Alexandra Bražinová

Epidemiological Methods I. Data Collection and Description – Principles and Use

CHAPTER OBJECTIVE:
This chapter aims to familiarize students with the basic methods of epidemiology – with data collection and data description – and with the application of the results of the description in practice.

In epidemiology, we study health and the factors associated with it. We observe and describe the **distribution** and **determinants** of health and disease in a population in order to address health problems of a population.
The basic processes in epidemiology are:
data collection -> description -> analysis
This chapter deals with the collection of data and its description.

Data collection

Data on population health can come from a variety of sources.
The most common data sources are the following:

- **Official statistics** at the national level
- **Local routine data collection** for a specific purpose (e.g. a hospital information system)
- **Individual collection** (e.g. questionnaires, collection of biological samples for seroepidemiological investigations, etc.)

The basic source of public health data is **official statistics** collected, processed and evaluated by national and international institutions.
In the Slovak Republic, the main institutions that routinely collect data on the health status of the population are the National Health Information Centre (NCZI), the Statistical Office of the Slovak Republic (SUSR), the Public Health Office of the Slovak Republic (UVZSR) and the Social Insurance Agency (SP).

National Health Information Centre(NCZI) is a state-funded institution whose main activity is the collection, processing and publication of statistical data on health. It involves information about the use of health services on the basis of reports (forms) transmitted to the NCZI by health care providers in the fields of outpatient and inpatient care as well as joint examination and treatment units (such as laboratories, imaging centers etc.). The NCZI processes and publishes this data in the form of regular or special yearbooks on its website www.nczisk.sk. The NCZI provides data on the state of health of the population in relation to the use of services – these are the numbers of people who have been examined or treated for individual

diseases. It is important to note that in the case of many diseases or medical conditions, this is not the actual occurrence in the population, but only those people who have used healthcare (Fig.1).

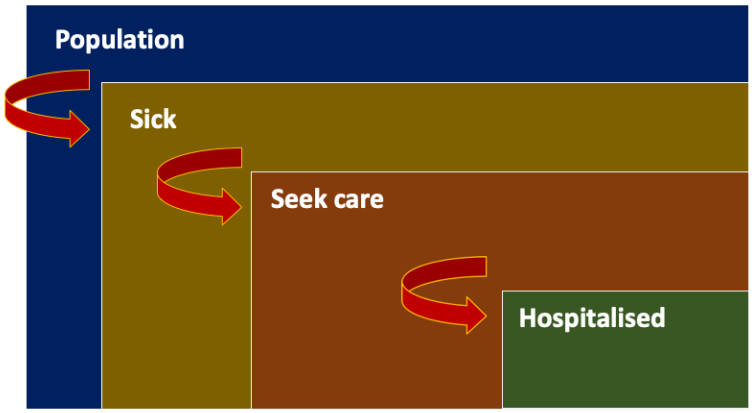

Fig. 1. The population: progression from health to varying degrees of disease severity. (Modified from White KL, Williams TF, Greenberg BG. The ecology of medical care. N Engl J Med. 1961;265:885–892.)

The NCZI also provides data on the network and the activities of healthcare providers and other organizations, on employees, on the economy of the healthcare system, including the financing of healthcare based on health insurance, and on medical technology.
The basic annual summary publication is the Health Yearbook. In addition, thematic statistical outputs from selected medical disciplines, such as surgery, diabetology, cardiology, immunology and allergology, are published annually.
The NCZI also manages national health registries. These are special information systems that collect, process and analyse the data on selected disease groups that cause high burden for the society in the Slovak Republic. The National Cancer Registry, the National Diabetes Mellitus Registry, the National Registry for Congenital Defects, and the National Registry for Circulatory System Diseases are just a few examples.
The main purpose of the registers is to monitor the trend of the disease in the population. They collect more detailed data than the routine data collection system where the healthcare providers send regular periodic reports to the NCZI database. The results of the registries can be used to optimize intervention measures aimed at improving the health of the population. One example is the use of data from the National Cancer Registry to set up a screening for the most common forms of cancers. The desired outcome may be earlier detection of the disease, timely treatment, improved survival, and lower mortality from the given disease in the population.

