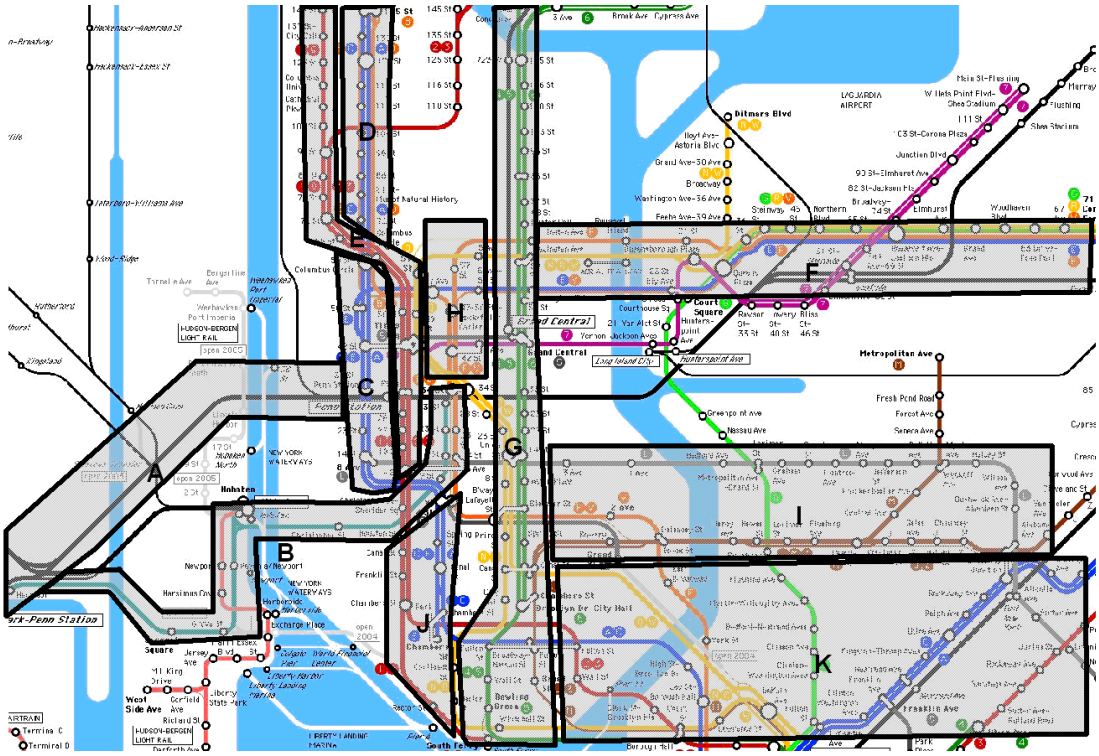
Computing with Epi Info
Part I
Introduction to Public Health Computing



Andrew G. Dean, MD, MPH
www.EpiInformatics.com

Revised March 2009

Table of Contents

Course Objectives

Course Title

Epi Info Course I (Introduction to Public Health Computing with Epi Info)

Course Description

Epi Info™ is a public domain software package freely distributed worldwide. Using Epi Info for Microsoft Windows®, students will develop and computerize questionnaires for a bioterrorism investigation, analyze data from acute and chronic disease studies to produce epidemiologic statistics, maps, and graphs, and construct a menu for a permanent information system.

Learning Objectives

At the conclusion of the course, students will be able to:

a. Define and explain computer and database concepts needed to set up public health information systems
b. Develop and computerize a questionnaire for an epidemiologic field investigation or survey
c. Enter and analyze data from both acute and chronic disease studies, producing epidemiologic statistics
d. Develop graphs and maps to explore time and place, and understand Geographic Information Systems
e. Access and use statistical calculators on the Internet
f. Understand how to incorporate questionnaires, data analysis programs, and documentation into a menu as the basis of a permanent information system

Schedule

Day 1

Time	Session	What Happens	Materials
Launching the Course			
	Introductions	We meet each other and go over the course objectives and materials.	Course notebook and schedule
	Review of data concepts	Introduction to database and Epi Info concepts.	Epi Info Intro.PPT on CDROM
	Epi Info Demo	A whirlwind tour of Epi Info to provide perspective on the software and what will and will not be covered in this course.	Watch the magic screen, and/or "Introduction to Epi Info for Windows" DOC file.
Designing a Study and a Questionnaire and Entering Data			
	Making a questionnaire	An anthrax attack on New York City suggests that we have an *immediate* need to develop a questionnaire for an epidemiologic study of New Jersey cases and controls. We jointly design a questionnaire and then construct the questionnaire view in Epi Info.	Anthrax in New York Exercise
	Entering data	We enter a few records as a pilot study	Anthrax in New York Exercise
	Adding Check Code	We add some conditions for entering cleaner data and also examine some fancier questionnaires	Anthrax in New York Exercise
	Doing the Case-Control Study	We organize to use copies of the same View and enter the data from the cases and controls	Anthrax in New York Exercise
Analyzing Data			
Time	Session		Materials
	Analysis	READing a data table. LISTing the data. Doing FREQuencies and TABLES. READing an Excel table. Merging tables from other team members into a single data table for analysis.	Anthrax in New York Exercise
	Seeing results	Analyzing the case control study, using FREQ, TABLES, MEANS, and MATCH. Saving, editing, and running a program.	Anthrax in New York Exercise

Day 2

Graphing and Mapping			
	Graphing	Analyzing Disease Reports	Analysis of Reportable Disease Surveillance Data
	Mapping	What if John Snow had had a Global Positioning System (GPS) instrument in London of 1854?	Demo
	Mapping	Showing the data from the Anthrax study on a downloaded image of New York	Anthrax in New York Exercise
Statistical Calculators			
	Statcalc	Convenient DOS-based epidemiologic calculator	Demo
	OpenEpi	Open Source Internet calculator with links to other programs, statistical tutorials, and searching	Demo www.openepi.com or local copy
Solving Other Epidemiologic Problems			
	An outbreak of serious Hospital Infections	Case control methods and microbiology uncover the source of an epidemic of Rhodococcus after cardiac surgery in a hospital. You use OpenEpi for analysis of summary data, sample size calculations, and random number generation. You analyze the descriptive study and the matched case-control study.	Postoperative Surgical Infections in a Large Hospital www.openepi.com
	General Analytic Methods	Analysis of a mortality database, and techniques for approaching an unfamiliar dataset	Mortality exercise
Constructing a Menu System			
	The Epi Info Menu	Demonstration of Epi Info menu programming as a tool for constructing systems	Epi Info menu and MakeMenu exercise
Finishing Up			
	Questions and suggestions	Discussion How will you use Epi Info? Suggestions for future versions.	
	Evaluation		Evaluation form
Notes			
	Yes, there will be snack breaks..But, please no food or liquids in the computer room. The owner of a fried motherboard or keyboard is not a happy person. Count on lunch around noon, depending on the flow of events.		

Anthrax Bioterrorism

Using Epi Info During an Epidemic Investigation

The Emergency

In your position as a New Jersey epidemiologist, you receive a call from the New York City Office of Emergency Preparedness. A case of anthrax has been diagnosed in a subway worker and several of his colleagues have respiratory illnesses consistent with the same diagnosis. They ask you to canvas New Jersey medical facilities in search of possible cases and to conduct a case-control study to determine the possible location and time of infection if cases are found.

