Section 6: Understanding Basic Analytic Concepts

A common understanding of key terms in data analysis and the methods to be used in developing the HIV/AIDS epidemiologic profile is critical for a planning group. This section presents basic terms and definitions and, when applicable, the methods you are encouraged to adopt when preparing your profile. See the glossary for other relevant terms.

Terms, Definitions, Calculations

case: A condition, such as HIV (e.g., an HIV case) or AIDS (e.g., an AIDS case), according to a standard case definition.

confidence interval (CI): A range of values for a measure that is believed to contain the true value at a specified level of statistical certainty (e.g., 95%).

convenience sampling: A technique that relies upon selecting people who are easily accessible at the time of a survey (e.g., a survey of clients who attend a group meeting or are in a clinic when a researcher happens to be there).

The advantage of convenience sampling is that it is easy to carry out. The weakness is that the findings may not represent the group you are trying to study.

cumulative cases: The total number of cases of a disease reported or diagnosed during a specified time. Cumulative cases can include cases in people who have already died.

> *Example:* Assume that 9,000 AIDS cases had been diagnosed in a state from the beginning of the epidemic through the year 2001. Among the 9,000 persons with AIDS, 4,000 had died. The cumulative number of AIDS cases diagnosed in that state through 2001 would be 9,000.

cumulative incidence rate: The total number of cases during a specified time period, among all people at risk for the disease.

A cumulative incidence rate is calculated by dividing cumulative incidence for a specified time period by the population in which cases occurred during the time period. A multiplier is used to convert the resulting fraction to a number (numerator) over a common denominator, often 100,000.

$$\frac{\text{Number of new cases in specified period}}{\text{Population at risk in specified period}} \times 100{,}000$$

Example: Assume that from 1990 through 2001, 19,000 AIDS cases occurred in a state. During the same time 1,900,000 people lived in the state.

$$\text{Cumulative incidence rate} = \frac{19{,}000}{1{,}900{,}000} \times 100{,}000 = 1{,}000 \text{ AIDS cases per } 100{,}000 \text{ persons}$$

estimate: When accurate data are not available, an estimate may be based on the data that are available and an understanding of how they can be generalized to larger populations. In some instances, national or state data may be statistically adjusted to estimate local conditions. Estimates should be accompanied by statistical estimates of error (a confidence interval), which describe the uncertainty associated with the estimate.

Example: The estimated HIV incidence in State X was 2.1% per year (95% CI, 1.4–2.6).

incidence: The number of new cases in a defined population in a certain time period, often 1 year, which can be used to measure disease frequency. It is important to understand the difference between HIV incidence and reported HIV diagnoses. HIV incidence refers to all persons infected with HIV during a specified period of time (usually 1 year). However, new diagnoses include cases in persons who have been infected for longer periods; they do not include cases in persons who were tested anonymously. Because anonymous test results are not included, HIV surveillance data may not represent incident cases.

Example: During the year 2001, a total of 1,100 AIDS cases were diagnosed in a given state. This is the incidence of AIDS for 2001 in that state.

incidence rate: The number of new cases in a specific area during a specific time period among those at risk in the same area and time period.

Incidence rate provides a measure of the effect of illness relative to the size of the population. Incidence rate is calculated by dividing incidence in the specified period by the population in which cases occurred. A multiplier is used to convert the resulting fraction to a number over a common denominator, often 100,000.

$$\frac{\text{Number of new cases in specified period}}{\text{Population at risk in specified period}} \times 100{,}000$$

Example: Assume that during the year 2001, a total of 1,100 AIDS cases were diagnosed in a given state. This is the incidence of AIDS for 2001 in that state. The population in the state was 2,200,000 in 2001.

$$\text{The incidence rate} = \frac{1{,}100}{2{,}200{,}000} \times 100{,}000 = 50 \text{ per } 100{,}000 \text{ persons in the state}$$

interpretation: The explanation of the meaning of available data. An example is examining a trend, such as the number of HIV cases diagnosed during a 5-year period. Interpreting a trend enables a planning group to assess whether the number of events is increasing or decreasing. However, groups should use caution in interpreting trends that are based on small increases or decreases.

mean: The sum of individual scores in a data set divided by the total number of scores. The mean is what many people refer to as an average.

> *Example:* Assume that people in a given service area in 2001 are the following ages at diagnosis of HIV: 18, 18, 19, 20, 22, 23, 26, 31, 41. The total of the 9 ages = 218 years.
>
> $$\frac{218 \text{ years}}{9} = 24.2 \text{ years}$$

median: The middle value in a data set. Usually, approximately half the values will be higher and half will be lower. The median is useful when a data set contains a few unusually high or unusually low values, which can affect the mean. It is also useful when data are skewed, meaning that most of the values are at one extreme or the other.

