Descriptive epidemiology

Involves observation, definitions, measurements, and interpretations	Dissemination of health-related states or events by a person, place, and time.
1. Providing information about a disease or condition.	2. Providing clues to identify a new disease or adverse health effect.
3. Identifying the extent of the public health problem.	4. Obtaining a description of the public health problem that can be easily communicated.
5. Identifying the population at greatest risk.	6. Assisting in planning and resource allocation.
7. Identifying avenues for future research that can provide insights about an etiologic relationship between an exposure and health outcome.	

The research problem, question, and hypotheses are supported by descriptive epidemiology.

Hypotheses are tested using appropriate study designs and statistical methods.

Describing data by person allows identification of the frequency of disease and who is at greatest risk.

Describing data by place (residence, birthplace, place of employment, country, state, county, census tract, etc.)

Descriptive statistics are a means of organizing and summarizing data.

Descriptive Study Designs

	Description	Strengths	Weakness
Ecologic	Aggregate data involved (no information is available for specific individuals	Takes advantages of preexisting data. Relatively quick and inexpensive. Can be used to evaluate programs, policies, or regulations implemented at the ecologic level. Allows estimation of effects not easily measurable for individuals	Susceptible to confounding, exposures and disease or injury outcomes not measured on the same individuals
Case Study	A snapshot description of a problem or situation for an individual or group; qualitative descriptive research of the facts in chronological order	In depth description provides cues to identify a new disease or adverse health effect resulting from an exposure or experience. Identifies potential areas of research	Conclusions limited to the individual, group, and/or context under study, cannot be used to establish a cause-effect relationship

Descriptive Study Designs (cont)

Cross-Sectional	All variables measured at a point in the time. No distinction between potential risk factors and outsomes	Control over study population. Control over measurements. Several associations between variables can be studied at the same time. Short time period required. Complete data collection. Exposure and injury/disease data collected from same individuals. Questions can be asked to obtain prevalence data.	No data in the time relationship between exposure and injury/disease development, no follow up, potential bias from low response rate, potential measurements bias, higher proportion of long term survivors, not feasible with rare exposures or outcomes, does not yield incidence or relative risk

Descriptive study designs include case reports and case series, cross-sectional surveys, and exploratory ecologic designs.
These designs provide a means for obtaining descriptive statistics without typically attempting to test particular hypotheses.
In an ecologic study, the unit of analysis is the population. In a case report, case series, or cross-sectional survey, the unit of analysis is the individual.

Ratios, Proportions, and Rates I

Ration	In a ratio, the values of x and y are distinct, such that the values of x are not contained in y. The rate base for a ratio is 100 = 1. The rate base for a ratio is 100 = 1. For example, in 2017 in the United States, the leading causes of death for the age group 15–24 were unintentional injury (9,746 in males and 3,695 in females), suicide (5,027 in males and 1,225 in females), and homicide (4,234 in males and 671 in females). Corresponding ratios indicate that males are 2.64 times more likely die from unintentional injury, 4.10 times more likely to commit suicide, and 6.31 times more likely to die from homicide.
Proportions	A proportion is typically expressed as a percentage, such that the rate base is 102 = 100. Thus, for the preceding data, we can say that of deaths involving unintentional injury, 72.5% were male; of suicides, 80.4% were male; and of deaths due to homicide, 86.3% were male.
Rates	Is a type of frequency measure where the numerator involves nominal data that represent the presence or absence of a health-related state or event. It also incorporates the added dimension of time; it may be thought of as a proportion with the addition that it represents the number of disease states, events, behaviors, or conditions in a population over a specified time period. An incidence rate is the number of new cases of a specified health-related state or event reported during a given time interval divided by the estimated population at risk of becoming a case.

In epidemiology, it is common to deal with data that indicate whether an individual was exposed to an illness, has an illness, has experienced an injury, is disabled, or is dead. Ratios, proportions, and rates are commonly used measures for describing dichotomous data. The general formula for a ratio, proportion, or rate is: X/Y *10^Z.

A mortality rate is the total number of deaths reported during a given time interval divided by the population from which the deaths occurred.

Ratios, Proportions, and Rates II

A mortality rate is the total number of deaths reported during a given time interval divided by the population from which the deaths occurred.

The attack rate is also called the cumulative incidence rate. It tends to describe diseases or events that affect a larger proportion of the population than the conventional incidence rate

Outbreak refers to more localized situations, whereas epidemic refers to more widespread disease and, possibly, over a longer period. The investigation of the outbreak involved first constructing a line listing of those at the picnic. Each line represented an individual, with measurements taken on age, gender, time the meal was eaten, whether illness resulted, date of onset, time of onset, and whether selected foods were eaten.

Another common measure for describing disease and health-related events is prevalence, which is the frequency of existing cases of a health-related state or event in a given population at a certain time or period.

Period prevalence is the frequency of an existing health-related state or event during a time period. For example, the period prevalence of arthritis in a given year includes existing cases the first day of the year, along with new (incident) cases diagnosed during the year. Period prevalence is less commonly used than point prevalence.

The crude rate of an outcome is calculated without any restrictions, such as by age or gender or weighted adjustment of group-specific rates; however, these rates are limited if the epidemiologist is trying to compare them between subgroups of the population or over time because of potential confounding influences, such as differences in the age distribution between groups.

A confidence interval is the range of values in which the population rate is likely to fall.