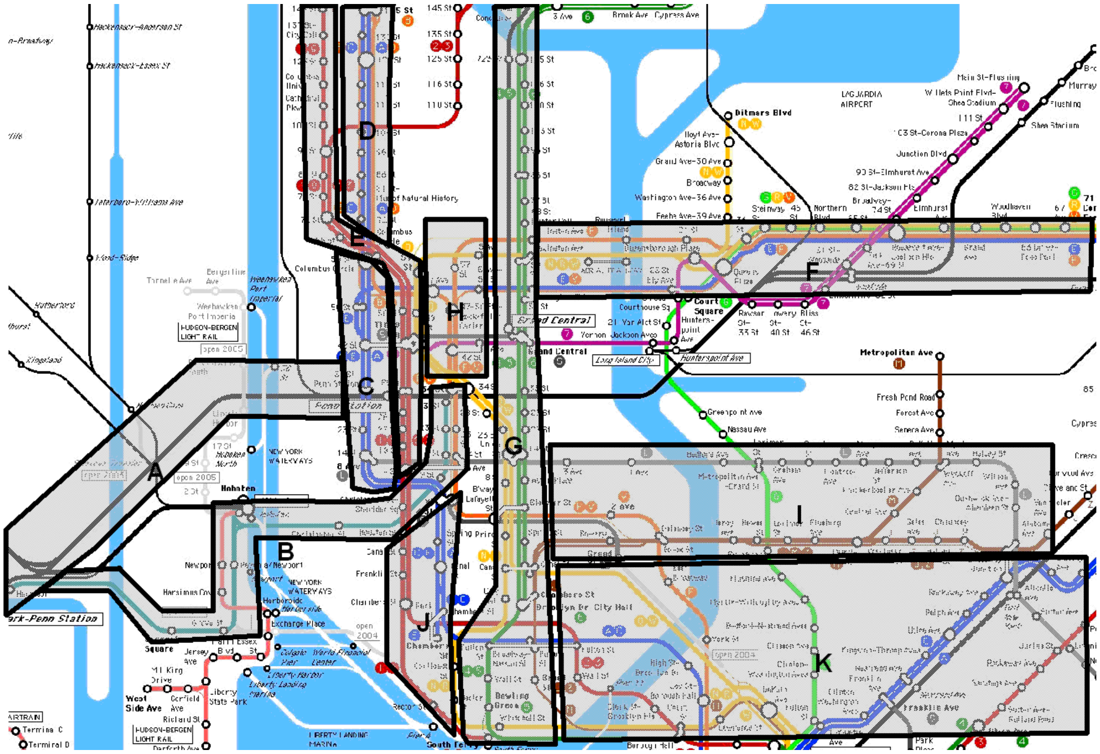
You immediately ask colleagues to begin calling medical facilities in a case-finding effort while you compose a suitable questionnaire for the investigation. Although you are familiar with anthrax, you use www.google.com to search for a recent reference data sheet and find one on the website of the Centers for Disease Prevention and Control (CDC) (Appendix 1).

Planning

The Epidemiologic Strategy

A few minutes later, more than 50 possible cases have been located in New Jersey hospitals, and you decide to do a case-control study. The major question concerns the geographic location of infection, with "somewhere in the New York City subway system" as the working hypothesis. Since persons in New Jersey have only limited contact with the New York subway, New Jersey is a good place to ascertain exposures of cases within New York. Pending confirmation, it is assumed that most of the cases had visited New York, and your first series of interviews establishes this to be true.

You download a map of the subway system and quickly sketch zones of interest on its surface, calling them "A" through "K". Interviewers will use the maps to identify one or more zones visited by the cases and controls within the incubation period prior to onset.



You and your colleagues choose the following elements for the questionnaire:

Person

- Case ID Number
- Interviewer
- Demographic information (Age, sex, zip code of residence)
- Occupation (Office worker, Police, Subway, Taxi, Other)
- Diagnosis (Inhalational, cutaneous anthrax, not anthrax)
- Case or control
- Smoked within the past 5 years (possible but unproven risk factor)

Place

- Visits to NYC within the incubation period?
- Zones visited within incubation period prior to onset

Time

- Date of Onset of case
- Last information date

It is tempting to include more information, but, since our goal here is to learn how to computerize a questionnaire rather than to construct the perfect investigation, let's keep it short and omit such details as antibiotic treatment or prophylaxis, previous immunization, etc. Certainly more detail about each of the zone visits would be important. Although we don't know yet, it is possible that there will be hundreds or thousands of cases.

Collecting Data

Calls to major hospitals in the area have disclosed many admissions with acute febrile disease consistent with anthrax. You convene your epidemiologic group to design the case-control study and decide how to select controls. The consensus is that randomly selected telephone controls of the same sex will be suitable. As interviews proceed, it appears that ALL of the cases had visited NYC within the incubation period. The team adds to the control-selection criteria, "Visited NYC at least once during the case incubation period, if the case did so." Why? What alternative methods might be used to select controls?

The Data Management (Informatics) Strategy

As informatics resource person for the investigation, you work with the epidemiologists to develop a plan for data management. Until preliminary results come back from interviews, some details of the plan are left unclear. Meanwhile, however, you plan to enter and analyze hundreds of case and control interviews as soon as they become available, and to produce results within a few hours. You have a team of data entry people (the present class, if you are in a classroom setting) and a laptop or desktop computer for each. The latest version of Epi Info has been installed on each computer and tested briefly by running each of the programs from the menu. You verify that each computer has at least 10 megabytes of free hard disk space after installation. Updates of Epi Info are available at www.cdc.gov/epiinfo if you do not have the latest version. Check the website for installation instructions, especially if you have Microsoft Vista.

For security and other reasons you decide not to enter the data in a Local Area Network, but to enter data on local computers and append data tables to a master database after copying to removable media. Removable media might be USB memory sticks, floppy diskettes, or writeable CDROMs. Whatever the choice, be sure to check that both source and destination computers can handle the medium chosen. The same medium will be used to back up and transport data.

To avoid complexity, you will not create related tables or grids in your first questionnaire. Each participating computer is given a number (Station 1…x) using sticky labels. You assemble floppy disks, markers, notebooks, and your own thoughts so that a systematic approach can be made to what might otherwise be a confused mess of lost or redundant data.

This is the time to write down your plan for data entry, validation, and analysis—the informatics strategy. Please sketch your own version before looking at the sample strategy in Appendix 2. Every investigation is different, and there are many right answers, (but even more wrong answers!!). One of the most difficult parts is the naming of files and tables, and identifying records, geographic areas, data entry sites, etc. It deserves as much attention as naming your children.

Implementation

Constructing the Questionnaire View in MakeView

Sketch your questionnaire rapidly on paper and then construct it in Epi Info, using the MakeView program. Use your Name or Initials for the MDB or Project name and name the View inside the project CASECONTROL. Instructions for making a View are given in detail in the "Introduction to Epi Info" supplied with the course materials or downloadable from www.epiinformatics.com .

Entering Data—Pilot Test

Interviewing teams come back from the hospitals and from the telephones with armloads of completed questionnaires. You select ENTER DATA in the FILE menu of MakeView, and enter a few to try out the computerized data entry system. Try this with your own questionnaire, inventing plausible pilot data.

Refining the Data Entry Process with Field Properties and Checkcode

So that the entire class will be working with the same questionnaire, use the File menu in MakeView to Open the Project called Station.MDB and the View AlmostNoCheckCode. Follow the steps below and test the results by entering data. A working example is provided in the View called CaseControl in the same MDB.

There are a several things you can do to improve the ease and accuracy of the data entry process.

1. For the Case ID number, check "Required" in the field dialog.
2. Using the blue button labeled "Program" on the left side of the MakeView screen, go to the check code view. Choose Page 1 as the Field Where the Action Will Occur, and note the checkcode we have used to set the Case ID automatically to UniqueKey, which is maintained by Epi Info. The EndBefore statement assures that this happens before the user enters Page 1.
3. Make the Interviewer field required and have it repeat the value in the preceding record unless changed
4. Insert Legal values for Diagnosis, ZipCodeofResidence, Sex, CurrentCondition, and other fields where the choices are limited. You may have done so already.
5. Set the Last Information Date automatically to SystemDate--the current system date--after page 1 is entered or edited
6. If the form is for a Control (Case=(-)) then, after the Case field, skip (GOTO) to LastInfoDate. This requires constructing Checkcode in the Case field with an IF statement, the condition, Case=(-), and the consequence GOTO LASTINFODATE. No else clause is needed, since the cursor will automatically go on to the next field if Case=(+).
7. Calculate age from DateofBirth and the current system date (hint: use the YEARS function with Systemdate). When do you want this action to occur?

