing a hypothesis which suggests an association between a population characteristic or agent and the occurrence of a disease. It involves a single examination of a cross-section of a population at a given point in time. Data is collected and analysed in relation to the disease and its population characteristics such as age, sex, occupation etc. Because data is collected at one point in time, cross-sectional studies yield information on the number of existing cases of a disease at the given point in time that the study took place. They are therefore sometimes referred to as prevalence studies. Only chronic conditions whose natural courses are usually long are suitable for study by this method.

Cross-sectional studies are very similar to descriptive studies except that it also examines the relationship between a population characteristic such as age, sex, occupation and disease occurrence. This means that cross-sectional studies describe as well as examine factors associated with the occurrence of a disease. Another
difference is that cross-sectional studies are population based while descriptive studies depend on the analysis of routinely available health statistics. Like descriptive studies, cross-sectional studies are cheap, easy and quick to conduct. Its major disadvantage however, is that data on the population characteristic or causative factor being investigated and the disease are collected at the same point in time. It is therefore difficult for an investigator to say with any degree of certainty which came first - the disease or the causative factor. Results from cross-sectional studies may therefore sometimes be very difficult to interpret. As an example, in a cross-sectional study on the association between smoking and lung cancer, it will difficult for an investigator to say with any appreciable degree of certainty whether smoking resulted to lung cancer or if smoking was commenced after the onset of lung cancer.

**CASE-CONTROL STUDIES.**

A case control study is a non-experimental or observational study aimed at testing an epidemiological hypothesis on the association between suspected aetiological factors and the occurrence of a disease by comparing the frequency of exposure to the aetiological factors in a group with the disease (cases) and another group without the disease (controls). The investigator selects cases (those with a disease under study) and controls (those without the disease under study) and looks at their past experiences in terms of their history of exposure to a suspected aetiological factor. The past history of exposure to the suspected aetiological agent may be ascertained by either interviewing them or referring to their case notes or other records such as death certificates. The investigator looks backwards in time or
retrospectively for exposure history of the cases and control group. Case-control studies are also referred to as retrospective studies.

By examining the past history of exposure among the two groups, the investigator aims at determining whether they differ in the proportion of those who had in the past been exposed to the suspected risk or aetiological factor. If it is found that a higher proportion of the cases had previously been exposed to the suspected aetiological factor compared to the controls, it could be reasonably concluded that there is a positive relationship between the presence or exposure to the risk factor and the occurrence of the disease. This can be illustrated by examining the influence of consumption of coffee and bladder cancer. The investigator starts by selecting cases of bladder cancer and a comparison or control group without bladder cancer. Although the control group may have other diseases other than bladder cancer, these should be diseases which do not have the same risk factor as bladder cancer. Both groups are interviewed to ascertain their past history of coffee consumption, including the duration and average quantity consumed per day. If there is a causal relationship between coffee drinking and bladder cancer, the researcher is likely to find that a higher proportion of the cases consumed coffee in the past compared to the controls. As demonstrated in Fig. 7.1, $\frac{A}{B} + \frac{C}{D}$ will be greater than $\frac{A}{B} + \frac{D}{B}$.
The accuracy and validity of results from a case-control epidemiological study will greatly be enhanced if the cases and controls are properly selected. It is therefore important that an investigator is aware of the various sources of cases and controls, including their limitations.

### Sources of Cases

<table>
<thead>
<tr>
<th>Medical care facilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>This includes Hospital inpatients and outpatients, disease registers such as cancer registers, death certificates and infectious disease notifications. In using a medical care facilities as a source of cases, it is assumed that all patients with the disease have an equal chance of appearing at the centre. This in effect means that cases in the medical care facility are representative of all cases of the disease in the...</td>
</tr>
</tbody>
</table>
community. In practice, this assumption is not true for all diseases. The probability that a case of disease will appear in a medical care facility differs for different diseases and is influenced by so many complex variables such as accessibility, severity of the condition, availability of effective treatment. All cases of breast cancer may for instance eventually appear in a teaching hospital, but this may not necessarily be true for all cases of peptic ulcer.

In spite of the limitations in using medical care facilities as a source of cases, it has a number of advantages. It is easy to identify and reach the patients for interview. It is cheap and quicker than going to the patients' homes. Hospital patients are relatively more cooperative than those in their homes. This may be a population in a defined geographical area. This source of cases is utilized if the investigator believes that hospital cases are not likely to be representative of diseased persons in the community. Individuals with the disease under investigation are therefore first identified from the community through prevalence studies. Although this is a more valid source of cases, it is time consuming and expensive.

Sources of Controls

Sources of controls should, as much as possible be similar to that of cases for ease of comparability. The assumption is that the risk of exposure to the suspected etiological
agent is likely to have been the same in the two groups. The sources of controls therefore include the following:

a. Population of a defined Area.

This population should be the same as that from where the cases originated. It is assumed that the controls have the same chances or probability of exposure as the cases.

b. Relatives, Spouses or Associates of the Cases.

These groups are easily accessible, co-operative and come from the same environment as the cases. Relatives and spouses are however not a suitable source of controls for the study of genetic diseases. Examples of associates of cases include those from the same age grade, school, club, and work place.

c. Medical Care Facilities.

Even if the diseases in a medical care facility are not representative of those in the community, it is assumed that if patients at the medical care facility are coming from the same geographical area, the same selection factors which determine which patients attend the hospital or health centre will be equally applicable to both cases and controls. In selecting controls from hospital sources, the investigator should ensure that the diseases they have do not share similar risk factors with the disease under study. If this happens, the strength of the association between the aetiological or risk factor and the disease will be severely weakened. If
for instance an investigator selects patients with bronchitis to act as controls in a study investigating the relationship between lung cancer and smoking, the result will be faulty because smoking is also a risk factor for bronchitis.

Distortions in the interpretation of the influence of causal or risk factors on disease occurrence can also arise as a result of the influence of some variables which may be related to both the occurrence of the disease and the causal or risk factors under investigation. These variables are called confounding variables and have to be taken into account in the design, analysis and interpretation of results of the relationship between the aetiological agent and the disease under investigation. As an illustration, a researcher may wish to examine the relationship between lack of ante-natal care and maternal mortality. An example of a variable which in this case influences both the attendance for ante-natal care and maternal mortality is socio-economic status. If the study is conducted in population of predominately mothers of low socio-economic status, a high maternal mortality will be recorded which may have arisen due to the low socio-economic status of the study population and not necessarily poor ante-natal attendance. Socio-economic status is therefore a confounding variable. The influence of confounding factors such as socio-economic status in the above example can be taken care of by a procedure called matching.

Matching is the process of selection of cases and controls
so that both groups are similar in specified confounding characteristics or variables. Matching is therefore aimed at removing the influence of confounding variables. These variables such as age, are known to be associated with both exposure to the aetiological agent under consideration and also with the disease under study. Other examples are sex and social class.

Matching may be by individual cases and controls or group matching such as by age groups. Matching individual cases and controls for age will mean that for each case with a given age, a control with the same age is selected. It is important to note however that once a given variable is matched for, an investigator cannot test the effect of that variable in disease aetiology. In other words the investigator should not match for the risk factor(s) under study as its effect can then no longer be investigated. As an example, if risk factors of cervical cancer are being investigated, and cases are matched with controls in terms of parity, the investigator can no longer test for the influence of number of children on cervical cancer.

Advantages and Limitations of the Case-Control Study Design

Case-control studies are very useful for the testing of specific aetiological hypothesis. They are quick, simple and cheap to conduct. Small numbers of cases and control are needed and subsequently variables can be studied in greater detail using expensive tests where necessary. It is useful for the study of rare diseases since an investigator starts with those already with the disease. In such situations a case control study may in fact be the only practical method of studying the disease. Ethical problems associated with the use of humans for studies are minimal when
I compared to cohort and experimental studies.

In deciding to use the case-control study design, a number of limitations should be borne in mind. A common limitation of case control studies is bias. Bias is any effect at any stage of an investigation that produces results that are systematically different from true values. One of those commonly encountered in case control studies is *Recall Bias* in which errors arise due to inaccurate or incomplete recall of past events by cases and controls. These may be differential recall among cases and controls in favour of a more accurate recall by cases compared to controls.

Another type of bias common in case control studies is *selection bias*. This is a systematic error arising from the fact that hospital cases are a selected group unrepresentative of all cases of the diseases in the community. Hospital cases may for instance be more severe, while those who are already dead from the diseases are not included. The poor who cannot afford the cost of hospital treatment are also unrepresented.

Other limitations of the case control study design are the inaccuracies or incompleteness of hospital case note recording. Finally, a case control study cannot measure incidence directly.

