Clinical Research Methods

Primer on Epidemiology 1: Building blocks of epidemiological enquiry

SHIVANI ANIL PATEL, POORNIMA PRABHAKARAN

INTRODUCTION

Epidemiology is the backbone of the science of identifying risk factors and also testing prevention strategies and treatment measures at the population level. The knowledge gained from epidemiological work guides individual treatment as well. The occurrence of illness and disease states is not a random phenomenon. Characteristics such as genetic makeup of an individual, social factors and environmental context interact to predispose individuals to illness. The factors that determine the distribution of illness in individuals and populations can be identified by systematic studies of occurrence and correlates of disease. The knowledge thus gained can be applied to the prevention and treatment of these conditions. This is the core principle underlying the discipline of epidemiology.

This article is the first of a six-part primer on epidemiology. The primer is divided as follows: (i) Building blocks of epidemiological enquiry; (ii) Elements of study validity and key issues in interpretation; (iii) An overview of observational study designs; (iv) Interventional or experimental designs; (v) Sampling methods and developing a research protocol and (vi) Statistical analysis of research data.

A BRIEF HISTORY OF EPIDEMIOLOGY: THE FOUNDATIONS

The early foundations of epidemiology lie in the careful observations of patterns surrounding births, deaths and diseases. The recognition that the development of disease may be due to personal attributes and external factors was suggested by Hippocrates, the father of modern medicine.[1] Nearly 2000 years later in the 17th century, John Graunt first depicted the importance of collecting routine data when he studied weekly reports of births and deaths in London, and published the analyses of disease patterns, seasonal distribution and gender variations. His work on life tables documenting survivorship remains a central tool in computing life expectancy to this day.

William Farr, a physician in charge of medical statistics in the Office of the Registrar General for England and Wales in 1838, further showed how routine data collected on population vital statistics, such as births and deaths, were crucial in the studies of the general health of people. His meticulous documentation of the temporal and geographical distribution of cholera deaths was combined with an analytical approach that continues to inform the basis of epidemiological methods today. This included defining the population at risk, choosing a comparison group and studying other factors that may be responsible.

John Snow, another British physician, followed up on Farr's ideas in 1854 when he analysed the reasons for a large number of deaths due to cholera in a particular residential locality—about 600 deaths occurring around the Broad Street area in London. Snow collected data to obtain the number of deaths and the company supplying water to each household by going from house to house, thus coining the term 'shoe-leather epidemiology'. These data allowed Snow to compare the relationship between the water source and the number of deaths due to cholera; the number of deaths was minimal in households supplied by the Lambeth company (water source from upstream Thames) compared with deaths in households supplied by the Southwark and Vauxhall company (water source from polluted downstream Thames). Importantly, he identified the water source as the cause of the cholera outbreak before the advent of germ theory through his careful data collection and analyses.

Around this time in Vienna, a lawyer-turned-physician Semmelweis, used a simple intervention to conclusively show that higher rates of maternal deaths from puerperal sepsis in two wards were the result of lack of simple hand hygiene. Semmelweis observed that the deaths were higher in obstetric ward 1 served by physicians and medical students who were also responsible for conducting autopsies of puerperal sepsis patients. Midwives conducted deliveries in obstetric ward 2 where the mortality rate was nearly half that in ward 1. Semmelweis introduced a mandatory handwashing and nail-scrubbing clean-up for doctors before attending to deliveries. There was a dramatic drop in the death rates, comparable to the rates in ward 2. The results were conclusively confirmed when Semmelweis's successor relaxed the rule and the mortality rate again rose in obstetric ward 1.

All the above-mentioned early examples laid the basic principles for the science of epidemiology, which is grounded in the careful observations of individuals in real-world (rather than laboratory) settings and point to the importance of applying epidemiological methods in the study of diseases.