The Statistical Office of the Slovak Republic (SUSR) is responsible for state statistics in all areas. For health assessment purposes, demographic and social statistics that provide

information on the status and movement of the population (births, deaths, migration, marriages, divorces, abortions, etc.) across the country in individual regions and districts, are most widely used. The Mortality Database provides information on the age, gender, cause of death, district and region of permanent residence of the deceased. Information on the number of inhabitants of the Slovak Republic in individual years by sex, age, region and residential district is used in assessing the state of health in relation to the population rates of disease incidence indicators (such as incidence, prevalence, mortality). The SUSR makes data available via a transparent online platform at http://datacube.statistics.sk/. Every European country has a similar statistical office.

The Public Health Office of the Slovak Republic (UVZSR) manages the monitoring and control of the occurrence of infectious diseases and for this purpose manages the epidemiological information system EPIS. We have the ***surveillance*** (i.e. epidemiological vigilance) of infectious diseases in mind. Surveillance is "an epidemiological procedure based on the continuous monitoring of all disease frequencies and distribution characteristics, i.e. for the systematic collection and evaluation of data on the incidence, morbidity, death rate, mortality, diagnostic reliability, treatment effectiveness, effectiveness of used vaccines, disinfectants, pest and rodent control on the selected disease, but also on living conditions (drinking water supply), working conditions and way of life, that is all the attributes associated with the occurrence and course of the disease and which have an impact on the health of individuals or population segments." We deal with the system of surveillance in the next part of this publication (Surveillance of Infectious Diseases).
EPIS collects data on the incidence of selected communicable diseases. Cases of individual illnesses are reported to EPIS by epidemiological employees of the regional health authorities, depending on the location. Based on this data collection, we monitor the current incidence of disease cases, identify the emergence and spread of epidemics, observe long-term trends, and assess the effects of vaccinations on the morbidity of the population.

The Social Insurance Agency (SP) provides data on the decreased work ability and disability to assess the health of the population.

Local routine data collection is, for example, hospital information systems that collect mandatorily reported data and make it available to the NCZI as well as for local research – such as monitoring the effectiveness of treatment, the cost-effectiveness of individual parts of a medical facility, and demographic and health parameters and trends in the catchment area.
When an epidemiologist uses existing data to research a particular aspect of health, such data is called secondary data.

Targeted data collection is used when the information of interest is not available from existing data. This is mainly done through research according to a specific design and with precisely

defined methods. The data is collected, for example, through a structured questionnaire or by taking samples of biological material (such as blood, urine, sputum, etc.). This data is called primary data (obtained by the researcher himself). Collecting primary data is time-consuming and costly.
The collection of primary data must be designed from the beginning in such a way that the data obtained enables the research question to be answered. If the data is not of good quality and informative, the results will be useless ("garbage in, garbage out").
It is therefore very important that the research team creates a procedural protocol prior to the actual primary data collection, which should include the following steps:

- Research objectives.
- Type of study (e.g. cohort, case-control, cross-section).
- The monitored population, case definition, sample size and selection.
- Data collection procedures, monitored variables, procedures for coordinating participants.
- Security and protection, data control technology.
- Analysis plan.
- Logistics, including a budget, staff, and schedule.
- Legal requirements, including rules and regulations.

Description

After the necessary data has been collected and organized in a database, the first methodological step of its processing is the description of it.

For example, we describe the distribution of diseases or determinants of health at a particular point in time, in a particular population, and in a particular geographical area – we describe the temporal, geographical and population characteristics of the distribution of diseases (the time, place and persons).

The description in epidemiology answers the questions **WHAT? WHO? WHEN? WHERE?**
When monitoring the occurrence of a particular disease or health-related phenomenon, we answer the following four questions:
WHAT: Which illness/condition/health determinant are we describing?
WHO: What population are we interested in? This can be accurately determined based on age, gender, education, employment and other demographic, socioeconomic or other indicators
WHEN: What period of time are we interested in? Years, months, seasons, calendar weeks, etc.
WHERE: What area am I describing? Region, country, urban/rural etc.