**PROSPECTIVE STUDY DESIGN**

This is also referred to as cohort studies and form an important method of investigating the relationship between a hypothesized aetiological factor and disease occurrence. This type of study derives its name from the term cohort, which means a group of persons who share a common experience within a defined period of time. Some examples include the following:

...
a. Birth cohort - those born over the same time period.
b. Marriage cohort - those who married over the same time period.
c. Occupational cohort - those who worked in a given occupation over a given specified time period.

In the cohort or prospective study design, the investigator compares the incidence of disease in a group of individuals exposed to the suspected risk factor with another group of individual who are not exposed. The investigator starts with a cohort of healthy individuals who are then classified in terms of their exposure to the suspected risk factor. This is achieved at the commencement of the study by collecting information from the study cohort on their history of exposure to the suspected risk or aetiological factor. The relevant information on exposure may for instance be collected using self-administered questionnaires, physical examination and laboratory or other investigations. On the basis of this base line data, individuals are then classified into two groups - exposed and non-exposed. It may be necessary to further classify those exposed by degree of exposure. After the classification, both groups are followed up forwards in time (prospectively) and compared for the development of the associated disease. (Fig. 8.1) Follow-up usually involves periodic history and physical examination or surveillance using routinely collected health statistics on the outcome of exposure which may be the onset of disease, death or disability. There may be the need for the evolving of a mechanism for notifying the investigator of the death of any member of the cohort in situations where adequate health statistics are non-existent. If the proportion of the exposed group who develop the disease in future, \( \frac{A}{A+B} \) is higher than the non-exposed \( \frac{C}{C+D} \), this is an indication of a positive association.
between the risk factor and development of the disease. This is further illustrated in Table 8.2.

FIGURE 8.1 Follow up and Outcome in a Cohort Study.

![Diagram showing follow up and outcome in a cohort study: healthy cohort, exposed, disease (A), not exposed, disease (C), healthy cohort, no disease (B), no disease (D).]

TABLE 8.2

<table>
<thead>
<tr>
<th></th>
<th>WITH DISEASE</th>
<th>WITHOUT DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPOSED</td>
<td>A</td>
<td>A + B</td>
</tr>
<tr>
<td>NON EXPOSED</td>
<td>C</td>
<td>C + D</td>
</tr>
<tr>
<td></td>
<td>B + D</td>
<td>A + B + C + D</td>
</tr>
</tbody>
</table>

It is easily noticeable that this situation is in contrast to that in retrospective studies where the investigator starts with those with the disease (cases), and then looks back (retrospectively) at their exposure history. Cohort studies
may however be retrospective in nature (Retrospective Cohort studies). In this type of study design, the sequence of events from the time the individuals were healthy to the time of development of a particular disease is traced. This is however possible only in situations where adequate records exist covering periods before and after the onset of the disease. Employment records of individuals spanning over several years is a very good source for the design of retrospective cohort studies.

In some situations, it is not feasible or easy to select a healthy cohort and classify them into exposed and non-exposed groups. In such situations two cohorts are chosen, one of which has been exposed to the risk factor and the other not exposed. In such situations, the two groups should be similar in most other respects and both should be followed up in the same manner and for the same duration of time.

Sources of Study Cohorts

a. Groups with a history of exposure to the aetiological agent under investigation. An example is a study of the effect of radiation on reproduction, where a good study cohort will be those occupational groups exposed to radiation such as radiographers.

b. Groups that are stable over a long period of time and have adequate records for follow-up. Examples include civil servants, life insurance policy holders, patients on a general practitioners register including obstetric populations, members of a refugee camp or other such stable populations within a limited
specified geographical area.

Sources of Controls

a. The general population: In this case it is assumed that the general population is similar in all respects as the study cohort including with respect to the normal level of risk of the disease. The comparison with the general population is made in relation to only the time period the study cohort was being followed up. Sometimes it may be necessary to compare the study cohort with only a similar subgroup of the general population for example comparing a black study cohort with only blacks in the general population.

Another method of comparing a study cohort with the general population is to apply the age-sex specific rate of the general population to the study cohort to obtain the expected number of cases of the disease which would have occurred in the study cohort.

b. External comparable cohort: This consists of another cohort similar in demographic characteristics to the study cohort but not exposed to the agent under investigation. An example is the selection of other secretaries as a control group in a study of the effect of radiation among computer data entry secretarial staff.

Advantages & Limitations of the Cohort Design

Cohort studies provide a stronger evidence of a causal
relationship compared to case control studies. Unlike case control studies, incidence rate of the disease can be measured directly. Cohort studies do not suffer from the inaccuracies arising from inaccurate or incomplete records used in case control studies or those arising from inaccurate recollection of past events among cases and controls.

Cohort studies are inappropriate in situations where resources are very small or when results are urgently required. Cohort studies are labour intensive and therefore expensive. One of the major problems associated with cohort studies is the identification of all those who had developed the disease. This is particularly tasking if the disease has a low incidence or has a very long incubation period. With time, members of the study cohort are lost to follow-up mainly due to emigration, change of address or occupation. If the rate of attrition is high, those left behind become a very selected group. Results based strictly on this group will be highly biased. In such situations, it is necessary to examine the demographic characteristics of those lost to follow-up to ascertain if they differ from the remaining members and the possible influence of these differences on the final result.
STUDY DESIGN 11: EXPERIMENTAL OR INTERVENTION STUDY DESIGN

INTRODUCTION

An experiment is a study in which the variables under investigation have been manipulated to suit the needs of the investigator. This manipulation is usually in the form of either an introduction or withdrawal of that variable whose effect the investigator wishes to study. Such manipulations are easily carried out with animals such as rats, rabbits or
even monkeys in laboratory experiments. The degree of lighting or noise for instance can, be altered or a drug and other chemicals introduced as required by the investigator.

In a study of the effect of smoking on disease, an investigator may pipe smoke into the animals and later dissect them to extract the lungs for study. Such procedures are ethically not possible with human beings. Experiments with human subjects have to follow very strict laid down procedures. An epidemiologist is expected to be able to design and conduct such studies adhering to the laid down procedures.

Experimental studies provide the strongest evidence of the relationship between a suspected etiological agent and disease occurrence. It is the last stage in the establishment of the risk factors or determinants of disease. As pointed out previously, the first step in establishing the relationship between an agent and disease occurrence is the descriptive study which allows an investigator to generate a hypothesis linking the risk factor to a specific disease. In the next stage, the investigator proceeds to test this hypothesis using initially analytical studies and finally experimental studies.

The difference between the analytical and experimental methods of testing a hypothesis is in the fact that in analytic studies the investigator observes events as they have naturally occurred while in experimental studies, the factor under investigation is deliberately introduced or removed by the investigator. In experimental studies, the investigator is therefore able to manipulate the factor(s) under investigation, including the dose or quantity introduced to suit the purpose of the study. Experimental studies are useful for the evaluation of the effectiveness of health care which could be either preventive (eg.
Experimental studies may be categorized into true or quasi-designs. In true experimental designs, individual subjects or the units of study are selected and allocated randomly into either a treatment or control group. The treatment group receives a therapeutic, preventive or prophylactic measure whose efficacy is under study while the control group receives the treatment in current use or no treatment where none was previously available. True experiments to evaluate the efficacy of medical care may therefore be therapeutic, preventive or prophylactic. A typical example of a true experimental study design is the randomised controlled trial.

Sometimes, it may not be ethically or practically possible to assign individuals randomly to either a treatment or control group. In such situations quasi-experimental study designs are adopted in which either randomization does not take place or a separate control group is lacking. Examples of such designs are community trials and the Before/After experimental design.

The various steps in conducting an experimental study are the same for both true and quasi-experiments except in the concepts of randomization and the availability of a separate control group. The methodology of the randomised control trial includes both randomization and the use of a control group.
1. Definition of the Study Population

This is sometimes referred to as a therapeutic trial and is aimed at evaluating the efficacy of a therapeutic agent such as a new drug or other forms of treatment such as surgery. A randomized controlled or therapeutic trial starts with a definition of the study population. This should be selected in such a way that as much as possible it is similar to the target population to which the findings will ultimately be applicable to. Criteria for the selection of the population to be studied include geographical factors, age, sex, occupation and prevalence or incidence of the particular disease or any other factor associated with the intervention which is to be studied. As an example, the result of a study on the effect of a new drug on adult males with a disease of specified severity cannot be generalized to apply to the entire population including children and women with varying stages of the same disease. It is therefore important to clearly state the eligibility or inclusion criteria from the onset.

2. Obtaining informed consent of participants

After defining the study population the investigator needs to obtain the informed consent of participants in the proposed study. Individuals can only be used in experimental studies after their consent is obtained. This informed consent involves adequate explanation of the purpose of the study, type of treatment or intervention to be administered and any possible side effect or complication. Participants also need to be made aware of
the fact that they may be assigned to either the study or control groups. Obtaining the consent of the subjects or volunteers is usually by having them sign a consent form. Although a volunteer population may not be ideal for an epidemiological study because they differ from those who refuse to participate, randomization into experimental and control groups to a large extent eliminates this source of bias.