Epidemiology is most often described as the study of the distribution of disease and its determinants in populations. The term 'epidemiology' is composed of the Greek roots 'epi' (which means 'upon'), 'dem' (which means 'people') and 'logy' (which means 'study of'). Epidemiology is the science of studying the frequency of disease, the distribution among populations and the factors and determinants that may be responsible. It provides a framework for systematic observation and data collection, assessing a possible causal association between factors studied and the disease, and testing interventions to promote better health. This analytical and logical reasoning process also involves evaluating a hypothesis using carefully collected quantitative data.

APPLICATIONS OF EPIDEMIOLOGY

Epidemiological studies are the foundation of evidence-base for

Rollins School of Public Health, Emory University, Atlanta, GA, USA
SHIVANI ANIL PATEL

Public Health Foundation of India, Gurugram, Haryana, India
POORNIMA PRABHAKARAN

Correspondence to SHIVANI ANIL PATEL, Emory University, 1518 Clifton Road NE, CNR 7037, Atlanta, GA 30322, USA; *s.a.patel@emory.edu*

public health, and their results have huge implications for clinical practice, preventive care and policy-making. The knowledge of causative factors and possible disease outcomes that is gained from epidemiological studies in turn aids in appropriate treatment decisions in clinical care. Epidemiology also contributes to the designing of preventive care at primary, secondary and tertiary care centres and feeds to appropriate policy-making.

The landmark Framingham Heart Study that established the link between cigarette smoking and heart disease and identified for the first time the major risk factors for cardiovascular diseases, the studies by Coburn and Pauli[2] and others that established the link between *Streptococcus haemolyticus* and rheumatic fever, the body of research establishing the link between air pollution and cardiovascular health and the developmental origins of health and disease paradigm with its plethora of research studies relating early life factors to later life cardiovascular disease are all classic examples of the contribution of epidemiology to public health and clinical practice. The remainder of this chapter focuses on some basic principles and terminology used in epidemiological literature and applications. Table I presents a brief list of key terminology that is described below in detail.

EPIDEMIOLOGICAL FOCUS: INDIVIDUALS AND POPULATIONS

In contrast to clinical medicine and training, which focus on health and disease in the individual, epidemiology focuses on events and causes in populations. The two are connected, however, because clinical medicine is informed by discoveries made through studies of groups of individuals, such as the Framingham Heart Study described earlier. Similarly, new epidemiological investigations are often motivated by clinical questions.

The backbone of this quantitative science is enumerating the population at risk for disease and enumerating the presence of disease and suspected causes of disease in the population. Individuals with the disease under study are referred to as 'cases' and hypothesized causes are often referred to as 'exposures'. Although epidemiology began with a focus on specific infections and their prevention, it currently evaluates many 'outcomes' related to health such as treatment uptake, risk factors for disease and acute events such as myocardial infarctions (MIs) and strokes. The degree to which there is a correlation between the exposure and the outcome (e.g. the likelihood that exposed individuals have a tendency to become cases) provides evidence of an 'association' between the exposure and disease outcome of interest. Epidemiological studies are designed to generate data to quantify the degree to which 'associations' between exposures and outcomes of interest exist in the population to better understand the causes of disease.

This chapter refers to the Centre for Cardiometabolic Risk Reduction in South Asia (CARRS) surveillance cohort study to provide examples.[3] In India, the study was conducted among a representative sample of non-pregnant adults residing in New Delhi and Chennai, India. The cohort was enrolled by going door to door in the community following standard sampling methods. The first round of CARRS study data collection, termed the baseline assessment, began in 2010. This was a comprehensive assessment of cardiometabolic history and risk factors. The CARRS is a longitudinal study, meaning that it is designed to track individuals over time to follow health outcomes. Annual data collection is ongoing with more detailed assessments only every other year.

