INT J TUBERC LUNG DIS 8(9):1151–1156

Choosing a research study design and selecting a population to study

D. A. Enarson,* S. M. Kennedy,† D. L. Miller‡

*** Scientific Activities Unit, International Union Against Tuberculosis and Lung Disease, Paris, France; † University of British Columbia, Vancouver, British Columbia, Canada; ‡ Emeritus Professor of Public Health Medicine, Emperor College, University of London, London, United Kingdom**

SUMMARY

Epidemiological studies have been standardised into a group of 'designs'. The descriptive study describes disease by time, place and person and can develop hypotheses about associations between disease and possible determinants. The analytic study tests these hypotheses. The cross-sectional study measures the disease and determinants at a single point in time. The cohort study identifies those within a group with or without a determinant, and observes the occurrence of disease in the two groups. The case-control study identifies a group of patients with a disease and selects a group of persons from the same population who do not have the disease, comparing the presence of a determinant in the two groups. The experimental study, a type of cohort study, is one in which the investigator 'assigns' the determinant (a treatment) to one subgroup in a population and compares the occurrence of a disease between those with and those without the determinant. All such studies must ensure that the comparisons made have relevance to a defined population. This is done by selecting a 'representative' sample from that population. Carefully selecting a study design and population facilitates the creation of new knowledge while avoiding, as far as possible, important errors.

KEY WORDS: research; protocol; lung; education

THE PROCESS of health research often begins with observations on a person with a disease (the case report). Assembling a number of cases of the same disease (the case series) then identifies the characteristics specific to the disease. Observational studies of the disease in larger populations using a standard approach (the descriptive study) can provide insights into possible determinants of the disease (hypothesis generation), while studies involving planned comparisons between groups provide stronger evidence of causation (the analytic study). Finally, the effects of an intervention are compared with those in another group to which the intervention has not been applied (the experimental study).

The basic framework of these study designs relates the presence or absence of a health-related state (usually disease) to the presence or absence of possible determinants (e.g., treatment, exposure, risk factor). The architecture of all the main types of epidemiological study derives from these two components (disease and determinants) (Figure).

STUDY DESIGN DETAILS

Descriptive studies can help to give perspective to the burden of disease and may assist in planning services.

Aims

- To measure the importance of, and monitor changes in, diseases in a community
- To describe the frequency of different diseases within a community

Methods

Descriptive studies analyse morbidity or mortality statistics or data on health-related variables looking for variations that correspond to patterns in the prevalence of possible determinants under three headings:

- Time, including: secular trends; cyclic changes; seasonal variation; epidemics
- Place, including: geographical; urban-rural; institution
- Person, including: age and sex; marital status; ethnicity; family; occupation and socio-economic status

Advantages of this design are that it is:

- Cheap and quick, cost-effective use of existing information
- A useful initial overview of a problem
- Useful to identify parameters for further study

Disadvantages include:

- Difficulty in identifying all cases, especially those that are rarely fatal or not usually medically managed

Correspondence to: Prof Donald Enarson, International Union Against Tuberculosis and Lung Disease (UNION), 68 boulevard Saint Michel, 75006 Paris, France. Tel: (+33) 1 44 32 03 60. Fax: (+33) 1 43 29 90 87. e-mail: union@iuatld.org

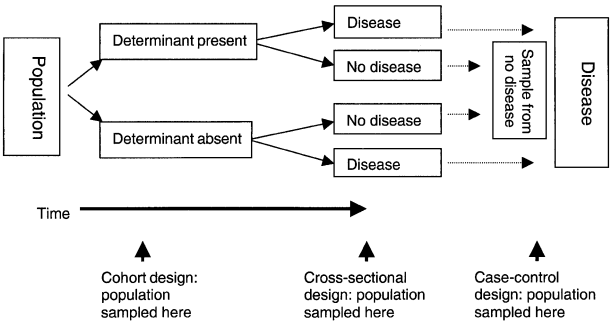

Figure Overview of study architecture.

- Data on related variables may not be available or not in required form
- Methods of data collection and diagnostic criteria often not standardised

ANALYTIC STUDIES

Hypotheses generated from associations between diseases and possible determinants in descriptive studies need to be tested by analytic studies. There are three types of analytic studies:

Cross-sectional studies

Definition

A study that aims to identify disease or health-related states and suspected determinants at a particular point in time.

