Computing with Epi Info
Part II
Intermediate to Advanced

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

Table of Contents

Epi Info II

Objectives

1. Understand common data management problems encountered in community public health or clinical computer systems and their solutions
2. Develop skills in managing, importing exporting, aggregating, merging, backing up, and encrypting data with Epi Info
3. Enhance analytic and statistical skills to include analysis of complex sample data and the use of logistic regression and survival analysis
4. Be able to understand and use relational tables for data entry and analysis in Epi Info
5. Build an Epi Info application with a newly constructed menu
6. Understand and use geographic information in Epi Info

Schedule

Day 1

Time	Session	What Happens	Materials
Launching the Course		**Files in …\EpiInfoCourseII\**	
	Introductions	Introduction of Instructors and Class Members	
	The Plan	Review of Objectives, Materials, and Schedule	Course notebook and schedule in EpiInfoCourseII.doc
	Review of concepts	Epidemiology from an Internet Café. Review of database and Epi Info concepts.	InternetCafe.PPT on CDROM

Solving Data Management Problems in a Community Health Department or Clinic Files in …\EpiInfoCourseII\exDataManagement\			
	Problem 1	Working with Excel Files	Restaurants.xls:AprilWK2
	Problem 2	Importing a Text File	HospitalCommas.txt HospitalTabs.txt HospitalFixed.txt READFIXED.PGM
	Problem 3	Monitoring and Managing Hard Disk Space	MyComputer program on your computer
	Problem 4	FoxPro System and Departed Contractor	Patients.dbf
	Problem 5	Computer Viruses	Discussion
	Problem 6	Backing Up and Restoring Data	Discussion
	Problem 7	Sending Confidential Files	Orders.Wk34.xls

	Problem 8	Adding a New Field and New Data to Existing Views	Surveillance.mdb:viewSurveillance
	Problem 9	Merging Data from Multiple Offices	Central.mdb Local1.mdb Local2.mdb

Statistics—Advanced Data Analysis		**Files in …\EpiInfoCourseII\exStatistics\**	
	Statistics 1	Obtaining Historical Data	Internet access to www.openepi.com or use the htm files in …\EpiInfoCourseII\exStatistics
	Statistics 2	Analyzing a Community Cluster Survey	Sample.mdb: viewEpi1 and viewEpi10
	Statistics 3	Finding Relationships among Three or More Variables	Sample.mdb: viewOswego
	Statistics 4	Logistic Regression and a Classical Heart Disease Study	Sample.mdb: viewEvansCounty
	Statistics 5	Followup Study of a Clinical Drug Trial—Survival Analysis	Sample.mdb:viewAnderson

Day 2

Related Views and Data Tables		**Files in …\EpiInfoCourseII\exRelated\**	
Time	Session	What Happens	Materials
		Creating Views and Entering Relational Data	What's a Record? Introduction to the Related View Exercise
		Analyzing Relational Data	Related.mdb
	Goal 1	Relating Mother and Child Views and Calculating Mother's Age at Child's Birth	Related.mdb Goal1.pgm
	Goal 2	Using SUMMARIZE to Find Date of First Positive HIV Test for Each Mother	Related.mdb Goal2.pgm
	Goal 3	Finding Rate of HIV Transmission to Child from HIV-Positive Mothers	Related.mdb Goal2.pgm
		Adding Alternative Keys in a Relational System (Optional)	Related.mdb Goal2.pgm

Menus and Permanent Applications			**Files in …\EpiInfoCourseII\exStatistics\**
	Menu 1	Making a Menu	ClickHereToRunMakemenu, MakeMenu.mnu
	Menu 2	Configuring the Menu to Include New Features	Files from exRelated

Geographic Information Systems (GIS)			**Files in …\EpiInfoCourseII\exGIS\**
	GIS 1	Mapping Points in Epi Map Animals Tested by County Displaying Positive Rabies Tests by Point Coordinates	GISData.MDB:Rabies2003
	GIS 2	Mapping Categorical Data in Epi Map	PositiveBirds2003.XLS
	GIS 3	Editing or Creating Shape Files Creating a New Shape File by Combining Polygons from Others Making an Entirely New Shape File Based on an Image	WV.shp PA.shp OH.shp Charleston1996.jpg

Finishing Up			
	Questions and suggestions	Class discussion.	
	Evaluation		Evaluation form

Data Management Problems in a Community Health Department or Clinic

Introduction

You find yourself in charge of informatics for a community public health agency. Although you have very little training in this area, your skills with Epi Info in computerizing surveys and questionnaires have convinced the head of the agency that you are equal to this task.

You spend a few weeks observing, getting acquainted, and collecting ideas. There are a number of computers and several computerized systems, but very little software related to the needs of the agency. Integration of the various systems is—shall we say—"under development."

Among your observations are:

1. The sanitarians are maintaining records of restaurant inspections in Excel, and would like to do more analysis for their year-end report.

2. The local hospital has computerized medical and financial records, and will provide reports of communicable disease as text files with columns separated by tabs or spaces if you can figure out how to use them.

3. The receptionist in the agency clinic has a computer and a program that looks up patient records or receives new-patient entries, but the system has not

functioned for the past 2 months because the hard disk on the computer is full. Attempts to enter new records produce an error message.

4. The receptionist's look-up system was written in FoxPro by a long-forgotten contractor and no one seems to know how to maintain it or has a FoxPro development system handy to make changes. You decide to convert the system to Epi Info and provide the same lookup and data entry functions, since you are more familiar with Epi Info than with FoxPro.

5. There are computer viruses in at least one of the agency computers, and you find that it is protected by Norton AntiVirus with an expired subscription; the virus definition files were last updated 2 years ago.

6. There is a large-capacity tape drive in the server for a small local area network, but no one knows how to make it work, and there is no regular system for backup of either the network or the several free-standing computers used for word processing or research.

7. The HIV/AIDS clinic in your agency is located on the other side of the city. They send orders for medication to the agency pharmacy in Excel format, but are afraid to attach them to emails since they contain patient identifiers. They are currently FAXing the orders to the pharmacy, but a secretary from Environmental Health brought you one of the FAXes last week that had been mistakenly sent to her FAX machine through miskeying a single digit in the similar phone numbers. You are grateful that the FAX is within the agency, but wonder how many other confidential FAXes could have gone astray.

8. Several offices of your agency are entering computerized surveillance data in the form of disease reports using Epi Info Views, but the data are sent to your office on paper and must be reentered. You would like to merge the files at the Central office, but the data from the local offices has overlapping ID Numbers and no indication of which office entered the data. You would like to add a new field to Surveillance data View called AGENCY so that the agency providing the data and the ID Number together will be unique for each case report.

9. You want to merge the data from other offices into a Central surveillance database, using both AGENCY and ID Number as the record identifier, since the ID numbers for the reports are only unique within a given agency, and there may be records with the same number from different agencies. One of the local offices has added extra variables for its own use, however, and no one is sure how to combine the dissimilar data tables.

Problem 1: Working with Excel Files

The sanitarians are maintaining records of restaurant inspections in Excel, and would like to do more analysis for their year-end report.

Data sample: Restaurants.xls

Objective: READ the Excel file in Analysis and do analysis

Instructions:

1. **Run ANALYZE DATA from the Epi Info main menu.**
2. **Click the READ command at the top of the command tree on the left.**
3. **For FILE FORMAT, choose EXCEL 8.**
4. **Click the button with the three dots and navigate to the folder called EpiInfoCourseII\exDataManagement.**
5. **Select the file called Restaurants.xls, and then choose the worksheet called APRILWK2. Click OK and wait for Epi Info to read the file.**
6. **You should see a notice telling the number of records in the file.**
7. **Use the LIST command to see the data.**
8. **Now that you have read the file, you can produce immediate tabulations with FREQUENCIES or TABLES. Try a FREQuency of VIOLATIONS.**
9. **If you prefer to have the file in another format, the WRITE command can be used to write either an native Epi Info/Access database or one of the other file formats offered.**

Problem 2: Importing a Text File

The local hospital has computerized medical and financial records, and will provide reports of communicable disease as text files with columns separated by tabs or spaces if you can figure out how to use them.

Data samples:

HospitalCommas.txt
HospitalTabs.txt
HospitalFixed.txt
READFIXED.PGM

Objective: Import a text file with columns delimited by tabs, spaces, or commas

The READ command is capable of reading both Delimited and Fixed format text files. In both cases, if you try to use the READ command dialog, you get a message about needing a "FileSpec".

The command line and the FILESPEC line(s) must be typed into the program window in the lower right of Analysis. From reading the HELP file, we discover several examples of "FILESPEC"s, but, unfortunately, they are not quite correct, and even if you type the recommended commands in the program window, the READ command "is not recognized." Placing single quotation marks around the file path and name corrects this situation, and the following command will import the comma delimited file if you make the necessary changes in the file location.

Instructions:

1. **Type the following command into the program window in Analysis; the command must be on two separate lines, as follows:**

```
READ "Text (Delimited)" 'c:\EpiInfoCourseII\exDataManagment\HospitalCommas.txt'
FILESPEC HDR="YES" FMT="CSV"  END
```

2. **If the text wraps around, allow it to do so, but insert only one end-of-line character—the one before FILESPEC. Place the cursor on the first line and click the button labeled RUN THIS COMMAND.**

This means "Read the indicated file as a text file with delimiters, assuming that the first record contains field names (HDR="YES") , that the delimiter is a Comma, and that text items are enclosed in quotes, each record being on a single line ("CSV" format).

Other commands that you might use are given in the HELP file for the READ command. We have corrected these examples by adding the necessary single quotes.

```
READ "Text (Delimited)" 'c:\EpiInfoCourseII\HospitalCommas.txt'
FILESPEC HDR="YES" FMT="TAB" END

READ "Text (Delimited)" 'c:\Epi2000\datadlm.txt'
FILESPEC HDR="YES" FMT="DLM(32)" END
```

Presumably this obscure format indicates that the delimiter (DLM) is the ASCII character #32, which happens to be a space. ASCII codes for other characters are fairly easy to find by searching with www.google.com for ASCII TABLE.

The next one is for Mainframe fans who are handed a file and a detailed "data dictionary" indicating a name, the position on the data line, and the data type of each field. The data in this case are identified by position (Fixed) and not by delimiters. Every location on the line must be accounted for, although you can choose to read a number of fields into a single variable (e.g., DUMMY) if their content is not of interest.

3. **You could type the following into the program window, then RUN the program, but we have provided a text file called READFIXED.PGM in the exDataManagement folder. Click OPEN in the program editor, then TEXT FILE in the dialog that pops up. Locate READFIXED.PGM and then click OK. If the location of your EpiInfoCourseII folder is not the same as the one indicated, make necessary changes. Now click the RUN**

button and you should see that the file has 8 records. Use LIST to see the data.

```
READ "Text (Fixed)"       (no end-of-line character)
’c:\EpiInfoCourseII\exDataManagement\HospitalFix.txt’
FILESPEC HDR="NO" FMT="FIX"
   Clinic 1-5 Text
   DateOfEntry 6-15 Date
   Age 16-18 Number
   DUMMY 19-21 Text
   DX 22-31 Text
   FirstVisit 32 YesNo
END
```

If something is not exactly right, the program text will turn red rather than green, and a message will pop up. This will happen, for example if there is an end-of-line character after the “Fixed)”, or if there IS NOT an end-of-line before the FILESPEC, or if one of the fields does not have a field type indicated. Keep trying until you find that the program works.

Problem 3: Monitoring and Managing Hard Disk Space

The receptionist in the agency clinic has a computer and a program that looks up patient records or receives new-patient entries, but the system has not functioned for the past 2 months because the hard disk on the computer is full. Attempts to enter new records produce an error message.

Data sample: The hard disk(s) on your computer

Objectives:

- **Find out how much space is left on a hard disk**
- **Discover which folders occupy the most space**
- **Reduce the amount of space occupied on your disk**
- **Discuss methods of backing up or moving large files or folders to prevent or resolve full disk problems.**

Instructions:
Let’s pretend that the computer on which you now work is the one with the “full disk”.

1. **How many hard disks does your computer have, and how much space is left on each?**
2. **What Folder in your system occupies the largest amount of disk space?**
3. **What are the options for backing up a 150 megabyte file? A 567,000 kb file? A 3 gigabyte folder? A 20 gigabyte folder?**

Answers are on the next page, but try to solve the problems before peeking.

HINTS:

1. **How many hard disks does your computer have, and how much space is left on each?**
   There are several ways to do this, but generally you can see the drives and other storage devices by running MY COMPUTER from an icon on the desktop or from the Windows START menu. If the capacity of the drives is not shown, try right clicking on the icon or drive name and choosing "Properties" from the menu that appears. You can also choose START | PROGRAMS | ACCESSORIES | Windows Explorer and navigate to My Computer to see the drives and their capacity. If you do not see capacities listed, try changing the VIEW to DETAILS and/or Right Clicking on the icons.