TABLE I. A brief reference to key terminology used in epidemiology

Term	Definition
Target population	The population that we wish to study or among whom we would like to target our intervention, for example, the target population may be adults with coronary heart disease in India, and a study sample would be designed to best represent that target population
Sample	The group of individuals actually selected for a study or observation
Exposure	The independent variable in an epidemiological study; often times, this is a potential risk factor for disease
Outcome	The dependent variable in an epidemiological study; this is usually a disease or an intermediary end-point prior to overt disease
Prevalence	The proportion of a population that has an exposure or outcome, for example, prevalence of smoking among adults in India is the proportion of adults who smoke in India
Risk	The proportion of a population that newly develops an exposure or outcome over a specified period of time
Rate and incidence	The number of newly developed cases of an outcome per unit of time. 'Incidence' is defined as the rate at which new cases appear over a period of time
Association	The statistical relationship between an exposure and an outcome; this is often measured as a ratio or difference
Prevalence ratio	This is defined as the ratio of the prevalence of a condition in the exposed sample relative to the prevalence in the unexposed sample. It is a measure that quantifies the relative association between an exposure and an outcome
Odds ratio	This is defined as the ratio of the odds of a condition in the exposed sample relative to the odds in the unexposed sample. It is a measure that quantifies the relative association between an exposure and an outcome
Risk ratio	This is defined as the ratio of the risk of a condition in the exposed sample relative to the risk in the unexposed sample. It is a measure that quantifies the relative association between an exposure and an outcome
Rate ratio	This is defined as the ratio of the rate of a condition in the exposed sample relative to the rate in the unexposed sample. It is a measure that quantifies the relative association between an exposure and an outcome
Prevalence difference	This is defined as the prevalence in the exposed sample minus the prevalence in the unexposed sample. It is a measure quantifying the absolute association between an exposure and an outcome
Risk difference	This is defined as the risk in the exposed sample minus the risk in the unexposed sample. It is a measure quantifying the absolute association between an exposure and an outcome
Rate difference	This is defined as the rate in the exposed sample minus the rate in the unexposed sample. It is a measure quantifying the absolute association between an exposure and an outcome

DEFINING A POPULATION

Target population

Defining the population of interest is the first task of any epidemiological study. The population informs the 'denominator' of epidemiological quantities of interest. The population of interest is often termed the 'target population'; the target population must be defined along the dimensions of person, place and time: Who do we seek to study, where and when? Ultimately, this is the group about which we wish to make an inference. Another defining feature of the target population is consideration of who is at risk for a disease outcome. For example, the target population for a study of heart disease during pregnancy would include only women who are pregnant.

Our target population may also change over time with increasing information or secular changes in disease patterns. For example, much initial research in cardiovascular disease focused on men, who more often presented with risk factors such as smoking and hypertension. In effect, studies that only enrolled men considered men to be the target population. With time, cardiovascular disease became recognized as an important concern among women, and mixed-gender cohorts were initiated. The target population in these subsequent studies was all adults. The early emergence of cardiovascular disease has become a concern. Some studies of paediatric and adolescent cohorts implicitly consider the adolescent age group as the target population.

Person, place and time

Person, place and time criteria are important not only for the design of a study but also for the interpretation of the results. Person criteria may include gender, age, race/ethnicity and any other conditions defining the criteria for inclusion in a study. In the CARRS study, men and women aged 20 and older of all ethnic backgrounds were the target population.[3] Place definitions refer to both geographical locale and the specific location for study recruitment. The CARRS was conducted in New Delhi and Chennai, and all residents of these cities (rather than a focus on patients in hospital) were of interest. Finally, time dimensions include both calendar year and follow-up plans. The CARRS is representative of the adult populations of these two cities in 2010. Follow-up in the CARRS occurs on an annual basis by phone or in person. All reports from these longitudinal data must describe the time interval since the baseline, that is, the duration of follow-up.

Static and dynamic populations

Populations can themselves change in composition over time. For example, when studying the population of New Delhi, we must consider new individuals entering the population each year through migration and birth; in contrast, there are individuals who leave through emigration and death. Therefore, on an annual basis, the population of New Delhi would be considered dynamic. If, however, we were interested in the population of New Delhi in 2011, that is a well-defined and static group of individuals; only those individuals who resided in the city in that particular year are of interest.

