

PRINCIPLES OF INFECTIOUS DISEASE EPIDEMIOLOGY

MODULE V – DISPLAYING AND INTERPRETING EPIDEMIOLOGIC VARIABLES

Portions of this module were adapted from the Centers for Disease Control and Prevention (CDC) “Principles of Epidemiology, Second Edition, An Introduction to Applied Epidemiology and Biostatistics, 1998.”

Note: You will need to access the course to view the many examples associated with this module, as they are not included in this outline.

I. INTRODUCTION

Module V is designed to prepare public health workers to meet the following objectives:

- Define the three categories of epidemiologic variables
- Identify the three main methods used to organize epidemiologic data
- Correctly interpret graphic presentations of epidemiologic data
- Choose appropriate display methods and formats for specific kinds of epidemiologic data

II. WHAT ARE EPIDEMIOLOGIC VARIABLES?

A. Epidemiologic variables are characteristics that can be observed and/or measured. They may be characteristics of:

- **Time** - the time of illness or of a relevant event.
Examples: date of exposure or onset of illness.
- **Place** - the environment in which illness occurs.
Examples: place of residence, of work, or of suspected exposure (such as a retail food establishment).
- **Person** - individuals who are infected, ill, or at risk. Examples: age, gender and occupation.

B. We look at epidemiologic variables to:

- identify characteristics that might be important, and
- to form hypotheses about the
 - source,
 - causative agent, and
 - mode of transmission of illness.

C. When we review surveillance data or investigate an outbreak, we are trying to answer these questions:

WHO? Identify individuals and sub-populations at risk of exposure or transmission

HOW? Identify the modes of disease transmission, especially any changes from known transmission patterns

HOW CAN WE INTERVENE? Identify factors or conditions that can be manipulated to modify or prevent disease occurrence and spread

WHAT ARE OUR PRIORITIES? Identify the usual patterns of disease in a population, so we can set priorities and respond more quickly to outbreaks

Remember, ***the ultimate goal of epidemiology is to prevent disease.*** This is done by finding associations between a disease and the characteristics of

- time,
- place, and
- person.

III. METHODS FOR ORGANIZING EPIDEMIOLOGIC DATA

- The field epidemiologist needs to organize data for several reasons:
 - It is a necessary step in data analysis.
 - It helps the epidemiologist visualize patterns and trends, and identify variations from those trends.
 - It provides a useful way to communicate information to others.
- There are three basic methods of organizing epidemiologic data
 - Tables
 - Charts and
 - Graphs

A. Tables

1. In General:

- A table is a set of data, organized into rows and columns.
- Tables are useful for identifying patterns, exceptions, differences and other relationships.
- Tables also serve as the basis for charts and graphs.

A table should be self-explanatory. It should convey all the information the reader needs to understand the data, including:

- A clear and concise title
- Row and column labels (clear and concise)
- Totals for rows and columns
- Footnotes to explain any codes, abbreviations, or symbols

- Footnotes that note any data excluded from the analysis
- Clear identification of the source of the data in a footnote

2. Types of Tables

a) *One Variable Tables.* The simplest form of table has only one variable. That is the frequency distribution, which was discussed in Module IV, Statistical Measures.

- To review briefly, in a frequency distribution table, the first column shows the values (or categories) of a variable, and the second column shows the number of people or records that fall into each value or category.
- Often, there is a third column that lists the percentage of persons or events in each category.
- Sometimes a one-variable table shows the cumulative frequency or cumulative percent.

b) *Two- and Three-Variable Tables.*

- Additional columns are added to a table to show counts by a second variable, for example age and sex.
- As you learned in Module IV, a two-by-two table is used to show cross-tabulated data. Another name for a two-by-two table is a contingency table. Such tables display two variables, each with two categories.
- It is usually better to use only one- or two-variable tables. Sometimes, though, a third variable is needed to show a set of data more completely, for example, age, sex and race. A three-variable table is hard for the reader to interpret. No table should attempt to show more than three variables.

c) *Tables of Other Statistical Measures.*

- Although all the examples used so far show counts (frequency) of the variables, tables can also be used to show other statistical measures.
- The cells can contain rates, means, relative risks or other measures.
- Just be sure the titles and row/column headings clearly identify what data is being presented.