2. **What Folder in your system occupies the largest amount of disk space?**
   The same techniques will show the capacities and space used or left in folders. There is no simple way to see the folders at different levels, but you can see and compare all those immediately below the root of the C: drive, for example.

3. **What are the options for backing up a 150 megabyte file? A 567,000 kb file? A 3 gigabyte folder? A 20 gigabyte folder?**
   a. A 150 megabyte (150,000,000 or 150,000 kilobyte—kb) file is too big for a floppy disk that holds 1.4 megabytes. It will fit easily on a CD ROM, which holds in the neighborhood of 600 megabytes, and also on a Flash Memory stick or card if it is of the larger type that holds 256 megabytes or more.
   b. The 567 megabyte file will just fit on a CD ROM (or a 1 gigabyte Flash Memory device.)
   c. Three gigabytes is too much for a CD ROM, but will fit on a recordable DVD (4 gigabytes). The problem is that there are several DVD formats, and it is important to be sure that the data can be restored to another computer with a reader of the same format if necessary. Testing is the best way to assure this, since DVD writing and reading is more sensitive to mechanical and satanic influence than the more robust CD ROM format.
   d. More than 4 gigabytes requires either a high-capacity tape drive or another hard disk. Large (> 180 gigabytes) hard disks are available that can be plugged into a USB or Firewire port on almost any modern desktop or laptop and then moved to another computer after backing up files.
   e. The size of almost any file or directory can be reduced by using ZIP compression, either with a program like WinZip (downloadable from the Internet) or by right clicking and using Windows (XP) facilities for ZIPping. Microsoft Access files, for example, can often be reduced 20-fold by zipping. It is important to have adequate space to Unzip the files again, or you find yourself in the position of the homeowner who found himself trapped in a closet by starting to paint the floor at the wrong end of the room and gradually reduced the area of unpainted floor space.
   f. There are many free or commercial programs available that claim to perform backups easily. If you use them, but sure to confirm that they are really

doing what you surmise, and that restoration is equally easy and effective, preferably on another computer.

The computer in the original example had 37 gigabytes of files on the C: Drive and very little space left on the generous 40 gig hard drive. Exploration using right clicks and choosing "Properties" for the various folders disclosed a folder called "Lily," that contained 27 megabytes of files. Most of these were either video clips or songs unrelated to the patient intake system. Erasing the files AND emptying the Recycle Bin folder provided enough disk space for future growth.

Problem 4: FoxPro System and Departed Contractor

The receptionist's look-up system was written in FoxPro by a long-forgotten contractor and no one seems to know how to maintain it or has a FoxPro development system handy to make changes. You decide to convert the system to Epi Info and provide the same lookup and data entry functions, since you are more familiar with Epi Info than with FoxPro.

Objectives:

- **Convert the FoxPro file to an Epi Info data table—the Microsoft Access format.**
- **Construct a View for entering more data into the converted data table.**
- **Program the View using Check Code so that entering a PatientNumber number or first and last name does an automatic search for a matching record in the database**

Data sample:
PATIENTS.DBF
Objectives:

- **Import the FoxPro file (.DBF format) into an Epi Info data table**
- **Use the data table to make a new View for data entry and lookup**
- **Use the FIND button to search for existing patient registrations**
- **Program the View using Check Code so that searching is automatic**

Instructions:

Now that you know where you stand regarding disk space and have an idea how files might be backed up, let's copy the FoxPro file to an Epi Info database and make it work for patient lookup. A sample file called Patients.dbf is provided.

1. **Run the Analysis program and READ the FoxPro format file called Patients.dbf. To do this, you need to know that FoxPro uses the dbf (for dBASE) format. If you are not sure which version of dBASE to specify, try several until something works. Since the file is a few years old, the older versions are probably the best guess, and dBASE IV will work.**

2. **After you have successfully read the file, WRITE it to the Epi Info 2000/Access format, specifying CLINC.MDB as the database and PATIENTS as the table.**
3. **This data table would be fine for analysis, but our goal is to enter data and find patients, so we need to create a View that matches the data table. To do this, run the MakeView program and choose NEW from the FILE menu. Navigate to CLINIC.MDB and open it. You will be asked to name the new View. Use PATIENTS as the View name (If this view already exists, use "Patients1" or another name.) Now you are ready to make the view, but, instead of creating fields, choose from the TOOLS menu the item, MAKE VIEW FROM DATA TABLE. Locate the data table called PATIENTS when asked. Then click the double arrow on the left side of the dialog to select all of the fields so that the field names move over to the right side of the dialog. Click OK and wait until Epi Info makes the new view. The fields are not be in the order you wish, but you can move them around on the screen by left clicking and dragging the prompts (the words preceding the blanks). Move PatientNumber to the top position, followed by First Name and then Last Name. When you have the fields arranged satisfactorily, choose from the FILE menu the ENTER DATA item and use the arrows on the lower left in the ENTER screen to review the data.**

- **Program the View using Check Code so that entering a PatientNumber or first and last name does an automatic search for a matching record in the database**

  1. Return to MakeView and Open the new Patient View within the Clinic.MDB. Click the PROGRAM button in the left panel. We will use the AUTOSEARCH statement in two places to do automatic searches when an item has been entered. In FIELD WHERE ACTION WILL OCCUR, choose PATIENTNUMBER. Then click the AUTOSEARCH command in the tree on the left or the tabs above. Choose PATIENTNUMBER as the field on which to search, then click OK and SAVE. Now choose LASTNAME as the FIELD WHERE ACTION WILL OCCUR and insert another AUTOSEARCH command, but this time, specify that the search will be of the fields FIRSTNAME and LASTNAME. Click OK, then SAVE and then the OK button at the top of the screen.

  2. Test your new functions by entering some data. First enter a PatientNumber that you know to be already in the database. The corresponding record will be listed, and you can choose it by double-clicking on the arrow at the left of its row.

  3. Click NEW and this time do not enter a Patient Number, but enter a first and last name that you know already to be in the database. Again the record will be found. Click CANCEL. Try entering a new Record

Number, but a name, Ali Essa, that is already found in the table. Save the record by clicking NEW, and then enter the First and Last Names as Ali Essa. This time, the system should find two records, and you can choose which one is correct. Try entering a new record for Amina Essa, Ali's wife, and then doing a search by entering only the last name, Essa.

4. Continue these experiments until you understand the possibilities in the context of the reception desk in a busy clinic. Note that you can do the most common searches using the two AUTOSEARCH fields, PatientNumber and LastName, but that you can search on any field or combination of fields using the FIND command in the left panel of ENTER. Experiment with searching by birthdate, for example, or First Name alone.

Problem 5: Computer Viruses

There are computer viruses in at least one of the agency computers, and you find that it is protected by Norton AntiVirus with an expired subscription; the virus definition files were last updated 2 years ago.

Objective: Make a brief plan for yourself and other staff members that will protect one or more computers from viruses. Include:

- **Antiviral software**
- **One or more Firewalls**
- **Policies that will help to avoid risk**
- **Necessary procedures to monitor and maintain a virus-free computing environment**

Suggestions are found on the next page, but don't peek until you have come up with your own.

Suggestions

- Antiviral software
    1. Choose an Internet Service Provider (ISP) that has a good reputation for security, since they must run some kind of security software on their own servers, and there seems to be quite a difference among providers. This does not refer to "freebie" security software for your own computer, but to control of the number of viruses and other threats reaching your Internet connection.
    2. Microsoft Windows does not provide antiviral protection, although many new computers come with trial versions of antiviral software from Symantec (Norton), McAfee, or other sources. Current reviews are available on the Internet by searching in Google.com for "antiviral software review" (without the quotes).
    3. Some programs are free; others require purchase and/or a subscription. All must be updated frequently to cope with current viruses.
    4. Microsoft Windows also offers updates, mostly security related, that can be downloaded automatically. The author prefers to set the updating program so that installation only occurs after a prompt gives the user choices and information.
- One or more Firewalls
    1. A firewall is a barrier to various kinds of hacker attacks. It works by shutting down some of the "ports" in computer connections, and by examining data coming and going and looking for possible problems. To find out more, search the Internet for FIREWALL WORKS. A hardware firewall is normally associated with a router, with which you can connect several computers to your Internet modem or DSL(fast telephone) connection. Some routers also serve as the basis of wireless connections to your own computers (and others in the neighborhood unless you implement secure, encrypted, wireless connections.
    2. A software firewall performs some of the same tasks, and is also recommended, but some also cause problems. Usually they can be turned off temporarily to test whether they are conflicting with other programs. Running two software firewalls on the same computer is not a good idea, and Internet searching will find discussions of problems caused by various products.
- Policies that will help to avoid risk
    1. Email messages from unknown sources should be treated with suspicion, and should preferably not be opened.
    2. Attachments to email should not be opened until their origin is clear. Antivirus software should be updated and enabled to scan attachments. Never open, save, or run an executable attachment with the extension

       .EXE .COM .VBS .LNK .PIF .SCR .BAT

Unless you are sure that you know the sender and he/she verifies intending to send it to you.

3. Never send personal information such as bank account numbers, social security number, or your email address in email messages. If you must do so, conceal some of the digits or the "@" sign by spelling them out or inserting other characters. Putting half the number in each of two emails might help. Do not label a number as "SSN", or "account number."
4. Be alert for "scams" such as helping people get access to large sums of money or chain letters.
5. Avoid pornography, videos, or illegal of questionable activities on a work computer. They open doors to "spam" suppliers and may take large amounts of disk space, for which you may be embarrassed to ask for assistance.
6. Floppy disks or other media from other computers should be carefully scanned for viruses before copying or opening files.
7. A full scan of the computer for viruses should be done with updated antivirus software at least weekly, and, in a public setting, daily.

- Necessary procedures to monitor and maintain a virus-free computing environment
    1. Frequent, and preferably automatic updating of antiviral software and virus definitions. (Weekly)
    2. Frequent, and preferably automatic or prompted updating of the Windows or other operating system. (Weekly)
    3. Weekly or more frequent scanning of the computer for viruses and other threats.
    4. "Epidemiologic" investigation of events that lead up to any virus infections that occur.
    5. Reporting of suspicious events such as strange messages, slow computers, virus-found messages, to a single coordinator who can shut down connections and isolate and disinfect infected computers, as well as consult the Internet for the latest news and fixes for current infections.
    6. Periodic testing of security through use of Internet sites that provide this service. Choose carefully, since some of these may be actually propagating security threats, and many are selling their own software answers.

Problem 6: Backing Up and Restoring Data

There is a large-capacity tape drive in the server for a small local area network, but no one knows how to make it work, and there is no regular system for backup of either the network or the several free-standing computers used for word processing or research.

Objective: Make a plan for regular backup. Include:

- **Type of medium**
- **What will be backed up**

- **Software**
- **Frequency**
- **Storage of backups**
- **Testing of backup to see if it can be restored**
- **Who will do it, and who will check that it is done**

Suggestions are found on the next page, but don't peek until you have come up with your own.

Suggestions

- Type of medium
    - Something that can be read and written to by any computer in your agency (CDROMs may be a good choice, or devices that connect easily via USB ports and do not require extra software).
    - Write-once media (CDR-ROM) have the advantage of not being overwritten during a panic-stricken attempt at restoration.
    - Rewritable media (CDRW, Flash memory, etc) can be rotated so that they are not overwritten until after 4 or 5 backups.
- What will be backed up
    - If you have appropriate "ghosting" software, the entire hard disk on each computer
    - If not, then selected folders containing work, settings, email, and other files. Programs must usually be reinstalled from the original disks or other sources.
- Software
    - Search the Internet for recent reviews of backup software
    - Be sure to try out your choice for backup, restoration, and also ease of use
- Frequency
    - Depends on the volume and importance of the work being done, as well as the likelihood of power failure or other problems.
    - Combinations can be used, e.g., every few hours to flash memory, daily to CDROM, and weekly for larger systematic backups.
- Storage
    - Against theft—not attached to computer or in its case
    - Against fire, earthquake, administrative upheavals—off site, perhaps in a bank vault.
    - Storage of encrypted files on a remote Internet site can also be considered.
- Testing of backup to see if it can be restored
    - Fairly easy for partial backup with test files.
    - For full backup, requires an identical "practice" computer, and may still require serial numbers, keys, and activation of software.
    - Programs for comparing files can verify that accurate copies were made and can be read from the backup medium.
- Who will do it, and who will check that it is done
    - Users should be involved in backup, preferably with the help of automatic prompting and easy menu choices.
    - In an agency, a "system administrator" should provide assistance and monitoring of backups, perhaps once per week. Reminders and confirmations might be done by email after skills have been acquired. Program and virus definition updating can be done at the same time.

- Of course Local Area Networks should be backed up and maintained by appointed staff (but beware of the unique medium/unique drive problem).