> *Example:* Assume the following ages at diagnosis of HIV in the year 2001 data for a given service area: 18, 18, 19, 20, 22, 23, 26, 31, 99. Although the mean age is 30.7, the median age is 22. In this instance, the median age better reflects the central value for age in the population.

no identified risk (NIR): Cases for which epidemiologic follow-up has been conducted, sources of data have been reviewed—which may include an interview with the patient or provider—and no mode of exposure has been identified. Any case that continues to have no reported risk 12 or more months after the report date is considered NIR.

no reported risk (NRR): Cases in which risk information is absent from the initial case report because the information had not been reported by the reporting source, had not been sought, or had not been found by the time the case was reported. Cases may remain NRR until epidemiologic follow-up has been completed and potential risks (exposures) have been identified. If risk has not been identified within 12 months of being reported as NRR, the case may be considered NIR.

percentage: A proportion of the whole, in which the whole is 100.

Example: Assume that 15 of the 60 cases of AIDS in a given year in a state occurred in women.

$$\frac{15}{60} = .25 \times 100 = 25\%$$

prevalence: The total number of cases of a disease in persons not known to have died in a given population at a particular time.

Example: By the end of 2001, if the cumulative number of persons with AIDS in State X is 1,900 and 1,000 have died, then the prevalence of AIDS in State X is 900 (1,900 persons who have ever had a diagnosis of AIDS minus 1,000 who have died).

Prevalence does not indicate how long a person has had a disease and cannot be used to calculate rates of disease. It can provide an estimate of probability that an individual in a population will have a disease at a point in time. For HIV/AIDS surveillance, prevalence refers to persons living with HIV or AIDS regardless of time of infection or diagnosis date. Note the difference between the prevalence of a condition in the population and the prevalence of cases, namely, that a case must be diagnosed according to a definition.

probability sampling: A technique that relies upon random selection to choose individuals from a defined population; all individuals have a known chance of selection. Types of probability samples include simple random sample, systematic random sample, stratified sample, and cluster sample.

proportion: A portion of a complete population or data set, usually expressed as a fraction or percentage of the population or data set.

Example: Assume that 12 of 20 HIV prevention programs in a given service area are school-based programs.

To calculate the proportion as a fraction,

$$\frac{12}{20} = .6 = 6/10 = 3/5$$

To calculate the proportion as a percentage,

$$\frac{12}{20} = .6 \times 100\% = 60\%$$

qualitative data: Information from sources such as narrative behavior studies, focus groups, open-ended interviews, direct observations, ethnographic studies, and documents. Findings from these sources are usually described in terms of common themes and patterns of response rather than numerically or statistically. For the purposes of epidemiologic profiles, qualitative data are useful as supplements to surveillance data to obtain information on risk behaviors and associated factors in specific locales or populations that may not be well represented in routine surveillance data.

quantitative data: Numeric information (e.g., numbers, rates, and percentages).

range: The values of the largest and smallest values in a data set.

> *Example:* Assume the following ages at diagnosis of HIV in the year 2001 in a given service area: 18, 18, 19, 20, 22, 23, 26, 31, and 41. The range is 18–41.

rate: A measure of the frequency of an event or a disease compared to the number of persons at risk for the event or disease. Usually, when rates are being calculated for an epidemiologic profile, the general population, rather than the population potentially exposed to HIV infection by various high-risk behaviors, is used as the denominator. The size of the general population is known from census data, whereas the size of the high-risk population is usually not known.

$$\frac{\text{Number of reported HIV cases occurring during a given period}}{\text{Population at risk during the same period}} \times 100{,}000$$

For ease of comparison, the multiplier (100,000) is used to convert the resulting fraction to number of cases per 100,000 population. The choice of 100,000, although arbitrary, is standard practice.

> *Example:* Assume that 16 cases of HIV were reported in a service area and that 400,000 persons live in the area.

To calculate the rate,

$$\frac{16}{400{,}000} \times 100{,}000 = 4 \text{ per } 100{,}000$$

This means that 4 of every 100,000 persons at risk have been reported.

sample: A group selected from a total population with the expectation that studying this group will provide relevant information about the total population.