Entering Data—Production Style

At this point, it is time to make copies of your MDB for each data entry person. If you plan to merge the data together into a single data table, it is important that the structure of all the copies is the same. Hence, the copies should be made after the pilot test and final editing of the View.

Let's assume that you have carefully made the copies of the MDB and that each data entry station now has STATION.MDB, containing a View with Check Code called CaseControl. We have provided this View so that the actual copying does not hold up class progress.

To begin entering data, OPEN the View CaseControl in the Enter program in the project STATION.MDB. You should be informed that it has no data table, and when you acknowledge this message, an opportunity is offered to make a data table and to specify the beginning value of UniqueKey. To make sure that each data entry station has different keys, develop a convention such as "Data station number times 1000" to produce non overlapping ranges of UniqueKeys. (*This is important for later use when your dataset might contain grids or related tables*.).

Enter several practice records, timing the entries so that you can calculate how long it will take to enter 600 records. From this you should be able to estimate how many data entry stations you need to do single or double data entry in a period of, say, three hours—the longest that CDC, the Press, and the Governor care to wait for results to be available.

Data Validation

With a simple questionnaire, the accuracy of data entry can be checked and improved by having one person read the screen and the other check the entries against the original paper form. The CaseID field is used to match the form to the record on the screen.

Another method of validation is by having the data entered twice in different MDBs (Projects) by different people. Then the Data Compare utility in Epi Info is used to find and correct differences in the two sets of entries. The Script called CompareEntries.TXT can be opened in Data Compare to compare two data tables called Entry1 and Entry2. ENTER1 and ENTER2 have about 100 records each, with some mistakes so that you can see how Data Compare works. Be sure to specify that CASEID is the match key.

Appending Data from More than One Station

Suppose that you have entered as many records as possible in your copy of STATION.MDB:CaseControl. It is now time to copy it to a flash memory stick, a floppy diskette, or other removable medium from whence it can be appended to a master dataset. First use Windows explorer or My Computer to copy STATION.MDB to the removable medium. Label the copy carefully with your Station number.

To create your own Master.MDB and CaseControl dataset, use Analysis to READ the copy on the storage device by clicking CHANGE PROJECT and then navigating to A:\STATION.MDB:CaseControl. After choosing the CaseControl view, click OK. There is only one more step—appending the data to the Master.mdb:CaseControl table. Use the WRITE command and specify the file as MASTER.MDB in a suitable directory and the table as CaseControl. Don't worry that neither exists yet; Analysis will create them. Since you want to APPEND to an existing dataset, choose APPEND rather than REPLACE and click OK.

Now trade portable storage devices with your neighbor and insert this second device. Since the names are the same, all you have to do is insert the device and RUN the small program in the program editor window. To do so, click on the RUN button above the window. Do so only once to avoid duplicating records. You could accept devices from the whole class and repeat this process until all had been appended to the Master, but you get the idea. READ the MASTER.MDB:CaseControl data table and confirm with LIST that it contains records from both the contributing stations.

To start the process again, you can use REPLACE in the first WRITE command, or, more conservatively, just give a new name to the next batch of appended datasets, such as CaseControl1 or CaseControl2PM. Your data entry shop is functioning as a team, and you are ready for high-volume data entry.

Of course, as soon as the Master dataset is complete, you make a backup copy onto a CDROM (preferably) or other medium. This can be done with the EpiLock program, which can compress (ZIP) the file. It can also encrypt files if they are to be carried elsewhere or sent over a telecommunications network. Of course, remembering and securely transmitting the password to the recipient through another channel is vital to success in decrypting the file.

Analyzing the Data

Now comes the moment we have been preparing for! Open the Analysis program and use the READ command and CHANGE PROJECT button to navigate to the SampleAnthrax.MDB and the table called FinalData. It contains 600 records, about half being cases and the other half controls. It is your job to describe the cases, and by comparing them with the controls, to find the risk factors associated with having Anthrax. You can use the LIST command to see the data, FREQ to ascertain the frequency of variable values in the dataset as a whole, and TABLES XXX CASE to test the association of individual factors with the outcome. XXX represents a candidate risk factor variable. Case="Yes" identifies a Case record and Case="No" means a control interview.

Some questions you might ask include:

1. How many case and how many control interviews are included?
2. How do cases and controls compare with regard to Sex and ResidenceZipCode?
3. Do the MEANS of Ages differ by Case status. Is the differnce significant?
4. Among the Cases, what are the diagnoses?
5. What was the biggest day for case onset?
6. Graph the number of cases by CaseOnsetDate. Is the epidemic over?
7. Graph the number of interviews by DateofLastInfo
8. How do cases and controls compare with regard to visiting zone A? Use the Odds Ratio, confidence intervals, chi square, p-value, and fisher exact p-value. Check each of the other zones for significant differences.
9. Is Smoking a risk factor for disease? For dying from the disease?
10. What conclusions can be drawn from this analysis? What conclusions *cannot* be drawn?

As you work, Analysis is generating a program that can be run again to do the same analysis, perhaps after adding more data. To save the program, click the SAVE button at the top of the program editor. Give it a name, such as AnthraxAnalysis. To SAVE it in the MDB, click OK. To save it as a text file that can be edited in Notepad, click the TEXT FILE button first and then choose a location and name to save it.

Programs can be run from inside Analysis using RUN SAVED PROGRAM, from the Program Editor with the RUN button, and from an external menu as we shall see later. You can also use the OPEN button in the program editor at the bottom of the screen and then choose the Text file or other saved program to load into the program editor, running it with the RUN button.

Maps

The map at the beginning of the exercise instructions was drawn quickly in Epi Map, using the New York Subway map image available from the Internet. The zones were arbitrarily drawn and labeled A through K to provide a simple method of recording visits to New York City areas. Much of the information has to be obtained from relatives of friends of dead or dying cases, and details are difficult to obtain.

Restructuring the Data by Visit Rather than by Person

Note that a person may have visited more than one Zone. We will map *visits* to zones rather than persons. Since all the persons in the study visited at least one zone, we will expect more than 600 visits.

Although placing the visit data in ten different fields (A-J) simplified the data entry, it is necessary to write a new data table containing a geographic field with values “A” through “K”, and one or more other fields including Case status. If the first Case record in the original dataset is a CASE and fields A, C, and E are checked, we want to produce three VISIT records as follows:

CASE	ZONE
(+)	A
(+)	C
(+)	E

The VISIT file will be processed to produce summary data of the following type:
CASE ZONE COUNT

Here's the recipe for writing the VISIT records to a table, using the WRITE APPEND command.

Start by READing the FinalData data table in SampleAnthrax.MDB after clicking the CHANGE PROJECT button in the READ command dialog.

DEFINE a new variable called ZONE.

Now select only the records for which field A is "Yes" by inserting the condition A=(+) in the SELECT command. Set the value of ZONE to "A" by using ASSIGN

ASSIGN ZONE = "A"

WRITE the selected records to a data table called MapData1 in the MDB called SampleAnthrax, choosing all the fields, including ZONE. To be sure that you are starting a new table, select REPLACE.