Problem 7: Sending Confidential Files

The HIV/AIDS clinic in your agency is located on the other side of the city. They send orders for medication to the agency pharmacy in Excel format, but are afraid to attach them to emails since they contain patient identifiers. They are currently FAXing the orders to the pharmacy, but a secretary from Environmental Health brought you one of the FAXes last week that had been mistakenly sent to her FAX machine through miskeying a single digit in the similar phone numbers. You are grateful that the FAX is within the agency, but wonder how many other confidential FAXes could have gone astray.

Objectives:
Use the EpiLock utility to encrypt the files before sending
Provide Epi Info to the pharmacy so that they can decrypt files received as email attachments

Data sample:
OrdersWk34.xls

Instructions for ENCRYPTING:

1. **In the Analysis program, use READ with the Excel 4.0 Format and LIST to see the contents of the sample file, OrdersWk34.xls.**

2. **Run Epi Lock from the Utilities menu of Epi Info. Choose the ENCRYPT tab. Click the button with three dots and locate the file called OrderWk34.xls. This is the Excel file containing simulated orders for medications with patient names, to be transmitted as an email attachment after encryption. Because we have placed an extra copy in the OriginalOrders directory, check the box called DELETE AFTER ENCRYPTION.**

3. **Now enter a password, and enter the same password again for confirmation. The password must be made known to the recipient so that the file can be decrypted after it is received and saved. Click OK to encrypt and note that the extension .ELH has been added. The H stands for “High” or 128-bit encryption. You now have two files:**
   OrdersWk34.xls and
   OrderWk34.xls.elh

The level of encryption, if 128-bits, is quite secure, and would require efforts usually associated with national defense to "crack." There are other risks that the confidential information could become public, however. What are the most likely ways in which others might gain access to the unencrypted or decrypted file? Consider:

1. Human error
2. Poorly chosen password
3. Password available to the wrong people
4. Change in status of someone who knows the password
5. Is it a good idea to keep a written record of the password?

There is also a chance that the recipient will not be able to decrypt the file due to:

1. Not receiving the attachment
2. Wrong password
3. Lack of decryption program or skills to use it
4. Computer not working

How can the sender know if the message is received or not? Successfully decrypted?

The benefits of privacy must be weighed against the cost of non-receipt. As lapses in HIV/AIDS treatment can lead to resistant virus strains, this is an important concern. What is your "backup" plan if the orders are not received by the pharmacy or the medications are not delivered?

If you have email access, you might try sending the encrypted file to yourself as an attachment. On receipt, save the file to disk as though you were the pharmacy.
If you prefer to skip this step, just imagine that the .ELH file is the one received, and that you are ready to decrypt it as an Excel file once again.

Instructions for DECRYPTING:

1. **Run the EpiLock program and choose the DECRYPT tab. Click the tab with the 3 dots and find the file OrdersWeek34.xls.elh. Enter the password and click OK. The file will be decrypted in the same folder with the .elh file.**

2. **Open the decrypted .xls file in Epi Info as before and verify that the contents are the same. You might compare it with the backup copy in the OriginalOrders folder. It should have exactly the same content and the same file size. If you are from Missouri (the "show me" state), experiment with trying to read the encrypted .elh file in Excel, Notepad, or Epi Info. Even if you are an experienced hacker, it is quite secure against revealing data without the password.**

Extra credit question: If you wanted to encrypt a file so that two different persons had to participate in decryption (like the systems for firing nuclear-armed rockets), how would you do it?

Answer: Encrypt the file first with one password ("CBA31999") and then encrypt the encrypted file with a different password ("YZM44422"). Give one password to each of the two participants in the opening process, and be sure that they have access to the file in the right order ("YZM44422" first). Such hocus-pocus is rarely necessary in the health field, and every added complexity carries the risk of not being able to open the file in practice. We have now doubled the number of people who can lose a vital password!

Problem 8: Adding a New Field and New Data to Existing Views

Several offices of your agency are entering computerized surveillance data in the form of disease reports using Epi Info Views, but the data are sent to your office on paper and must be reentered. You would like to merge the files at the Central office, but the data from the local offices has overlapping ID Numbers and no indication of which office entered the data. You would like to add an AGENCY field to the Surveillance View so that the agency providing the data and the ID Number together will be unique for each case report. After adding the field, you would like to insert a value representing the name of the local agency in each existing record in the data table.

Data samples:

Surveillance.MDB

There is a View called Surveillance with several related Views for specific diseases. We will ignore the latter in this exercise.

Objectives:

- **Add a field called AGENCY to the SURVEILLANCE View with Legal Values of "Central", "Local1", and "Local2". The field should have the property, REPEAT LAST, and both it and the IDNumber field should be REQUIRED fields.**
- **Use Analize Data (Analysis) to insert new data in the AGENCY field for all existing records. We will insert the value, "Central".**

Instructions:

1. **Use the MakeView program to Open Surveillance.mdb and the Surveillance view. Make a larger space after ID Number by clicking and dragging the prompts of other fields. Right click in the space and create a field called Agency. Make it a text field with Legal Values of "Central", "Local1", and "Local2" (without the quotation marks). Check the boxes beside REQUIRED and REPEAT LAST and click OK. In order to create the field in the data table, it is necessary to run the ENTER program. Choose from the FILE menu the ENTER DATA item. When you see the new field in data entry mode, exit from the Enter program.**

2. **Now we have a new field, but it is empty. It is possible to enter values by going back to each existing record and entering, but this is a pain, even for 40 records. Let's use Analysis to insert the data.**

3. **Run Analyze Data from the main Epi Info menu and READ the Surveillance View from Surveillance.mdb. Use LIST to see the contents of the AGENCY field (none).**

You might think that merely ASSIGNing the value "Central" to the variable AGENCY and WRITEing all the variables or just the AGENCY variable would accomplish the desired result. If you try this, however, you get a message that the MERGE command should be used instead. MERGE what? Here is the strategy…

- Make the ASSIGNment as suggested, but WRITE a new table called TEMP1 containing only the AGENCY variable and a unique identifier for each record. Unfortunately MERGE will not allow you to use UniqueKey as the identifier, so you have to use IDNumber. (In a real system, you would go back and provide unique values for IDNumber in the Surveillance data table where these are missing. LIST with Update enabled would be a reasonable way to do this.)
- MERGE the TEMP1 table with the Surveillance View, using IDNumber as the key

And now the dirty details…

1. **Click the ASSIGN command and choose AGENCY as the variable. Insert "Central" as the value on the second line (with quotes). Click OK.**
2. **Click the WRITE command and choose the AGENCY and IDNUMBER fields. Leave the data type as Epi 2000 and the MDB as SURVEILLANCE. Specify TEMP1 as the data table and click OK. The data table will be written, although there is no confirmatory indication.**
3. **Click the MERGE command and locate the TEMP1 table as the source of data to merge. At the bottom of the dialog, type IDNUMBER :: IDNUMBER to specify the key for merging (upper/lower case are not important for variable names. You can also use the Key Builder function, but this is a simple case.) Leave the UPDATE function checked so that the values in TEMP1 will be inserted into corresponding fields in SURVEILLANCE. APPEND does not matter, since there is a one-to-one match of the records in the two tables. Click OK.**
4. **To verify that values for the new field have been permanently merged into the view, READ the SURVEILLANCE view again (thus canceling your temporary ASSIGNment) and use LIST to see the AGENCY field. It should contain the word "Central" for each record that had an IDNumber.**
5. **Whew!!**

Problem 9: Merging Data from Multiple Offices

You want to merge the data from other offices into a Central surveillance database, using both AGENCY and ID Number as the record identifier, since the ID numbers for the reports are only unique within a given agency, and there may be records with the same number from different agencies. One of the local offices has added extra variables for its own use, and no one is sure how to combine the dissimilar data tables.

Data samples:

All have had an AGENCY field added and populated with the name of the Agency that entered the records, using the techniques of the previous exercise.

Central.MDB
Local1.MDB
Local2.MDB

Each has a View called Surveillance and several related Views for specific diseases. We will ignore the latter in this exercise.

Objectives:

- **Merge data from the two local agencies into the Central data table, using both the IdNumber and Agency fields together as a unique identifier for the merge.**
- **Examine the merged data to see what happened to the extra fields in Local1.MDb:Surveillance**
- **Make changes in the data of one of the Local agencies, inserting new or corrected information. Run the merge again, and assess the results. (Showing how Updates can be done automatically.)**

Instructions
Suppose that Central is the main database, and that you wish to merge data from Local1 and Local2, the local offices, into Central. In each case the View is called Surveillance.

1. **Run Analyze Data and READ the Surveillance view in Central.MDB. Note that there are 41 records.**
2. **Click the MERGE command, locate Local1.mdb and its Surveillance view, and then click BUILD KEYS. In each table, you want to use both IDNumber and Agency as keys. To do this, type IDNumber in each of the two larger spaces. The entry boxes behave strangely if you use the Enter key alone, so place the cursor after the word “IDNumber,” and press Enter while holding down the Ctrl key (Ctrl-Enter). This will put the cursor on the next line so that you can type, “Agency” (without quotes). When you have both**

field names in both of the large boxes, click OK. Note that the key expression is constructed as "IDNumber::IDNumber AND Agency::Agency". Click OK to do the merge. How many records are there now in the Central database?

3. **The same process will work to merge Local2's records, but it is easier to copy the MERGE command in the program window and change Local1 to Local2 in the command. Do this by selecting the line(s) of the Merge command and then copying with Ctrl-C. Place the cursor on the next free line, and press Ctrl-V to paste the statement. Change Local1 to Local2 in the command and then click RUN THIS COMMAND to do the merge.**

Local1:Surveillance contains three more fields than the other two Views—FollowupBy, Result, and DateofFU. Find out if these were merged with the Central database by looking for the fields in a LISTing. What would happen if some of the fields in Central were missing from the View in Local1?

1. **The program window should now contain a short program that merges both Local1 and Local2 data with the data in Central. Click the SAVE button, choose TEXT FILE, and then save the file with the name WEEKLYMERGE.PGM.**

2. **Make some changes in Local1 data using Enter Data. Add a new record and amend some of the existing ones. Make the changes distinctive or write down what you have changed. Let's pretend that a week has passed and that Local1 has now sent you another copy of it's MDB. Run WEEKLYMERGE.PGM again and use LIST to see what happened with the new data. Note that submitting the same records twice did no harm, since they match and only changes cause updates. Assure yourself that the corrections were inserted and correctly overwrote the older entries.**

3. **You have the beginnings of a functional 3-office surveillance system, although it will need further refinement for serious use.**

Advanced Statistics

Introduction

Having solved some of the data management problems, and feeling more confident that you can import almost any file type into Epi Info, you turn your attention to statistics.

The following challenges arise:

1. You need to know (immediately, of course) if 4 cases of Salmonella enteriditis and 3 cases of Listeriosis are statistically in excess of normal rates for the population for which you are responsible.

2. A local non-profit organization has done a community survey of prenatal care and immunization status using the WHO cluster survey technique. They ask your help with the analysis.

3. You agree to teach public health computing in a statistics course at a local community college. Using the Oswego exercise, you successfully lead the class through analysis of the epidemic and identify vanilla ice cream as the most likely risk factor for illness. Two of the advanced students, however, have additional questions about the protective effect of chocolate ice cream and want to explore the statistics further.

4. Your teen-aged daughter is taking an advanced placement course in Statistics in her high school and wants to use Logistic Regression to analyze a classical heart disease dataset from Evans County. You agree to help.

5. A local hospital is participating in a national drug trial. Researchers there want to use the Kaplan-Meier and Cox Proportional Hazards statistics in Epi Info to analyze their own data, and ask you how to find instructions for using these routines. You do not have time to participate, but refer them to a manual that gives examples.

Statistics 1: Obtaining Historical Data

You need to know (immediately, of course) if 4 cases of *Salmonella enteriditis* and 3 cases of Listeriosis are statistically in excess of normal rates for the population for which you are responsible.

Objective:
Compare case rates for the two infections with previously reported rates from the same geographic area, using the Internet to find both the population and the previous rates. Use the Compare Rates program in OpenEpi to see if the difference is statistically significant.

Data sample: To be found on the Internet using Google.com and/or a MEDLARS search of the National Library of Medicine bibliographic database. The OpenEpi program has links to the two search engines.

Instructions:

1. Run the OpenEpi program either from the hard disk or from www.openepi.com. If there are problems you may have to adjust the security settings in your browser and/or popup prevention, as OpenEpi requires both running JavaScript to do the statistics and popup windows to enter data or present results.

2. In the Net Links section of the menu are items for Internet Search (using Google.com) and Medline Search (using Pub Med). Using these two items, perform searches to find case rates for Salmonella enteriditis and for Listeriosis. While doing so, find information about these two organisms and the infections they cause, unless you are already an expert in this area.

3. (If the Internet is not available, choose a geographic area and use the rates from the .HTM and .PDF pages in the class exercises, in the exStatistics\InternetSearch\ folder. General information on the two organisms is also found in .HTM files in this folder.)