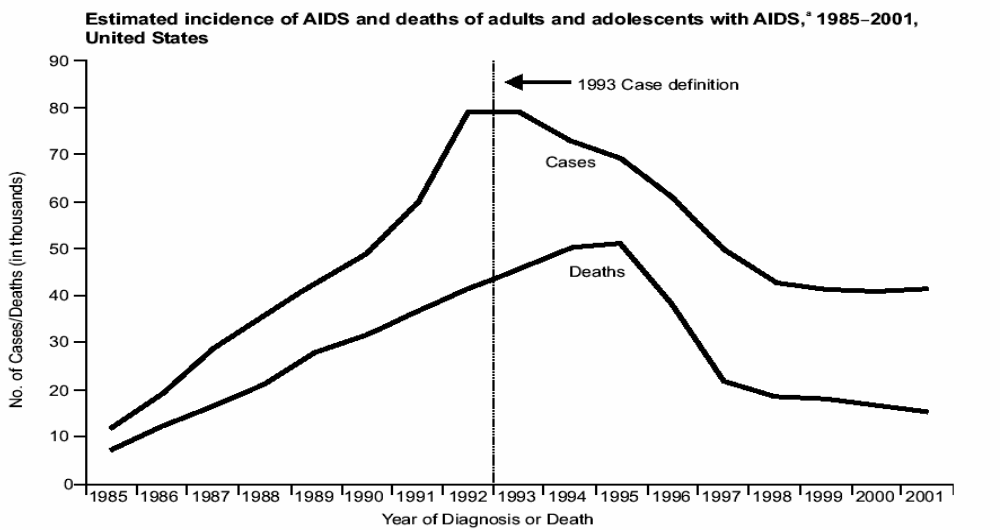
seroprevalence: The number of persons in a defined population who test positive for HIV infection (based on HIV testing of blood specimens). (Seroprevalence is often presented either as a percentage of the total specimens tested or as a rate per 100,000 persons tested.)

stratification: The separation of a sample into subsamples according to predetermined criteria, such as age group, gender, socioeconomic status. Stratification is used to control confounding effects and to detect modifying effects.

trend: A long-term change in frequency, usually an increase or a decrease. A simple linear trend could be described by calculating how much the quantity being measured increased (or decreased) from the beginning value (at the beginning of the period) to the ending value (at the end of the period). The trend could be further described by calculating a time-rate of change in the quantity measured. This is the difference between the beginning and ending values, divided by the number of time units (e.g., years) for which the trend is measured. This calculation yields the amount of increase (or decrease) per time unit. Another key factor is the statistical significance of the trend, which could be a problem if the annual values fluctuate widely from year to year, which would be likely for small numbers.

Trends can be illustrated graphically, by plotting the number of events by time, as shown in Figure 2-2.

Figure 2-2
Example of trend graph



Source. CDC.

[a]Adjusted for reporting delays.

Introduction to Analysis and Interpretation

Collecting and presenting HIV/AIDS data are only part of the task. To be useful to planning groups and others, the data must be analyzed and interpreted.

Analysis is the application of logic in order to understand and find meaning in the data. It involves identifying consistent patterns and summarizing the relevant details.

The purposes of analysis in an HIV/AIDS epidemiologic profile are to:

- identify populations that are infected with HIV and describe their key characteristics
- understand the trends and the impact of HIV/AIDS in a service area
- identify groups or populations at risk of acquiring or transmitting HIV and identify their prevention needs
- identify emerging populations and their needs

The following are a few general guidelines for analyzing and interpreting data for the HIV/AIDS epidemiologic profile.

Descriptive analysis

Descriptive analysis is concerned with organizing and summarizing health-related data according to time, place, and person. An example of descriptive analysis might be "The exposure category for 44% of men reported with AIDS in the United States in 1999 was male-to-male sex."

To carry out an effective descriptive analysis, become familiar with the data before applying analytic techniques. This initial examination should progress to summarizing the data with descriptive statistics, such as frequencies and percentages, in a table to explain the distribution of the HIV/AIDS epidemic in your service area.

As you analyze and interpret your data, keep the following cautions in mind:

- Be aware of the strengths and limitations of the data source. For example,
  - AIDS data do not include those who have been infected most recently.
  - Not all areas report HIV data.
  - EMA service areas may have dissimilar HIV reporting systems (e.g., EMA geographic boundaries cross state lines of 2 states that have different HIV reporting requirements).

- Surveillance data reflect where a person lived when the diagnosis of HIV or AIDS was made, which may or may not be where the person currently lives.

- Confidentiality of public health data is a special concern when dealing with small numbers of cases because of the potential that a person can be identified.

- Interpret surrogate or proxy data with caution (e.g., using STD data as a marker for HIV exposure or infection).

- Concerns about lack of reliability mean that you should be careful about overinterpreting large percent changes (increases or decreases) based on small numbers.