3. *Creating Class Intervals.*

- Some variables such as sex or “ate potato salad?” have a limited number of possible values.
- Others have a broader range of possible responses, and categories or “class intervals” are needed to group them.
- The following rules are important when creating class intervals:

- Create categories that are mutually exclusive and include all of the data. For example, if your first category is 0-5, the next one must start with 6, not 5.
- Use a relatively large number of narrow categories for the initial analysis, since you can always combine them later.
- Try to use standard groupings if they are available – for example, age categories used by CDC for a particular disease.
- Create a category for unknowns, since there will usually be missing information for some of the cases.

If there are no standard or natural class intervals for a particular variable, there are several ways to create class intervals. These are discussed in more detail in the CDC Principles of Epidemiology home study course.

B. Charts

1. In General

Charts:

- Are a method of organizing and illustrating data using only one coordinate.
- Are best used for comparing data with discrete categories.

Several types of charts may be produced using common spreadsheet software such as Excel.

2. Types of Charts

a) Bar Charts. Bar charts are used to create a visual display of the data from a table. The bars may be either horizontal or vertical.

- ***Simple bar charts***
 - Used to display the data from a frequency distribution (one-variable table).
 - Each bar represents one value of the variable.
 - This makes it very easy to compare the relative magnitude of the different values.
- ***Grouped bar charts***
 - Used to illustrate data from two- or three-variable tables.
 - The bars must be shaded or colored differently and described in a legend.
 - It is best not to use more than three bars per group.
 - Leave a space between adjacent groups of bars, but not between bars in a group.
- ***Stacked bar charts***
 - Another way of showing two variables.

- o The values of the second variable make up segments of the bars that represent the first variable.
- o These charts can be hard to interpret, since only the first segment rests on a flat baseline.
- **Deviation bar charts**
 - o Used to show how a variable deviates from a baseline, in both positive and negative directions.
 - o The bars are usually positioned horizontally.
 - o Each week, CDC's *Morbidity and Mortality Report* (MMWR) uses a deviation bar chart to show the number of cases of several diseases reported during the last four weeks, compared to the number reported during the same four weeks for the past five years.
- **A few simple rules for constructing bar charts:**
 - o Arrange the categories that define the bars in a natural order (for example, alphabetically or by increasing age), or in an order that produces increasing or decreasing bar lengths.
 - o Make all of the bars the same width, whatever looks good.
 - o Make the length of the bars in proportion to the frequency of the event.
 - o Code different variables by differences in bar color or shading, and include a legend that interprets your code.

b) Pie Charts.

- A pie chart is simple and easily understood.
 - o Very useful for showing the component parts of a single group or variable.
 - o The size of the pie "slices" shows the percentage for each component part of the whole.
 - o Pie charts are easily generated using spreadsheet software such as Excel.
- A few simple rules for constructing pie charts:
 - o Start at "12:00 o'clock" (straight up) and arrange the component slices from the largest to the smallest, going clockwise.
 - o Put the categories "other" and "unknown" last.
 - o Use different colors or shading for each "slice."
 - o Show somewhere on the graph what 100% of the pie represents (for example, the total number of cases).
 - o Indicate the percentage that each "slice" represents.
- Multiple pie charts are sometimes used to compare the same components in two different groups or variables. However, it is hard to accurately compare two or more pie charts visually.

c) Maps.

- Maps are a very widely used type of chart.
- They are also called geographic coordinate charts.
- Spot maps and area maps are commonly used in field epidemiology.

- **Spot maps** use dots or other symbols to show where an event took place, or where a disease condition exists.
 - Spot maps are good for detecting clusters of disease cases.
 - However, we must remember that a spot map does not take into account the size of the population at risk.
 - So it does not show the risk of the event occurring in that particular place.
 - A heavy clustering of dots may simply mean that more people live in that area and therefore more cases appear there.

- **Area maps**, however, can be used to illustrate differences in risk in different areas.
 - An area map uses shaded areas to show either the incidence of an event, or the distribution of some condition over a geographic area.
 - We can show either rates or numbers with an area map.
 - When we calculate and show a specific rate for each sub-area, we can make direct comparisons of risk between them.

- Mapping technology has grown very sophisticated in recent years. The widespread availability and use of Global Positioning Systems (GPS) devices and Geographic Information Systems (GIS) software has made it easier to produce accurate, up-to-date maps to aid in epidemiologic investigations.