3. Randomization

Random allocation or Randomization refers to the allocation of individuals to experimental and control groups in such a way that the process of allocation is determined purely by chance. This ensures that any differences between the two groups can only occur by chance. Randomization is a means of ensuring that the two groups are comparable in every way except the intervention or treatment under consideration. Randomization of subjects into experimental and control groups should not be carried out before the final stage of recruitment of subjects into the study. This is to minimize the chances of attrition after the inclusion of subjects into the study.

Randomization can be performed using a table of random numbers or by tossing a coin. In using the random number table, even and odd numbers from the table may be regarded as treatment and control groups respectively. This can then be used to prepare a list in advance which is held by another worker other than the investigator and showing the sequence to be followed eg T T C T C C C. (T = Treatment Group; C = Control group). To further minimize the chances of bias in the study, this allocation list is copied out on pieces of paper and only opened as
each sequential subject is admitted into the trial. The effectiveness of the process of randomization needs to be checked for initial comparability of the two groups under study in relation to selected variables such as age, sex etc. Statistical tests of significance on the comparability of the groups should not show any difference. To further enhance the comparability of the two groups the study population may first be stratified in terms of age, sex, or disease severity before randomization of each strata to treatment and control groups.

4. Administration of the Treatment or other Intervention

It is important that the dose, duration of treatment and mode of administration of a drug or other interventions is clearly stated before it is instituted. Unless there is a need for variation of the drug regimen during the trial as a result of ethical or other reasons, it must be strictly adhered to, for the results to be accepted as valid. It should not be taken for granted by the investigator that the subjects in the experimental group have taken the prescribed drug or other intervention. The subjects must either be seen to have taken the drug or a method of monitoring for compliance devised. Testing urine specimens for by-products or counting the number of remaining tablets are some examples of methods for monitoring compliance.

5. Assessment of Outcome

A valid assessment of outcome in both the treatment and control groups depend on the absence of bias arising from either the investigator or the subjects. The measures of outcome must be agreed upon before the trial commences.
and should apply equally to both the experimental and control groups. A clear guideline on how, when and how often these measures will be assessed and the total number of measurements to be made also need to be clearly stated.

Some of the measures of outcome such as the presence or absence of headache are subjective. They may however be more objective as in the measurement of temperature, respiratory rate, pulse rate or blood pressure. Knowledge of the group to which a subject belongs is likely to introduce bias in the analysis of outcome. This can arise as a result of a preconception by the investigator or the participants as to the efficacy of the drug under investigation. If for instance, the participants are aware that they are in the treatment group and the believe the drug is efficacious, they are likely to exhibit an improvement especially if the outcome measures are subjective.

The sources of bias highlighted above can be removed by ensuring that the participants, investigator and those analyzing the data are unaware as to which group a participant belongs. This is called blinding. A study is said to be single blind if the participants are not aware as to which group they belong to. This is achieved by giving them placebos. In double blind studies, neither the participants nor the investigator is aware to which group the participants belong. A third party not involved in the study is invited to interpret the results.

COMMUNITY TRIALS

In this type of design, communities rather than individuals form the unit of allocation to experimental and control groups for purpose of the introduction and evaluation of a therapeutic or preventive measure. It is important however
that the communities are comparable to a reasonable extent for example in terms of population distribution. As an example, a researcher may wish to examine the effect of improved sanitary disposal of sewage on the incidence of diarrhoeal diseases in a given community. One village in which ventilated improved pit latrines are constructed is selected as the experimental group while another comparable village is used as a control group. The control group is allowed to continue with their usual methods of sewage disposal. The investigator then proceeds to assess the incidence of diarrhoeal diseases in the two communities.

BEFORE/AFTER EXPERIMENTAL DESIGN

In this type of design, one single group acts as both the experimental and control group. A baseline assessment is carried out on the study group followed by the proposed intervention. After the intervention, another assessment is made on the group to determine the effect of the intervention. As an illustration, the effect of improved toilet facilities on diarrhoeal diseases in a community can be assessed by first obtaining baseline data on incidence of diarrhoeal diseases, followed by the building of sanitary latrines. After the intervention another survey is carried out to assess the impact of the intervention on the occurrence of diarrhoeal diseases. An investigator however needs to exercise caution in the interpretation of the results of such studies because of the influence of factors other than the intervention on the results. A decrease in the incidence of diarrhoeal diseases in the above example may in fact arise as a result of improved water supply or health education which may have been going on at the same period without the knowledge of the investigator.
The importance of data collection in the epidemiological research is amply illustrated by a phrase commonly used by computer programmers which states, "Gabbage In, Gabbage Out". The quality of the research output or results therefore ultimately depends on the input. Inappropriate
data collection techniques will invariably adversely affect the quality of the output and this can never be camouflaged or corrected through the adoption of complex statistical analysis and presentation of results. The type and method of data collection should be systematically planned in advance to suit the objectives and design of the proposed research study.

The concept of planning for data collection is best illustrated in relation to the construction of a building. A builder may begin from scratch by moulding the blocks for the construction or he may purchase already moulded blocks. A decision as to which of the alternatives is better will depend on the type of house being envisaged, the time available to build the house and the amount of human, material and financial resources necessary for the construction. A similar situation exists in epidemiological research. An investigator may collect data directly from the subjects under investigation. In the alternative, he could make use of already existing data which have either been collected routinely from health and related institutions or have been collected from the field for other purposes. The former type of data is called primary data while the later is referred to as secondary data.

Research involving primary data collection is more labour intensive, takes more time and is more expensive compared to that based on the use of secondary data. Primary source of data however, provides the researcher with more accurate and up to date information which is complete and directly relevant to the study compared to the later whose accuracy will depend on the dedication and commitment of those who did the collection. Such data may sometimes have already become out of date by the time it is needed. In addition, having been collected for other uses, some of the important variables necessary to answer the
research question may be missing. In developing countries where routinely collected health data are nonexistent or grossly inadequate, investigators often have no choice but to resort to primary data collection for their research.

SECONDARY DATA COLLECTION

To be able to effectively make use of secondary sources of data, an investigator needs to be aware of where to locate them and also their limitations. The major types of secondary health data include: Population census, vital statistics, data related to morbidity, mortality or disability.

POPULATION CENSUS DATA

A census is the complete enumeration of the entire population residing in a defined geographical area at a defined point in time. It is conventionally carried out every ten years for a country. Some important health related data emanating from the census are fertility patterns and the age-sex composition of the population. These are important for the planning of health services, especially those for special groups such as maternity, paediatric and geriatric services. Other useful health related information from the population census is occupation which forms the basis for the classification of the population into various social classes. The population census forms the denominator for the calculation of mortality and morbidity rates. One major limitation of census data is undercoverage of the population, especially children and the homeless. This ultimately affects the accuracy of rates calculated using the census figures as denominator. In developing countries, the validity of census results is also grossly affected by
inaccurate recording of age and occupation due probably to the influence of either religious or cultural reasons. The competence of those collecting the data and the fact that a census takes place only once in ten years is another source of limitation of the reliability of figures.

MORTALITY DATA

The main source of mortality data is the death certificate and death registration. Death is easily recognizable, occurs once and can be recorded even by lay people. By law every registered medical practitioner who has attended any person during his or her last illness is required to complete and sign a death certificate on the patient stating the cause of death. This contains the patient's name, age, date and place of death and the immediate cause of death. The immediate cause of death is the actual disease or injury that resulted to the death of the patient not the mode of dying. A classification of the causes of death is published by the World Health Organization and acts as a guide for researchers to enable international comparisons of causes of death. After death certification, a relative of the deceased is required to register the death with the local registrar of deaths and births. Mortality data is easier to collect than morbidity data. One of its major disadvantages however, is that it will cover mainly diseases which result in death with an under-coverage of those whose outcome is either recovery or disability. Mortality data also suffers from other deficiencies. Accuracy of the data depends on the level of competence of the persons who recorded it. Facilities for diagnosis, especially post mortem may also affect its accuracy. In developing countries, mortality data
contains a gross under representation of death occurring outside the hospital, especially the rural dwellers, infants, the poor and the elderly.

MORBIDITY DATA

When compared to mortality data, morbidity data is more difficult to collect for a number of reasons. First, it needs trained manpower and special facilities for its recognition. Secondly, an individual can get sick more than once in any given period of time.

One of the main advantages of morbidity data is that it includes health problems that do not necessarily lead to death and so provides a more detailed information on the health of the community. Morbidity statistics are very useful for the planning, monitoring and evaluation of health services. Morbidity data can be obtained from the following sources:

Data From Health Institutions
This consists of records of outpatient attendances, admissions, and discharges from hospitals, health centres, and laboratories. Information normally recorded include: age, sex, occupation and those relating to the disease such as diagnosis, treatment given, duration of admission and outcome. In developed countries, these information are summarized periodically by hospitals and geographical area.