Defining a case

Case data are the 'numerators' of epidemiological quantities. Cases are the individuals who already have developed or go on to develop the disease under study. Case definitions are needed to clearly identify those who have the disease and those who do not. For example, there are many ways of defining MI (the WHO definition and the universal definition).

Samples

Since it is impossible to study the entire population, we study just a subset called a sample. The method of selecting the subset is called sampling. Sampling methods, or systematic approaches to defining the subset of individuals who will be invited to participate in a study, have been extensively developed for the community setting by survey researchers who are particularly concerned with the representativeness of the sample. Representativeness refers to the extent to which a sample reflects the characteristics of the target population. The term 'random sampling' is thought of as a cornerstone of 'representativeness'; a pure random sample of the population would be expected to provide the same picture as the population itself. Similarly, there are parallel methods and concerns when selecting samples from a patient population. These are described in later sections.

It is impossible to randomly select from the entire population, so, we define an intermediate pool of potential participants—this is called the source population. The source population is in fact the group of individuals used to actually facilitate sampling. For example, the target population may be all individuals living in Delhi. However, to sample individuals, the source population becomes household listings based on census enumeration blocks. If an individual lives in Delhi but does not exist in the household listing, then he/she is not part of the source population for the study and has no chance of being sampled. Therefore, great care is taken to ensure that the source population is well defined and represents the target population.

MEASURES OF OCCURRENCE

We have already described epidemiology at its heart to be a field of 'counting'. Measuring the occurrence of disease events of interest is at the heart of epidemiological research. We describe the ways in which disease occurrence is quantified.

Prevalence

This is among the most common ways of describing the current burden of disease; it is a measure of disease status in the population. Prevalence is the proportion of individuals in a population with a particular disease at a particular point in time:

$$Prevalence = \frac{m}{N}$$

where m refers to the number of existing cases and N refers to the number of individuals in the population. Prevalence is also used to describe the proportion of individuals with a particular exposure or risk factor in the population; for example, the prevalence of smoking is reported using data from national surveys. Another example is the prevalence of hypertension in India, which was estimated to be 28% in 2014.[4] Technically speaking, the term prevalence is reserved for situations when the population denominator is well defined.

Prevalence is not technically a rate, although the term 'prevalence rate' is often used to refer to prevalence. Similarly, incidence is not a proportion, although annual incidence will be described in units of percentage of new cases. While not technically correct, it is still possible to deduce the meaning of these quantities in published papers using the surrounding text.

Risk

Risk refers to the probability of developing a disease ('event') in the population:

$$Risk = \frac{I}{R}$$

where I refers to the number of new events accumulated over a specified time frame in the population and R refers to the population at risk for the event at baseline. These new events that comprise the numerator are sometimes referred to as 'incident events' or 'incident cases'. The denominator is all individuals at risk for developing the event at the beginning of the follow-up period (or the age). The numerator and the denominator are both persons, so risk has no units. The time frame over which risk is computed must be specified clearly—for example, a 5-year risk of diabetes. The time frame specification for risk is extremely important because cases are accumulated over time: a longer time frame will allow for the accumulation for more new cases, whereas the population at risk at the beginning of the time frame is generally fixed. For example, consider the importance of the time frame when evaluating the risk of death. The 3-year risk of death in the CARRS cohort was <2%, whereas we all know that the lifetime risk of death is 100%. Similarly, while the risk for an atherosclerotic cardiovascular disease event is 2.1% over 10 years for a 50-year-old, healthy, American, white man, the risk of an event is 5% over the full lifetime.

As an example, suppose an investigator is following a sample of 100 individuals who are aged 50 years at baseline. In the first year, she observes five new cases, so the 1-year risk of developing diabetes among 50-year-olds is 5/100=5%. At the end of the second year, she observes another three cases of incident diabetes, so the 2-year risk will be (5+3)/100=8%. In the third year, she observes four cases of incident diabetes, so the 3-year risk will be (5+3+4)/100=12%. Note that the denominator is fixed, but the numerator increases with increasing duration of follow-up and the period at risk must be specified.

