Data (information) are the building blocks of epidemiological research. Careful measurement, recording and handling of information are crucial to the research process. The nature of the data required must be precisely defined and its collection requires the use of reliable tools or instruments. Standardisation of definitions and procedures used in measurement helps to ensure comparability. Care in undertaking measurements and rigorous training of personnel carrying out the measurements help to ensure precision. Special care is necessary to ensure completeness and accuracy in recording the data. Information may be obtained from routine or especially designed forms and evaluated by counting deaths or cases of disease or measures of dysfunction. Questionnaires are a frequently used instrument for collecting information. They should include data identifying the individual being studied, followed by specific questions concerning diseases and determinants. To be useful, such questionnaires must be carefully structured. Diseases are usually compared by calculating ‘rates’ or ‘risks’, consisting of numerators (cases, events) and denominators (population). Prevalence and incidence are standard measures of ‘risk’. The collection of high quality information (data) and its use to construct estimates of ‘risk’ or risk ratios form the basis of epidemiology.

KEY WORDS: research; protocol; lung; education

TYPES OF DATA
Measurement produces numbers or creates categories that can be arranged in such a way as to identify patterns from which logical conclusions can be drawn. It is these numbers or categories (data) that are the ‘building blocks’ of research. If the building blocks are not sturdy (precisely defined and carefully gathered), the whole structure of the research is weak.

Continuous data are points located on a continuous scale of values. The numbers reflecting continuous data have no ‘discrete’ reality and flow into one another along a scale (e.g., age or height).

Discrete (categorical) data represent groups of unique descriptors (e.g., sex). Information as to whether an individual or object falls into one or other category is usually indicated by an answer of ‘yes’ or ‘no’ to that category. The categories are mutually exclusive and do not overlap. There may be more than two discrete but independent possibilities (e.g., number of children in a family).

INFORMATION COLLECTION
Research protocols must set out as precisely as possible the definitions of the diseases or determinants they are proposing to study. Failure to do so will impair the quality of measurements and jeopardise the ability to draw conclusions from the information collected.

Information is gathered using tools or instruments such as:

- Questionnaires and routine records
- Measuring devices such as spirometers and air sampling devices
- X-ray films
- Microbiological techniques.

Some measurements are very precise (e.g., sex), but when categories overlap (e.g., between diseases) or judgement is involved (e.g., reading radiographs), measurements are less precise.

Ensuring comparability: standardisation of definitions and procedures
Standardisation is one means of minimising methodological differences and allowing comparisons between studies. In some topic areas, definitions and techniques/instruments of measurement are established by international consensus. Where such recommendations exist, investigators should follow them or indicate the precise nature and rationale for departure from the recommendations.
Ensuring precision: minimising error in measurement

Errors in measurement may reflect carelessness on the part of study personnel, inadequacy or incorrect use of measurement tools and instruments, or failure to follow standard procedures.

To ensure maximum precision, the measurements taken by each observer should be compared with those of others and with a ‘gold standard’ (inter-observer error). If the measurements of an individual research worker vary greatly from those of others, the worker must either be trained to reduce variation or excluded from making the measurements. Measurements of the same values should also be taken more than once by each individual to ensure that the measurements taken on different occasions do not vary greatly (intra-observer error).

Completeness and accuracy of recording

No matter how careful personnel are in carrying out measurements or how precise the instruments are in providing information, if the data are not carefully compiled and analysed they are of no value for scientific purposes. No amount of statistical manipulation will overcome problems created by sloppy handling of information.

The amount of error introduced in the stage of recording and compiling of data may be surprisingly large. Great care must therefore be taken to minimise such error and to estimate its extent.

Sources of information

Measurements used to categorise health status include death, disease incidence and prevalence, functional capacity, pre-morbid states and risk group. Relevant information may be obtained from a variety of sources, but to be useful for scientific purposes it must be reliable, complete and accessible.

 Routinely collected information used in the care of patients, while usually relatively accessible and relevant, is frequently neither complete nor reliable. The most useful routinely collected information is that which is collected on a field-tested standardised form such as are frequently internationally recommended. Information collected in this way is more likely to be complete and comparable from one patient to another.

Information collected specifically for study purposes, while more expensive, is frequently the only way to obtain reliable and complete information.

Mortality data

Information on deaths is usually obtained from routine notification of deaths in the Vital Statistics Register. This information is collected on a standardised form using an internationally recommended classification of causes of death. All rates referring to deaths are ‘incidence’ rates, as death is a discrete and time-limited event.