Infectious Disease Notifications
Medical practitioners and other health workers have a legal obligation to notify the appropriate authority of every case of a list of diseases which have been designated as notifiable. Each local government area and country have
their list of diseases designated as notifiable, though certain diseases such as plague, smallpox, yellow fever and cholera are internationally notifiable.

Notification of Industrial Diseases and Accidents
This is similar to infectious disease notification, the only difference being that this deals with certain prescribed industrial diseases and accidents.

Disease Registers
Most developed countries maintain a register of some diseases for ease of identification and follow up of patients. Common diseases in this category are cancers and congenital malformations. New cases seen in hospitals are notified to the appropriate authorities keeping the register. The information is used for the calculation of incidence, prevalence, survival rates, prognosis and mortality rates.

Records of Occupational Health Services
This includes records of pre-employment and periodic examination of workers. Sickness absence records also provide useful information on the health status of workers. One of its major limitations however is that it covers only a segment of the entire population. The working population is a selected group in that unhealthy and elderly people tend to leave the work force.

School Health Services
This is made up of records of examinations of children at school entry and graduation, periodic examinations, immunizations, sickness absence and treatments.

Insurance Company Records
Insurance life policies contain records of clinical
examinations. Some of the records span over so many years and are very useful for the follow-up of patients.

In using any of the above sources of morbidity data, an investigator needs to take into account their inherent inadequacies, some which are as follows:

1. Accuracy of Diagnosis

The accuracy with which disease diagnosis are made will depend on the calibre of personnel and facilities available. A community health worker in a rural area without adequate facilities for diagnosis could lump all cases of dysuria as gonorrhea. In a sexually transmitted diseases clinic, cases of gonorrhoea will include only those confirmed through laboratory investigations. Using the first source will tend to over-estimate the incidence or prevalence of sexually transmitted diseases.

2. Criteria for Defining a Case or Diagnosis

Criteria for the diagnosis of any given disorder may vary from one institution to the other or even from one doctor or health worker to the other. When using routine sources of data for a study, it is therefore important for the investigator to enquire into the criteria for diagnosis. As an illustration, a case of diabetes may either have been defined in terms of sugar in the urine or blood glucose levels. Unless this is clearly stated, an investigator may unintentionally compare his findings based on sugar in the urine to another study in which diagnosis was based on blood sugar levels. Such comparisons between communities or even at different time periods in the same community will be fallacious. This problem has been simplified by the
publication by the World Health Organization of an international classification of diseases as a basis for uniform diagnosis.

3. Accuracy of Case note Records

Most physicians and other health workers do not appreciate the importance of accurate and complete recording of health information. Investigators are quite often frustrated by the number of cases with missing data on age, occupation and residential area. Measures of morbidity from such records can only be as accurate as the records used for the study. Incomplete data on diagnosis will lead to a gross undercoverage of the disease under study.

4. Limited Coverage Of Diseases

Hospital patients are a very selected group in relation to all diseases in the community. There are so many complex factors which determine the probability that any given illness will be brought to the attention of hospital workers. Some of these factors include distance, finance, time, severity of symptoms, interpretation of the symptoms and cultural factors. A phenomenon whereby only a small proportion of diseases are brought to the attention of professional medical care is referred to as symptom or disease iceberg.

PRIMARY DATA COLLECTION

In epidemiological research, the main techniques for the
collection of primary data are the use of questionnaires, interview schedules or guides and observation. A questionnaire is a set of written questions which the respondent may read and fill in the relevant answers himself or which the investigator elicits and writes down the responses from the respondent. The former is said to be self administered and therefore has the pre-requisite that the respondent should be able to read and write. Compared to interviewer administered questionnaires, it is cheaper, takes less time to complete and can even be sent and collected indirectly through another worker or by mail in areas where postal services are efficient. It is therefore particularly useful if the study population is very large and probably scattered over a wide geographical area. In developing countries where the literacy rate is very low, this method of data collection has a very limited scope. Self administered questionnaires have the disadvantage of a low response rate. This is a major problem in situations where the respondents have not been provided with adequate time within which to complete the questionnaire especially if they happen to be individuals with very busy work schedules.

**QUESTIONNAIRE DESIGN**

Whether a questionnaire is interviewer or self administered, a number of factors which need to be taken into consideration in its design include the following:

a. Will the questionnaire be structured or unstructured?
b. Are the respondents able to read and write?
c. What information does the investigator need to gather with the questionnaires in relation to the research objectives?
Types of Questionnaires

Questions in a self or interviewer administered questionnaire may be structured or open-ended. Structured questionnaires are similar to multiple choice examination questions in which the respondent is required to choose the appropriate answer from a limited list of answers. As an illustration, a structured questionnaire for the study of "Factors affecting the non-utilization of primary health care services," with questions on age and sex will require the respondent to circle or tick the appropriate answer as shown below.

1. What is your age?
   (1 = 16-20, 2 = 21-24, 3 = 25-29, 4 = 30-34, 5 = 35-39)

2. What is your sex?
   (1 = Male, 2 = Female)

3. Did you have your last baby at the rural health centre?
   (1 = Yes, 2 = No).

In contrast to structured questionnaires, in open ended questionnaires, the respondent is given the opportunity to provide any answer in his own words to a particular question. Enough space is normally provided for the respondents' answer which may be filled by either the respondent or the interviewer. Compared to open-ended questionnaires, structured questionnaires give room for a standardization of the answers and thereby making it quicker and easier to analyse. It is useful for the investigation of sensitive or embarrassing issues where the respondents will like to be assured of a certain degree of
anonymity and confidentiality. Structured questionnaires however have the disadvantage of restricting the answers of the respondent to only those contained in the questionnaire and thereby losing responses that may be very relevant to the research question.

Contents of Questionnaires

Many students who have sought for advice on how to design their questionnaires are perplexed when asked to produce their study proposal where the general and specific objectives of the study have been stated. The objectives of a study describe the questions which the study is expected to answer and therefore form the basis for the framing of the relevant questions for the questionnaires. There will usually be a number of objectives in a given study. For each objective, the investigator constructs a number of questions suitable for the provision of answers to a particular research question. In most cases, the first part of a questionnaire deals with background variables such as age, sex, marital status, educational status, occupation etc. The subsequent sections are then based on the research objectives. As an example, in a study of the influence of the use of health services on perinatal mortality, the objectives were stated as follows:

1. To determine the perinatal mortality rate.
2. To examine the proportion of women using antenatal services.
3. To determine the proportion of pregnant women whose delivery took place in a medical care facility.
Based on these objectives, the following set of questions could be compiled as a section of a questionnaire used for the study.

### ANTE-NATAL CARE OFFICE

**USE ONLY.**

1. During your last pregnancy, did your health cause you any worry?

   - 1 = Yes
   - 2 = No
   - 3 = Don’t know.

2. During your last pregnancy, did you ever seek medical care (Western or Traditional) because you were ill?

   - 1 = Yes
   - 2 = No
   - 3 = Don’t know.

3a. During your last pregnancy, did you visit a health worker (Western or Traditional) for a general check up of your pregnancy and not necessarily because you were sick?

   - 1 = Yes (Go to Q3B)
   - 2 = No (Go to Q4)
3b. Whom did you see?

( ) 1 = Doctor
( ) 2 = Nurse/Midwife
( ) 3 = Other health workers
( ) 4 = Patent medicine dealer
( ) 5 = Traditional Medical Practitioner
( ) 6 = Others: Specify

4. Where did your last delivery take place?

( ) 1 = Health centre
( ) 2 = Hospital
( ) 3 = Home
( ) 4 = Home of traditional med. practitioner
( ) 5 = Others: specify

5. Who assisted you during delivery?

( ) 1 = Physician
( ) 2 = Nurse/Midwife
( ) 3 = Trad. Med. Practitioner
( ) 4 = Relative
( ) 5 = Did not receive any assistance
( ) 6 = Others: specify
( ) 7 = Don’t know

6. What is the status of the baby since delivery?

( ) 1 = Alive
( ) 2 = Stillbirth
( ) 3 = Died within first week
In formulating questions for each variable or research objective, care should be taken to ensure that the questions are clear and unambiguous. They should convey the same meaning to different respondents. Questions which suggest a specific answer need to be avoided as this will tend to bias the results. This is particularly common in situations where a researcher has set out with a fixed perception of the nature of the results of his study. The investigator therefore unconsciously tends to ask leading questions which will provide him with the expected answers to confirm the predetermined results.

In developing countries, it can sometimes be quite difficult eliciting accurate answers on age, number of children and some other variables due to cultural or other related reasons. In such situations it may be useful to have one or more questions to act as a means of verifying the authenticity of the answer to the main question. As an illustration, the question: "What is your age?", can be followed up with another question: "What is your date of birth?". The second question is used to verify the answer to the first.