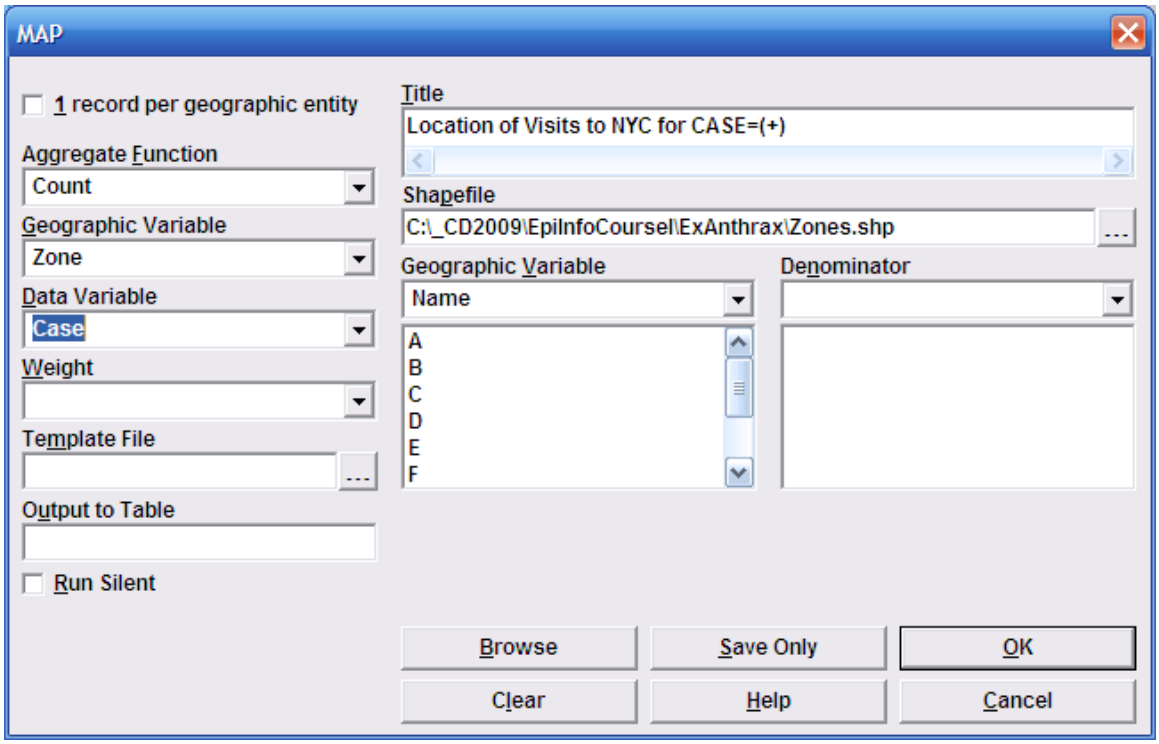
Now UNSELECT the selection and use SELECT again with the condition B=(+). Then ASSIGN ZONE = "B" and use WRITE again to append the records. Be sure that APPEND is selected when you do the WRITE.

Repeat this for all the Zone fields through K, and you will have about 2000 records in the MapData1 table. The task has been completed for you in the table called MapData in SampleAnthrax.mdb. Verify this with the READ command and a FREQuency of ZONE.

Linking Data to a Shapefile Using the MAP Command

Now it is time to link the data to the Zones shapefile and produce a map. To do so from within Analysis, first READ the MapData table. It is a Data table and has no VIEW. Hence to see it in the READ dialog, you must choose ALL rather than VIEWS in the SHOW box. Use LIST to see its structure.

First use the SELECT command to SELECT Case=(+), and note that the number of visit records is now 1441, each one containing the zone visited. Click on the MAP command, and complete the selections as follows:



When you click OK, a map of the zone should appear with the counts for cases shown as colors, darker zones having more cases. Sample images are provided in Appendix 4, but don't peek yet!

If you would like to have the subway map behind the zone image, use Map Manager from the FILE menu to load the image NYSubwayMap.jpg as an additional layer, after selecting STANDARD IMAGE FORMATS within the file dialog.

To produce a dot density image rather than the color/pattern (choropleth) map, select the Map Manager from the FILE menu in Epi Map and then click the PROPERTIES button and the Dot Density tab. Change the dot size and color if you wish and click OK. Now case visits are displayed as individual dots randomly scattered inside the zones to which they belong.

Exit from Epi Map to return to Analysis, and now cancel the previous selection (Case=(+)) with CANCEL SELECT and use SELECT to select only controls, the records having Case=(-). Repeat the mapping process with the MAP command, changing the title to "Zone Visits by Controls."

Now it's OK to look at Appendix 4. If you want to make your own report, you will find images of the maps you made as .BMP files in the working directory, and these can easily be imported into a Microsoft Word document using the INSERT, PICTURE, and FILE menu choices.

Although the maps give a dramatic picture for the newspapers, you should consider whether the statistics obtained in the risk factor analysis may be more effective in showing differences between cases and controls. The single map of visits by Cases =(+) is misleading without the map of Case=(-) visits. By itself, it might indicate only the popularity of certain zones with all New York visitors.

Conclusion

That completes the exercise. You have had a rapid tour of some of the informatics challenges that arise during an emergency investigation. The same problems arise in any investigation, although the time schedule may be different. Review and discuss the steps in Appendix 2 to see where we have been.

It is clear that a complicated and environment-dependent process like epidemic (or terrorism) investigation cannot be practiced occasionally and kept in a can labeled, "Open only in case of bioterrorism." The procedures used must be put into use for more mundane tasks on a routine basis. A tool designed only for bioterrorism will accumulate rust and fail to function if used only occasionally or in sporadic drills.

We have emphasized the use of Epi Info in this exercise, because our purpose was to teach how to use Epi Info. In a longer exercise, or a real-life setting, we would have exploited the use of computer industry standards, such as the Microsoft Access database format, Web Pages (HTML), and the ESRI ArcView shapefile standard to exchange data between Epi Info and commercial programs, and allow each to do its job with the least possible resistance, and with the shortest learning curve for the participants.

We hope you enjoy the new skills you have acquired and will be able to use them to improve public health information management.

Notes

The data table in this exercise was generated by using the RND (Random) function in the Analysis program. Any resemblance to a real epidemic should be taken with a grain of salt. In particular, the epidemiologic approach taken here can answer certain questions, but, by virtue of matching by Sex, Age, and having visited New York City, is prevented from producing information on these potential risk factors. In other words, the evidence is so overwhelming from a descriptive study that visits to New York were involved, that the study was designed to move on from there and ask the question, "Where in New York?" Although not every epidemic investigation proceeds in this way, it is common for a descriptive study of a foodborne outbreak, for example, to show that all cases "ate at Al's Barbecue," and for the case-control or cross-sectional study then to address the questions of, "What food at Al's Barbecue?" and " In what time period?"

Please send comments and suggestion about the exercise to the author at:
andy.dean@gmail.com

Postoperative Surgical Infections: An Exercise in Hospital Epidemiology

The Problem

Hospital A.is a 900-bed tertiary care center, performing over 15,000 surgical procedures per year. The infection control committee, composed of physicians, nursing supervisors, and administrators, is headed by an infectious disease specialist, who also serves as hospital epidemiologist.

She became concerned about surgical site infections in seven patients who had undergone open heart surgery (OHS) between May 1 and December 31. Rhodococcus bronchialis, a rare, gram-positive aerobic actinomycete, was recovered from blisters (two patients) or purulent drainage (five patients) from the sternal wound.

Instructions for this exercise are contained in the tutorial Rhodococcus.htm in the folder exRhodo. A similar, but older version is supplied with Epi Info.

Since you have already made a questionnaire in this course and entered data, we suggest that you skip the data entry part of the exercise and do the parts that pertain to sample size calculation, random number generation, and analysis of the 6 cases and the case-control study. Instructions are included for use of Statcalc, the original DOS program supplied with Epi Info, and also for OpenEpi. We recommend using OpenEpi from your class materials or from the Internet (www.openepi.com) for two-by-two tables and random number generation.