The description of the data begins with the observation of the **absolute values** of the monitored variables. In addition to absolute values, we often describe a variable using the mean or median.
We get the mean by adding up all values and dividing by the number of values. Example: We have 5 people in the sample aged: 55, 10, 43, 67, 23 years. The mean age in this group is: (55+10+43+67+23)/5 = 198/5 = **39.6** years
The median is the middle number. We get it by sorting the values in the sample by size and finding the middle value: 10, 23, **43**, 55, 67. The age median in the set is 43 years. If the number of values in the sample is even, the median is the average of the two middle values. The median has the advantage of not being so easily influenced by extreme values.

Frequency of occurrence indicators

We monitor the incidence and distribution of diseases primarily based on the **incidence rates**. According to the International Epidemiology Association Dictionary (2008), the rate is "an expression of the frequency with which a phenomenon occurs in a defined population, usually over a period of time".
In addition to the frequency, we are often interested in a pattern, the pattern of occurrence of a disease or a medical condition which, depending on the time, place or population, can have certain characteristics: time period (annual, seasonal, monthly, weekly); territorial disparities (rural/urban); population factors (demographic, socioeconomic, behavioural, etc.).
The fundamental incidence rates are morbidity and mortality. These are so-called relative indicators – they express the number of cases related to the population. It is important to remember that the denominator must always be the population at risk, that is, the population who may get the disease. For example, when we study cervical cancer, the denominator will be only the population of women living in a given area in a given year.

Morbidity comes in two forms: incidence or prevalence

- **Incidence** is the number of **new cases of a disease** in a given population at a given point in time, related to the population and then multiplied by a multiple of 10, e.g. 100,000 inhabitants. For example, we calculate the incidence of Type 2 diabetes mellitus (T2DM) in the Slovak Republic in 2018 as the number of newly diagnosed cases of T2DM in Slovakia in a given year, divided by the number of people living in Slovakia in 2018. Since the number we get by dividing this is less than 1, interpretation on its own is difficult. Therefore, we usually multiply it by the power of ten (e.g. 10, 1,000, 1,000,000) and interpret the result verbally: "The incidence is x (number obtained by division) per 100,000 inhabitants".
  In our example, T2DM was diagnosed for the first time in 18,177 patients in 2018. According to the SUSR, that year, 5,450,421 people lived in Slovakia. The incidence is

thus 18,177/5,450,421 = 0.003335. Multiplied by 100,000 this is 333.5; the verbal interpretation is: "In 2018 the incidence of T2DM was 333.5 cases per 100,000 inhabitants."

- **Prevalence** is the number of **all cases of a disease** in a given population at a given point in time, related to the population multiplied by the power of 10.
  For example, we calculate the prevalence of Type 2 diabetes mellitus in the Slovak Republic in 2018 as the number of all people treated for T2DM in 2018 in Slovakia, divided by the number of people living in Slovakia in 2018. Again, the number we get by this division is less than 1, so we usually multiply it by the power of 10 (e.g. 10, 1,000, 1,000,000) and interpret the result verbally: "The prevalence is x (number obtained by division) per 100,000 inhabitants".
  In our example, 323,897 people were treated for T2DM in the Slovak Republic in 2018. That year, 5,450,421 people lived in Slovakia. The prevalence is thus 323,897/5,450,421 = 0.05943. If it is multiplied by 100,000, then it is 5,943 per 100,000, or just multiply the result by 100, which is 5.9 per 100, or 6%. The verbal interpretation is: "In 2018, the prevalence of T2DM was 6%, or: 6% of the population were being treated for the disease."

Mortality is also expressed as a population rate, that is, related to the population. In 2018, for example, a total of 54,293 people died in the Slovak Republic. That year, 5,450,421 people lived in Slovakia. The mortality is thus 54,293/5,450,421 = 0.00996. Multiplied by 100,000 this is 996; the verbal interpretation is: "In 2018 the all-cause mortality rate was 996 cases per 100,000 inhabitants."

Morbidity (incidence, prevalence) and the mortality rate can describe the condition in the entire population or be **specific** to, for example, sex or age (men/women, children under 18/adults, etc.) – then they are usually given for certain age groups, or a disease, or region etc.

Another indicator that determines the severity of the disease **case fatality**. This is an indicator of how deadly a disease is. It is expressed as the number of deaths per number of patients for a given diagnosis.
For example, the lethality of acute myocardial infarction in the Slovak Republic in 2018 is calculated as the number of deaths caused by heart attack in the Slovak Republic in 2018, which was 2,521 people, divided by the number of all people who had a heart attack in the Slovak Republic in 2018 (15,405 people). It is expressed as a percentage, so let's multiply the result by the number 100. In our example 2,521/15,405 = 0.16. Multiplied by 100 = 16%. The verbal interpretation is: The case fatality of acute myocardial infarction in Slovakia was 16% in 2018. Or: Of those who had an acute myocardial infarction in 2018, 16% died.

