Chapter 17: Analysis of Surveillance Data

Melinda Wharton, MD, MPH; Sandra W. Roush, MT, MPH; Siiri Bennett, MD

I. Background

Ongoing analysis of surveillance data is important for detecting outbreaks and unexpected increases and decreases in disease occurrence, monitoring disease trends, and evaluating the effectiveness of disease control programs and policies. This information is also needed to determine the most appropriate and efficient allocation of public health resources and personnel.

Analyses should be performed at regular intervals to identify changes in disease reporting. These analyses can be performed using standard approaches (e.g., tabulating reports manually and filling in a summary data sheet, or running a standard computer program to generate a summary report). Findings of these analyses should be reviewed regularly, and provided as feedback to medical providers and others in the community who are asked to report cases. Additional special analyses are often needed to answer specific questions that arise;[1] these analyses may require additional customized approaches beyond what are routinely performed.

II. What computers can do

In many health departments, surveillance data are routinely entered into a computerized database program. Use of computers can greatly facilitate analysis of surveillance data, especially for large and complex datasets.

Analyses can be done using any one of a number of database and statistical programs. In many health departments, Epi-Info, a public domain word processing, database, and statistics package for IBM-compatible computers, is used for data entry, analysis, and generating reports.[2] Mapping capability is an important adjunct to data analysis. Although mapping of public health surveillance data may be performed using a variety of software packages, some are quite expensive and complex to use. Many health departments use Epi-Map, a public domain mapping program.[3]

Contact your state health department for information about recommended software and to identify support for setting up a surveillance database at your local health department. The state health department may also give assistance in setting up useful analyses and reports that can be generated as needed.

III. What computers cannot do

Although computers can greatly facilitate analysis of surveillance data, especially if the dataset is large and the analyses complex, most analyses of surveillance data are simple (see examples included in this chapter) and can be readily performed with the assistance of an inexpensive pocket calculator. Likewise, data can be graphically presented with only graph paper, a ruler, and colored pencils. **There is nothing that needs to be done routinely that requires a computer, and in fact, there are many things that must be done routinely that the computer cannot do.**

Computers cannot contact physicians and laboratories and obtain missing information. Computers cannot interpret laboratory tests or make judgments about epidemiologic linkage. Computers cannot make judgments about duplicate records or identify and correct mistakes in data entry. Computers cannot even tell you if there is an outbreak in progress; they can provide information that may help you make a decision, but even a sophisticated trend analysis is no substitute for familiarity with the people and the disease patterns in your community and with your reporting system.

The mistake most commonly made in analysis and use of public health surveillance data is not related to statistical testing, improper presentation of data, or failure to perform complex multivariate analyses; the most common mistake is not looking at the data. Computer hardware and software can facilitate the epidemiologist’s task, but are no substitute for looking, thinking, discussing, and taking action.

IV. Analyzing surveillance data

Analyses of surveillance data begin with characterizing the pattern of disease reports by person, place, and time. Compare patterns of disease reports at different times (e.g., the number of mumps cases reported in 2001 compared to the number of mumps cases in 2000); in different places (e.g., the number of pertussis cases reported in one district compared with the number of pertussis cases in another district); and among different populations (e.g., the number of measles cases reported among infants, pre-school age children, school age children, adolescents, and adults). Vaccination status of cases should also be examined; if there is disease transmission in the community, lack of vaccination is likely to be a factor most strongly associated with illness. Analyses looking at delays in reporting, completeness of reporting of critical variables, and applying case definition criteria also are useful in evaluating the quality of case investigation and reporting and should be undertaken regularly.

Missing or inaccurate data may limit the usefulness of any analysis. Erroneous or incomplete data cannot be corrected through statistical procedures.

V. Suggested analyses: description of cases by person, place and time

The following analyses should be regularly performed as part of routine analysis of surveillance data. Additional analyses may be needed under special circumstances; the state health department can provide you additional guidance in routine and special analyses of surveillance data. The interpretation and possible action steps are only examples, to indicate some of the information that may be gained from the analysis.

By person

Describe the persons (cases) with vaccine-preventable diseases who were identified by your surveillance system. Attributes of the cases include age group, sex, and race or ethnicity.

It may be appropriate to divide age groups based on recommended ages of vaccine administration (e.g., separating those too young to be vaccinated from those eligible for vaccination), as well as the age distribution of reported cases. Age groups should span a narrower age range for ages in which disease incidence is highest and a broader age range in which disease incidence is lower.