Layout of the Questionnaire

After formulating the questions in relation to the objectives of the study, the next step will be to arrange the questions in a logical order. This helps to maintain the interest of a respondent. As a guide, background questions on age, sex, and occupation usually appear in the first section of the questionnaires. Difficult or sensitive questions are left until the tail end of the questionnaire. In order to further improve the response rate, the layout of the questionnaires should not be overcrowded. Enough space should be provided between questions and most
importantly, the questionnaires should not be too long as this tends to put off many respondents even before they have had the opportunity to know the contents or reasons for the questionnaire.
It has been highlighted in previous chapters that a researcher expects that at the end of an epidemiological investigation, his findings will unearth new facts, truths or knowledge which will be of benefit to the improvement of health and health care. The data collected during the course of his
investigation is usually in a raw form, consisting of a mass of numbers or responses to questionnaires both of which at face value are of little help in providing answers to the research question. The raw data will still need to be analysed to enable the researcher highlight easily and at a glance the important features of the results of the study. It is on this basis that answers to the research question are provided.

A plan for data analysis will include a careful consideration of the method to be adopted which may vary from manual analysis to the use of a microcomputer. Although desktop computers are widely available today and can perform various complex analysis in a relatively very short period of time, the researcher should not lose sight of the fact that data entry into the computer can be very time consuming, expensive and error prone, especially when the task is performed by poorly trained data entry clerks. Use of computers for data analysis can therefore only be cost-effective when the sample size and total number of variables are relatively large.

The types of statistical analysis open to an epidemiological researcher are quite wide and range from very simple calculations such as computing a percentage to those that need complex statistical procedures. Analysis of data in epidemiological studies will therefore invariably require at least a very rudimentary knowledge of biostatistics. A book of this nature can only be confined to common statistical analytical procedures. For details of more advanced statistical methods, readers should refer to any of the numerous standard textbooks of medical statistics.

Most researchers with basic knowledge of biostatistics still realise that there is a need to seek the advice of a biostatistician on the type of analysis suitable.
for their study. In such instances where a biostatistician needs to be consulted, this should be done during the planning phase of the study. Supposing for instance that the analysis is to be done with a computer, consulting the statistician at the planning stage of the study gives him enough opportunity to advise the investigator on the layout and coding of the questionnaire to ensure that the data is in a form that can be entered into and analysed by the computer. Consulting a statistician at the planning stage of the study also ensures that the investigator does not collect data which at the end turns out to be difficult or impossible to analyse. For researchers with some knowledge of statistics, dummy tables may be constructed as a basis for further discussion with the statistician as regards to the best form of analysis suitable for the study. Dummy tables are also required by some agencies considering a research proposal for funding. It enables them to assess whether the results of the study are likely to meet its objectives.

The type of analysis ultimately adopted by a researcher will to a large extent be determined by the objectives and type of study being contemplated. In a descriptive epidemiological study for instance, the analysis will be limited to the generation of summary statistics. This is in contrast to analytical studies where in addition, evidence and strength of associations will also be explored. The type of data collected by a researcher is another major determinant of the nature of analysis required for a study to reach valid conclusions. Data may be classified as quantitative or qualitative.

Quantitative Data

Quantitative data include all observations which can be...
expressed as numbers or quantities. In other words the data is measurable. Examples include age, weight, height, number of children or pulse rate. Quantitative data can be further classified as either quantitative continuous or quantitative discrete. The former include those data forms which can be expressed both as whole numbers or fractions of whole numbers such as weight, height, and temperature. These variables can therefore be measured on an uninterrupted scale. Quantitative discrete data on the other hand can only be expressed as whole numbers or integer values. It is obtained by counting rather than measurements and also can only take positive values. An example is the number of children per family which cannot be expressed as fractions of children.

Qualitative Data

Qualitative data as the name implies is an expression of the quality of a variable. It is further classified into Qualitative Nominal and Qualitative Ordinal. Qualitative nominal data is represented as categories or names which cannot be arranged in any logical decreasing or ascending order. Examples are sex, occupation or race. In situations where the nominal data can only take either of two values such as present/absent, positive/negative or yes/no, it is referred to as quantal, binary or dichotomous data. In qualitative ordinal data, although the variables are not measurable, the categories can be expressed or arranged in a logical order in terms of their magnitude. A proviso however is that the number of categories should not be less than three. Examples are educational level and social class.
THE FREQUENCY DISTRIBUTION TABLE

The simplest form of data analysis involves the production of a frequency table for each variable. A frequency table shows the number of occurrence of each variable. The categories of the variable are listed in one column and the number of times of their occurrence in another column as shown below.

**TABLE 10.1**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>6</td>
</tr>
<tr>
<td>Female</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
</tr>
</tbody>
</table>

Manual analysis of data to produce a frequency table may be accomplished by either hand tallying or hand sorting. Each variable is usually considered and analysed individually. In hand tallying, each variable is listed showing its various categories. Against each category of the variable, the frequency of occurrence is indicated. Each frequency is first represented as a stroke, the fifth stroke crossing the preceding four strokes. One such set of strokes therefore represents a count of five which simplifies the exercise of adding up the total frequency for each variable. An illustration is shown below.
The hand sorting method of manual analysis as opposed to hand tallying is used when each study unit has its own separate record form or sheet. Taking each variable at a time, the records are arranged in piles corresponding to the various categories of the variable. As an illustration, in analysing sex, the record forms are sorted into two piles representing males and females respectively. The total number in each pile is counted and recorded against the appropriate sex in the table.

A table should always contain a title which describes its contents, including the source of the data. The row and column headings should be clearly stated with the appropriate totals shown at the end. The simplest type of table is the univariate table which consists of two columns made up of the variable and its frequency of occurrence. An illustration of a univariate table showing the frequency of occurrence of different types of sexually transmitted diseases (STD) is shown in Table 10.3.
Absolute total numbers are unsuitable for the presentation of results in an epidemiological study because they do not take into account the population in which the event occurred. In addition to the total number of events, more information can be derived from a frequency table by generating summary statistics such as percentages, ratios, rates, measures of central tendency and measures of dispersion.

**RATIOS:**

A ratio is a relative number expressing the magnitude of one event relative to another and is a useful statistic for comparing two events. The numerator and denominator are separate quantities and neither is included in the other. An example is the ratio of males to females in a rural community which can be expressed as 1:3; 2:6; 4:12; or 8:24; all of which mean that for every male, there are three females.
PROPORTIONS

This is a type of ratio in which the numerator is included in the denominator. A typical proportion is the percentage. As an illustration, the proportion of males in a rural area relative to the entire rural population may be expressed as:

\[
\frac{\text{MALES}}{\text{MALES} + \text{FEMALES}} = \frac{5000}{5000 + 7000} = \frac{5}{12} \approx 40\%
\]

RATES

This is the basic tool or unit of measurement in epidemiological studies. It is more suitable than ratios and proportions because it takes into consideration both the total population and a specified time period. The denominator is multiplied by a co-efficient to obtain a unit of population. This forms the basis for comparison of events in similar units of different populations over the same period of time. The rate can therefore be expressed as:

\[
\frac{X \text{ events in 1 year}}{Y \text{ population in 1 year}} = \frac{X}{1000} \times \frac{1000}{Y}
\]

When \( Y \) is restricted to only those at risk of acquiring a disease, it is referred to as the population at risk. Some examples of morbidity and mortality rates commonly used in epidemiological studies are as follows:

1. Incidence Rates:

This is a measure of the rate of occurrence of new cases of a disease. It is defined as the number of new cases of a
disease per unit of population at risk over a given period of
time (usually one year). Mathematically this is expressed as:

\[
\text{Number of people who develop a disease} \times \frac{1000}{\text{Population at risk}}.
\]

Incidence rate measures the risk of healthy people
developing a particular disease. It can relate to either the
number of people developing the disease or to the number
of episodes or spells of the disease. One individual may
therefore contribute more than one episode of the same
illness. Incidence rate is therefore sometimes expressed in
terms of spells or episodes of illness as follows:

\[
\text{Number of spells or episodes of illness} \times \frac{1000}{\text{Population at Risk}}.
\]

The population at risk in a year is usually taken as the mid-
year population because the actual population at risk is
constantly changing due to births, deaths and migrations.

2. Prevalence

Prevalence is a measure of the number of existing cases of
a disease at a given point in time. This is also referred to
as the point prevalence rate as opposed to period
-prevalence which is a measure of the number of existing
cases over a period of time. It is worthy of note that unlike
the incidence rate, the denominator for prevalence rate

includes the entire population and not just those at risk. This is because the numerator contains both old and new cases. Prevalence rate is an expression of the burden of disease on the community and is most appropriate for chronic disorders. If the study is however on an acute illness, of short duration, the incidence rate is a better measure to use. Non survivors of the disease will be missed out if a researcher used the prevalence rate to express acute disorders such as cardiovascular accidents.