Use of the description

Descriptive epidemiology provides information about a population's health status, the occurrence of diseases and health risk factors, and identifies health problems and potential threats in society. Its results are therefore used for targeted prevention programmes. For example, if we describe the suicide rate or the incidence of cervical cancer by gender and age group, we know which population groups are most at risk and can suggest appropriate preventive measures.

In addition to monitoring health indicators, descriptive epidemiology is also used to evaluate intervention programmes (by monitoring population health indicators before and after the intervention) and to formulate priorities for public healthcare – any decision on building a health network should be based on existing data on population health and trends in the development of individual indicators.

Standardization

The above-mentioned incidence, prevalence and mortality rates are so-called crude rates. In order to be able to compare such indicators between populations, we need to standardize this data. This need arises from the different compositions of the population in different age groups. It would not be "fair" to compare, for example, the crude death rate from colon cancer in the US and Nigeria, where each of these countries has a markedly different age structure of the population and we know that the incidence of certain diseases is age-related (Fig. 2).

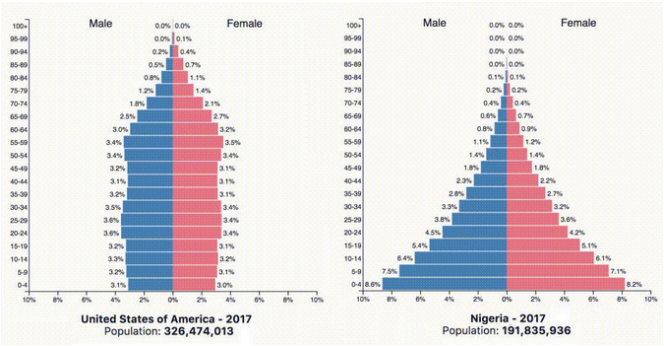

Fig. 2. Age structure of the population of the USA and Nigeria in 2017

Standardization (adjustment) is a classic epidemiological method that removes the confounding effect of variables that we know to differ in the populations being compared. The most common factor that we have to adjust (standardize) for is age. The age structure of the population varies from country to country. The crude incidence/prevalence or mortality rate of age-typical diseases – such as cardiovascular diseases and cancer – will therefore be higher in countries with a higher proportion of older population. It is not correct to compare crude rates in different countries; we have to look at age as a confounding, distorting factor, and as such, we need to standardize the crude rates.

There are two basic methods of standardization – direct and indirect.

Direct standardization uses a population with a certain age structure as a standard, mostly a fictitious world or European standard population.[1]. This has an accurately set representation of the population in each age group. We then standardize the crude morbidity or mortality rate in the observed population for the selected European or world population – we calculate the incidence/mortality rate in each age group of the observed population multiplied by the number of inhabitants in the age groups of the standard population and we get the expected incidence/prevalence/mortality figures in the standard population. We divide the sum of the expected numbers from the individual age groups of the standard population by the total number of inhabitants of the standard population and thus get the so-called standardized incidence/prevalence/mortality rate of our observed population.

The indirect standardization

Used less often and implemented when we don't have an age-specific number of illnesses or deaths. We need to know the specific incidence/mortality rates in the compared (standard) population. We first calculate the crude rate in the population we observed and from this we calculate the expected number of illnesses/deaths in the compared population. Dividing the number of observed and expected illnesses/deaths, we get the standardized mortality ratio (SMR).

What is the purpose of descriptive epidemiology?

Descriptive studies are very useful for estimating the burden of disease in a population (e.g., incidence, prevalence, mortality). This information is important for resource planning and setting up a network of medical facilities. Data from descriptive studies obtained in different populations or in the same population over time help to identify differences in disease occurrence, and this allows hypotheses to be generated about the cause or impact of the disease, which we then verify in analytical studies.