Data files are included with the exercise.

Analyzing a Mortality Database

Developed by Juan Carlos Zubieta, MD, MPH
Edited by Andrew G. Dean, MD, MPH

NOTE
To complete this exercise, please copy the files ESMortality.MDB, ES.shp, ES.shx, and ES.dbf into your C:\Epi_Info folder or wherever Epi Info is located in your computer.

This exercise is based on Epi Info, 3.2.x. or later. If you do not have the latest version installed, please notify one of the instructors.

You have received a database with information about deaths. The table called **FiveWeeks2002** in **ESMortality.MDB** contains data from individual death certificates in a limited period in 2002. You are asked to analyze the data. Please use the Epi Info ***Analysis*** program to answer the following questions

1. How many deaths are in the database? Answer

Hint: In the READ command select ALL to see the data table, as it does not have a View.

2. What is their distribution by sex? Answer

3. What is the mean age at death in the country? Answer

Answer

5. Which province has the greatest number of deaths?

Answer

6. Which province has the greatest number of female deaths?

Answer

7. Which province has the greatest number of male deaths?

Answer

8. What is the leading cause of death in the country?

Answer

9. How can I sort the list in descending order? Answer

Hint: To produce a data table containing the frequencies for each cause of death:

- Read the database as explained and click on the command FREQ
- Select the variable Diag and specify output to a table called Diagnoses

10. How many cases of stroke were reported? Answer

Hint: Use the FindText() function in an IF statement. If FindText("Stroke",Diag)>0 then Diag contains the word "Stroke", and you can set either Diag or a newly defined variable to "Stroke".

11. How many cases of Cancer were reported? Answer

Hint: Same idea as for Stroke, but you should also include some IF statements for categories like "Leiomyosarcoma" and "Leukemia" that do not contain the word "Cancer." This could best be done by creating another variable called CANCER and setting it to (+) or (-) using IF statements. Your instructor may choose to skip this and the next question.

12. What is the most frequent cancer? Answer

13, What is the mortality rate?

Answer

Hint: There is a dBASE file called ES.DBF that is part of the shape file for mapping the country. Population figures by province are found in this file.

14. How can I create a map with counts and rates?

Hint: First do a Frequency by PROVINCE, routing the output to a table called PROVINCE, as we did above for Diag. READ the PROVINCE table and Assign COUNT=Count * 1000. This will allow calculating rates per thousand population. Now use the MAP command and choose the ES.shp file with Geographic variable Admin_Name and Denominator Pop_Admin.

15. How can I save the program to use again as the weeks pass?

Answer

Analysis of Reportable Disease Surveillance Data with Epi Info

The database called Surveillance.MDB contains a table of sample surveillance data from a 5 month period in a data table called CaseReports. The data have been artificially generated to protect confidentiality and names and addresses have been removed.

Considering that disease surveillance depends on detection of changes over time, try to develop a set of analytic procedures that will allow you to monitor trends in this kind of data. The graphing features of Epi Info allow graphing counts over time (MMWR Week for example), and producing a separate graph for each EVENTNAME (disease) and/or geographic entity.

Graphing Surveillance Data

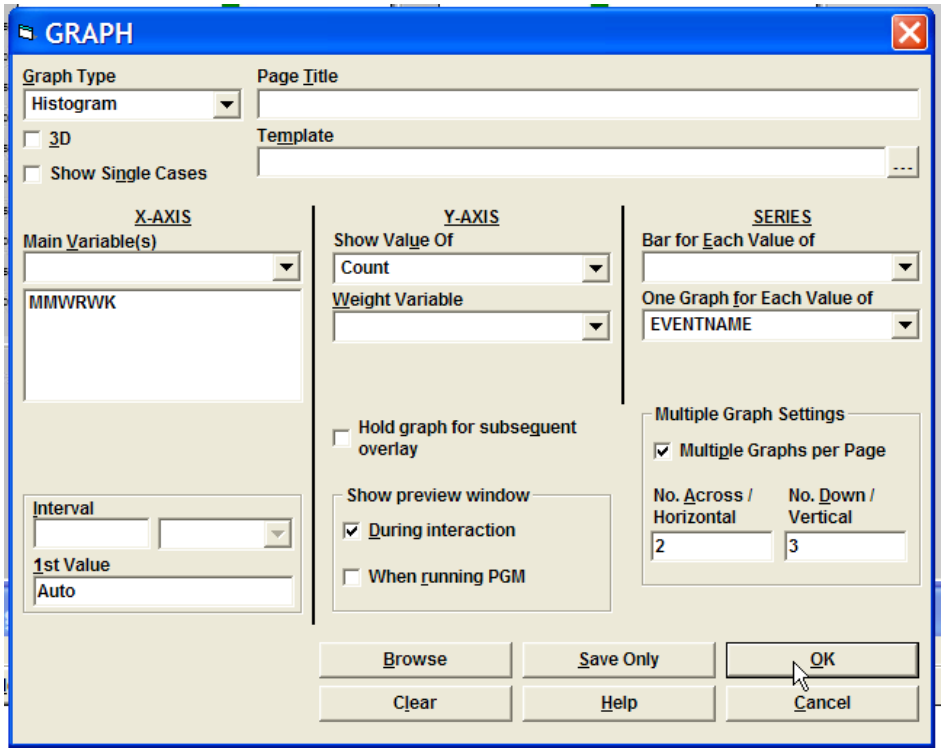
The first step in working with a new dataset is to discover the file format for the data. In this case, the file name, SURVEILLANCE.MDB, tells us that the file is a Microsoft Access (MDB=Microsoft Data Base) file, the native format of Epi Info. It may have come from Epi Info or from Microsoft Access, or may have been exported in this format by another program.

In Epi Info, choose Analyze Data from the main menu and then click on the READ command, the first item in the command tree on the left. Choose CHANGE PROJECT and navigate to the directory where SURVEILLANCE.MDB resides. Since there are no Epi Info Views, click on the ALL option above the blank window. The data table is called CASEREPORTS. Select it with a click and then click the OK button. You should see that more than 1100 records are available.

To see the records, use the LIST command and scan for interesting variables that have non-missing values. MMWRWK is a code for the week in the U.S. epidemiologic year and is a good proxy for time of report. EVENTNAME in this table contains the names of reportable diseases.

Graphing surveillance events over time can be done with the GRAPH command. Click on the command, and choose HISTOGRAM for graph type rather than BAR. The difference between the two types is evident if you graph data containing missing or zero values for some of the weeks. HISTOGRAM will include these weeks on the horizontal axis, but a BAR graph will omit the weeks with zero or missing values, giving the wrong impression of the passage of time.

Choose MMWRWK for the X-AXIS and Count for the Y-AXIS to show the number of records or reports for each week. Although the total number of reports per week is an interesting administrative figure, we are more interested in the number of reports for particular diseases. Hence, under SERIES, specify that you want ONE GRAPH FOR EACH VALUE OF the variable EVENTNAME. This will make quite a few separate graphs, but you can make them smaller by checking the box, MULTIPLE GRAPHS PER PAGE, and requesting 2 graphs horizontally and 3 vertically on each page. Your completed dialog should look like this:



Click OK, wait a moment, and you should see a number of graphs, one for each disease. The index in the upper right corner of the page allows navigation to the multiple pages of graphs. Note that the Histograms do not compress the horizontal axis when values are zero or missing, making it easy to compare the graphs with each other and with the normal calendar.

When you close the graph window by clicking on the X in the upper right corner, the graph images are automatically placed into the HTML output page named at the top of Analysis, and they can be accessed later from this page. They are separate images, however, and must be copied with the HTML if they are moved elsewhere.