Risk is technically a proportion because risk refers to the number of new cases of disease in the population divided by the total population at risk for the disease. Despite the technical definition of 'risk' in epidemiology, it is important to know that 'risk' is also used as a more general term for 'chance' of an event and is often also used interchangeably with 'rate'. Another usage of the term is describing individuals as being 'at risk' for a disease. Being 'at risk' implies the group of people who are currently 'eligible' to newly develop the disease; we typically exclude those who already have the disease because they are no longer eligible to newly develop the condition. For example, individuals who have experienced an MI in the past would be excluded in a study of first-incident coronary artery disease. We may also exclude from the population at-risk individuals who for some biological reason are unlikely to experience the exposure or outcome; for example, men would be excluded from a trial investigating the impact of hormonal replacement therapy on coronary heart disease.

Rates

Rate refers to the number of new events in a group of people over follow-up time:

$$Rates = \frac{I}{T}$$

where *I* refers to events and *T* refers to the follow-up time measured in person-time. 'Person-time' is the total amount of accumulated follow-up time (e.g. days, months and years) across all individuals followed in the study. For example, if 10 individuals were followed for 8 months in a study of secondary events after an MI, the total person-time accumulated in the study would be 10 persons×8 months=80 person-months. If one person were to experience an event, the rate would be 1/80 person-months. In general, epidemiologists use rates to describe the incidence of events or mortality. Unlike prevalence and risk, incidence or mortality rates are not a proportion. The numerator is the number of new cases, but the denominator refers to both persons at risk and the amount of follow-up time accumulated. Usually, the incidence rate is reported for 1 year, such as an annual incidence rate of new cases. Person-time is indifferent to whether individuals or follow-up time contribute to the quantity: 30 person-years could refer to 30 individuals being followed for 1 year, or alternatively 10 individuals followed up for 3 years each. Person-time is a better way of accounting for differing levels of follow-up in a cohort. For example, if there is a cohort of 100 individuals designed to be followed for 10 years, we may still have some individuals dropping out of the study after 3 years. If we were to consider event status of these five individuals at 3 years, we may underestimate the total events that actually occurred in the full 10 years because our window of observation was truncated. As we can count only what we can observe, person-time allows us to adjust the denominator to the time frame we observed, so we can include the numerator data from all participants at baseline despite truncated follow-up status. In addition to dropouts, rates can be used to account for deaths in the cohort, or other competing events that cut the window of follow-up short of what was planned.

Rates are frequently reported measures of disease occurrence by public surveillance systems and often have specific technical definitions. For example, death rates are often reported as the total count of deaths per 100 000 population per year, and cardiovascular mortality (the death rate from cardiovascular disease) is the average number of cardiovascular disease-related deaths per 1000 adults aged 20 years and older in the population per year. Although the 'time' unit may be missing in the reporting, it is implicitly factored into the actual calculation and unless otherwise specified, the unit of time is generally 1 year.

Relationship between risk, incidence rates and prevalence

It may be difficult to grasp the difference between risk and incidence rates initially. Risk is an accumulated probability of disease over a specified time frame, whereas incidence is the number of new cases observed in the population defined by time and number of people being followed. Thus, risk is a function of incidence, and sometimes risk is referred to as cumulative incidence. Revisiting the earlier example of 100 individuals aged 50 years at baseline, the annual rates of diabetes can be computed as follows: In year 1, the rate of diabetes is 5/100=5 cases/100 person-years. In year 2, the rate of diabetes is 3/(100–5)= 3/95=3.16 cases/100 person-years. In year 3, the rate of diabetes is 4/(100–5–3)=4/92=4.35 cases/100 person-years. The average incidence rate over the years is then (5+3+4)/(100+95+92)=4.18 cases/100 person-years. Notice that individuals who have developed the disease in the previous year are removed from the denominator for the current year of follow-up.