[1] https://apps.who.int/healthinfo/statistics/mortality/whodpms/definitions/pop.htm

Epidemiological methods II. Analysis, epidemiological cycle, hypothesis, measures of association (OR, RR), statistical significance

CHAPTER OBJECTIVE:
This chapter aims to familiarize students with the importance of analytical epidemiology in identifying the cause or effect of a disease.

Analysis is used in epidemiology to elucidate the aetiology and risk factors of a disease. It monitors the association between (at least) two variables: One is always a disease/medical condition and the other is a risk factor (or factors). Analysis in epidemiology answers the questions WHY? and HOW? (Why did this group of people get sick and others not? How did they get sick?), and determines the **causality** of the disease.
The analysis thus follows on the description, since the descriptive data provide us with basic information about the distribution of the observed disease/health status in the population, but do not talk about the causes, occurrence and spread.
Epidemiologists assume that a given disease or condition does not occur randomly in the population, but only when the necessary risk factors or determinants accumulate among individuals. The analytical method in epidemiology is used to reveal these factors. We use it to determine whether population groups with a different disease incidence differ in various characteristics, such as demographics, behaviour, environmental exposure, immunological profile, and socioeconomic status.

Use of the analysis

While descriptive methods in epidemiology serve to describe the occurrence of a disease or its determinants in a population, analytical epidemiology provides information about the quality and extent of the influence of determinants on the development and course of the disease.
The most commonly used method to get this information is to compare groups. Such a comparison begins with a hypothesis (one or more) about how a particular determinant affects the incidence of a disease. Analytical epidemiology also tries to uncover the cause of an epidemic. A case-control study allows the epidemiologist to examine the factors that preceded the disease. This often involves comparing a group of people who had the disease with a group without the disease, but who are similar in terms of composition by age, gender, socioeconomic status, and other variables, such as genetic or environmental factors that may be associated with the epidemic.

What are the uses of analytical epidemiology?

Analytical methods are used in various fields in epidemiology to elucidate the causes of a disease and the relationship between the disease and risk factors. This is necessary for many areas of medicine:

For example, it is used to detect the source and transmission factor of a disease in the field of infectious diseases. An example is an investigation of whether the consumption of a certain food presents an increased risk of diarrhoea (e.g. in a dining room, at a wedding or a party), or whether the source of drinking water represents an increased risk of diarrhoea.

In the field of epidemiology dealing with non-communicable diseases, an example of the use of analytical methods is to study whether exposure to a particular risk factor represents an increased risk of developing a particular disease – such as smoking causing a heart attack.

Comparison is the basis of analysis in epidemiology. We can observe a high incidence of a particular disease in the population and consider whether it is higher than we would expect based on, for example, its incidence in other population groups. In another case, we observe that in one group of cases of a community outbreak (e.g. people who contracted diarrhoea at the same time), several patients reported eating at a particular restaurant. Is that just a coincidence or did more people really got sick after eating there than we would expect? The answer to these questions is the comparison of the observed group to another group that represents the expected values (people that ate at different restaurants).

Association measures enable to study the **relationship between exposure and disease** to be quantified.

Exposure means not only exposure to a certain risk factor, such as contaminated food, infected mosquitoes, partners with sexually transmitted diseases or environmental toxins, but also personal characteristics (age, gender, race), biological characteristics (immune status), acquired characteristics (partnership status), activities (employment, leisure activities) or living conditions (socioeconomic status, access to healthcare).

The basis for the analysis is the division of the study group into groups that define the presence or absence of the monitored disease (sick/healthy) – or the effects of the disease (death/survival) – and whether or not the patients were exposed to the observed factor (exposed/unexposed). A suitable tool for the transparent representation of the study group according to these parameters is the so-called 2x2 Table (it has two columns and two rows) (Tab. 1):

Tab. 1. A 2x2 contingency table showing the relationship between the disease and the exposition factor

	Sick	Healthy	Total
Exposed	a	b	a + b
Non-exposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

The basic and most frequently used **association measures** that assess the relationship between the disease and a determinant are the risk ratio (RR), also known as the relative risk, and the odds ratio (OR).

Risk ratio

The strength of the relationship between exposure and disease **in the cohort study** is calculated using the **risk ratio.**
We will present the known data from the cohort study in a so-called 2x2 table (also called a contingency table), depending on whether the persons are sick/healthy or exposed/unexposed to the observed risk factor (Tab. 2).
The output is the RR, which is the risk ratio of disease due to exposure to a given factor.