3. Attack Rate

This is a form of incidence rate but is used in the study of epidemics whose durations are usually very short. It is important to note that the unit of population is one hundred instead of one thousand and is expressed as:

\[
\text{Number of new cases of a disease} \times \frac{100}{\text{Population at risk}}.
\]

4. Crude death Rate

This is defined as:

\[
\frac{\text{Total number of deaths in a year}}{\text{Average or mid year population}} \times 1000.
\]

The crude death rate is a measure of the probability or risk of dying for an individual in a given community. The crude death rate is influenced by a number of factors, the most important being the age distribution of the population. The higher the proportion of elderly people in the population,
the higher the crude death rate for that population.

To compare death rates of two populations that differ in terms of age distribution, it will be more accurate to examine their age-specific death rates or in the alternative adjust their death rates for differences in age distribution. Adjustment of rates is carried out through a statistical procedure called standardization. In spite of its limitations, analysis of data by crude death rate is widely used in epidemiology for comparison of the health status of communities because it is relatively easy to calculate. It requires only a knowledge of the number of deaths and the population size. It also constitutes just one summary measure which therefore makes comparisons easy.

5. Infant Mortality Rate

\[ \text{No. of deaths in children less than 1 year of age} \times \frac{1000}{\text{No. of live births in 1 year}} \] 

6. Maternal Mortality Rate

\[ \frac{\text{No. deaths ass. with pregnancy, childbirth & puerperium}}{\text{No. of live & still births}} \times \frac{1000}{\text{per 1000 of the population}} \]

7. Perinatal Mortality Rate

\[ \frac{\text{No. of deaths in infants less than 1 week old}}{\text{No. of live & still births}} \times \frac{1000}{1} \]
8. Case fatality rate

\[
\frac{\text{No. of deaths from a given disease}}{\text{Total number of cases of the disease}} \times 100
\]


\[
\frac{\text{Number of deaths from a given cause}}{\text{Total number of deaths in a given period}} \times 100
\]

It should be noted that the denominator for proportional mortality ratio is the total number of deaths and not the population at risk. It is therefore called a ratio and measures the proportion of deaths from a given cause in relation to all other causes of death.

**AVERAGES:**

Average is a term which is quite common in everyday expressions, such as average speed, average height or average weight. It is a single value which is typical of a series of values. Averages are sometimes referred to as measures of central tendency because they tend to lie centrally within a set of values arranged according to magnitude. Averages represent the most useful summary statistics for continuous data. Many other statistical computations for continuous data are based on the average. Three commonly used averages are the arithmetic mean, the median and mode.
The arithmetic mean of a set of measurements is the sum of all the values divided by the total number of observations. As an example, if the ages in years of five children suffering from diarrhoea are 10, 15, 6, 4, 5, the arithmetic mean of their ages will be $\frac{10+15+6+4+5}{5}=8$.

The median is the midpoint of a set of values which have been arranged in order of magnitude. When there are an even number of observations, the mean is obtained by computing the arithmetic mean of the middle two values. As an example, the median of the ages of the five children with diarrhoea is computed by first arranging the ages in order of magnitude as follows: 4, 5, 6, 10 and 15. The value at the midpoint is 6 and this represents the median age. The median is a better measure of central tendency than the arithmetic mean when the distribution contains outlying values which are much higher or lower than the rest of the values. Consider a situation in which the ages in years of five patients are as follows: 10, 14, 16, 20, and 75. The arithmetic mean will be 29 compared to the median value of 16. The relatively higher arithmetic mean is as a result of the effect of the outlying value of 75.

The last measure of central tendency, the Mode, is the value with the highest frequency of occurrence. It can be easily identified at a glance after the values have been represented in a frequency distribution table.

**DIAGRAMMATIC PRESENTATION OF DATA**

In addition to tables and summary statistics, data can also be presented as diagrams. Pictorial presentation of data is a useful means of drawing quick attention to important characteristics of the study results such as trends, associations and clustering of variables. It is also a useful
tool for comparison of different variables. The commonly used types of diagramatic presentation of data are: Graphs, histograms, bar charts, frequency polygons and pictograms. Details of how to construct these diagrams can be found in standard statistical textbooks. Only the histogram, pie chart and bar chart will be described briefly here. (See Fig. 10.1-10.3)

The Histogram

This is a diagram constructed on the X and Y axis typically used for a graph and is based on a frequency distribution table. The X-axis represents the independent variable while the Y-axis represents its frequency of occurrence. Bars or rectangles are drawn in such a way that their base rest on the X-axis while the height of the bars is proportional to the frequency of the variable under consideration. There are no breaks between the bases of the rectangles. Histograms are used to present data involving continuous variables. When the midpoints of the tops of the bars in a histogram are joined to one another, a FREQUENCY POLYGON is produced.

The Bar Chart

This is similar in shape to a histogram. A major difference however is that unlike the histogram there are gaps between the bases of the bars or rectangles. This is because the bar chart is used for the presentation of nominal, ordinal and discrete data.
Pie Chart

The pie chart is a useful tool for the comparison of various events in a study population. As the name depicts, a pie or circle which represents the total frequency of occurrence of an event is divided into segments in such a way that the angle enclosed at the centre of the circle represents a proportion of the frequency relative to the whole circle. If for instance, out of a total number of 100 children with malaria, 60 are males, the corresponding angle enclosed at the centre of a pie chart as a representation of male children is computed as $\frac{60}{100} \times 360 = 216$. 

ASSOCIATION BETWEEN VARIABLES

As indicated earlier in the chapter, the next step after the computation of descriptive statistics is to look for evidence of association between variables. Most studies will start by looking for evidence of association between dependent and independent variables such as age, sex, occupation and marital status. For qualitative data, this is achieved through a cross-tabulation of variables while a scatter-gram or line graph serves a similar purpose for quantitative data. A two by two table is the simplest cross tabulation used mostly in analytical studies to examine for evidence of association between variables (See Table 10.4).
Fig. 10.1 A Histogram

Fig. 10.2 A Bar Chart

Trichomoniasis 32%
Syphilis 25%
Gonorrhea 43%
If $A/A+C$ differs significantly from $B/B+D$, it can be said that there is an association between the suspected aetiological factor and the occurrence of the disease. As an example, a 2 X 2 table for the examination of the relationship between smoking and lung cancer is shown below in Table 10.5.

**Table 10.5**

<table>
<thead>
<tr>
<th>WITH LUNG CANCER</th>
<th>WITHOUT LUNG CANCER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMOKERS</td>
<td>80</td>
</tr>
<tr>
<td>NON SMOKERS</td>
<td>20</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100</td>
</tr>
</tbody>
</table>

If $A/A+C$ differs significantly from $B/B+D$, it can be said that there is an association between the suspected aetiological factor and the occurrence of the disease.
The Relative Risk

The strength of the association between a suspected aetiological factor or characteristic and the occurrence of a disease is measured by a ratio called RELATIVE RISK (RR). It is defined as the ratio of the incidence rate of a disease in the exposed group to that in the non-exposed group:

\[
\text{RR} = \frac{\text{Incidence rate of a disease in the exposed group}}{\text{Incidence rate of a disease in the non-exposed group}}
\]

The relative risk is commonly used in analytical studies. If the relative risk of cancer in smokers compared to non-smokers is 3, this means that the risk of lung cancer among smokers is 3 times that of non-smokers. Unlike in case-control studies, since incidence rate can be derived directly in cohort studies, it follows that RR can also be calculated directly in this type of study. This is as opposed to case-control studies where RR is calculated indirectly using a cross-product of the frequencies in the 2 x 2 table (Table 10.4):

\[
\text{RR} = \frac{\text{AxB}}{\text{BxC}} = \frac{\text{AD}}{\text{BC}}
\]

This cross-product is also referred to as the ODDS RATIO (OR) and approximates to the relative risk. An indirect estimation of the relative risk or odds ratio is however valid under the following conditions:
1. The disease has a low incidence in the general population.

2. Both the cases and controls selected for study must be representative of all cases and non cases in the population.

Details of the mathematical derivation of the odds ratio or relative risk in a retrospective study can be found in the more advanced textbooks of epidemiology suggested for further reading.

Having determined that there is an association between a suspected factor and a disease, however strong this association happens to be, it could still have occurred by chance. An assessment of whether the observed association occurred by chance is determined using a statistical test of significance. A level of significance is a measure of the probability that an association occurred by chance. An arbitrary level of 5% is commonly chosen for statistical computations. A 5% level of significance means that in 5 out of every 100 times, the observed association could have occurred by chance. In other words, in 5 times out of 100, an association will be said to be significant when actually none exists. Commonly used statistical tests of significance are the Chi-square ($\chi^2$) test for enumeration data and the $t$-test for continuous data. Details of their computations and that of other significance tests used in the analysis of epidemiological studies can be found in the biostatistical textbooks suggested for further reading.