In a closed population with no migration, prevalence is a function of incidence (new cases entering the population) and the cure rate or mortality rate from the disease (cases exiting the population). The relationships can be defined as prevalence= incidence×duration of disease. A high prevalence could indicate a high incidence of disease, or a long-expected duration of disease. For example, a high prevalence of obesity may be observed even in low-incidence areas due to the fact that people who become obese seldom lose weight sufficient to go back to the normal-weight category, yet they live a long life. On the other hand, a high point prevalence of the common cold may be observed simply because there are many cases occurring in certain seasons.

The study design that generates the data is critical for determining which measure of disease occurrence is appropriate. In a cross-sectional survey, we can measure only prevalence because diseases are measured only at one point in time. Incidence and risk can be determined only through a cohort study. Prevalence, incidence and risk measures may be applicable only to the target population for which the cohort study was designed. For example, the prevalence, risk and rate of obesity from a rural village in Maharashtra will not accurately describe the prevalence, risk and rate of obesity in Mumbai.

Acute, recurrent and chronic conditions

There are additional considerations when quantifying the prevalence, risk and incidence of disease related to the nature of the disease/health outcome. We generally have a clear set of criteria to define the onset of a condition, and once a person meets those criteria, they are said to have the disease. Acute conditions also have a clear end; they are transient. Some acute conditions are also recurrent, in that there is a possibility that they can be repeated. An MI is an example of an acute condition that can be recurrent; for example, it is possible to have a second or third MI. Other acute events happen at one point in time and cannot occur again—for example, death. Finally, there are chronic conditions that exist for long periods of time and often never go away. One example is diabetes; once an individual has diabetes, we do not describe him/her as 'cured' even when his/her blood sugar is controlled again. Each of these factors must be considered when determining how to analyse the disease of interest, and what time frame should be used as the reference, or measures of disease occurrence.

Person, place and time

Disease occurrence is not uniform over person, place or time––therefore, each of these dimensions for the measure of disease occurrence must be specified. Our audience must know among whom the measure was computed (e.g. men aged 18 and older), where (e.g. in New Delhi) and calendar time (e.g. 2015) for completeness of interpretation. The statement that 'the prevalence of diabetes is 10%' has no meaning unless it is contexualized with 'among men aged 18 and older in New Delhi in 2015'.

CAUSATION

'Cause' is a complex concept. Under the 'counterfactual' paradigm, we say that an exposure 'causes' disease if the disease occurs because of the exposure, and the disease would not have occurred without the exposure. For example, if we were studying the relationship between a high-salt diet and stroke, the ideal experiment is one in which we could observe the stroke event rate in a group of individuals who were all exposed to a high-salt diet and compare that with the stroke event rate in that exact same group of individuals in the absence of the high-salt diet. Observing such a scenario, of course, is impossible in the real world (and therefore 'counterfactual') because we cannot simultaneously expose participants to a high-salt diet and withhold a high-salt diet at the same time. Epidemiologists therefore are often searching for a study design that allows us to identify groups of individuals who are similar in all characteristics other than the potential exposure studied, and then compare future outcomes. In practical terms, epidemiologists consider the statistical association between an exposure and an outcome, after controlling for differences in the groups, as evidence for a causal effect. The approach is described below.

Studying associations as a path to understanding causation: The 2×2 table

We quantify effects by looking at associations between exposures and outcomes, or the frequency of co-occurrence of an exposure and an outcome. In plain words, preliminary evidence of a cause exists if the outcome tends to occur in the presence of the exposure. The 2×2 table is the simplest way of assessing this co-occurrence (Table II). A 2×2 table cross-classifies the population or sample under study by exposure and outcome status. These frequencies are used to compute differences and ratios between the risk of disease among the exposed compared with that among the unexposed (Table III). 'Risk' can be replaced with the prevalence of disease if we are working with prevalent rather than with incident cases.

Risk difference is the absolute difference between the risk in the exposed and the risk among the unexposed. A negative risk difference implies a protective association, or that having the exposure reduces the chance of disease. A positive risk difference implies a direct association, or that exposure increases the chance of disease. A risk difference of 0 implies that there is no difference in absolute risk between the groups compared; 0 is therefore referred to as the 'null value' for risk difference computations.