Tab. 2. 2x2 table for recording the cohort study data and calculating odds ratio

	Sick	Healthy	Total
Exposed	a	b	a + b
Non-exposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

$R_{(exp.)} = a / (a+b)$
$R_{(nonexp.)} = c / (c+d)$

$$RR = R_{(exp.)} / R_{(nonexp.)} = [a/ (a+b)] / [c/ (c+d)] = [a.(c+d)] / [c.(a+b)]$$

The lines “Exposed” and “Unexposed” divide the entire cohort into those who were exposed to the observed factor and those who were not. The columns “Sick” and “Healthy” divide the cohort into those who are monitored for the disease and those who are not.
We calculate the risk for the Exposed: Rexp. = a/(a + b) and the risk for the Unexposed: Rnonexp. = c/(c + d). The risk ratio (relative risk) is Rexp./Rnonexp.

The relative risk (risk ratio), since it is a ratio, can either be >1, = 1 or <1.
If it is **>1**, then exposure to the given factor represents a risk, i.e. the observed factor is a **risk factor**.
If it is **<1**, then exposure to the given factor reduces the risk of disease, i.e. the observed factor is **protective**.
If it is **=1**, there is no connection between the factor exposure and disease.

The Framingham long-term cohort study, which for the first time identified risk factors for the cardiovascular system, is well-known and important in the field of cardiovascular health. It was initiated in 1948 in Framingham, Massachusetts, USA, with more than 5,000 residents involved – location was their common characteristic. The participants have been monitored

over time and the study is still ongoing with several generations of offspring from the original participants. We said that in a cohort study we start with a known exposure and watch the development of diseases that are affected by the exposure. In the case of the Framingham Study, there were multiple exposures and these were previously known factors that were considered (but not scientifically proven) to be harmful to the heart and blood vessels. The Framingham Study was the first to define risk factors for cardiovascular health and showed an increased risk of their cumulative effects.

Odds ratio

We calculate odds ratio usually **in case-control studies**.
One example is the international INTERHEART study, which aimed to determine the effects of several known risk factors on the incidence of heart attacks. The study examined nearly 30,000 people from 52 countries: 15,152 from the case group and 14,820 from the control group. The cases were those who had suffered a heart attack; the controls were those who did not have a heart attack. The risk factors monitored were smoking, high blood pressure, diabetes, waist-to-hip ratio, diet, physical activity, alcohol consumption, cholesterol, and psychosocial factors. Data from this case-control study are also processed and presented in the form of 2x2 table (Tab. 3).

Tab. 3. 2x2 table for recording the case-control study data and calculating odds ratio

	Sick	Healthy	Total
Exposed	a	b	a + b
Non-exposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

The chance of being exposed to a certain risk factor in a group of patients $O_{(cases)} = a / c$
The chance of being exposed to a certain risk factor in a group of healthy $O_{(healthy)} = b / d$

$$OR = O_{(Sick)} / O_{(healthy)} = (a/c)] / (b/d)] = (a.d) / (b.c)$$

The group of patients with a heart attack are all marked as “Sick” (a + c). The control group are people without a heart attack; therefore they are “Healthy” (b + d). As mentioned earlier, this study examined several risk factors and their association with myocardial infarction. Let’s take one of them as an example – diabetes. Of the total group monitored, all those who had diabetes are those exposed to the risk factor = group a + b. Those who did not have diabetes are not exposed = c + d. We are interested in the connection between diabetes and heart attack, i.e. how much diabetes contributes to the development of a heart attack. In this case, we are using the association rate called the odds ratio (OR). The calculation is as follows: We

calculate the likelihood of being exposed to a risk factor (diabetes) in a group of patients (those who have had HA): O1=a/c.
Then we calculate the likelihood of being exposed to a risk factor (diabetes) in a group of healthy people (who have not had HA): O2=b/d.
We calculate the odds ratio OR as follows:
OR = O1/O2 = (a/c) / (b/d) = (a x d) / (c x b)

The odds ratio, since it is a ratio (similarly as risk ratio), can either be >1, = 1 or <1.
When OR > 1 → the exposition factor contributes to the development of the disease
When OR < 1 → the exposition factor is protective
When OR = 1 → there is no relationship between exposure to RF and the development of the disease

The INTERHEART study showed an OR = 2.37 for diabetes. The interpretation of this result is: People who have had a heart attack are 2.37-times more likely to develop diabetes than people without a heart attack.
Just out of interest, the odds ratio for other risk factors was as follows:
Smoking OR = 2.87; Hypertension OR = 1.91; Abdominal obesity OR = 1.12; Psychosocial factors 2.67.