For many purposes, this single operation would produce all the graphs that are needed to monitor reported diseases over time. In other settings, it may be necessary to further stratify the data by geographic area, using either BAR FOR EACH VALUE OF (say, CITY, if there are only a few), the SELECT command in Analysis, or ONE GRAPH FOR EACH VALUE OF (say, COUNTY) with one BAR FOR EACH VALUE of

EVENTNAME. The program in the lower window of Analysis can be edited, saved, and run each week as new data are added to the data table.

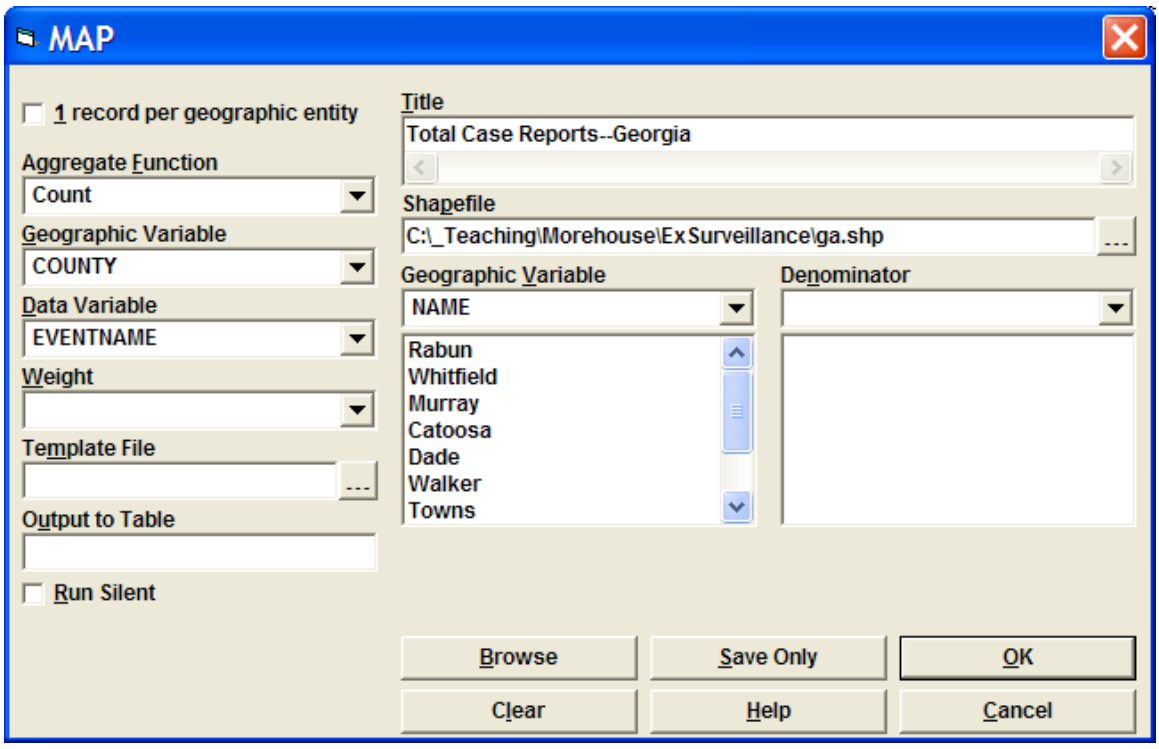
Mapping Surveillance Data

The CaseReports data table used in the preceding section contains geographic information--CITY, COUNTY, COUNTYCODE, and DISTRICT. To show the data on a map, we need to have a shape file or map of the territory that matches the database. The artificial database is constructed for the U.S. state of GEORGIA, and we can download a map of Georgia by county from the Epi Info website (www.cdc.gov/epiinfo) in the MAPS section. Many other sites offer free or inexpensive maps in the popular ESRI/ArcView/ARCInfo shapefile format. The company that produces ArcView and ARCInfo (Environmental Sciences Research Institute or ESRI) provided a special license to use and distribute their MapObjects software in Epi Map, and shape files can easily be read and manipulated in Epi Map.

We have provided the GA.SHP file in the same directory with SURVEILLANCE.MDB. The first thing to note, however, that shape files are not single files, but consist of at least three files and sometimes five. GA has separate files called ga.dbf, ga.sbn, ga.sbx, ga.shp, and ga.shx. Why the collection that makes a single map is referred to as "a shape file" is a mystery perhaps known to Geographic Information Systems (GIS) techies. Whatever the source of the shape file and the Epi Info or Microsoft Access data, the identifiers for polygons (counties) must have the same spelling. Epi Map will list those that do not match, and we could change the spelling later.

In Analysis, READ the same SURVEILLANCE.MDB, data table CASEREPORTS, that was used in the graphing exercise. Click on the MAP command. In the dialog that comes up, leave COUNT as the aggregate function. Choose COUNTY as the geographic variable. In the DATA variable item below it goes the variable to be counted. Choose EVENTNAME so that the result will be a count of all records in which EVENTNAME has a value. Note that this will merely count reports, and will not distinguish between the various values of EVENTNAME—the individual disease names.

On the right side of the dialog, set up the shape file to be used. Click on the button with the three dots and find GA.SHP. Choose NAME as the geographic field to be matched, and click OK.



You will see a dialog showing which names do not match. Most of those on the left did not match because there were no records in the data table that specified these counties. On the right-hand side there are only three entries. DEKALB differs only in capitalization, and MC DUFFIE should not have a space in the middle. The word UNKNOWN is not a NAME in the shape file. But all the other records matched a county, so click OK. You should see a map of Georgia representing numbers of EVENTNAMEs reported during the period covered by our data table.

At the top of the window is a toolbar. Click on the "i" (for Information) button and then click on one of the counties having reports. Look for the variable called COUNT_EVENTNAME to see the number of reports in the county.

The number of reports will be more graphically displayed as dots. Choose MAP TYPE from the menus at the top of the window and then DOT DENSITY. Note that the property window that pops up offers many options for customizing the display. For now, simply choose the tab called DOT DENSITY and click OK. Now the case reports are represented as dots randomly distributed within the counties to which they belong. This feature is useful for items such as caseload, and has the advantage of retaining confidentiality, since cases are only shown at county level, and not by individual address or geographic coordinates.

Mapping works best if the data table contains "Only one record per geographic entity," in this case, per COUNTY. Fortunately we can aggregate the data by COUNTY and by EVENTNAME with the TABLES command and send the output to a new data table for linking with the shape file. Here's how:

READ the CASE REPORTS table

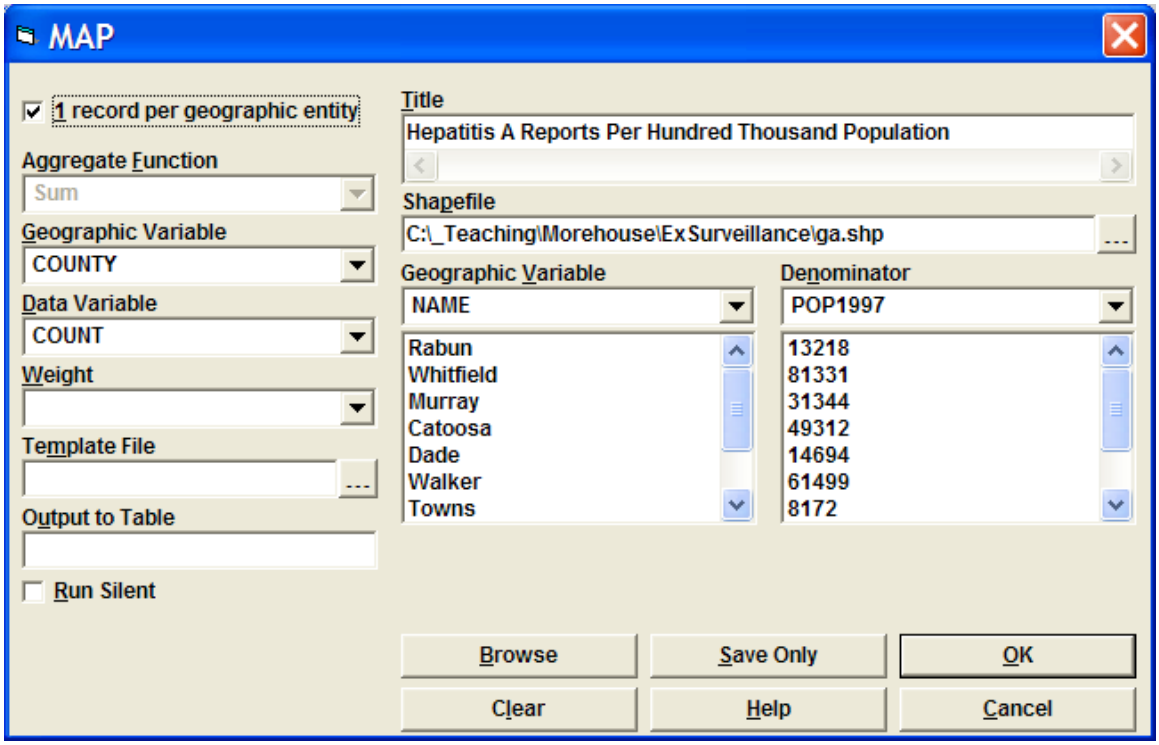
The objective is to make a data table of COUNTY by EVENTNAME in Epi Info/Access format. Click on the TABLES command and choose COUNTY as the "Exposure" and EVENTNAME as the "Outcome". To obtain a new data table put the name CountyByEvent in the OUTPUT TABLE field. Click OK to make the table. It may happen rapidly and without much notice.

Now READ the new table CountyByEvent with the READ command. Use LIST to see its contents. It has a new field called COUNT that gives the number of records having the stated values of COUNTY and EVENTNAME, and it is COUNT that we want to show on the map for a given EVENTNAME in each COUNTY.

There is one more detail that needs attention. If we calculate rates using population data in the shape file, some trial and error shows that a rate of a few cases divided by the population (thousands or millions) gives numbers like 0.0000001 that are rounded off to two decimal figures in the map legend. Hence, it is a good idea to multiply COUNT by 1000000 and express the rates as "Per Hundred Thousand Population". This can be done with the ASSIGN command as ASSIGN COUNT=COUNT*100000.

Since we want a map of Hepatitis A reports, use the SELECT command with EVENTNAME="Hep/A", the value used for Hepatitis A.

Finally we are ready to do the map. Choose MAP and "One record per geographic entity". The Geographic field for the data table is COUNTY, and the Data Field is COUNT. Since we have selected only the hepatitis A cases, there is no need to use EVENTNAME.



Add a title to indicate that we are showing hepatitis A rates per 100000 population. On the Shape file side, choose GA.SHP as before and NAME as the geographic field. Then for Denominator choose POP1997, the latest available in this venerable shape file. Click OK and you should have a nice map of Hepatitis A over the time period of the dataset. Of course you could make further use of SELECT to confine the display to a particular time period, and after Canceling SELECT, could choose another disease. After doing the interaction a couple of times, you may prefer to edit the file in the program window, using copy, paste, and editing to repeat the steps for a number of diseases. As before, if you save the program file, you can run it again in subsequent sessions after more data become available.

Making Tables

The classical items for analysis in epidemiology are Time, Place, and Person (meaning everything else). We have made graphs to study Time and maps to display Place. The remaining items can be displayed and studied in tables. Let's look at a few examples.

1. Produce a Frequency of AGE for Hepatitis A reports. This would also make a nice bar graph or histogram.
2. Do a Table of Disease by County (Don't forget to CANCEL SELECT first), and remember that Disease is called EVENTNAME. If each cell contains percentage values along with the count, bring up the SET command and uncheck the Show

Percents feature. You may also want to set STATISTICS to NONE for large tables, but be sure to turn it back on for your next outbreak analysis!!

3. For each COUNTY, obtain a frequency of reports by EVENTNAME.
4. For each COUNTY, obtain a table of EVENTNAME by MMWRWK.

For the last two, separate frequencies or tables for each COUNTY can be produced with a single command by putting COUNTY in the STRATIFY BY box in the Frequency or Tables dialog.

As surveillance coordinator for a state, do you see how you might produce a weekly report to send back to the counties that contained Graphs, Maps, and Tables of data from that county? You might choose to do this over a longer interval, since many counties have only small numbers of reports in a week. More frequent events, such as ambulance runs or home visits would be better candidates for weekly output.

Graphs and maps are probably better modalities for displaying and understanding the data, but tables are more convenient for reporting counties to check the figures and see if all their reports were received by Data Central.

Making a Menu

The Epi Info menu (EpiInfo.EXE) is a flexible tool for organizing a project. Rather than having to remember the names and locations of files and repeat dozens of steps interactively each time new data become available, you can automate much of the process using menu commands, Analysis programs, and other features that are designed for repeated use. The same features can be used to develop applications for others to use, for example, in an office routine.

The menu has its own set of commands, accepts DOS batch file commands, and can run Windows or DOS programs. Once commands have been encapsulated in menu and Analysis programs, they can be run with the click of a button or a menu choice, and no longer depend on human memory and informatics skills.

To begin this exercise, examine the text file program behind the Epi Info menu by choosing **Edit** and **Edit This Menu** from the Epi Info main menu. Note that the menu structure is represented at the top with POPUP and MENUITEM commands, and that these correspond to the items on the menu. Each MENUITEM calls a corresponding block below by name and finds instructions between the BEGIN and END statements about what to do when you click on the corresponding menu item.

What happens, for example, when you click on the MakeView item under Programs? Find the menu code that does this.

Other menus can be made by copying and editing the EpiInfo.mnu file or writing a new one. We have made one called MAKEMENU.MNU for teaching purposes. To run MAKEMENU.MNU, find the ExMakeMenu folder in your disk copy of the course materials, and click on the file called **_ClickHereToRunMakeMenu.**

Examine the MNU file using Tools|Edit This Menu. The menu contains examples for entering data, running Analysis programs, and linking documentation to a button or menu item so that you can easily find it. It also shows how you can create a button to access an Internet address, essentially creating your own web portal with a couple of lines of Menu code.

MakeMenu has been programmed to reproduce itself and construct a new menu. (If you are a programmer, have a look at the last part of the menu file to see how this is done.)

Try making your own menu for the reportable disease surveillance system that we worked with earlier.

1. In MakeMenu, click on CREATE NEW MENU under TOOLS.
2. Give GEORGIA as the name of the menu to create.
3. After a couple of dialog boxes, you will be advised to look on the desktop for the new GEORGIA icon. Close MakeMenu, find GEORGIA, and click on it.