4. Similarly, use Google to find the population of the area for which you are responsible.

5. As a first test, use both sexes and all ages as the population at risk, although you might later wish to deal with specific age groups or sexes if all the cases fall into a particular category. If you do not have Internet access, you can obtain the population from more local sources, such as the public library or town clerk's office.

6. Use the Compare-2-Rates program in OpenEpi to compare the reported rates from your search with the calculated rates for your 4 and 3 recent cases. You might consider the previous rates to be the "Unexposed" population and the current rates to be the "Exposed" group. "Person-time" can be the population multiplied by the number of years for which the number of cases is reported—often 1 year. Is there a significant difference in the two rates?

Statistics 2: Analyzing a Community Cluster Survey

A private non-profit agency has done a survey of immunization rates and the influence of prenatal care on immunization, using the World Health Organization cluster survey technique. They ask your advice in doing the analysis, since they have heard that Sudaan, the recommended software is rather expensive and takes a while to learn.

Data sample:
Sample.mdb in the Epi_Info directory
Epi1 and Epi10 datasets. These are based on cluster samples with fields for Primary Sampling Unit (PSU). Epi10 also has geographic (perhaps ZIP code) Strata, and Weights to adjust for the sampling ratio and perhaps other factors.

Objective:
Determine rates of immunization among children in the cluster survey for those with prenatal care and those who did not have prenatal care.
Determine if there is a significant association between prenatal care and immunization rate

Instructions: These are taken from:
Introduction to Epi Info (Version 3.2.2) Analysis Module
Kevin M. Sullivan, PhD, MPH, MHA and Minn Minn Soe, MD, MCTM, MPH
Department of Epidemiology
School of Public Health of Emory University
October 2004
Available from www.sph.emory.edu/~cdckms/

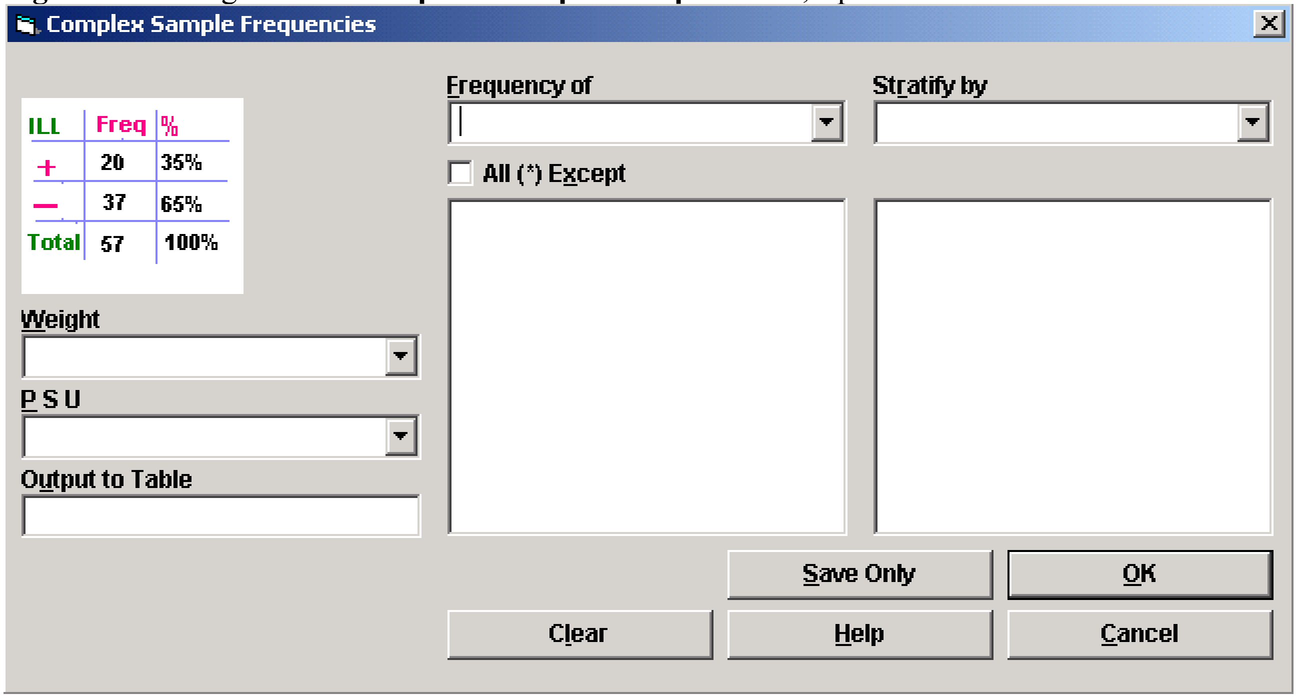
[Additional comments or omissions […] are marked by square brackets]

Complex sampling saves time in the data collection phase of a survey because those interviewed do not have to be randomly scattered through the population, but can be located in clusters that are randomly selected. Imagine the cost of traveling to 210 random points in a large area to interview one person, versus going to 30 points and interviewing 7 persons at each site. The latter is called a *cluster* survey and each randomly-selected cluster is called a Primary Sampling Unit (PSU). If the survey is designed to provide separate results for several geographic areas, it is said to be *stratified*. *Weights* are assigned to correct for differences in population, response rate, or other factors among the strata.

Complex Sample Frequencies

Like the **Frequencies** command, the **Complex Sample Frequencies** provides the frequency of a variable. The dialog box for **Complex Sample Frequencies** is shown in Figure 60.

Figure 60. Dialog box for **Complex Sample Frequencies**, Epi Info.



To use the **Complex Sample Frequencies** command in this example, read the file **viewEpi1** which can be found in the **Sample.mdb** file. These data are based on an Expanded Program for Immunization (EPI) cluster survey (see Appendix 1 for the details of this survey). In general, the EPI method selects 30 communities (i.e., clusters) from a selected geographic area, a survey team then visits each of the 30 communities from which seven children in an appropriate age range are selected and each child's immunization status is determined.

Complete the dialog box as follows and click **OK** button. The results are shown in Figure 61.

Frequency of `VAC`
PSU `CLUSTER`

Figure 61. Example output from **Complex Sample Frequencies**, Epi Info.

```
VAC             TOTAL
1                 155
          Row %   100.000
          Col %   73.810
           SE %   4.599
          LCL %   64.795
          UCL %   82.824
2                  55
          Row %   100.000
          Col %   26.190
           SE %   4.599
          LCL %   17.176
          UCL %   35.205
TOTAL             210
  Design Effect   2.298 Sample Design Included:
Weight Variable: None
PSU Variable: CLUSTER
Stratification Variable: None
0 records with missing values
```

`VAC` denotes whether a child is vaccination or not; 1=Yes and 2=No

To receive all of the output as shown in Figure 61, please make sure the **Set** command, **Statistics** option is set to **Advanced**. Information provided in the output includes:

Row %	For a frequency will always be 100%
Col %	The column percent; in the above example, 73.8% were vaccinated
SE %	The standard error, which takes into account the complex sample design
LCL %	Lower Confidence Limit
UCL %	Upper Confidence Limit
Total	Total number of individuals/elements surveyed

[Design Effect is the ratio of the variance for the complex design divided by variance assuming simple random sampling. Think of it as the ratio by which the sample must be increased to achieve the same precision as a random sample.]

Additional information is provided on the **Sample Design Included** at the bottom of the output. The interpretation of the results in Figure 61 would be that 73.8% (155/210) of the children were vaccinated with 95% confidence limits of (64.8%, 82.8%). Note that had the **Frequencies** command been used and therefore ignoring the cluster design, the proportion immunized would also be 73.8% but the confidence interval would be too narrow (67.3%, 79.6%).

As another example using the **Complex Sample Frequencies**, read the **viewEpi10** file in **Sample.mdb**. These data are similar to viewEpi1 except that this is a stratified cluster survey with a separate 30 cluster survey completed in each of 10 strata […]. As in **viewEpi1**, there is a variable for whether or not a child is vaccinated `(VAC,1=yes,2=no)`. To correctly analyze this data set, we need to take into account the variable for which stratum each child lives (`LOCATION`) and a variable to add statistical weights to account for differences in population sizes between strata (`POPW`). Complete the dialog box as follows and the results are presented in Figure 62.

Frequency of	`VAC`
Stratify by	`LOCATION`
Weight	`POPW`
PSU	`CLUSTER`

Figure 62. A second example output from **Complex Sample Frequencies**, Epi Info.

VAC	TOTAL
1	1242
Row %	100.000
Col %	55.263
SE %	2.620
LCL %	50.128
UCL %	60.398
2	910
Row %	100.000
Col %	44.737
SE %	2.620
LCL %	39.602
UCL %	49.872
TOTAL	2152
Design Effect	5.975

Sample Design Included:
Weight Variable: POPW
PSU Variable: CLUSTER
Stratification Variable: LOCATION

0 records with missing values

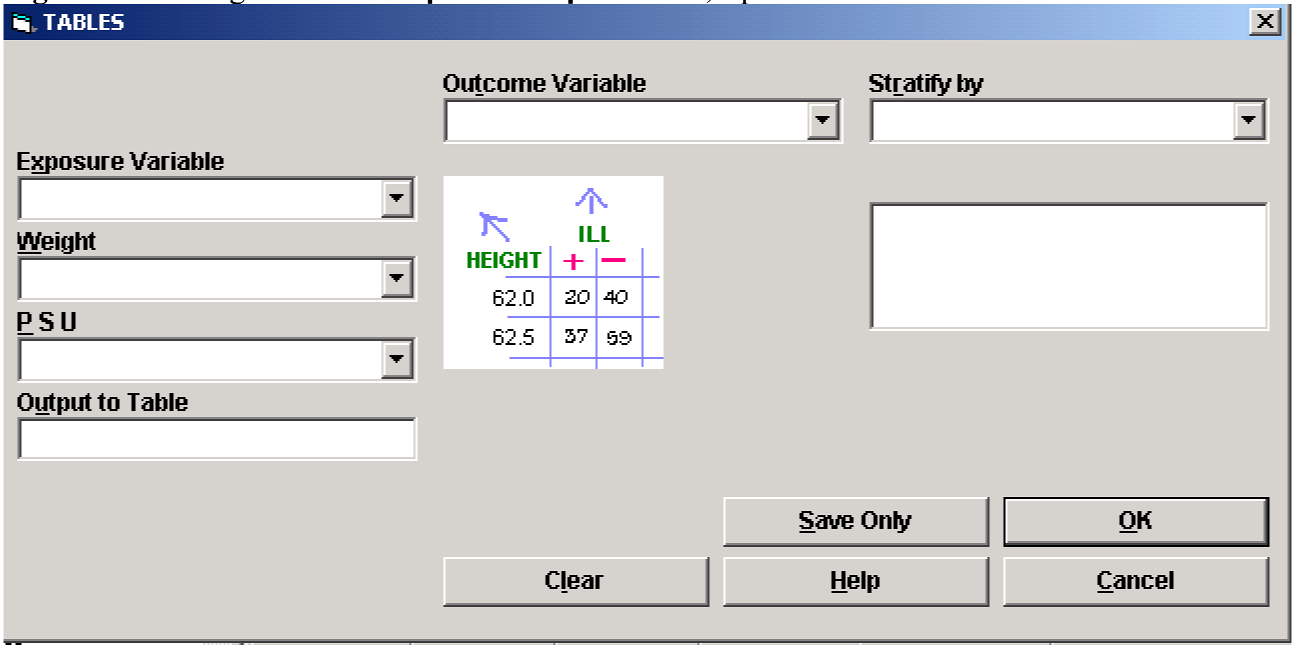
The interpretation for Figure 62 would be that the overall estimate of the percentage of children vaccinated is 55.3% with 95% confidence limits (50.1%, 60.4%) taking into account the stratification and population weights.

Complex Sample Tables

The **Complex Sample Tables** command is similar to the **Tables** command in that you specify a row and column variable. The dialog box for this command is shown in Figure 63. Using the **viewEpi10** data, let's analyze the data using whether or not the mother received prenatal care `(PRENATAL)` as a predictor of the child's vaccination status. If [the child] had received prenatal care, then `PRENATAL=1;`otherwise `PRENATAL=2.` The dialog box should be completed as follows:

Outcome Variable	`VAC`
Stratify by	`LOCATION`
Exposure Variable	`PRENATAL`
Weight	`POPW`
PSU	`CLUSTER`

Figure 63. Dialog box for **Complex Sample Table**, Epi Info.



Note inconsistency between the command name **Complex Sample Table** and the dialog box name **TABLES**.

The results are shown in Figure 64 and while they appear similar to the output for the **Complex Sample Frequencies** command, there are a number of important differences. First, with the goal of assessing whether or not children whose women had received prenatal care were more likely to be immunized compared to those with mothers who had not received prenatal care, the important proportions are the Row% in the first column. Among children whose mothers had received prenatal care, 60.7% were immunized compared to 42.3% among those whose mothers did not receive prenatal care. The confidence limits (LCL and UCL) are for the Row% values.