Apart from the relative isk, we can also calculate attributable risk (also called risk difference) (Tab 4) and population attributable risk (Tab. 5). These we calculate when we want to know to what extent does the exposition factor contributes to the development of the disease (in noncommunicable diseases the cause is usually multifactorial, therefore we have to have in mind that there might be also other contributing factors, not just our exposition factor).

Tab.4. Attributable risk (AR)

	Ill	Healthy	Total
Exposed	a	b	a + b
Non-Exposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

$R_{(exp.)} = a / (a+b)$
$R_{(nonexp.)} = c / (c+d)$

$$AR = R_{(exp.)} - R_{(nonexp.)} = [a/(a+b)] - [c/(c+d)]$$

Tab.5. Population attributable risk

	Ill	Healthy	Total
Exposed	a	b	a + b
Non-Exposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

% of cases out of exposed that are attributable to the exposure factor

$R_{(exp.)} = a / (a+b)$ $R_{(nonexp.)} = c / (c+d)$

$$\% \text{ AR} = \frac{R_{(exp.)} - R_{(nonexp.)}}{R_{(exp.)}} \; x \; 100$$

In addition to determining the strength of the association itself, we also need to determine the so-called **statistical significance**, i.e. how great is the probability that the result of the association is only random or if it is statistically significant (see below in the section **Statistical Significance**).

Epidemiologic cycle

We have already mentioned that epidemiology studies the distribution and determinants of health and disease in a population in order to address health problems. We also mentioned that this study will be carried out in several steps based on the order Data Collection -> Description -> Analysis.

The entire so-called epidemiologic cycle for solving a health problem in the population can be depicted as follows (Fig. 1):

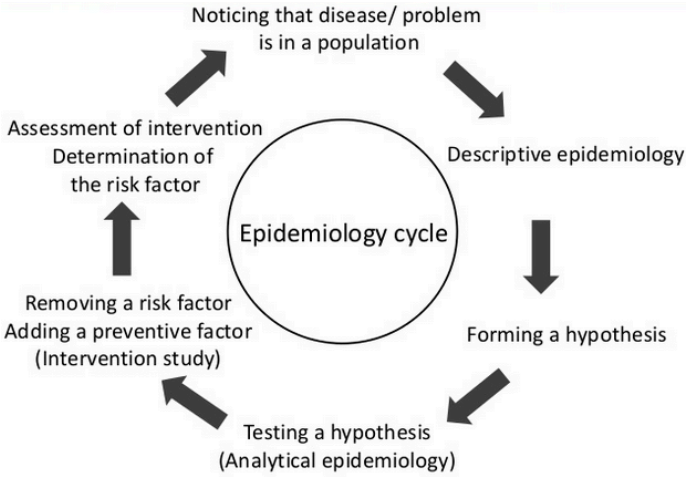

Fig. 1. Epidemiologic cycle for resolving a health problem in a population

Step 1. A health problem in a population means that we must first identify what problem we want to solve in our research. We define this as a research question. For example, we know from the international scientific literature how high the incidence of Type 2 diabetes mellitus is in the world population in individual age groups. We want to know how Slovakia is doing in this regard. We formulate a research question, in this case, for example: What is the current incidence of Type 2 diabetes mellitus (T2DM) in the Slovak Republic?

Step 2. Description: at this point, we will make an observation and describe the situation. In our example, this means that we get data on the number of people treated for T2DM in the Slovak Republic in the last reporting year from the database of the National Health Information Centre. Since absolute numbers are not meaningful for the situation in the population, we calculate crude incidence and prevalence rates. If we want to compare the situation in Slovakia with foreign countries, we need to standardize these measures.

Sep 3. Formulation of a hypothesis. Based on the description from the previous point, several questions may arise: Why is the incidence of T2DM higher in the Slovak Republic than in other countries? What is the observed incidence and prevalence of the disease associated with? If we want to find answers to these and other questions, we continue in our research by hypothesizing.