Aims

- To test hypotheses on disease causation by showing the degree of correlation between possible determinants and disease
- To assist health service planning by measuring the burden of disease in subgroups and identifying those in greatest need of services

Methods

The research protocol should:

- Define the target population and how the subjects for study will be sampled
- Describe how the various subgroups for comparison will be identified
- Provide clear definitions for measurement and classification of disease and possible determinants
- Prescribe the study instruments and the training of the research team
- Set out plans for data management and analysis

Advantages of this design include the following:

- Results can be obtained relatively quickly and cheaply
- Large numbers of possible associations can be explored
- Methods of measurement of both determinants and outcomes are standardised

Disadvantages are that:

- Temporal relationship between determinants and disease is not always clear
- When disease is rare a large study population is required
- Recall of past events may be unreliable
- Population being studied comprises survivors of a cohort

Cohort studies

Definition

A cohort is a group of persons who share a common experience. A cohort study is one in which a group of persons who are free of disease is classified according to the presence or absence of a determinant, and then observed over a period of time to identify the appearance of disease.

Aims

- To show if the presence of a particular suspected determinant predicts a greater risk of developing disease
- To measure the excess risk (or rate ratio) attributable to the determinant

Methods

The cohort may comprise a sample of the general population or a group known to have a high risk for the disease under study or persons with special attributes that facilitate their study.

All participants are examined using standard methods to record potential determinants at the outset and during follow-up. After a period of time the incidence rates for one or more disease(s) (or causes of death) are calculated. Rates are compared among groups within the cohort with varying degrees of exposure to the determinant being studied, or between cohort members and the general population (often as rate ratios).

Advantages of this design include the following:

- Temporal sequence of events can be observed
- Incidence rates can be calculated
- Several possible outcomes can be studied simultaneously
- Determinants and outcomes can be measured precisely

Disadvantages are:

- Large population required if incidence is low
- Long time before results emerge especially if incubation period is prolonged
- Relatively expensive in resources
- Losses from population during study may bias results
- Standard methods and criteria may drift over prolonged follow-up

Case-control studies

Definition

A study in which the frequency of a determinant in a group of persons with a disease (cases) is compared to the expected frequency of the determinant in the population that gave rise to the cases. The 'expected frequency' is usually determined by studying a group of persons without the disease (controls).

Aims

- To compare the frequency of the determinant among those with and those without the disease
- To estimate the relative risk of any excess frequency using the 'odds ratio'

Study methods

Data on past exposures or personal risk factors may be obtained directly or from records.

To avoid bias it is essential to elicit data and make observations on controls in exactly the same way as for cases.

Selection of cases

- Ideally, select all cases in a defined population, but it is usually only practical to recruit a sample
- All persons with the disease in that population must have an equal chance of being identified and selected
- Commonly recruited from attendees at a health facility or persons on a disease or death register

Selection of controls

- Should be representative of the population from which cases were recruited
- Must not be discarded or replaced unless erroneously included
- To ensure similarity with cases (for factors not under study), may be 'matched' on potentially confounding variables
- To increase statistical power, select two or three per case
- Commonly recruited from persons living in same locality, population registers, hospital patients with unrelated conditions, random digit telephone dialling

Advantages

- Results can often be obtained more quickly and cheaply than with cohort studies
- The size of population required is economical
- Often easy to identify a relevant case group
- The only practical method for study of rare diseases

Disadvantages

- Temporal sequence of events not always clear
- Cannot measure incidence of disease as the population size is not known
- Difficult to ensure controls are representative of the population giving rise to cases
- Incompleteness of records and unreliability of recall of past events and past exposures

Analysis

As neither incidence nor prevalence rates of disease can be calculated in a case-control study, the frequency of exposure in the diseased and non diseased groups is compared using the odds ratio.

Experimental (intervention) studies

The efficacy and safety of relevant interventions need to be formally tested. Such studies normally take the form of intervention studies.

Definition

A study comparing the outcome in an experimental group receiving an intervention with that in a comparison group receiving 'conventional' treatment, placebo or an alternative intervention. The classic form is the randomised control trial (RCT), in which individuals in the trial have been allocated at random to an intervention or comparison group.