Example 1

Pertussis Cases by Age Group

AGEGRP	FREQ	PERCENT	CUM.
< 6 MO	57	36.1%	36.1%
6–12 MO	41	25.9%	62.0%
13–18 MO	6	3.8%	65.8%
19–23 MO	6	3.8%	69.6%
2–5 YR	18	11.4%	81.0%
6–9 YR	17	10.8%	91.8%
10 YR+	12	7.6%	99.4%
AGE UNK	1	0.6%	100.0%
Total	158	100.0%	

Interpretation

Pertussis cases were clustered among infants, with more than 60% of reported cases among those 12 months of age and younger (**Figure 1**). The occurrence of pertussis among infants < 6 months of age is extremely worrisome, because these children are too young to have received 3 doses of pertussis vaccine. Note that it is difficult to draw any conclusions about disease incidence from these data; although these age group divisions are logical for analysis of pertussis data, presentation of data in such unequal age groups may obscure important differences in disease incidence. **Figure 2** shows the incidence of pertussis, by age group.

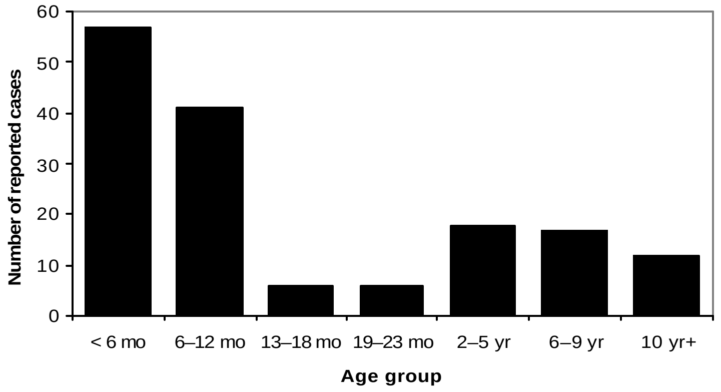

Figure 1: Pertussis Cases by Age Group, 1995

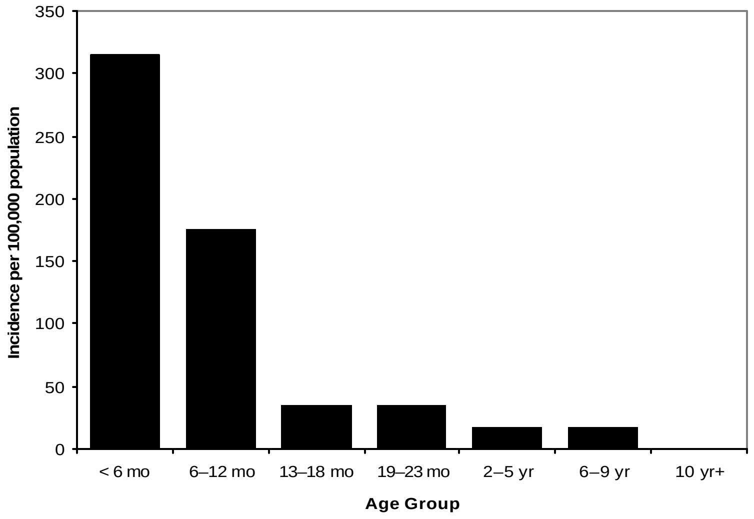

Figure 2: Pertussis Incidence by Age Group, 1995

Example 2

Rubella Cases by Sex

SEX	Freq	Percent	Cum.
FEMALE	27	69.3%	69.3%
MALE	12	30.7%	100.0%
Total	39	100.0%	

Interpretation

Of the 39 cases of rubella, more than two-thirds were among females. Assuming the population under surveillance includes approximately equal numbers of males and females, the female predominance among cases may reflect a real difference in disease incidence among females, due to differences in susceptibility or exposure or differences in ascertainment, e.g., due to concerns about rubella in women of childbearing age. The occurrence of rubella among women of childbearing age is of great concern because of the risk of congenital rubella syndrome (CRS) among infants born to women infected with rubella during the first trimester of pregnancy. Because many cases of rubella are asymptomatic or mild, there likely were many more cases than were reported. Subsequent surveillance for CRS in this community is essential.

Next steps

Look at cases among women by age group, to identify women of childbearing age.

Example 3

Pertussis Cases by Hispanic Ethnicity

ETHNIC	Freq	Percent	Cum.
HISPANIC	32	20.3%	20.3%
NOT HISP	77	48.7%	69.0%
UNKNOWN	49	31.0%	100.0%
Total	158	100.0%	

Interpretation

Of the 158 cases of pertussis, one-fifth occurred among persons of Hispanic ethnicity and almost half were among non-Hispanics. However, ethnicity was unknown for almost one-third of cases, suggesting incomplete case investigation.