Estimates of the odds ratio, risk ratio, and risk difference are provided for 2x2 tables. In order to assure that these parameters are estimated correctly, the table setup must be the same as described for the **Tables** command (i.e., outcome as the column variable, exposure as the row variable, etc.). Note that complex sample designs are most frequently applied to cross-sectional data and that cross-sectional surveys usually estimate prevalence, ***not*** risk. Therefore, in many situations the correct names for the epidemiologic parameters would be the prevalence odds ratio, the prevalence ratio, and the prevalence difference.

Figure 64. Partial output from **Complex Sample Table**, Epi Info.

PRENATAL	**VAC**		
	1	**2**	**TOTAL**
1	675	413	1088
Row %	60.734	39.266	100.000
Col %	76.817	61.349	69.897
SE %	3.375	3.375	
LCL %	54.118	32.650	
UCL %	67.350	45.882	
Design Effect	5.198	5.198	
2	567	497	1064
Row %	42.560	57.440	100.000
Col %	23.183	38.651	30.103
SE %	2.414	2.414	
LCL %	37.828	52.708	
UCL %	47.292	62.172	
Design Effect	2.537	2.537	
TOTAL	1242	910	2152
Row %	55.263	44.737	100.000
Col %	100.000	100.000	100.000
SE %	2.620	2.620	
LCL %	50.128	39.602	
UCL %	60.398	49.872	
Design Effect	5.975	5.975	

```
CTABLES COMPLEX SAMPLE DESIGN ANALYSIS OF 2 X 2 TABLE

Odds Ratio (OR) 2.088
Standard Error (SE) 0.307
95% Conf. Limits (1.50, 2.901 )

Risk Ratio (RR) 1.427
Standard Error (SE) 0.110
95% Conf. Limits (1.23, 1.660 )
RR = (Risk of VAC=1 if PRENATAL=1) / (Risk of VAC=1 if PRENATAL=2)

Risk Difference (RD%) 18.174
```

```
Standard Error (SE%) 4.021
95% Conf. Limits (10.26, 26.089 )
RD = (Risk of VAC=1 if PRENATAL=1) - (Risk of VAC=1 if PRENATAL=2)
Sample Design Included:
Weight Variable: POPW
PSU Variable: CLUSTER
Stratification Variable: LOCATION

 records with missing values: 0
```

The prevalence odds ratio in the example data in Figure 64 is 2.088, the prevalence ratio is 1.427, and the prevalence difference is 0.182 or 18.2%. The interpretation of the prevalence ratio is that children born to women who had received prenatal care are 1.4 times more likely to be immunized (60.7%/42.3%).

Statistics 3: Finding Relationships among Three or More Variables

You agree to teach public health computing in a statistics course at a local community college. Using the Oswego exercise, you successfully lead the class through analysis of the epidemic and identify vanilla ice cream as the most likely risk factor for illness. Two of the advanced students, however, have additional questions about the protective effect of chocolate ice cream and want to explore the statistics further.

Data sample: SAMPLE.MDB:viewOswego in the Epi_Info directory

Objective: Explore the relationships between ILL, CHOCOLATE, and VANILLA in the Oswego dataset

Instructions:

1. **READ the dataset.**
2. **Do TABLES of VANILLA by ILL**
3. **Do TABLES of CHOCOLATE by ILL**
4. **Do TABLES of CHOCOLATE by VANILLA**
5. **Are chocolate and vanilla associated with illness? Positively or negatively?**
6. **Is chocolate associated with vanilla?**
7. **What is the epidemiologic definition of a confounder?**
8. **Do TABLES of VANILLA by ILL stratified by CHOCOLATE. Observe the difference between the crude and adjusted risk ratios. Is CHOCOLATE a confounder for the VANILLA ILL association?**
9. **Do TABLES of CHOCOLATE by ILL stratified by VANILLA. Observe the difference between the crude and adjusted risk ratios. Is VANILLA a confounder for the CHOCOLATE ILL association?**
10. **Is there interaction in the CHOCOLATE ILL relationship? In the VANILLA ILL relationship?**
11. **Using Logistic Regression, with ILL as the outcome variable and CHOCOLATE and VANILLA as the other variables, which one of the two seems to be strongly associated with illness?**
12. **What overall conclusion do you draw about the relationship of CHOCOLATE and VANILLA?**

Answers are on the next page, but don’t peek until you have completed the exercise.

2 **Do TABLES of VANILLA by ILL**
   a. The Risk Ratio(RR) of 5.57 and very small mid-p exact value are evidence of a significant positive association.

3. **Do TABLES of CHOCOLATE by ILL**
   a. This time there is a modest negative or protective association, with RR=0.72 and mid-p exact of 0.04, although the confidence limits for RR include 1.0.

4. **Do TABLES of CHOCOLATE by VANILLA**
   a. The RR of .64 with confidence limits excluding 1.0 and the mid-p value of 0.00099 show a negative association**.**

5. **Are chocolate and vanilla associated with illness? Positively or negatively?**
   a. See above

6. **Is chocolate associated with vanilla?**
   a. See above

7. **What is the epidemiologic definition of a confounder?**
   a. "A condition or variable that is both a risk factor for disease and associated with an exposure of interest. This association between the exposure of interest and the confounder (a true risk factor for disease) may make it falsely appear that the exposure of interest is associated with disease."
      www.epa.gov/iris/gloss8.htm

8. **Do TABLES of VANILLA by ILL stratified by CHOCOLATE. Observe the difference between the crude and adjusted risk ratios. Is CHOCOLATE a confounder for the VANILLA ILL association?**
   a. Since the crude and adjusted risk ratios (and odds ratios) are within 10% of each other, we would conclude that CHOCOLATE is not a confounder in the sense that stratification removes the confounding.
   b. We note, however that the zero cell in the CHOCOLATE=NO stratum makes it impossible to calculate the RR and OR, but does not remove the strong association that can be observed from the counts. Unfortunately, neither the RR nor the OR can be calculated when associations are perfect (all those with the factor were ill and none of those without the factor were ill). Statistics seem to be most useful in imperfect situations!

9. **Do TABLES of CHOCOLATE by ILL stratified by VANILLA. Observe the difference between the crude and adjusted risk ratios. Is VANILLA a confounder for the CHOCOLATE ILL association?**
   a. According to the criteria we set out, stratifying on VANILLA removes the relationship between CHOCOLATE and ILL, and both the OR and RR are adjusted to 1.0, changing by more than 10% in the process.

10. **Is there interaction in the CHOCOLATE ILL relationship? In the VANILLA ILL relationship?**
   a. What is interaction, also known as Effect Modification? In these cases it would be a significant difference in the strength of an association (RR or OR) in the different strata. With CHOCOLATE ILL stratified by VANILLA, interaction would mean significantly different OR or RR values for the VANILLA = YES table and the VANILLA=NO table.

Unfortunately, since one of the tables in each case has an unknown OR and RR (zero in the denominator), this determination cannot be made.

11. **Using Logistic Regression, with ILL as the outcome variable and CHOCOLATE and VANILLA as the other variables, which one of the two seems to be strongly associated with illness?**
    a. The odds ratios of 23 and 1 are quite clear. When controlled for VANILLA, the protective effect of CHOCOLATE disappears.
12. **What overall conclusion do you draw about the relationship of CHOCOLATE and VANILLA?**
    a. What we can say is that there is a strong relationship between VANILLA and ILL whether CHOCOLATE was consumed or not.
    b. When Tables are made of CHOCOLATE by ILL, there appears to be a protective effect, which disappears when the tables are stratified by VANILLA.
    c. Biologically, it makes sense that eating chocolate ice cream might reduce the quantity of vanilla consumed by the same person, either to zero or to a lesser quantity.
    d. Logistic regression seems to give the clearest results, with VANILLA clearly the strong risk factor, and CHOCOLATE, when controlled for VANILLA consumption, a neutral factor with an OR close to 1.0.

Statistics 4: Logistic Regression and a Classical Heart Disease Study

Your teen-aged daughter is taking an advanced placement course in Statistics in her high school and wants to use Logistic Regression to analyze a classical heart disease dataset from Evans County. You agree to help, again using the excellent instructions written by Drs. Kevin Sullivan and Minn Minn Soe at Emory University.

Logistic Regression

Epi Info can perform either ***un***conditional logistic regression for ***un***matched case-control, cross-sectional, cohort, and randomized clinical trial study designs, or conditional logistic regression for matched case-control study designs. The outcome variable for this command must be dichotomous, i.e., either the individual had the outcome of interest or they did not. Also note that with the current version of Epi Info, the outcome variable must be a “Yes/No” type variable rather than 1/0 coding. Predictor variables can be categorical (2 or more categories) or continuous. The dialog box for logistic regression is shown in Figure 49. First unconditional logistic regression will be described followed by conditional logistic regression.

Sample Data: Sample.mdb:viewEvansCounty - Evans County Heart Disease Study Data

The data are based on the Evans County heart disease study concerning the seven-year incidence of coronary heart disease among a cohort of 609 white males. The variable CAT (endogenous catecholamine level) was fabricated for illustrative purposes and dichotomized into categories "high" (the top quintile of cohort values) and "low." There are no missing values in this dataset. Thanks to Dr. David Kleinbaum for making the data available.

Reference: Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and quantitative methods. Lifetime Learning Publications, Belmont, California, 1982.

File name: viewEvansCounty Project: Sample.mdb Number of records: 609

Variable	Label	Values/Description	Freq
Identification Number	**ID**	Range: 21-19161	

Coronary Heart Disease	**CHD**	No = not a case
		Yes = case
			538
			71

Age (years)	**AGE**	Range: 40-76 Mean: 53.71 SD: 9.26	

Catecholamine Level	**CAT**	No = low
		Yes = high
			487
			122

Serum Cholesterol (mg/100 mL)	**CHL**	Range: 94-357 Mean: 211.74 SD: 39.83	
Diastolic Blood Pressure (mmHg)	**DBP**	Range: 60-170 Mean: 91.18 SD: 14.50	

Electrocardiogram	**ECG**	No = normal ECG
		Yes = abnormal ECG
			443
			166

Hematocrit (percent)	**HEM**	Range: 29-58 Mean: 46.26 SD: 3.47	

Marital Status	**MAR**	No = not married
		Yes = married
			64
			545

Occupation	**OCC**	1 = ? 2 = ?	365 244
Pulse (beats/min)	**PLS**	Range: 45-120 Mean: 74.59 SD: 12.67	
Quetelet Index*	**QTI**	Range: 2.121-6.041 Mean: 3.62 SD: 0.59	
Systolic Blood Pressure (mmHg)	**SBP**	Range: 92-300 Mean: 145.48 SD: 27.50	
Socioeconomic Status (McGuire-White index)	**SES**	Range: 20-84 Mean: 57.86 SD: 13.62	

Cigarette Smoking	**SMK**	No = never smoked
		Yes = smoker
			222
			387

Age Group 1 (Years)	**AGEG1**	No = LT 55
		Yes = GE 55
			358
			251

Age Group 2 (Years)	**AGEG2**	1 = 40-44 2 = 45-49 3 = 50-54 4 = 55-59 5 = 60-64	109 138 111 92 63

		6 =	65-69	52
		7 =	70+	44
Cholesterol Group	**CHLG**	No =	LT 250	504
		Yes =	GE 250	105
QTI Group	**QTIG**	No =	LT 3.57	306
		Yes =	GE 3.57	303
SES Group	**SESG**	No =	GE 57	330
		Yes =	LT 57	279
Hypertension	**HPT**	No = SBP<160 & DBP<95		354
		Yes = SBP>159 or DBP>94		255

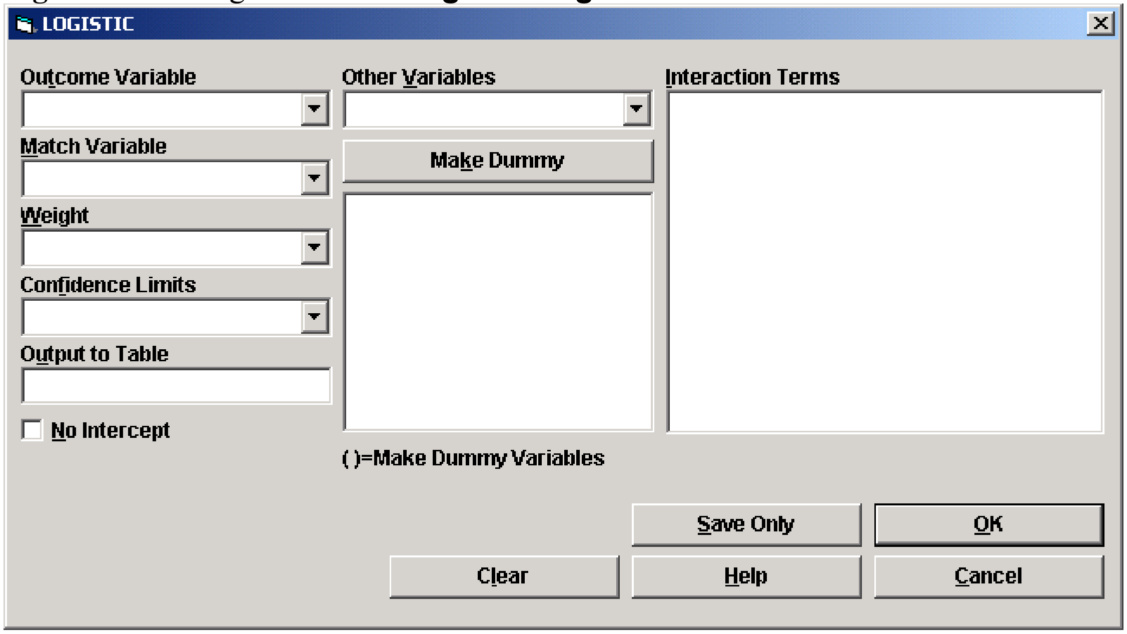
GE=greater than or equal to; LT=less than
*100[(weight in pounds)/(height in inches)]

Unconditional Logistic Regression

First, let's perform an unconditional logistic regression using the cohort study data **viewEvansCounty,** the outcome variable being CHD (Coronary Heart Disease) and the primary exposure variable CAT (catecholamines—not the feline sort). Complete the logistic regression dialog box as follows and click on the **OK** button:

Outcome Variable: `CHD`
Other Variables: `CAT`

Figure 49. Dialog box for the **Logistic Regression** command



Note slight inconsistency between the command name **Logistic Regression** and the dialog box name **LOGISTIC**.