The risk ratio is the relative risk in the exposed compared with the risk in the unexposed. A risk ratio <1 implies a protective association, or that having the exposure reduces the chance of disease. A risk ratio >1 implies a direct association, or that exposure increases the chance of disease. A risk ratio of 1 implies that there is no difference on the multiplicative scale in risk between the groups compared; 1 is therefore referred to as the 'null value' for risk ratio computations.

Both risk differences and risk ratios require cohort data, which by definition include follow-up information because prospectively measured disease risk is needed for these computations.

The odds ratio (OR) compares the odds of disease of the exposed with the odds of disease of the unexposed. We can pose the question as follows: 'What is the odds of having MI (disease) in a smoker (exposure)'? The OR in this example would be odds of MI in a smoker divided by the odds of MI in a non-smoker. While odds is not reported as a measure of disease occurrence, the OR is a useful quantity for comparing disease outcomes, especially in certain study designs such as case–control studies. This is because it does not require information on the population denominator of the numbers exposed or unexposed. Similar to the risk ratio, an OR <1 implies an inverse (protective) association, or that having the exposure reduces the chance of disease. An OR >1 implies a direct association, or that exposure increases the chance of disease. An OR of 1 implies that there is no difference in the odds between the groups compared. One is therefore the null value for OR computations.

TABLE II. Sample or target population 2×2 table

	Outcome	No outcome	Total
Exposed	A	B	N_1
Unexposed	C	D	N_0
	M_1	M_0	N

TABLE III. Common measures of effect or association

Risk difference	$\frac{A}{N_1}-\frac{C}{N_0}$
Risk ratio	$\frac{A/N_1}{C/N_0}$
Odds ratio	$\frac{A/B}{C/D}$

An example: Obesity and diabetes in a subset of cardiometa-bolic risk reduction in South Asia surveillance participants

As an example, let us consider baseline obesity and new cases of diabetes in participants of the CARRS study. Based on the 2×2 table shown in Table IV, we compute that the prevalence of obesity at baseline is 1708/9186×100=18.6% and the risk of diabetes is 2052/9186×100=22% in this subsample. We multiply all the proportions by 100 because prevalence and risk are typically reported as a percentage rather than as a proportion.

We will treat obesity as the 'exposure' and diabetes as the 'outcome' in computing the measures of association. To estimate the measures of association, we compare the risk of the outcome in the exposed and in the unexposed. We compute the risk of diabetes in obese participants as 622/1708×100=36.4% and the risk of diabetes in non-obese participants as 1430/7478×100= 19.1%. Table V shows the numbers used to compute the difference in the risk and the ratio of the risk. The positive risk difference indicates that the risk of diabetes is 17.3 absolute percentage points higher in obese participants compared with that in non-obese participants. The risk ratio of 1.90 indicates that the risk of diabetes is relatively 90% higher in obese participants compared with that in non-obese participants. Similarly, the OR is >1, indicating a direct association between obesity and diabetes. Note that the OR is much higher than the risk ratio; this is often the case in real-world data.

Because we are using data regarding the risk of newly developing diabetes, the 2×2 table provides preliminary evidence of causality. If we had computed associations based on prevalence, we could not infer causality. The temporal ordering of the association is suspect when we use prevalent data because they do not allow us to comment on whether diabetes causes obesity or obesity causes diabetes. From other studies, we know that these associations are 'bidirectional', or go in both directions. In reality, however, we more often work with cross-sectional data and thus prevalence ratio is the more appropriate term; the cross-sectional associations are also often our starting point for more detailed longitudinal studies.