C. GRAPHS

1. In General:

- A graph is a way to show numerical data visually, using a system of coordinates.
- A graph can help us see patterns, trends, aberrations, similarities, and differences in the data.
- People usually understand and remember the important aspects of data much more easily from looking at a graph than a table.
- Graph format:
 - Most graphs used in epidemiology have two lines, one horizontal and one vertical.
 - The horizontal line is called the x-axis
 - The vertical line is the y-axis.

- o The x-axis (horizontal) is used to show the values of the method of classification, for example, time in years, which is called the independent variable.
- o The y-axis (vertical) is use to show the dependent variable, usually a frequency measure such as number of cases or rate of disease.
- o Each axis must be clearly labeled and the scale of measurement marked.

2. Types of Graphs

Arithmetic-Scale Line Graphs

- Show patterns or trends over some variable, usually time.
- In epidemiology these graphs are often used to show the history of incidence of a disease over time.
- Arithmetic-scale line graphs are also good for comparing two or more sets of data.

Here are a few simple rules for constructing arithmetic-scale line graphs:

- Mark off each axis at equal intervals.
- Use a scale on the x-axis that matches the intervals used when collecting the data (for example, days, weeks, months or years). If very small intervals were used, combine them into larger ones.
- Make the y-axis shorter than the x-axis, so the graph appears horizontal.
- Always start the y-axis with 0.
- Pick a range of values for the y-axis that is slightly higher than the largest number you will be plotting.
- Select an interval size for the y-axis that will give you enough intervals to show the data in enough detail for your purposes.

A **histogram** is another type of graph that is very important in field epidemiology. We will learn about histograms later in this module.

IV. INTERPRETING EPIDEMIOLOGIC DATA: TIME, PLACE AND PERSON VARIABLES

As we look at tables, graphs and charts to draw inferences and form hypotheses, we often make comparisons between:

- Different **time** periods,
- **Places**, and
- Groups of **people**.

It is very important to use comparable data when making such comparisons. Be sure the data are comparable with respect to:

- Case definitions
- Level of effort in case-finding and data collection
- Time periods (compare weeks with weeks, months with months etc.)
- Populations (if two populations differ in age distribution or density, this should be taken into account)

Apparent differences in disease incidence can be influenced by any of these factors.

A. Time

1. In General

Variations over time in the frequency of a disease can tell us a lot about the determinants of that disease in a given population.

We may look at trends over many years (called secular trends), or seasonal variations, or variations over the short time period of an outbreak.

- As we look at disease data over time, we should ask these questions:
 - What is the pattern?
 - What factors might explain it?
 - What is the most likely future pattern?
- Variations over time may result from:
 - True increases or decreases--actual changes in the frequency of the disease in that population
 - Changes in sensitivity or specificity of the surveillance system
 - Mistakes made in collecting or organizing the data
 - Changes in the perceptions of the public or the health care community about the importance of diagnosing and reporting that particular disease
- A long-term increase in incidence of a disease may reflect changes such as:
 - Introduction of a new disease agent into a population
 - Decreasing effectiveness of control measures
 - Changes in the environment (for example, climatic conditions affecting the tick population)
 - Changes in societal practices (for example, urbanization leading to greater population density)

- Many diseases are subject to **cyclic changes** over time. Diseases that are strongly influenced by environmental factors may show seasonal variation.
- Another form of cyclic change is called **secular trends**. These are marked changes over long time periods that are caused by changes in environmental factors or host susceptibility.

Example of secular trend:

- Hepatitis A incidence has historically cycled up and down over periods of ten or more years.
 - When the disease is widespread, the number of susceptible people (especially children) goes down, so it begins to wane.
 - When enough susceptibles build up, it increases again.
 - This pattern may be changing now that an effective vaccine is available to prevent hepatitis A.
- Finally, time trends for a disease may simply show erratic change due to chance variations, sometimes called “noise.” This is true of many diseases with low endemic levels. It is also a common pattern when looking at very localized data, since the number of cases may be too low to exhibit a strong pattern.