Simply put, a hypothesis is the answer that we think we will get through research. In our example, if we want to answer the questions asked, we need information from the literature on the most common risk factors for T2DM. We know that these includes diet, physical activity and smoking. Based on these facts, we can make a hypothesis, such as: The incidence of T2DM in Slovakia is related to the smoking rate in the population.

Step 4. Testing a hypothesis. An analytical method comes into play where we test a hypothesis against a comparison group in order to quantify the relationship between exposure and disease. In our example, we would need to know how many people with T2DM smoke and how many people in the general population smoke. We can then use an analytical method to compare whether the incidence of T2DM in individual age groups is higher in smokers than in non-smokers.

Step 5. Intervention. So far we have only observed, i.e. worked with data that were collected in the population without our intervention in the course of the disease in the observed group. However, if we want to actively influence the state of health of the observed group, we can propose an intervention, that is, an intervention through which we eliminate the risk factor or introduce a protective factor and monitor the effect of this procedure on the health of the population group. In our example with a selected population group in which we found the highest association between smoking and the incidence of T2DM, we can intervene in the form of an intensive anti-smoking campaign.

Step 6. Intervention evaluation. In this step, we again use the analytical method to assess the effectiveness of the intervention on the health of the observed group. In our example, this would be, for example, a comparison of the occurrence of T2DM in the observed group before the intervention and after its implementation, with some hindsight.

The epidemiological cycle corresponds to the universal scientific basic method for gaining knowledge established in the 19th century:
Observation -> Hypothesis -> Prediction -> Experiment

Hypothesis

Let's talk a little more about forming a hypothesis.
A hypothesis, as stated earlier, is a formulated answer that we expect to find through research. It is used to explain causality or to predict the relationship between variables. The hypothesis must meet four evaluation criteria: 1. It must determine the expected relationship between the variables. 2. It must be testable. 3. It must be consistent with existing knowledge. 4. It must be formulated as simply as possible.
We speak of a null or an alternative hypothesis.
The null hypothesis (labelled as H_0) assumes that there is no difference between the groups being compared.
The alternative hypothesis (H1) is actually a negation of H_0; it assumes that we will find a difference between the groups being compared.

Statistical Significance

The final step in testing the hypothesis is to determine the likelihood that the outcome of the study is just random or, conversely, statistically significant and therefore generalizable to the whole population. **The statistical significance** of the result (for example, the relationship between exposure and disease) is determined through statistical tests, the use of which depends on the types of variables we are working with. Statistical significance is also important for the inference, i.e. the generalizability of the results obtained in the sample to a larger population.
It is expressed by the **p-value**. The p-value expresses the strength of the evidence against the null hypothesis H_0. It is the probability that the same results will be obtained after repeated testing, that is, evidence that the results obtained were not a product of chance.
The p-value can be between 0 and 1. The closer it is to 0, the less likely it is the result is just a product of chance. The closer to 1, the more likely the result is just random and therefore it cannot be generalized to a larger population.
The level of significance by which we judge the acceptance or rejection of the null hypothesis is artificially determined. The most common significance level used when publishing research results is 5% – this means that the probability that the result obtained is only part of the coincidence is less than 1 in 20, i.e. less than 5% (in this case, the p-value is <0.05). Significance levels of 1% (p-value <0.01) and 0.1% (p-value <0.001) are also used.
Another term that needs to be mentioned in connection with statistical significance is the **confidence interval** (CI), which represents the range in which the calculated result lies. If we

set a significance level of 5% and the p-value is less than 0.05 (5%), we have 95% confidence that the result is within the confidence interval.

When the CI crosses the number 1, our result (eg. OR or RR)
is **not statistically significant** (e.g.: OR 1.2, 95% CI: 0.79 – 1.5).

When our result (eg. OR or RR) is >1 and the whole CI is > 1, our result (eg. OR or RR)
is statistically significant (e.g., RR 2.89, 95% CI: 1.18 – 4.39).

When our result (eg. OR or RR) is <1 and the whole CI is < 1, our result (eg. OR or RR)
is statistically significant (e.g. OR 0.45, 95% CI: 0.12 – 0.83).