**Attributable Risk. (AR)**

This is another analytical measure of risk commonly used in epidemiological research. It is defined as the difference
between the incidence rate in the Exposed and Non-exposed groups. As an example, if in a study, the incidence rate of cancer among smokers is 5 per 1000 and that for non-smokers is 3 per 1000, the attributable risk will be 2 per 1000. This represents the rate of the disease in the exposed individuals that can be attributed to the exposure. In the above example, 2 deaths per 1000 of the study population are occurring as a result of smoking and therefore represents the number of deaths that can be prevented by eradicating smoking.
Writing of the research report is the culmination of the entire research process. The research report is the only objective evidence that a research project had been
successfully completed. Until a research is documented in the form of a report, all the efforts, resources and time expended in conducting the study would have been a colossal waste. Calnan has described an unpublished research as sterile (1).

It will be recalled that in the first chapter of this book, the research process was described as the search for new knowledge. This new knowledge can only be useful when it is shared with others. Although this can be achieved orally, the most effective and enduring means of communicating and disseminating one’s research findings is through a research report. The report affords the researcher an opportunity to present details of what was done with the important findings. This information will be useful to other health professionals who are seeking ways of improving health care. Fellow researchers reviewing the literature and who wish to learn from the experiences of previous researchers on a similar subject will also find the report both useful and stimulating. The research report also offers its readers especially other researchers an opportunity to evaluate the entire process leading to the research findings with a view to assessing the reliability and validity of the results.

A research report differs slightly in style and format from other types of published reports such as is found in magazines, newspapers and novels. Whereas these other types of publications are mainly for entertainment, a research report on the other hand is a scientific document which sets out to present only relevant information in an accurate but concise and objective form. This does not however mean that the report should be full of technical terms as this tends to make it less interesting and sometimes even difficult to read and understand by majority of its target audience. If highly technical terms
need to be used in the report, their meaning need to be explained. This is because though some technical terms are universal among health professionals, others may be too localised to a particular profession and therefore when used in a report without explanations may create barriers in communication. A good report is written in the third rather than the first person. A research report will therefore contain the statement "Data was collected from the records of two hospitals", instead of "I collected data from two hospitals".

Although a researcher can learn to write a research report by reading as many published reports as possible, this chapter aims at providing a general guide to the nature of a scientific research report. An aspect of the style of a research report that cannot be covered here is on methods of expression of the English language. Suffice to say that an appropriate and correct use of verbs, nouns, adverbs, adjectives, present and past tenses are all essential for the writing of a readable and interesting report. The use of too many unduly long sentences and inappropriate use of paragraphs can also make a research report dull and uninviting. A good dictionary and thesaurus need to be kept handy while writing the report.

The writing of a research report is a chore that cannot be rushed. A number of drafts have to be written and modified before the final draft is ready. Each draft is usually kept for a while before returning to it for modification. This process can be quickened if colleagues or supervisors are requested to criticise and comment on the initial drafts.

Research reports may be submitted as a thesis or dissertation for an examination, or as an article for publication in a medical journal. The main difference between the three types of research report is essentially in
their length. The journal article is the shortest of the three and in addition, the sections on introduction and literature review are usually integrated into a single section. Due to the slight variation in the arrangement of various sections of a report, it is always advisable to first review the guidelines of a relevant examining body, a funding agency or the editorial policy of a journal before writing the report. Some guidelines may for instance require that the aims and objectives be incorporated into either the introduction or the literature review. Even when the section on the aims and objectives is standing on its own, some examination bodies require that it is placed before the section on literature review while others prefer that it is placed after the literature review.

Scientific research reports usually follow a standard format with only very slight variations in the arrangement of the various sections. This standardization enables readers to quickly identify the various types of information contained in the report. It also makes for ease of comparison of findings of different researchers. Finally, this standard format forms the basis for an initial outline of the research report. A close examination of the general format of a research report will show that some sections would have been completed at the time the proposal was written. These already completed sections are therefore simply incorporated into the main report, sometimes with very minor modifications. While incorporating these sections into the main report, it should be remembered that the proposal was written in the future tense while the main report is in the past tense. The usual outline of a research report is as follows:
THE TITLE

The title of the report is the first part that a prospective reader comes across. It is therefore like a gateway to the report. The title should be attractive and at the same time should convey the contents of the report in the most effective and efficient way. Vague and very long titles should be avoided. One way of shortening titles is to avoid such common phrases as "A study of", "An investigation of", and "An examination of". Note however that other phrases such as "An evaluation of" are necessary because they may convey some information on the nature of the study, in this instance, an evaluative study.

THE SUMMARY OR ABSTRACT

Although this section appears at the beginning of the report, it is usually written after the entire report is completed. This is because as the name indicates, it is a summary of the entire report and can be likened to a miniature photograph of the entire report. The abstract is
the first section that a reader will usually go to after being attracted to it by its title. Some readers who have a very busy work schedules will infact confine themselves to this section especially when they find it uninteresting. It therefore becomes very important that the summary or abstract should incorporate in a concise form all the important points in the report. An abstract is usually between 150 to 300 words and should contain information on how a research problem was investigated, the important findings and the possible implications of the results. This is illustrated in the example below in which the author has presented a precise account of the type of study, the study population, the most important findings and their implications.

Eg. A survey of obstetrical complications associated with singleton breech presentation and delivery (2).

A retrospective analysis of obstetric complications associated with singleton breech presentation and delivery over a ten year period in a university hospital is presented. The incidence of premature rupture of membranes (6.2%), cord prolapse (2.1%), caesarean section rate (28.1%), severe asphyxia at birth (13.7%), low birth weight babies (26.2%), and perinatal mortality (20.4%), were higher than the corresponding figures for the general obstetric population. A closer surveillance on all breech pregnancies and perhaps the use of external cephalic version may reduce these complications.
THE INTRODUCTION

Since the summary is written after the completion of the entire report, the introduction is in reality the starting point of the report. The introduction usually contains all or some of the following: a statement of the problem, the theoretical framework of the study, its rationale and significance. The aims and objectives or purpose of the study is sometimes incorporated into this section. Technical terms used in the study are also defined or clarified. Although the introduction is one of the sections that would have to a large extent been completed after the proposal is written, it is being discussed again here because it contains the opening sentences and paragraphs of the report. These are sometimes referred to as the lead. The lead is very important because it is expected to draw the interest of the reader to the rest of the report. The opening sentences should therefore be captivating and strongly appealing.

Any of the above component parts of the introduction can be used as a basis for the opening sentences or paragraphs. A number of journal articles will be used to illustrate the various ways of starting the research report.

1. Statement of the purpose of the study.

Eg. "Nutrition in the cancer patient"(3).

The interactions between nutrition and cancer are many and varied ranging from the role of nutrition in the aetiology of disease to its supportive role in its treatment. This paper briefly...
reviews some of these interactions. In the above example, after a general statement on the interaction between nutrition and cancer, the author goes straight to a statement of the purpose of the paper. A reader is told at the outset what the article is all about and the scope of the problem that will be covered.

This is one of the simple ways to start a research report. It is quite useful in the discussion of disease entities or aspects of medical practice where there may be a need to remind readers of the meaning of the subject under study. This is particularly important in situations where the subject could be open to different interpretations. Starting the report with a definition will therefore enable the reader to put subsequent related aspects of the subject in their proper perspective.

Fig. The Epidemiology of Subacute Sclerosing Panencephalitis in England and Wales 1970-1989. (4).

Subacute sclerosing panencephalitis (SSPE) is a rare degenerative central nervous system disease characterised by personality change, progressive loss of cerebral function, myoclonus, paraplegia, coma and death. A viral aetiology was postulated.
by Dawson in 1933 from his observation of intranuclear viral inclusion bodies in the grey matter from fatal cases.

3. Statement of the Importance of the Problem.

This is one of the most common methods of commencing a research report. The potential of attracting and sustaining the interest of a prospective reader is further enhanced if statistics are used to highlight the problem. As previously discussed in chapter two, a health problem may be important because of any of the following reasons: high mortality, morbidity or disability rate, severity, its socio-cultural and economic impact. One other common reason for researching on a problem is inadequate or lack of information on the subject. The two examples below illustrate various ways a report could be started based on different aspects of the importance of the problem.

Eg. “Bacterial Aetiology of Childhood Meningitis in Port Harcourt”. (5).

Bacterial meningitis remains a major cause of morbidity and mortality in the paediatric age group. Several reports from different parts of Nigeria suggest that the commonest organism isolated from cerebrospinal fluid in patients with pneumonia is *streptococcus pneumoniae.*
Ranked third among malignant tumours affecting patients in Nigeria (after cervix and breast cancer) primary hepatocellular carcinoma remains the most common malignancy affecting men in much of Africa and Asia, and several recent surveys continue to show rising death rates with no evident reduction in annual incidence rates. So common is this condition that diagnosis is readily made in most hospital centres even in the absence of tissue biopsy.