In measuring mortality, various rates are used, all referring to a specified period of time:

- **Crude death rate** is the probability of death from any cause in a defined population
- **Cause- (or age-) specific death rate** is the probability of death from a particular condition (or in a specific age group) in the defined population
- **Case fatality rate** is the probability of death from a condition in individuals with the given condition (patients).

Morbidity data

Morbidity is any departure, subjective or objective, from a state of physiological or psychological well-being. It usually refers to a specific disease or functional state. Measures of morbidity may be:

- Persons who are ill (cases)
- Periods or spells of illness that the persons experience (events)
- Impact (extent of functional impairment, service utilisation) of illnesses.

Data features

**Frequency** of a disease is important in selecting a study design. Cohort studies are inefficient for studying rare diseases. The case control design may be more practical.

**Severity** of a disease or dysfunction must be taken into account in comparing differences between groups.

**Duration** of a disease has an important influence in selecting a study design. Diseases of short duration cannot be studied efficiently using cross-sectional study designs.

**Latent period** is a feature of many chronic diseases caused by environmental exposures and infectious diseases which, where it exists, must be taken into account in the design of the study.

Routine records, if they are to be used in research, must be:

- simple, to ensure accuracy
- useful to the practitioners to encourage completeness
- based on agreed operational definitions with systematic training of recorders to ensure consistency of response
- regularly submitted, with the results fed back to recorders
- regularly monitored for completeness and accuracy.

Features governing the choice of study methods or instruments include:

- existence of standardised, recommended methods for research
- accuracy of the instrument or method
- durability of the instrument
- acceptability to participants
- potential hazards associated with their use
- maintenance of confidentiality.
QUESTIONNAIRES

These are frequently used to obtain and record information in epidemiological studies. The aim is to ensure that information is obtained in a systematic manner in accordance with pre-determined, standardised definitions and methods.

Questionnaires may be administered by face-to-face interview or self-completed by the study subjects.

Questionnaires should begin with a short statement, indicating:

- the auspices and purpose of the study
- how subjects (cases and controls) have been selected
- the procedure being followed (interview or self-completion)
- the right of subjects to decline to participate or to answer specific questions
- arrangements for safeguarding confidentiality, and
- the subject's signed consent to participate, when appropriate.

The subject's name and address and identification number should be recorded on a cover sheet. This is then detached and kept in a secure file separate from the remainder of the data collected to which only the investigator has access. The subject identification number alone is used to label all other data.

Introductory section

This usually records:

- Title of the project and subject identification number
- Date, time and place of interview and name of interviewer
- Age, sex, marital status, occupation and social data related to the subject.

All questions must be strictly relevant to the aims of the study. The number of questions should be kept to a minimum consistent with the need to ensure that all the information required to answer the study question is obtained.

The wording of questions is crucial to obtaining good quality information. The questions should be:

- Simple, short and uncomplicated
- Intelligible, using words from the normal vocabulary of the subject
- Unambiguous, with meanings of terms precisely specified
- Unbiased, avoiding emotive words and leading questions
- Interesting, relevant and interesting to subjects to sustain their concentration and willingness to participate.

The order of questions may crucially affect the answers, and needs careful thought. It is important that:

- Questions start with those that are easy to answer
- The sequence of questions flows easily and logically
- The answers to early questions do not prejudice answers to later questions
- ‘Check questions’ are well separated.

Response options are of three general types:

- ‘Yes’, ‘No’ or ‘Don’t know’
- Multiple choice, including ‘Don’t know’
- ‘Open-ended’ format used when exploring uncharted territory such as personal beliefs and attitudes (‘ethnographic studies’). The main disadvantage of this type is the difficulty of analysis and reaching general conclusions.

Questionnaire structure

Completion is facilitated by:

- Clear instructions
- Numbering of questions for reference purposes
- Indication of where questions may be skipped after a negative answer.

Analysis is facilitated by:

- Discrete questions and answers
- Use of ‘boxes’ for answers that can be numbered and pre-coded
- Aligning boxes in the margin to facilitate coding and computer entry
- Providing boxes for all possible responses to each question.

Finally:

- Always use good quality paper and print
- Use coloured paper to distinguish different questionnaires
- Use short questionnaires to achieve a high response rate.