Aims

Experimental studies are used to:

- Assess the efficacy and safety of a new intervention compared with a control
- Compare alternative treatments or interventions
- Evaluate the effectiveness and efficiency of different forms of service provision
- Provide direct evidence that exposure to a suspected agent causes disease or that its removal prevents or reduces the frequency of disease

Design

- Similar in principle to that of a cohort study
- The population under study should be representative of the target population
- Subjects must be allocated at random to test or control groups
- To avoid bias in reporting illness or other relevant events, neither the subject nor the assessor should know to which group the individual participant belongs (double blind)
- Procedures and outcomes must be clearly defined using the same criteria in both treatment and control groups, using standardised and rigorously defined methods
- Outcomes should always include adverse events as well as beneficial effects
- Follow-up starts at allocation and continues for long enough to determine outcome in all subjects
- All losses to follow-up must be reported and every effort made to minimise them

Analysis

All randomised patients must be included in the analysis—this is called 'intention to treat analysis'. It means that:

- Persons are analysed in the group to which they were originally assigned
- All events throughout follow-up are counted
- All outcomes specified in the protocol, both beneficial and adverse, are analysed

DEFINING A STUDY POPULATION

Carrying out studies in an entire population is nearly impossible, so studies are usually carried out in a sample of available individuals. The population from which participants are to be drawn must be carefully selected in relation to the purposes of the study (Table 1). The target population is the collection of individuals about whom we want to draw conclusions (make inferences). The sample (study) population is the group of individuals chosen for study from an accessible population.

The sampling process

This aims to yield a population for study that is representative of the target population, is large enough to minimise the effects of random variation and adequately represents all groups of interest. Comparisons between characteristics of the target and sample populations and between participants and non-participants will identify possible differences that might bias the results.

The usual means of selecting study populations include:

- Population-based samples drawn from population registers, census databases or direct contact methods, such as telephone sampling
- Institution-based samples drawn from work places, professional associations, schools or lists of health services users

How do I ensure that the sample is representative?

Commonly used methods are:

- Random sampling: each sample unit has the same probability of being selected
- Systematic sampling: subjects are selected at regular intervals from a list
- Cluster sampling: a random sample from clusters of individuals
- Stratified sampling: the population is divided into subgroups or strata and separate random samples are drawn from each stratum
- Multi-stage sampling: a combination of two or more of these methods

How do I ensure that conclusions are correct? (precision and validity of the study)

There are several different issues related to drawing the correct conclusions from a study that have particular importance to sampling:

- Precision: 'the quality of sharp definition. It is a function of the extent of random error which may be attributable to sampling, subject or measurement variation. It can be expressed in terms of the confidence interval around a rate and may be enhanced by increased sample size'[2]
- Study validity: 'the degree to which the inference drawn from a study, especially generalisations extending beyond the study sample, are warranted when account is taken of the study methods, the representativeness of the study sample, and the nature of the population from which it is drawn'[2]

How big should the study population be?

Results from a sample population may not always reflect the 'truth' about relationships between disease and determinants in the target population because of both random and non-random (or systematic) variation in the way subjects are selected or measurements are made.

Table 1 Characteristics and definition of populations in a study

Research question Truth in the universe		Study plan Truth in the study
Step 1	Step 2	Step 3
Target population	Accessible population	Sample population
Specific clinical and demographic characteristics	Specific temporal and geographic characteristics	Defined approach to sampling
	Criteria for selection	
Suited to the research question	Representative of target population	Representative of accessible population
	Easy to study	Easy to do

Table 2 Possible conclusions based on results from a study comparing lung cancer rates in these two groups

		Truth about the population	
		Passive smoking IS related to lung cancer	Passive smoking is NOT related to lung cancer
Conclusion, based on results from a study of a sample of the population	Reject the null hypothesis (i.e., rates in the study appear to be different)	OK	Type I error Probability = α
	Accept the null hypothesis (i.e., rates in the study appear similar)	Type II error Probability = β	OK

The larger the study, the lower the chance of reaching an erroneous conclusion because of random variation. Similarly, if the true difference in disease rates between groups is large, it should be possible to detect this with a relatively small study. If there is a true difference, but measurement is subject to significant error, a larger study will be needed to ensure that differences found are not incorrectly attributed to error.

It is important to determine, in advance of carrying out the study, how large the study population should be in order to avoid drawing incorrect conclusions.