Interpretation

- SMR = 1: The health-related states or events observed were the same as expected from the age-specific rates in the standard population.
- SMR > 1: More health-related states or events were observed than expected from the age-specific rates in the standard population.
- SMR < 1: Fewer health-related states or events were observed than expected from the age-specific rates in the standard population.

Tables, Graphs, and Numerical Measures

A frequency distribution is a complete summary of the frequencies, or number of times each value appears.

Relative frequency is derived by dividing the number of people in each group by the total number of people.

Bar charts are often used for graphically displaying a frequency distribution that involves nominal or ordinal data.

A histogram shows a frequency distribution for discrete or continuous data. The horizontal axis displays the true limits of the selected intervals.

A frequency polygon is a graphical display of a frequency table.

An epidemic curve is a histogram that shows the course of an epidemic by plotting the number of cases by time of onset.

A stem-and-leaf plot is a display that organizes data to show their distribution. Each data value is split into a “stem” and a “leaf.”

Tables, Graphs, and Numerical Measures (cont)

A box plot has a single axis and presents a summary of the data.	A two-way (or bivariate) scatter plot is used to depict the relationship between two distinct discrete or continuous variables.	A spot map is used to display the location of each health-related state or event that occurs in a defined place and time.	A line graph is similar to a two-way scatter plot in that it depicts the relationship between two continuous variables.	Measures of central tendency refer to ways of designating the center of the data. The most common measures are the arithmetic mean, geometric mean, median, and mode.	Arithmetic mean is the measure of central location that one is most likely familiar with because it has many desirable statistical properties; it is the arithmetic average of a distribution of data.	The geometric mean is calculated as the nth root of the product of n observations. It is used when the logarithms of the observations are normally distributed.
Median is the number or value that divides a list of numbers in half; it is the middle observation in the data set. It is less sensitive to outliers than the mean.	Range is the difference between the largest (maximum) and smallest (minimum) values of a frequency distribution.	Interquartile range is the difference between the third quartile (75th percentile) and the first quartile (25th percentile). Note that the distribution of data consists of four quarters.	Variance is the average of the squared differences of the observations from the mean.	Standard deviation is the square root of the variance. The standard deviation has mathematical properties that are useful in constructing the confidence interval for the mean and in statistical tests for evaluating research hypotheses.	The coefficient of variation is a measure of relative spread in the data. It is a normalized measure of dispersion of a probability distribution that adjusts the scales of variables so that meaningful comparisons can be made.	

Measures of Association

Correlation coefficient	Represents the proportion of the total variation in the dependent variable that is determined by the independent variable. If a perfect positive or negative association exists, then all the variation in the dependent variable would be explained by the independent variable. Generally, however, only part of the variation in the dependent variable can be explained by a single independent variable.

Measures of Association (cont)

Spearman's rank correlation coefficient (denoted by rs)	An alternative to the Pearson correlation coefficient when outlying data exist, such that one or both of the distributions are skewed. This method is robust to outliers.
Simple regression model y = b0 + b1x1	A statistical analysis that provides an equation that estimates the change in the dependent variable (y) per unit change in an independent variable (x). This method assumes that for each value of x, y is normally distributed; that the standard deviation of the outcomes y do not change over x; that the outcomes y are independent; and that a linear relationship exists between x and y.
Multiple regression y = b0 + b1x1 + … + bkxk	An extension of simple regression analysis in which there are two or more independent variables. The effects of multiple independent variables on the dependent variable can be simultaneously assessed. This type of model is useful for adjusting for potential confounders.
Logistic regression Log(odds) = b0 + b1x1	A type of regression in which the dependent variable is a dichotomous variable. Logistic regression is commonly used in epidemiology because many of the outcome measures considered involve nominal data.
Multiple logistic regression Log(odds) = b0 + b1x1 + … + bkxk	An extension of logistic regression in which two or more independent variables are included in the model. It allows the researcher to look at the simultaneous effect of multiple independent variables on the dependent variable. As in the case of multiple regression, this method is effective in controlling for confounding factors.

A contingency table is where all entries are classified by each of the variables in the table. For example, suppose we were interested in assessing whether exposure to a dietary intervention (yes vs. no) is associated with a decrease in low-density lipoprotein (yes vs. no). A 2 × 2 contingency table could represent the data.

Conclusion

Descriptive epidemiology is used to assess and monitor the health of communities and to identify health problems and priorities according to person (who?), place (where?), and time (when?) factors. It also involves characterizing the nature of the health problem (what?). Selected descriptive study designs, statistical measures, and graphs and charts were presented for describing the frequency and pattern of health-related states or events.

Descriptive analysis is the first step in epidemiology to understanding the presence, extent, and nature of a public health problem and is useful for formulating research hypotheses. Descriptive studies are hypothesis generating; they provide the rationale for testing specific hypotheses. The analytic study design, which is the focus of a later chapter, involves evaluating directional hypotheses about associations between variables. Some of the same measures and statistical tests used in exploratory and descriptive studies are also used in analytic studies. After a hypothesis is statistically evaluated for significance and an association between variables is deemed to not be explained by chance, bias, or confounding, then an investigator can use this information as part of the evidence for establishing a cause–effect relationship. Other criteria to consider in making a judgment about causality must also be considered, including temporality, dose–response relationship, biologic credibility, and consistency among studies.