4. You should now have a menu quite similar to MakeMenu, but with a sky and clouds theme, and in a new directory of its own.
5. Use EDIT THIS MENU under EDIT and you will see the text program that configures the menu. The entire menu and the commands that it executes can be changed and saved. Note the POPUP and MENUITEM lines in the first portion. These indicate the menu structure. Let's change the first menu, ENTER DATA, to ENTER. Edit the POPUP item and change the words. Save the result with Ctrl-S or from the File menu, and you will immediately see the change. How did the menu know that it had been edited? It checks several times per second to see if the file date and time has changed, and reloads if so.
6. Click on CHANGE PICTURE under TOOLS, and find the file called SURVEILLANCE.JPG in the ExSurveillance folder (probably up one level). The menu should now have a different background image.
7. Note that the picture you have chosen is too dark for the text items on the screen. To fix this, click on EDIT THIS MENU under TOOLS. Find three lines beginning with SCREENTEXT about mid-way in the text, and change the color MAROON to YELLOW and the two BLACK items to WHITE. Save the file with Ctrl-S, or from the FILE menu, and you should see the new colors.
8. On the ENTER menu, choose ENTER CASE REPORTS, and note that this item already works and brings up the Surveillance database. Have a look at the menu block called ENTER CASES that does the work. What menu command is used to run the ENTER program? How does the menu know where to find ENTER.EXE?
9. Now click on ANALYZE DATA and RUN ANALYSIS STATISTICS. Wait patiently and confirm that it runs the sample program called STATISTICS that illustrates nearly all the statistical functions of Analysis.
10. Running an Analysis program from the menu is a little complicated, as Analysis, the PGM, and one or more MDB databases may be in different places, and Analysis has its own idiosyncratic syntax. We have placed the necessary commands to run HEPATITISARATES in the SURVEILLANCE.MDB in ExSurveillance in the menu, but they are commented out (disabled) with asterisks. To make them active, first place an asterisk in front of the EXECUTE line that runs the STATISTICS program in the menu block called RunAnalysisPGM. Then remove the asterisks from the two lines further on that will run the HepatitisARates program. We used TWO asterisks on these lines for clarity; remove them both. Save (Ctrl-S) your menu and test the Analyze Data item. It should run the hepatitis program and produce a Georgia map of rates (not real ones).

Test the menu and make adjustments as needed. Because the menu has commands similar to those in Analysis and Checkcode, and can also execute any of the DOS batch file commands for copying files, running programs, etc., and can also run other menus, you can make a complete system with more than one menu. The REPLACE command in the menu can be used to customize programs before they are run by searching and replacing words or phrases such as "%FileLocation%" that you have inserted as placeholders.

Appendices to Anthrax Exercise

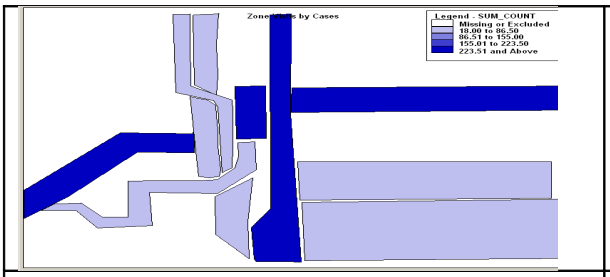
1. Anthrax Exercise Materials

File	Table(s)	Description
SampleAnthrax.MDB	viewCaseControl	Data entry View including Check Code
	FinalData	(Artificial) data from about 300 cases and 300 controls for analysis
	MapData	Zone Visit records generated from FinalData (2000+ records)
	MapSummary	Summary data from Zone Visits for linking to Map
	viewEntry1, viewEntry2	Two datasets with errors, to be compared with Data Compare
Station.MDB	viewCaseControl	Data entry View including Check Code for practice in entering data
	viewAlmostNoCheckCode	Data entry View for inserting your own Check Code
Zones.shp, .dbf, .shx		Shapefile, zones A-K
NYSubwayMap.jpg		Background image for map
Zones.MAP		MAP file or template combining shape file with image
AnthraxExercise.DOC		The document you are reading
CompareEntries.TXT		Script file for Data Compare
CDCAnthraxFactsheet .PDF		Separate background document

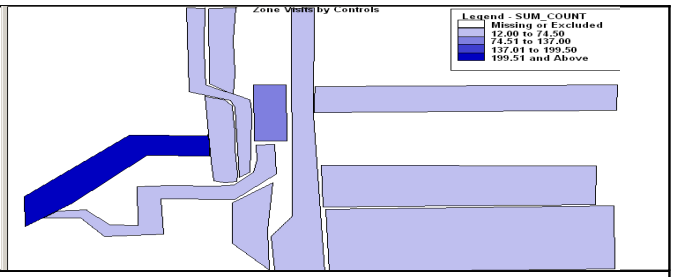
2. Sample Informatics Strategy on the First and Second Days of the Anthrax Investigation

1) Design a paper questionnaire and print copies from Microsoft Word
2) Develop an Epi Info questionnaire View in MakeView
    a. MDB=<YourInitials>.MDB View=CASECONTROL
3) Pilot test data entry with plausible but invented data
4) Add Check Code to control errors and facilitate data entry
    a. MDB=<YourInitials>.MDB View=CASECONTROL
    b. Working sample in SampleAnthrax.MDB View=CaseControl
5) Pilot test data entry again
6) Copy the Project MDB to each data entry computer
    a. Pretend this has been done and that STATION.MDB:CaseControl is the local copy
7) Enter data
    a. In STATION.MDB:CaseControl
8) Check data quality manually (Entering twice and using Data Compare is an alternative)
    a. In STATION.MDB:CaseControl
    b. Use Entry1 and Entry2 Views to see Data Compare work
9) Copy each dataset to a floppy diskette, transport to master computer, and append to the master data table
    a. Copy STATION.MDB to Diskette. Change file name and/or label with Station #
    b. WRITE APPEND each to MASTER.MDB:CaseControl
10) Make backup copies of the master table, using EpiLock to accomplish compression ("zipping") if needed, and encryption if they are to be transmitted or transported.
11) Analyze case and control data to find differences, and print tables, statistics, and graphs of the results
    a. Use SampleAnthrax.MDB:FinalData—300 Cases and 300 Controls
12) Make a new table with one record for each visit to a zone of New York City by a case or control
    a. Use SampleAnthrax.MDB:FinalData to make SampleAnthrax.MDB:MapData1
    b. Completed task is in SampleAnthrax.MDB:MapData
13) Write summary data to a new file for mapping
    a. Use SampleAnthrax.MDB:MapData to make SampleAnthrax.MDB:MapSummary1
    b. Completed task is in SampleAnthrax.MDB:MapSummary
14) Link the MapData table to the map of New York City zones
    a. Use SampleAnthrax.MDB to link to Zones.SHP with Subway.jpg loaded as first layer
15) Construct a menu to provide access to the programs developed and to documents that are received or developed during the investigation
    a. Use Core.MNU to develop a new menu NJANTHRAX.MNU
    b. Completed task is in SampleANTHRAX.MNU

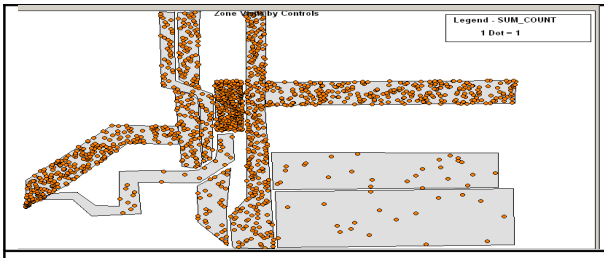
3. Epi Map Images of Case and Control Visits to New York Prior to Case Onset of Anthrax



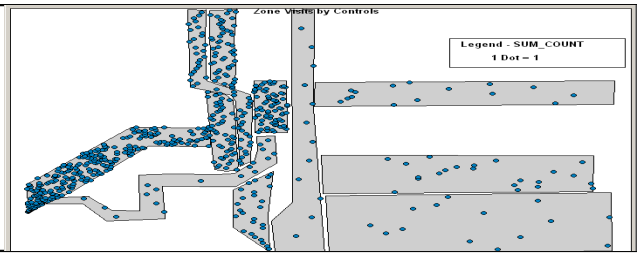
Case Visits (Choropleth)



Control Visits



Case Visits (Random Dot Density)



Control Visits

4. Background Information on Anthrax

From the Centers for Disease Control and Prevention, Atlanta
http://www.bt.cdc.gov/agent/anthrax/anthrax-hcp-factsheet.asp
December 2002. Reviewed 2004.
Consult original for current information. No change as of March 2009.
Provided as a separate .PDF document, CDCAnthraxFactsheet.PDF

5. Interviewer's Map, Anthrax Investigation