Sample size calculations are specific to the hypothesis being tested. Therefore, the hypothesis must be clearly stated. Often the sample size calculations may be made for the major hypothesis of the study, or for the hypothesis that will be tested using the smallest subgroup in the study.

In estimating the required sample size, we should be as certain as possible not to draw wrong conclusions. The notion of 'certainty' comprises two different concepts, illustrated in Table 2.

If, for example, lung cancer rates appear to be different in the study but there really is no biological association, you will 'reject the null hypothesis' incorrectly (Type I error or alpha). The confidence level is how certain you want to be that you don't make a type I error ($1 - \alpha$).

If the rates appear similar in the study but passive smoking really does increase lung cancer, you will 'accept the null hypothesis' incorrectly (type II error or beta). The power is how certain you want to be that you don't make a type II error ($1 - \beta$).

The expected magnitude and variability of the study results (effect size) critically influence the sample size calculations. The magnitude reflects the size of the expected difference between the groups, and the variability reflects the extent of variability in the measure you plan to use to evaluate the groups.

For further technical details on how to calculate sample size in a given study, readers are referred to the book from which this article has been abstracted.[1] Alternatively, many epidemiologists find it helpful to involve a statistician colleague in this aspect of study design.

Reference

1 Enarson D A, Kennedy S M, Miller D L, Bakke P. Research Methods for Promotion of Lung Health. A guide for low-income countries. Paris, France: International Union Against Tuberculosis and Lung Disease, 2001: pp 55–61.

2 Last J M, ed. A dictionary of epidemiology. 3rd ed. New York, NY: Oxford University Press, 1995.

RÉSUMÉ

Les études épidémiologiques ont été standardisées dans un groupe de « schémas ». L'étude descriptive décrit la maladie en fonction du temps, du lieu et de la personne et peut développer des hypothèses concernant des associations entre la maladie et ses déterminants potentiels. L'étude analytique teste les hypothèses. L'étude transversale mesure la maladie et ses facteurs déterminants à un seul moment donné. L'étude de cohorte identifie ceux parmi un groupe qui ont ou n'ont pas de facteur déterminant et observe la fréquence de la maladie dans les deux groupes. L'étude cas-contrôle identifie un groupe de patients atteints d'une maladie et sélectionne un groupe de sujets provenant de la même population mais n'ayant pas la maladie pour comparer la présence d'un agent déterminant dans les deux groupes. L'étude expérimentale, un type d'étude de cohorte, est une étude dans laquelle l'investigateur « impose » un agent déterminant (un traitement) à un sous-groupe d'une population et compare l'apparition de la maladie entre ceux qui ont ou qui n'ont pas subi le déterminant. Toutes ces études doivent s'assurer que les comparaisons faites ont un sens pour une population déterminée. Ceci est réalisé en sélectionnant un échantillon « représentatif » de cette population. Une sélection soigneuse du type d'étude et de la population facilite l'apport de nouvelles connaissances, tout en évitant autant que possible d'importantes erreurs.

RESUMEN

Los estudios epidemiológicos se han estandarizado en un conjunto de 'diseños'. El estudio descriptivo describe la enfermedad en función de variables como el tiempo, el lugar y la persona y puede formular hipótesis sobre la asociación entre la enfermedad y sus posibles determinantes. El estudio analítico verifica estas hipótesis. El estudio transversal mide la enfermedad y sus factores determinantes en un momento único temporal. El estudio de cohorte identifica, en un grupo, aquellos que poseen o carecen de un determinante y observa la aparición de la enfermedad en ambos grupos. El estudio de casos y testigos identifica un grupo de pacientes con una enfermedad y selecciona un grupo de personas de la misma población, sin la enfermedad, para comparar la presencia de un determinante en ambos grupos. En el estudio experimental, un tipo de estudio de cohorte, es el investigador quien 'asigna' un determinante (un tratamiento) a un subgrupo de la población y compara la aparición de enfermedad entre aquellos que recibieron y quienes no recibieron el determinante. Todos estos estudios deben garantizar que las comparaciones presentadas son aplicables a una población definida. Para conseguirlo se escoge una muestra 'representativa' de esta población. La selección cuidadosa del diseño de un estudio y de su población facilita la formulación de conocimientos nuevos y limita al máximo los errores considerables.