2. Time during an outbreak or epidemic period

- When investigating an outbreak, monitoring time trends becomes critical.
 - An outbreak or epidemic period is the time during which the number of cases of a disease exceeds the expected number.
 - Time is usually measured in hours, days, weeks or months.
- A special kind of graph, called a **histogram**, is used to plot cases according to the time of onset of symptoms. This is called an **epidemic curve**. Interpretation of epidemic curves will be covered in more detail in the workshop portion of this course.
 - o A histogram looks somewhat like a bar graph, but with several important differences. In an epidemic curve:
 - The x-axis is always made up of equal time intervals, and should begin just before the outbreak
 - The time intervals should be appropriate to the disease in question (hours, days, weeks, months)
 - The y-axis is the number of cases
 - Each case is represented by one square, and all squares are of equal size
 - There are no spaces between the columns

- There may or may not be horizontal lines between the squares
 - You may show a second variable in a histogram through the use of shading.
 - For example, you may want to look at the distribution of hepatitis cases who are county residents vs. visitors to the county.
 - This can be done either by shading the squares representing visitors, or by constructing two separate histograms.
 - The most common spreadsheet software, such as Excel, does not generate proper histograms. They may be constructed by hand, or using specialized software such as Epi Info, which was produced by CDC.
- A **frequency polygon** may be used instead of a histogram to show an epidemic curve. In this type of graph, the squares are replaced by a line. However, a frequency polygon is not the same as a line graph, as shown in the next example.

B. Place

- Place is a specific geographic area that can be described by latitude, longitude and altitude. As used in epidemiology, place:
 - May be a street address, city, state, region, or country, or
 - May be expressed as a dichotomous, “either-or” variable such as
 - urban/rural
 - domestic/foreign
 - institutional/non-institutional
 - lower vs. higher socioeconomic areas
- The association of a disease with a place implies that the most important causative factors are in the environment or people of that place. For example:
 - Population density (urban vs. rural) can affect how rapidly an airborne disease is transmitted, or determine risk of exposure to a vectorborne disease.
 - The incidence of many diseases increases as socioeconomic status decreases, due to the effects on immune status, quality of the environment, overcrowding, etc.
 - Regional variations can reflect the specific ecologic requirements of a disease agent or vector

- Country of origin or of exposure can be important, because diseases may be imported from endemic areas (for example, malaria)
- Rates, not counts, must be used to compare disease incidence in different places. Otherwise the difference in population size would make it impossible to interpret any differences.
- Place comparisons are most useful if we look at the data over time. Remember, no one point in time can give us all the information we need.

C. Person

As we learned in Module I, people can be described in terms of many inherited or acquired characteristics such as:

- Age
- Sex
- Race
- Immune status
- Marital status
- Educational level

They may also be described in terms of their activities, such as:

- Occupation
- Recreational activities
- Religious practices
- Customs

Or, they may be described by the circumstances in which they live, such as:

- Social conditions, for example housing
- Economic status
- Environmental conditions

These variables are important since they determine, to a large degree, who is at the greatest risk of acquiring specific infections.

Age is the single most important personal characteristic. To a large extent, it determines:

- The physiologic activity of the disease organism
- The level of immunity or resistance, and
- The potential for exposure to a disease agent

Our behavior, and therefore our risk of exposure, differs markedly at different life stages. Examples are the mouthing behavior of toddlers and increased sexual activity during adolescence and young adulthood.

Sex can also influence the risk of disease.

For many diseases, both sexes have about the same level of risk. However, if there is a gender difference in a particular disease it usually means either males or females had a greater opportunity for exposure. This could be due to differences in:

- occupation (for example, child care, agricultural work)
- recreational activities (for example, hunting), or
- social behaviors (for example, intravenous drug use)

Race and/or ethnicity can also be a factor in disease risk.

- Such disparities most often result from differences in exposure or immunity status (for example, immunization rates).
- With most diseases, racial and ethnic differences are a consequence of socio-economic and cultural differences, rather than physical differences.

Summary

Epidemiologic variables are characteristics that can be observed and/or measured. They may be characteristics of

- **time,**
- **place,** or
- **person.**

Tables, charts and graphs are good tools for organizing epidemiologic data. They make it possible to identify, explore, understand and present data distributions, trends and relationships.

After we organize the data, we can look at epidemiologic variables to:

- identify characteristics that might be important, and
- form hypotheses about the source, causative agent, and mode of transmission of illness.