The results for this simple analysis are presented in Figure 50. The odds ratio in Figure 50 could be described as a "crude" odds ratio because the model does not control for any other variables. Say the investigator wants to assess whether another variable (i.e., a "third" variable) modifies or confounds the relationship between `CAT` and `CHD`. As an example, use the `ECG` variable (electrocardiogram results). To determine if `ECG` modifies the `CAT -> CHD` relationship, we need to create an interaction term which can be done using the dialog box. Using the dialog box, do the following:

Outcome Variable: `CHD`
Other Variables: `CAT`
Other Variables: `ECG`

Figure 50. Example output for unconditional logistic regression

```
Unconditional Logistic Regression

Term              Odds Ratio   95%     C.I.   Coefficient  S. E.  Z-Statistic  P-Value
CAT (Yes/No)          2.8615 1.6878 4.8513        1.0513 0.2693       3.9033   0.0001
CONSTANT                   *      *      *       -2.3094 0.1581     -14.6103   0.0000
Convergence:              Converged
Iterations:                       5
Final -2*Log-Likelihood:   424.4271
Test               Statistic D.F. P-Value
Score                16.2465    1  0.0001
Likelihood Ratio     14.1312    1  0.0002
```

The variables `CAT` and `ECG` should appear in the middle of the dialog box; click on each variable to highlight them. Note that ***after*** highlighting the two variables, the button above these two variables will say **Make Interaction**. Click on this button and in the right side of the dialog box below where it says **Interaction Terms** you should see **CAT*ECG,** the interaction term. Click on the **OK** button and the results will be as shown in Figure 51. To determine whether or not there is a statistically significant interaction between `CAT` and `ECG`, use the P-value for the `CAT*ECG` interaction term, in this example, p=0.4196, which would lead to the conclusion that there is no statistically significant interaction.

Figure 51. Example output for unconditional logistic regression with an interaction term, Epi Info.

```
Unconditional Logistic Regression

Term                           Odds Ratio   95%     C.I.   Coefficient  S. E.  Z-Statistic  P-Value
CAT (Yes/No)                       3.0743 1.4002 6.7502        1.1231 0.4013       2.7988   0.0051
ECG (Yes/No)                       1.7278 0.8523 3.5027        0.5469 0.3605       1.5168   0.1293
CAT (Yes/No) * ECG (Yes/No)        0.6276 0.2025 1.9452       -0.4658 0.5771      -0.8071   0.4196
CONSTANT                                *      *      *       -2.4314 0.1844     -13.1868   0.0000

Convergence:              Converged
Iterations:                       5
Final -2*Log-Likelihood:   422.2477
Test               Statistic D.F. P-Value
Score                18.1738    3  0.0004
Likelihood Ratio     16.3106    3  0.0010
```

[Note that neither the p-value nor the confidence limits for the CAT*ECG variable suggest that it has an odds ratio different from 1.0.]

With no statistically significant interaction, the next question would be whether `ECG` ***confounds*** the `CAT -> CHD` relationship. To determine this, run another model with:

Outcome Variable: `CHD`
Other Variables: `CAT`
Other Variables: `ECG`

This time do not create an interaction term; press the **OK** button to run the model. The output from this model is shown in Figure 52. The odds ratio for `CAT` is 2.4483; this would be interpreted as the odds ratio for the `CAT -> CHD` association *controlling* for `ECG`. The crude odds ratio for the `CAT -> CHD` association was 2.86, and, controlling for `ECG`, the adjusted OR is 2.45. Are these values different enough to say that `ECG` confounds the `CAT -> CHD` association? One approach is to use the following formula; if the crude and adjusted estimates vary by some value, say 5% or 10%, we could conclude that there is important confounding:

$$\frac{\left|\hat{\mathrm{OR}}_{\mathrm{crude}} - \hat{\mathrm{OR}}_{\mathrm{adjusted}}\right|}{\hat{\mathrm{OR}}_{\mathrm{adjusted}}} x100$$

In this example, by applying the formula, we find a value of 18%, and therefore conclude that `ECG` is a […] confounder of the `CAT -> CHD` association in this population.

Figure 52. Example output for unconditional logistic regression to assess for confounding, Epi Info.

```
Unconditional Logistic Regression

Term            Odds Ratio  95%     C.I.    Coefficient  S. E.  Z-Statistic  P-Value
CAT (Yes/No)        2.4483 1.3677 4.3828       0.8954 0.2971       3.0139    0.0026
ECG (Yes/No)        1.4393 0.8147 2.5427       0.3641 0.2904       1.2540    0.2098
CONSTANT                 *      *      *      -2.3860 0.1723     -13.8455    0.0000
Convergence:            Converged
Iterations:                     5
Final -2*Log-Likelihood:  422.8874
Test              Statistic D.F. P-Value
Score               17.8952    2  0.0001
Likelihood Ratio    15.6709    2  0.0004
```

Statistics 5: Followup Study of a Clinical Drug Trial—Survival Analysis

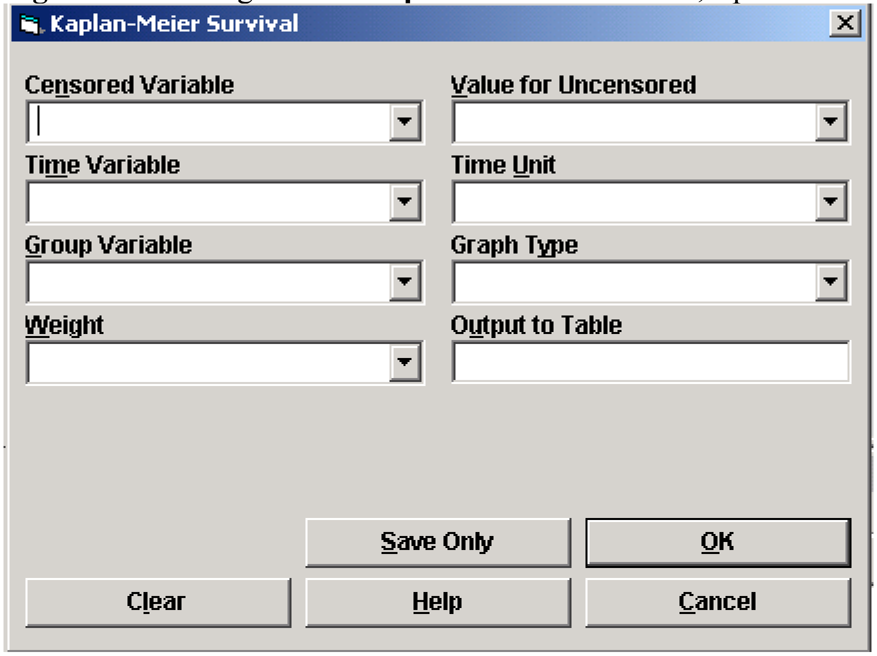
A local hospital is participating in a national drug trial. Researchers there want to use the Kaplan-Meier and Cox Proportional Hazards statistics in Epi Info to analyze their own data, and ask you how to find instructions for using these routines. You do not have time to participate, but refer them to a manual that gives examples. Again Drs. Sullivan and Soe come to the rescue, and their text is reproduced below.

There are two different commands that can perform a survival analysis in Epi Info, **Kaplan-Meier survival** and **Cox proportional hazards**. Each is described in the following pages.

Kaplan-Meier Survival

The Kaplan-Meier method is used for simple survival analysis of censored data in a longitudinal follow-up study. Epi Info can display a Kaplan-Meier survival function curve for one or more groups. The dialog box for the **Kaplan-Meier Survival** command is shown in Figure 54. In general, there is one variable for whether or not an individual developed an outcome or event, a variable for how long each individual was followed (person-time), and when comparing two or more groups, a "group" variable.

Figure 54. Dialog box for **Kaplan-Meier Survival**, Epi Info.



Censored Variable is for the name of the variable that contains information as to whether or not an individual has developed an event during the study, and **Value for Uncensored** is the code identifying those who ***developed*** the event (a little confusing). The **Time Variable** is the follow-up time for an individual until an event (success/failure) occurs or the subject is "censored", i.e., they did not have the event of interest while being followed. **Time Unit** is optional and allows the user to specify the unit of follow-up time, such as hours, days, weeks, months, or years. A **Group Variable** must be provided and must be categorical (1 or >1 categories). **Graph Type** is optional with a default survival probability plot.

As an example, let's perform a Kaplan-Meier Survival analysis using a data set from a clinical trial of leukemia patients, named **Anderson** within the file **Sample.mdb**. Complete the Kaplan-Meier dialog box as follows and click the **OK** button.

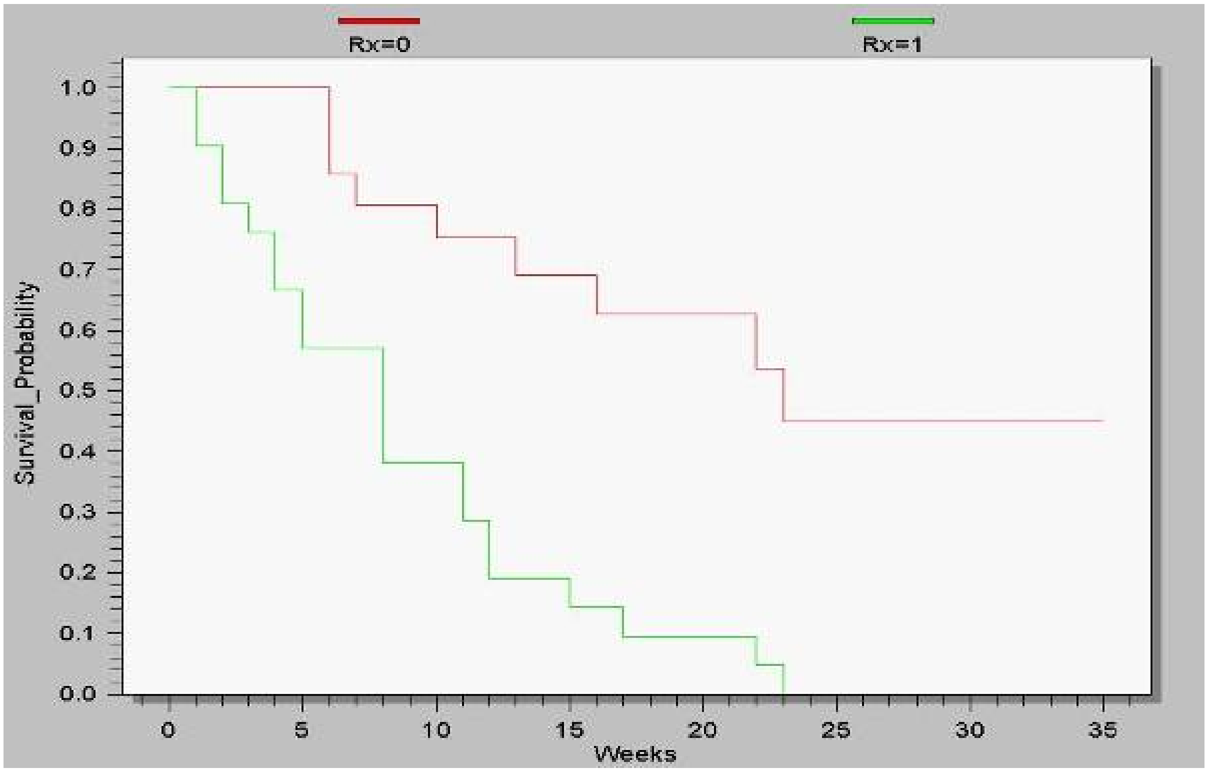
Censored Variable	`STATUS`	**Value for Uncensored**	`1`
Time Variable	`Stime`	**Time Unit**	`Weeks`

Group variable `Rx` **Graph Type** `Survival Probability`

(Note: for the treatment/grouping variable `Rx`, 0 = treatment group, 1 = placebo group.)