Even if the data were complete, we would need more information to know how to interpret these proportions. What proportion of the population under surveillance is of Hispanic ethnicity? Do the data suggest a disproportionate burden of disease in one group? Disproportionate reported disease burden could result from low rates of vaccine coverage, increased disease incidence in certain neighborhoods or communities, or different levels of reporting, due to differences in access to medical care, diagnostic testing, or differences in provider reporting practices (public clinics may be more likely to report cases than private physicians, for example).

Next steps

Obtain missing data, if possible; calculate incidence rates by ethnicity; look for geographic clustering.

By place

Describe the persons (cases) with vaccine-preventable diseases detected by your surveillance system by geographic location. Location may be defined as the place where the case was first reported, place of residence of the case, or place of hospitalization. Location may be a city, county, or health district.

Example 4

Pertussis Cases by Health District

DISTRICT	Freq	Percent	Cum.
1	10	6.3%	6.3%
2	12	7.6%	13.9%
3	2	1.3%	15.2%
4	67	42.4%	57.6%
5	10	6.3%	63.9%
6	57	36.1%	100.0%
Total	158	100.0%	

Interpretation

The data demonstrate marked clustering of reported pertussis cases in District 4 and District 6 (**Figure 3**). The number of reported cases in those two districts is of concern regardless of the distribution of population in this area, but comparing disease occurrence in the six districts requires knowing the district population and calculating rates. The differences in reported cases by district in this example may be due to differences in population, disease incidence, or case ascertainment.

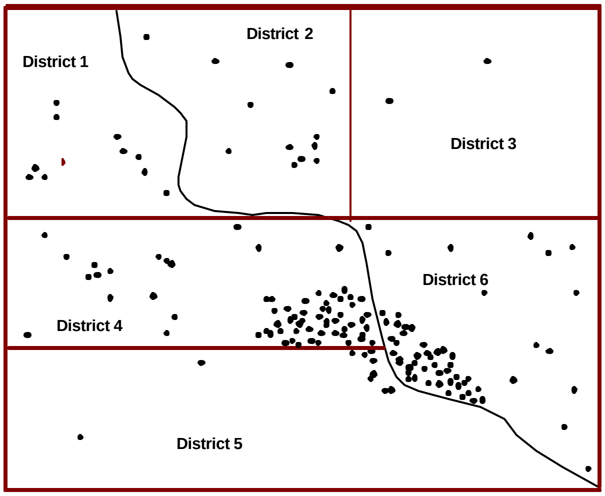

Figure 3: Pertussis Cases by Health District

By time

Describe the distribution of cases over time. Look for changes in the number of cases over time. Time intervals may be in years, months, weeks, or other unit of time. Date may be defined as date of onset of illness, date of diagnosis, or date of report to health department. Analysis of date of onset gives the most accurate representation of disease occurrence. Distribution of cases over time is most clearly presented as a graph with time on the x-axis and number of cases on the y-axis.

Compare the number of cases occurring in a current time period with the number reported during the same time period in each of the last 5 years. Compare the cumulative number of cases year-to-date with the cumulative number of cases year-to-date of previous years.

Example 5

Reported Pertussis Cases, 1998, by Month of Onset

MONTH	Freq	Percent	Cum.
A OCT97	3	1.9%	1.9%
B NOV97	1	0.6%	2.5%
C DEC97	1	0.6%	3.2%
D JAN	2	1.3%	4.4%
E FEB	3	1.9%	6.3%
F MAR	2	1.3%	7.6%
G APR	9	5.7%	13.3%
H MAY	13	8.2%	21.5%
I JUN	38	24.0%	45.6%
J JUL	35	22.2%	67.7%
K AUG	18	11.4%	79.1%
L SEP	14	8.9%	88.0%
M OCT	8	5.1%	93.0%
N NOV	6	3.8%	96.8%
O DEC	5	3.2%	100.0%
Total	158	100.0%	

Interpretation

There is marked temporal clustering beyond the expected seasonal increase in pertussis, suggesting that a large outbreak occurred during the summer of 1998. Note that in this dataset of cases reported during 1998 there are a number of cases with onset during 1997. Reports in 1999 should be reviewed to look for cases with onset in 1998, because of apparent delays in reporting. The magnitude of these delays can be monitored by tracking the interval between onset of disease and initial report. **Figure 4** demonstrates the reported cases of pertussis in 1998 by month of onset, deleting the cases with onset in 1997, and including the few additional cases reported in 1999, but with onset in the latter months of 1998.