Several other ways in which a research report could be started are illustrated in the numerous articles published in various Biomedical journals. Further development of the research report is achieved by establishing a logical link from one idea to another. A report may for instance start with a definition and then progress to discussions on the size of the problem, its causes and other related aspects.

MATERIALS AND METHODS.

The section on materials and methods usually would have been developed to a very advanced stage at the time the
A research proposal is written. Writing up this section involves a detailed description of both the target and study population, the type of study design adopted, sampling method and the sample size, methods of data collection and the variables studied. The technique of data analysis is also described. A description of the study population will include information on its geographical location, vegetation, main occupation of the people, climatic conditions, types of social and health facilities available, population size and distribution. This can be illustrated diagrammatically as shown below:

<table>
<thead>
<tr>
<th>Target Population</th>
<th>Study Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type of Study</td>
</tr>
<tr>
<td></td>
<td>Sampling</td>
</tr>
<tr>
<td></td>
<td>Data Collection</td>
</tr>
<tr>
<td></td>
<td>Variables</td>
</tr>
<tr>
<td></td>
<td>Data Analysis</td>
</tr>
</tbody>
</table>

**THE RESULTS.**

In addition to the use of tables, figures and diagrams to present the results, it should also be reported in this section in such a way that it can be understood by the reader even without referring to the tables and figures. They are only there to enhance and assist the reader in appreciating the important findings in the study. The tables may be placed...
in this section or as an appendix at the end of the report. Wherever they are placed however, they should be clearly labelled. The totals in the tables should add up correctly especially percentages. Results should be presented logically in relation to the objectives of the study.

DISCUSSION

In this section, the results are interpreted and their significance or implication highlighted. It is also during the discussion that the findings in the study are compared with those of previous ones, pointing out similarities and differences. The researcher is also expected to give his opinions on the reasons for these differences and similarities in findings and also its implications especially in relation to either health care or areas needing further investigation. It is important to note however that in commenting on differences in findings, not all differences are real. Some may be artificial arising from differences in research methodology especially in relation to variations in study populations and data collection techniques. In discussing the findings it is also important that the researcher acknowledges the limitations of these findings. This is important in order that the interpretations of the findings can be put in their proper perspective. Some reports insist start the section on discussion with a description of these limitations. As an instance, a very low response rate or difficulty in ascertaining the ages of respondents in a study will grossly limit the extent of generalizability of the findings.

CONCLUSIONS AND RECOMMENDATIONS

This is the last section of the body of the report. It
describes the extent to which the aims and objectives of the study have been achieved. A common method of concluding a research report is by summarizing the most important points contained in the report. The summary acts as an evidence justifying the achievement of the study objectives. A report can also be concluded by highlighting the practical implications or applications of the research findings. Many researchers have been known to make very wide generalizations in their conclusion which are not supported by evidence from the study. As an example, a study based on fifty undergraduates in Enugu could not be used as basis for a statement on all undergraduate students in Nigeria. Aspects of a study which were not conclusive and merit further investigation may also be described in this section. Recommendations for solving the problem under investigation follow logically from the research findings. They are also presented in this section, preferably as a numbered list.

REFERENCES.


Suggested further reading


INDEX

Abortion, 6
Anasmin, 47
Attributable Risk (AR), 128
Bar Chart, 123
Bias, 81, 82
Bladder Cancer, 75, 76
Budget, 27, 28
Cancer, 97
Cervical Cancer, 80
Chi-square (X²), 128
Cholera, 15, 68. See also Snow, John and Kock, Robert
Coffee Cancer, 75
Community diagnosis, 9, 10
problem of, 10
Cruze death rate, 119, 120
Data, 4, 96-99, 112, 122, 133
Census, 94
Collection, 95-97, 102, 103, 110
interpretation of, 4
morbidity, 99
mortality, 98
qualitative, 113
Diabetes, 101
Diagnosis, 9, 13, 98, 101, 102
Death certificate, 98
Definitions, 137
Descriptive Epidemiological Studies 68, 69
Doll and Hill, 16
Education, 8
Epidemics, 6, 13
Epidemiology, 5, 118, 128

definition of, 6, 9


contributions of, 8, 9

reductions of, 7

Excerpta Medica, 34

Frequency polygon, 60, 123

Frequency table, 114

Gaussian Distribution Curve, 60

Gantt Chart, 26

Health, 7, 10

budget, 7

definition of, 10

improvement of, 6

problems of, 8, 20, 21-22, 40, 67, 99

Health Care, 7, 8, 9, 24

concept of 7

Hippocrates, 6

Histogram, 60, 123

Hypothesis, 4, 12, 13, 43, 44, 70, 73, 88

aetiological, 80

Null, 43

Index 34, 36

Kock, Robert, 15

on Vibro Cholera, 15

Lab-medicine, 7

Literature review, 24, 31-38, 44, 133

Lung Cancer, 16, 74, 79, 126. See also Doll and Hill

Malaria, 6, 13, 69, 71

meaning of, 6

Matching, 79, 80

Medical research, 5, 8

"Mission theory", 6

Mortality, 30, 22, 29, 41, 69, 70, 98, 105, 122

infant, 23, 24, 41, 43, 118

maternal, 21, 46, 118

Robert Kock on Vibro Cholera, 15

Index 34, 36

Kock, Robert, 15

on Vibro Cholera, 15

Lab-medicine, 7

Literature review, 24, 31-38, 44, 133

Lung Cancer, 16, 74, 79, 126. See also Doll and Hill

Malaria, 6, 13, 69, 71

meaning of, 6

Matching, 79, 80

Medical research, 5, 8

"Mission theory", 6

Mortality, 30, 22, 29, 41, 69, 70, 98, 105, 122

infant, 23, 24, 41, 43, 118

maternal, 21, 46, 118
Murine typhus, 13
Null hypothesis, 43.
See also hypothesis
Occupational medicine, 15
Father of, 15. See also
Ramazzini
Odds Ratio (OR), 127, 128
Onchocerciasis infection, 71, 72
Pathology, 10
Pie chart, 124
Pneumonia, 69
Population census, 97
Prospective Study Design, 81
Quarterly Bibliography of Tropical Diseases, 34, 35
Questionnaires, 103, 104, 112
Ramazzini, 15. See also
Occupational medicine
Randomization, 91, 92
Randomized Controlled Trial (RCT), 90
References, 28, 29
Harvard system of, 28, 29
Vancouver system of 28, 29
Relative Risk (RR), 127
Religion, 43
Research, 1, 2-5, 17, 18, 20, 26
as an elitist activity, 2
benefits of, 4, 5
definition of, 3, 4,
edemiological, 5, 8, 10, 11, 12, 19, 46
medical, 5, 8

negative perception of, 2
planning a, 17, 18, 31
proposal, 18, 23, 24, 25, 26
report, 130-134, 137
topic, 19-22, 33
Researcher, 3, 17, 18, 23, 24, 132
motivating factor of a, 3
Sample, 47-50, 55, 58, 61, 64
accidental and convenience, 55
Sampling, 47-49, 61, 62
multistage, 54
non probability, 49
simple random, 49, 50
Stratified, 52, 53
Systematic, 51, 52
quota, 55
Sampling Distribution of the means, 62
Sampling frame, 49, 50, 51-54
Sampling interval, 52
Scurvy, 15
Smoking, 16, 42, 74, 79, 88, 129
See also Doll and Hill
Snow, John, 15
on Cholera, 15
Studies: epidemiological, 68, 72, 76
analytical, 72, 73
Cross-sectional, 73
Case-control, 74, 75, 80, 81
Study design, 66
Case-control, 80
experimental, 88, 89, 94
prospective or cohort, 81-83, 86
Study population, 46, 49, 58, 90
definition of, 46, 90
Epidemiological research involves an application of the basic knowledge from three distinct subjects: Epidemiology, Biostatistics and Scientific Research Methodology. Experience has shown that because these subjects have traditionally been taught as distinct disciplines in isolation to each other, students and even some health professionals have a lot of difficulty in integrating the theory from these three subjects into a practical framework for the design and execution of an epidemiological research project. This practical manual is therefore aimed at bridging this gap by providing an integrated and coordinated approach to epidemiological research. Some of the topics covered in the book include:

- The Research Proposal
- The Literature Review
- Formulating the study objectives
- Choosing the Study Population and Samples
- Sample Size Determination
- Study Designs
- Data Collection
- Analysis and Presentation of Results
- The Research Report

This book has been written primarily for undergraduate and post graduate students in the medical and nursing sciences. It is however also intended for use as a self-instructional manual by young practitioners in the following professions: Medicine, Nursing, Dentistry, Medical Laboratory Science, Public Health and Primary Health Care.