The results are shown in Figure 55. The Kaplan-Meier Survival curves are a visual comparison between treatment and placebo groups. A small table at the bottom of Figure 55 provides information on two statistical tests for comparing the survival curves: the **Log-Rank test** and **Wilcoxon test**. Based on these tests, for this example, we would conclude that the treatment and placebo groups have statistically significant different survival curves with the treated group having a longer survival.

Figure 55. Example Plot of Kaplan-Meier Survival curves, Epi Info

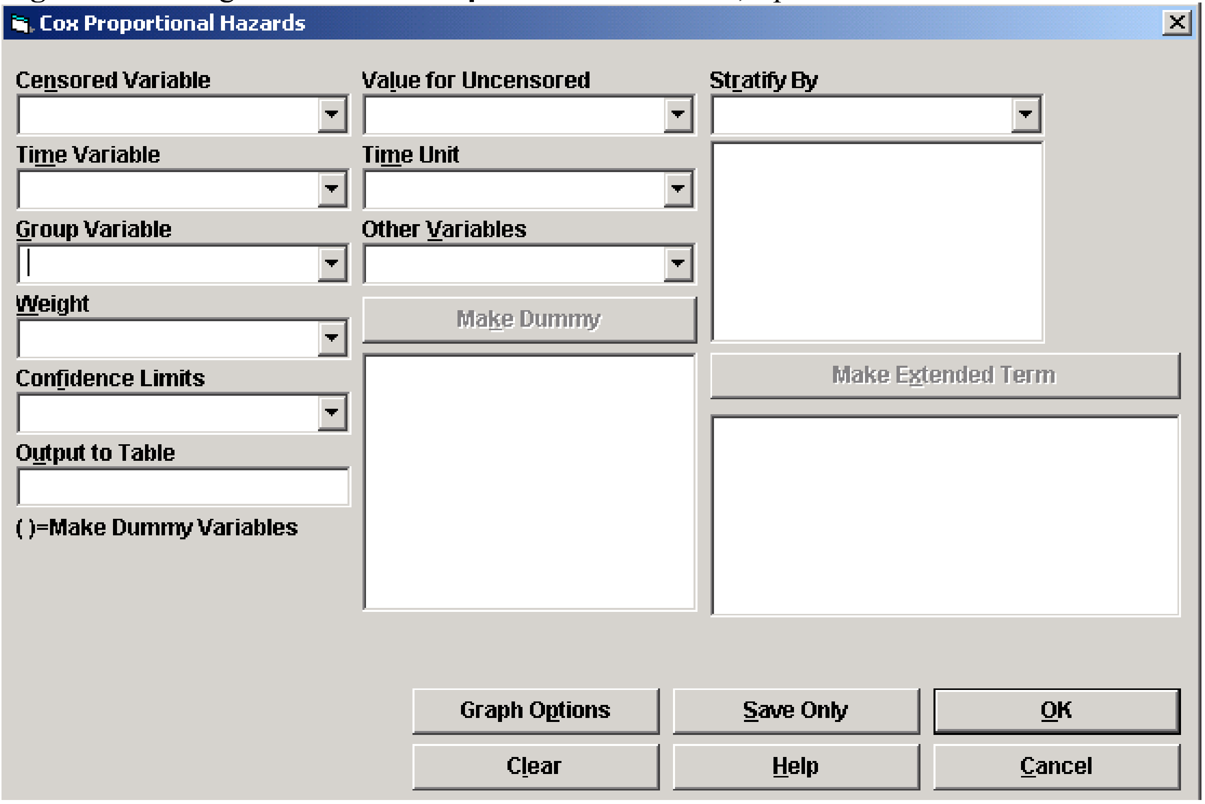


Test	Statistic	D.F.	P-Value
Log-Rank	16.7929	1	0.0
Wilcoxon	13.4579	1	0.0002

Cox Proportional Hazards

The **Cox proportional hazards** model is used with the assumption that predictor variables are time-independent, that is, the values of a given individual do not change over time (e.g., race, sex, country of origin). Cox proportional hazards model is more powerful than Kaplan-Meier Survival approach in the sense that Cox model not only compares the groups in terms of hazard ratio, but can also assess whether other variables modify or confound the relationship between the main predictor variable and time to event. The dialog box for **Cox proportional hazards** model is shown in Figure 56.

Figure 56. Dialog box for **Cox Proportional Hazards**, Epi Info.



The variable options are more or less the same as the **Kaplan-Meier Survival** procedure: **Censored Variable**, **Value for Uncensored, Time Variable, Time Unit,** and **Group Variable**. The **Graph Options** button can be used to hide or display different forms of unadjusted Kaplan-Meier Survival curves in the output screen. Its default is the plot of survival probability.

Let's perform a simple Cox proportional hazards model using the same data set Anderson from a clinical trial of leukemia patients, in order to compare the treatment and placebo group survival. Complete the Cox proportional hazards dialog box as follows and click on **OK** button.

Censored Variable `STATUS` **Value for Uncensored** `1`
Time Variable `Stime` **Time Unit** `Weeks`
Group variable `Rx`

The results are shown in Figure 57. The hazard ratio for treatment vs. placebo is 4.5231, which is a crude ratio as it simply determines the relationship between treatment variable (`Rx`) and time to event without taking into account the effect of other variables. The p-value from the Z-statistics is 0.0002, which denotes the significance of treatment effect. In conclusion, we can see that the hazard for placebo group is 4.5 times the hazard for the treatment group.

Figure 57. Example output for **Cox proportional hazards**, Epi Info.

```
Cox Proportional Hazards
Term          Hazard Ratio  95%     C.I.     Coefficient  S. E.   Z-Statistic  P-Value
Rx(Yes/No)          4.5231  2.0269  10.0932       1.5092  0.4095       3.6851   0.0002



Convergence:          Converged
Iterations:                   4
-2 * Log-Likelihood:   172.7592
Test                Statistic  D.F.  P-Value
Score                 15.9305     1   0.0001

Likelihood Ratio      15.2109     1   0.0001
```

Note: Rx=”Yes” for placebo group; Rx=”No” for treatment group

Say the investigator wants to assess whether another variable (e.g., a “third” variable) modifies or confounds the relationship between Group variable and time to event. As an example, let’s use the `Log_wbc` (log-value of white blood cell counts) as a third variable. To determine if `log_wbc` modifies the treatment effect (`Rx)`, we need to generate an interaction term (`Rx_logwbc`) by using the **Define** and **Assign** commands. First, **Define** the `Rx_logwbc` variable, then **Assign** it the following value: `Rx_logwbc=Rx*log_wbc`. This is a little more tedious than creating an interaction term in logistic regression where one can get it directly from the **Logistic Regression** dialog box.

Complete the Cox proportional hazards dialog box as follows and click on **OK** button.
Censored Variable `STATUS` **Value for Uncensored** `1`
Time Variable `Stime` **Time Unit** `Weeks`
Group Variable `Rx` **Other Variables** `Log_wbc, Rx_logwbc`

The results are shown in Figure 58. To determine whether or not there is a statistically significant interaction between `Rx` and `Log_wbc`, use the p-value for the `Rx_logwbc` interaction term. In this example, p=0.5103, and thus, it can be concluded that there is no statistically significant interaction. Because the interaction term was not significant, the next question would be whether `Log_wbc` *confounds* the relationship between `Rx` and time to event. To determine this, run another model without an interaction variable and complete as follows in the dialog box:

Censored Variable	`STATUS`	**Value for Uncensored**	`1`
Time Variable	`Stime`	**Time Unit**	`Weeks`
Group Variable	`Rx`	**Other Variables**	`Log_wbc`

The output from this model is shown in Figure 59.

Figure 58. Example output for **Cox Proportional Hazards** model with an interaction term.

```
Cox Proportional Hazards
Term          Hazard Ratio   95%      C.I.    Coefficient   S. E.   Z-Statistic   P-Value
Rx(Yes/No)         10.5375  0.3907  284.1802       2.3549   1.681        1.4009    0.1612
log_wbc             6.0665  2.5277     14.56       1.8028  0.4467        4.0359    0.0001
rx_logwbc           0.7102  0.2565    1.9668      -0.3422  0.5197       -0.6584    0.5103


Convergence:           Converged
Iterations:                    6
-2 * Log-Likelihood:    144.1314

                   Statistic  D.F.  P-Value
Test
Score                45.9021     3      0.0
Likelihood Ratio     43.8387     3      0.0
```

Figure 59. Example output for **Cox proportional hazards** model to assess for confounding.

```
Cox Proportional Hazards
Term          Hazard Ratio   95%     C.I.   Coefficient   S. E.   Z-Statistic   P-Value
Rx(Yes/No)          3.6476  1.5948  8.3426       1.2941  0.4221        3.0658    0.0022
log_wbc             4.9746  2.6088  9.4859       1.6043  0.3293        4.8716       0.0


Convergence:           Converged
Iterations:                    5
-2 * Log-Likelihood:    144.5585
Test               Statistic  D.F.  P-Value
Score                42.9382     2      0.0
Likelihood Ratio     43.4116     2      0.0
```

The hazard ratio for `Rx` is 3.6476; this would be interpreted that the hazard for placebo group is 3.65 times the hazard for the treatment group, adjusting for `Log_wbc`. Using the same formula for assessing the confounding effect at 5-10% cut point as in logistic regression, the crude hazard ratio 4.5231 is 24% higher than the adjusted hazard ratio of 3.6476. Therefore, it can be concluded that `log_wbc` is an important confounder of the treatment effect (`Rx`) in the study population.

In addition to Cox proportional hazards model, Epi Info can also perform the extended Cox model for survival analysis when the assumptions for Cox proportional hazards are not met (not discussed here).

Related Views and Data Tables

Concept

In a questionnaire view, each set of entries—whether one page or many—creates one record in the data table. The record corresponds to one instance of whatever the questionnaire or form represents, whether it is a person, a visit, an interview, a health provider, a case report, a hospital or clinic record, a medication order, a laboratory test, a surgical instrument tray, or an entire hospital.

A data system often needs more than one View to represent different entities that are related to one another. Clinical records, whether paper based or electronic, usually contain a single record for each patient, giving name, address, date of birth, etc., but a patient may have multiple clinic visits, operations, medications, or laboratory tests. These in an Epi Info system are separate views, with records related to the patient record by numeric keys.

Epi Info offers methods of creating and linking such different views so that they function as a single data system. In the system to be created here, a Mother record has two related views, for Children and Lab Tests. The views are linked by identifiers created automatically by Epi Info.

Relational file systems created by data managers often consist of many related files, and most public health professionals require the services of programmers to put the pieces back together for analysis. While Epi Info does not remove the need for careful design and understanding of such systems, the RELATE command in the Analysis program allows several Views to be assembled (related) for analysis with minimum difficulty if the automatic key structure of Epi Info is used, and with a little more effort if the keys have been set up by another system.

Introduction to the Related View Exercise

This exercise presents a simple data system with mother and child records. Each "mother" has zero to several children. The main View contains mothers and the related View contains children. During data entry, identification numbers ("keys") are generated automatically behind the scenes and stored in fields with the special names UniqueKey and FKey (for Foreign Key). A distinct UniqueKey number is generated for each mother's record.

Each child entered for a mother receives a value for the FKey field that is the UniqueKey of the mother, and can thus be related automatically to the mother in the Analysis program. The child knows its mother's number, and Analysis can thus answer questions that require data from both the Mother and Child records, for example, "What was the mother's age when each child was born, and what was the average of these ages for a group of children?"

Other data systems may use explicit keys such as a patient number rather than automatically generated keys. Analyzing such data is done in Analysis by specifying the keys needed to relate two data tables or Views. Using Checkcode, such explicit keys can be created during data entry as an alternative to the UniqueKey and FKey system. An optional section at the end of this exercise will show how to do this.