MEASURING EXPOSURES

When collecting information on determinants, including exposures to potentially harmful (or helpful) substances, the same provisos apply as for collecting information on diseases. They include using standardised forms to record information, following standardised procedures, and recording as completely as possible the methods used to obtain the data.

Features of exposure to be considered

- Intensity: the level or magnitude of exposure
- Duration: length of time over which exposure has occurred
- Profile: whether exposure is roughly constant or a series of peaks with periods of no exposure
- Time period: biologically relevant period taking latency into account
- ‘Setting’ vs. agent: identifies environments that have suspected harmful effects but no known specific agent.
Instruments for collecting exposure information
• Interviews, questionnaires, diaries, records
• Measurements of the macro-environment
• Measurements of the personal environment
• Individual dose measurements
• Measurements of concentrations in tissue
• Biological markers of the direct effect of exposure.

Choosing an exposure instrument
A content expert, if available, should be consulted when addressing environmental exposures or, say, nutritional aspects. The power of a simple question-naire (e.g., a standardised smoking habit questionnaire) should not be overlooked as a tool to derive quantitative or semi-quantitative results, such as intensity of smoke exposure, degree of employment risk or poverty status.

RATES AND INDICES OF RISK
Epidemiological data refer to specified populations and often to defined periods of time. They are usually expressed as rates, which implies the presence of a numerator and a denominator. Errors may arise in measuring both numerators and denominators.

Numerator errors in routine statistics may arise from variations in:
• Measurement instruments (e.g., diagnostic tests)
• Case ascertainment procedures
• Recording systems.

Denominator errors may arise from:
• Population migration
• Variation in population structure
• Changes in administrative boundaries.

Errors, when they are systematic, can lead to bias which is ‘any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth.’

Calculation of rates
Prevalence refers to the number of events of a specific disease or condition in a given population at a designated point in time (point prevalence). It may refer to the number of events over a specified period of time (period prevalence). When the number of events is divided by the population at risk, it is termed the prevalence rate (ratio).

Prevalence is most frequently used to measure relatively common chronic diseases with a prolonged clinical course, such as asthma and chronic airflow obstruction.

Incidence is the number of new events of a specific disease or condition commencing during a given period in a specified population. When this is expressed per unit of time and of population at risk, it is referred to as the incidence rate.

Incidence rates are generally used to measure the frequency of diseases that are either relatively uncommon or which have a short clinical course. They are frequently estimated by using notification rates. This is appropriate only when case ascertainment is relatively high and an efficient information system is in place.

Service utilisation rates are frequently used to estimate the economic costs of disease or the general burden to health in the community. To reflect the latter adequately, services must be both accessible and equitable and reporting needs to be complete and accurate.

Measures of risk
Risk may reflect the probability of developing disease or of being cured related to the presence or absence of a defined determinant or therapy. The results of epidemiological studies are often expressed in terms of such probabilities or risk ratios per unit of population. Some commonly used expressions of risk (see Figure) are:
• Absolute risk: the observed or calculated probability of an event in a population under study. Indicated by the crude incidence in an exposed population.
• Relative risk (RR): the ratio of the incidence of disease or death in those exposed to a determinant versus the risk among those not exposed (synonym = risk ratio). Alternatively, the ratio of the cumulative incidence rate in the exposed to the cumulative incidence rate in the unexposed (the cumulative incidence ratio).
• Odds ratio (OR): the ratio between odds in favour of exposure among cases to odds in favour of exposure in non-cases. The odds ratio is a good approximation of the risk ratio when the incidence of the disease is low.

\[
\text{Figure} \quad \text{Indices for estimating risk. Absolute risk} = \frac{a}{a+b}; \text{relative risk} = \frac{(a/a+b)/(c/c+d)}{odds ratio} = \frac{(a/b)/(c/d)}{odds ratio}; \text{attributable risk (AR)} = \frac{(a/a+b) - (c/c+d)}{odds ratio}; \text{population attributable risk} = \frac{a+b}{n} \text{AR}/(a+b/n). \]
• Attributable risk (AR): the rate of a disease or other outcome in exposed individuals that can be attributed to the exposure. It is derived by subtracting the rate of the outcome among the unexposed from the rate among the exposed individuals.

• Population attributable risk (PAR): the incidence of a disease in a population that is associated with (attributable to) exposure to a risk factor. This reflects the proportion of all the cases in the population that is due to the exposure.

References