  Example: You observe a 200% increase in cases in one group versus a 5% increase in another. However, the 200% increase represents a change from 2 cases in 1999 to 6 cases in 2000; the 5% increase represents a change from 1,000 cases to 1,050 cases. This is an absolute difference of 4 versus an absolute difference of 50. The 200% increase could be due to fluctuations typical of small numbers. Or perhaps 2 of the 6 cases in 2000 should have been reported in 1999. If so, then 4 cases would have been diagnosed in each of the 2 years, and there would have been no increase.

Triangulation

Triangulation, or data synthesis, refers to comparing and contrasting the results of different kinds of research that address the same topic. For example, you may want to see whether the same methods lead to similar findings (e.g., do biologic data and surveys indicate similar patterns in HIV prevalence?). The similarity of results from very different data is referred to as *convergent validity*.

When research findings from different studies or different methods are robust (i.e., not very sensitive to departures from assumptions, for example, that the data are normally distributed), profile writers have an empirical basis for making stronger statements about the validity of their findings and conclusions. If HIV prevalence data, AIDS prevalence data, STD prevalence data, and surveys of risk behavior show consistent evidence of higher HIV risk in a population, then you can be much more confident in saying that this population should be given a high priority for prevention services than you could be if you have only one kind of data. This is why multiple indicators of risk that address different aspects of HIV risk and use different methods are useful. Besides providing another index of validity, convergent findings may be clearer and more convincing to planning group members, service providers, policymakers, and others.

By the same token, different data may suggest contradictory findings. When this occurs, it is important for epidemiologists to account for the reasons that different studies have arrived at different conclusions. This process can be important in terms of identifying problems in data collection or previously undetected differences within populations. Surveys collected under poorly monitored conditions may yield results that are different from those in which the population is well characterized and sampling procedures are rigorously followed. Recent data such as HIV case reporting may reveal emerging populations at risk that are not evident from AIDS case reporting. Survey studies of drug use may suggest that methamphetamine injection may be increasing in a particular population, but no change has yet been seen in HIV prevalence. This may mean that HIV infection has not yet entered the population, which would suggest the need to look specifically at risk practices of this population that have protected them from HIV infection and also look at "mixing patterns" (persons with whom they share drugs and persons with whom they have sex). The use of rapid assessment in such a population

could lead to a better understanding of the epidemiology of a potential new epidemic. Divergent patterns like these also may suggest areas that should be investigated during the prevention needs assessment.

The simplest way to triangulate, or synthesize, data in the profile is to look at the main demographic categories and see how they differ according to data sources. Hence, you may want to look at similarities or differences across data sources by race/ethnicity, gender, geographic area, and age group. Summary statements based on triangulation of the data will be helpful to profile users in understanding how to integrate the large number of tables, figures, and findings that are typically included in an epidemiologic profile.

Where to Get Technical Assistance

If a state or local HIV/AIDS surveillance coordinator is not preparing the profile or is not part of the team preparing the profile, you may want to seek that person's assistance. The HIV/AIDS surveillance coordinator will be able to provide technical assistance in acquiring, analyzing, and interpreting core HIV/AIDS surveillance data. Also consult with the HIV prevention or care programs in the health department about remaining questions or needs for technical assistance.

If your technical needs cannot be addressed at the local level, technical assistance is available both from HRSA and CDC.

For CARE Act grant requirements

For technical assistance needs that relate directly to CARE Act grant requirements, contact HRSA. All technical assistance requests must go through the project officer:

HIV/AIDS Bureau
Division of Service Systems
Health Resources and Services Administration
5600 Fishers Lane, Room 7A-07
Rockville, MD 20857
301-443-9086

For prevention grant requirements

For technical assistance needs that relate to prevention cooperative agreement requirements, contact the Prevention Program Branch at CDC:

Chief, Prevention Program Branch
Division of HIV/AIDS Prevention
National Center for HIV, STD, and TB Prevention
Centers for Disease Control and Prevention
Mailstop E-58
1600 Clifton Road, NE
Atlanta, GA 30333
404-639-5230

For developing epidemiologic profiles for HIV prevention community planning

For technical assistance needs that relate to developing epidemiologic profiles for HIV prevention community planning, contact the HIV Incidence and Case Surveillance Branch at CDC:

Chief, HIV Incidence and Case Surveillance Branch
Division of HIV/AIDS Prevention
National Center for HIV, STD, and TB Prevention
Centers for Disease Control and Prevention
Mailstop E-47
1600 Clifton Road, NE
Atlanta, GA 30333
404-639-2050

Other sources

Other sources of technical assistance include researchers at local universities (such as those at schools of public health, programs in community health and education, and social science departments) and organizational entities, such as the American Psychological Association's Behavioral and Social Scientist Volunteers Program.