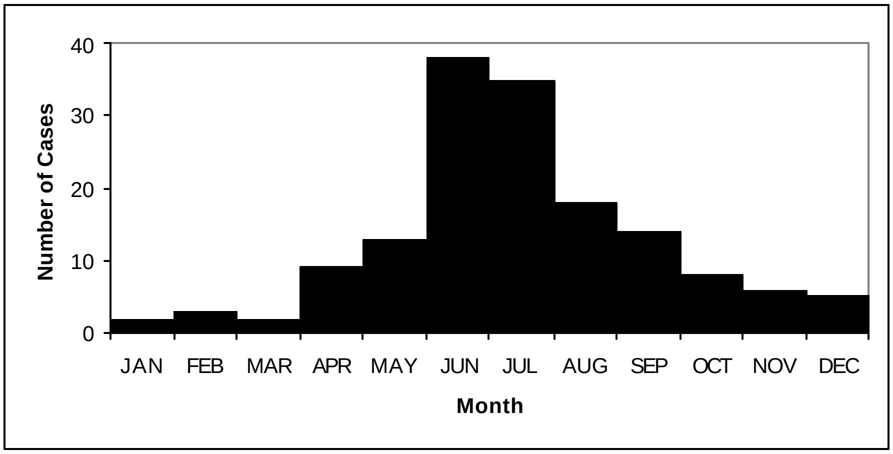

Figure 4: Reported Pertussis Cases by Month of Onset (1998)

Example 6

Pertussis Cases by Age Group and DTaP/DTP Doses

	DTP DOSES							
AGEGRP3	0	1	2	3	4	5	9	Total
A 0-2 MONTHS	7	1	0	0	0	0	0	8
B 3-4 MONTHS	7	6	1	0	0	0	0	14
C 5-6 MONTHS	2	6	1	0	0	0	1	10
D 7-18 MONTHS	5	6	9	10	4	0	0	34
E 19 MO-6 YRS	1	2	4	8	10	2	0	27
F 7 YEARS +	1	0	1	1	0	10	9	22
Total	23	21	16	19	14	12	10	115

Interpretation

Many of the children reported with pertussis were undervaccinated; cases among children < 6 months of age are not preventable by vaccination, because they are too young to have received 3 doses of pertussis vaccine, the minimum number of doses needed to confer protection. In order to be up-to-date, children 3–4 months of age should have received at least 1 dose, 5–6 months at least 2 doses, 7–18 months at least 3 doses, 19 months to 3 years of age 4 doses, and those ≥ 7 years of age should have received 5 doses. Many of these cases were among children who were not age-appropriately immunized, suggesting that there may be a wider problem with immunization coverage among young children in this community. It is often extremely difficult to verify vaccination of adults, which may account for the high proportion of cases with unknown vaccination status among cases ≥ 7 years of age.

Example 7

Pertussis Cases by Case Definition

CATEGORY	Freq	Percent	Cum.
A CX + COUGH	57	36.2%	36.2%
B COUGH = 14D +DFA	18	11.5%	47.7%
C COUGH = 14D	46	29.2%	76.9%
D DFA + COUGH < 14D	10	6.4%	83.3%
E LINKED CX + CASE	1	0.0%	83.3%
F INSUFF INFO	26	16.7%	100.0%
Total	158	100.0%	

Interpretation

Some reported cases were based on positive results by the direct fluorescent antibody (DFA) test. Because of both false-positives and false-negatives, this test should not be relied on for confirmation for purposes of national reporting. DFA positive cases with cough of < 14 days duration are particularly suspect. In areas using DFA to evaluate suspected pertussis cases, care should be taken to monitor the proportion of cases with positive DFA and negative cultures. If this proportion increases significantly, it may reflect problems with interpretation of the DFA test (false-positives).[4] Note that there was insufficient information to classify 26 cases; this likely reflects lack of information sufficient to classify duration of cough as < 14 days or ≥ 14 days. This information is essential and should be obtained in the course of case investigation of every pertussis case.

References

1. Chen RT, Orenstein WA. Epidemiologic methods in immunization programs. *Epidemiol Rev* 1996; 18:99-117.

2. Dean AG, Dean JA, Coulombier D, et al. EPI INFO, Version 6: A word processing database and statistics program for epidemiology on microcomputers. Centers for Disease Control and Prevention: Atlanta, GA, 1994.

3. Dean JA, Burton AH, Dean AG, et al. EPI MAP: A mapping program for IGM-compatible microcomputers. Centers for Disease Control and Prevention: Atlanta, GA, 1993.

4. Ewanowich CA, Chui LWL, Paranchych MG, et al. Major outbreak of pertussis in Northern Alberta, Canada: Analysis of discrepant direct fluorescent antibody and culture results by using polymerase chain reaction methodology. *J Clin Microbiol* 1993; 31:1715-25.