TABLE IV. Example 2×2 Table: Baseline obesity and diabetes cross-tabulation in a subset of Cardiometabolic Risk Reduction in South Asia Surveillance Cohort participants

	Diabetes	No diabetes	Overall
Obese	622	1086	1708
Non-obese	1430	6048	7478
Overall	2052	7134	9186

TABLE V. Example of measures of association: Baseline obesity and the development of diabetes in Cardiometabolic Risk Reduction in South Asia Surveillance Cohort

Risk difference (risk of diabetes in obese–risk of diabetes in the non-obese)	$100\times\frac{622}{1708}-\frac{1430}{7478}+17.3\%$
Risk ratio (risk of diabetes in obese divided by the risk of diabetes in the non-obese)	$\frac{622/708}{1430/7478}=1.90$
Odds ratio (odds of diabetes among obese divided by the odds of diabetes among non-obese)	$\frac{622/1086}{1430/6048}=2.42$

POPULATION ATTRIBUTABLE FRACTION

'Population attributable fraction (PAR)' is defined as the fraction of cases of a disease that occurred due to a particular exposure or risk factor in a specific population. To understand this, let us take the example of smoking and MI. If by some means we were to eliminate smoking completely from the population, a fraction of MI can be eliminated from the population. This fraction is called PAR. It is a function of the prevalence of the exposure and the strength of the association between the exposure and the outcome such that the higher the prevalence and the stronger the association, the higher is the PAR. The PAR is important in prioritizing public health interventions such as screening or a policy measure such as tobacco control. Let us consider the example of two risk factors with contrasting prevalence and strengths of association with MI. Familial hypercholesterolaemia increases the risk of MI in an individual 10-fold (relative risk of 10); however, its prevalence is only 1 in 500 in the US population (likely to be similar in India). While it is important at an individual level to identify and treat the disease aggressively, it will not make sense for studying the gene mutations in the whole population given its low prevalence. By contrast, some risk factors with modest associations may be highly prevalent, and thus exert a substantial impact on population health. Physical activity is an example. The relative odds for MI are 30% to 50% (OR=1.5-1.5) higher among those with the lowest physical activity, but low physical activity levels are widely prevalent in urban populations. If we were to eliminate low physical activity, it would have a measurable impact on the number of MIs in the population. Therefore, from a policy perspective, interventions that target low physical activity will provide larger health gains than interventions that target familial hypercholesterolaemia in the Indian population. The PAF is also highly variable across populations because it will depend on the prevalence of the exposure/risk factor in a given population.

CONCLUSIONS

In this article we have discussed the concepts related to populations, individuals, risks and causation. In the forthcoming articles we will discuss measurement, confounding, bias and interactions, which are important in understanding causation.

ACKNOWLEDGEMENTS

We are grateful to Dr D. Prabhakaran, Vice President for Research, Public Health Foundation of India (PHFI) for critical inputs in developing this manuscript. We also thank Sanjana Bhaskar, Research Assistant, Centre for Environment Health, PHFI, and Radhika Asrani, Emory University, for editorial assistance and referencing.

Reprinted with permission from *Tandon's Textbook of Cardiology Vol. 1*, by Dorairaj Prabhakaran, Raman Krishna Kumar, Nitish Naik, Upendra Kaul, 2019, Wolters Kluwer India, New Delhi. Copyright © 2019 by Wolters Kluwer Health (India).

Conflicts of interest. None declared

REFERENCES

1 Microsoft Encarta Online Encyclopedia. Microsoft Corporation; 2006.

2 Coburn AF, Pauli RH. Studies on the relationship of *Streptococcus hemolyticus* to the rheumatic process: I. Observations on the ecology of hemolytic *Streptococcus* in relation to the epidemiology of rheumatic fever. *J Exp Med* 1932;**56:**609–32.

3 Nair M, Ali MK, Ajay VS, Shivashankar R, Mohan V, Pradeepa R, *et al*. CARRS Surveillance study: Design and methods to assess burdens from multiple perspectives. *BMC Public Health* 2012;**12:**701.

4 Anchala R, Kannuri NK, Pant H, Khan H, Franco OH, Di Angelantonio E, *et al*. Hypertension in India: A systematic review and meta-analysis of prevalence, awareness, and control of hypertension. *J Hypertens* 2014;**32:**1170–7.