Creating Views and Entering Relational Data

1. **Using the MakeView program, from the FILE menu choose NEW and create a new MDB Project in the exRelated folder, using your initials for the file name--for example, ABC.MDB. Name the View MOTHER (in capital letters to indicate the main View) and create data entry fields for Mother Record Number (of type Number with pattern ######) and Mother Birthdate (a Date type). Of course, a working system would have other details like address and medical history, but we will keep this simple to illustrate principles. It is not necessary to include the word Mother in each field name, but it will make the rest of the exercise easier to follow.**

2. **Right click on the screen to create the next field and give the prompt as "Children". Then click on the button called RELATED VIEW. In the next dialog, leave the entries as they are, with the View accessible "any time", and click OK. Choose the default value in the next dialog, "Create a New Related View", and click OK. A button will appear on the form with the caption "Children."**

3. **Move the mouse cursor over the button to see instructions for accessing, moving, editing, and resizing the button. In this case we want to access the new blank view and add some fields. Hold down the control key and left click the button. You should see a blank view similar to the one on which you created the mother's view. Right click to create a field, and make a text field called Child Name. Create additional fields for "Child Birthdate" as a Date field, and "Child HIV Positive at 18 months?" as a Yes/No field. When the form is complete, click on BACK to return to the MOTHER form, agreeing to make the data table when asked.**

4. **Now repeat the same process to create a button called "Lab Tests" and develop a View that contains fields "Lab Test Date" as a Date and "Lab HIV Antibody Present?" (Yes/No). When you are satisfied with the new View, click the BACK button to return to the MOTHER View, again agreeing to make the data table. EXIT from MakeView (FILE menu|EXIT) to the main Epi Info menu. If asked, agree to make the main data table.**

5. **Let's enter some data! Click ENTER DATA on the main Epi Info menu. Choose OPEN from the FILE menu and find the MDB with your initials, and within it, the View called MOTHER. Enter data for the first mother from the tables below, and then click the CHILD button. Now enter the data for her 3 children. After you enter the last field in each record, a new blank record appears. Click the BACK button on the left panel when you are ready to return to MOTHER record number 1. Click LAB TESTS in the first MOTHER's record and use the same techniques to enter her 4 laboratory tests. Use the BACK button to return to the Mother's record. Click NEW to enter the second mother. Noting that mother number 3 has a LAB TEST, but no children, enter the rest of the data, using NEW to create new mother's as needed. (We have made it easy to enter the dates by using a lot of 1's for days and months.).**

Mother		Child		
Record Number	BirthDate	Name	Birthdate	HIV Positive at 18 months?
1	11-11-1980	Tom	11-11-2000	No
		Mary	11-11-2002	No
		Ralph	01-01-2004	Yes
2	11-11-1982	Rene	11-11-2001	No
3	11-11-1979			
44	11-11-1990	Leila	01-01-2004	No
555	11-11-1985	Jose	01-01-2003	No
		Matilda	01-01-2004	No

Mother	Lab Tests	
Record Number	Test Date	HIV Antibody?
1	11-11-2000	No
	11-11-2002	No
	11-11-2003	No
	01-01-2004	Yes
2	11-11-2001	No
3	11-11-2000	Yes
	11-11-2001	Yes
44	01-01-2004	Yes
555	01-01-2003	Yes
	02-02-2004	Yes

When you have finished, return to the MOTHER view and use the navigation keys on the lower left panel to check your data against the paper copy. (With a larger dataset, you could also us the LIST command in Analysis to display all the records for review.) Make changes if necessary. When you are satisfied, choose EXIT from the FILE menu or click the red X at the top of the screen to return to the Epi Info main menu.

Analyzing Relational Data

Analyzing a dataset of this size is most efficiently done with tick marks on a sheet of paper, but our purpose is to illustrate how it can be done for 10,000 or a million records, where the tick-mark approach becomes tedious.

Children of HIV positive mothers can have positive HIV serology after birth from passively transferred maternal antibody, but, if the child has negative serology at 18 months, transmission is assumed not to have occurred. Without antiretroviral medication to prevent transmission, about 25 to 30% of children born to HIV positive mothers will be positive at 18 months—that is, HIV infected.

Let's set some goals for the analysis. Before consulting the detailed instructions, you might write down or discuss the steps that you think are necessary to accomplish each task. Use the spaces below to record your preliminary ideas. In a real situation, it may take considerable trial and error to arrive at a suitable series of steps.

Goal 1: For each child calculate the mother's age when the child was born, i.e., the difference between the mother's date of birth and the child's birthday.

Goal 2: For each mother, find the date of the first positive HIV Lab Test.

Goal 3: Calculate the rate of transmission of HIV infection from seropositive mothers to their children.

Goal 1: Relating Mother/Child Views and Calculating Mother's Age at Child's Birth

The two variables are in different Views (and Data tables), which requires that we RELATE the two views before doing the calculation. To illustrate how the tables are related, we will also LIST the data before and after using the RELATE command.

1. **Run ANALYZE DATA from the main Epi Info menu and use READ and the Change Project button to read the RELATED.MDB project that is similar to the one you have just created and named. You can use your own MDB instead if your data entry went well. Use the CHANGEPROJECT button to navigate to the exRelated folder, choose RELATED.MDB or your own MDB, and then viewChildren within the MDB. LIST (choosing ALL) the data and confirm that the variable ChildBirthdate is present, but that the Mother's age or birthdate is not. There are 7 children. Note that the FKey field contains numbers that provide a link to the Mother's record for each child.**

2. **Now use the RELATE command to link the MOTHER View to the CHILD View. Simply click on RELATE, choose the MOTHER View, and click OK. Since you have not specified the keys on which to relate, Epi Info will automatically use UniqueKey in the parent file and FKey in the related file. If we wanted to include the MOTHER who has no children (perhaps because the first one is on the way), we would check the box USE UNMATCHED (ALL), but there is no reason to do so for our present objective.**

3. **Use LIST again and note that there are still seven records, one for each child, but that now each row of the table contains both ChildBirthdate and MotherBirthdate. You can expand the column width in the grid by clicking and dragging the vertical line between two column headings.**

4. **The relationship will persist until we either READ another table or exit from Analysis. Use the DEFINE command to create a new standard variable called MOTHERAGE. Click on the ASSIGN command and then on the FUNCTIONS button. Familiarize yourself with the syntax of the YEARS function that calculates the difference between two dates in years. Returning to the ASSIGN**

dialog, choose MOTHERAGE as the variable to which a value will be assigned. In the second blank, type YEARS(MotherBirthdate, ChildBirthdate) and then choose OK. (Since this is a function, the initial parenthesis must follow YEARS without a space.) Use LIST to see that each child now has a calculated MotherAge.

5. **The result is satisfying, but not permanent, until we write a new table to preserve the new calculations. Use the WRITE command to write a new file in the Epi Info 2000 format, naming it NewRelated.MDB and writing a table called MotherChildAge with all the variables shown. Set the Output Mode to REPLACE, in case a table by this name already exists.**

6. **Confirm that the new data table exists. Click on READ, then Change Project and find NewRelated.MDB. Click the ALL choice to see data tables, choose MotherChildAge, and use LIST to see the records.**

7. **Save the program commands generated during this interaction by clicking the SAVE button in the program editor (lower right window) and then the TEXT FILE button in the dialog that appears. Name the program GoalOne and click OK.**

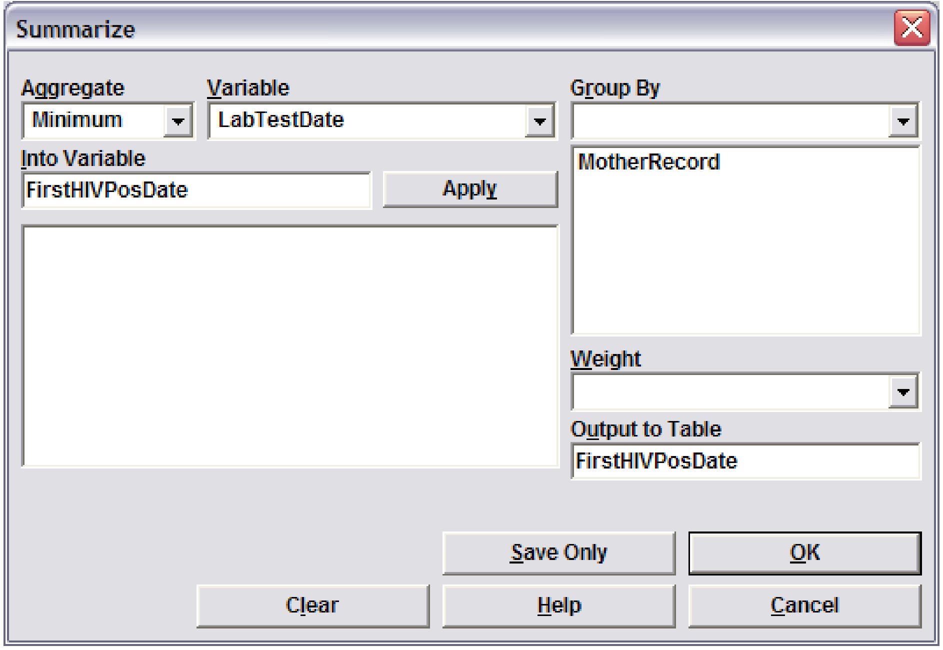
Goal 2: Using SUMMARIZE to Find the First Positive HIV Test for Each Mother

Relational database tables are, strangely enough, not processed in any particular order, even after sorting, and the SUMMARIZE command in Analysis is required to do the work of picking out the first (minimum) date with a positive HIV test for each mother. Here we use SUMMARIZE to find the first date on which the Mother had a positive lab test and write it and the Mother's UniqueKey identifier to a new data table called FirstHIVPosDate, using the same name for the variable. Once having obtained the FirstHIVPosDate for each mother, the new information can be related to the Mother's record and written to a new table containing the enhanced records.

1. **Run the Analysis program. If there are already program statements in the lower right window, click the NEW button and YES to remove them. Use READ with the ChangeProject button to read viewMOTHER in Related.mdb. Then RELATE viewLabTests. It is not necessary to specify a key when doing so. Use LIST to see that there is now one record for each lab test, and that each of the 9 lab tests**

has the related Mother information.

2. **Click on the SELECT command and insert the condition, LabHIVAntibody=”Yes”. Now only the records containing a Lab test positive for HIV antibody will be processed. There should be 4 records.**
3. **Click on the SUMMARIZE command, fill it in as in the image below, click the Apply button. (This dialog has one peculiar feature—after filling in all the information, you must click APPLY before clicking OK.)**



Summarize

Aggregate: Minimum

Variable: LabTestDate

Into Variable: FirstHIVPosDate

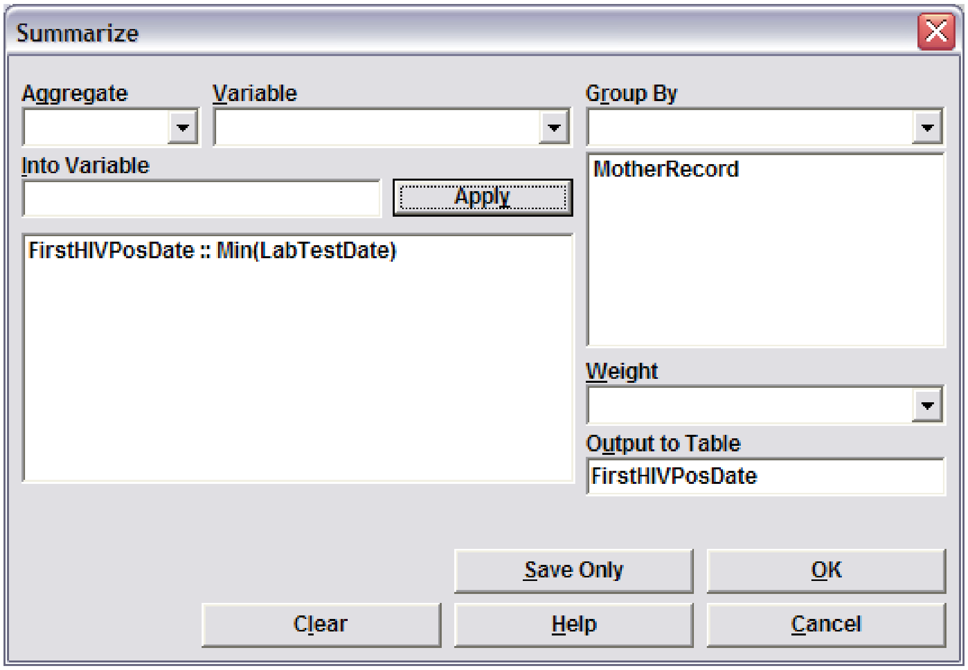
Group By: MotherRecord

Weight:

Output to Table: FirstHIVPosDate

(We are telling SUMMARIZE to find the Minimum value of LabTestDate for each group of records with the same MotherRecord number, and write it to a new variable called FirstHIVPosDate in a new table called FirstHIVPOsDate. There will be one record for each MotherRecord number in the new table, and it will contain the FirstHIVPosDate value.)

4. **The necessary text appears in the larger box as in the next picture. After clicking APPLY, the dialog should look like this:**



FirstHIVPosDate :: Min(LabTestDate)

Group By: MotherRecord

Output to Table: FirstHIVPosDate

5. If this goes well, then click OK. Agree to replace the existing output table, if any. A new table will be written containing one record for each Mother and the value of her FirstHIVPosDate, if any.
6. Use the READ command to read the viewMOTHER table without the Lab Test View. This also cancels the previous selection of HIV positive records.
7. Use LIST to see the currently active records, and ascertain that the new table is not yet connected with the original Mother records with which we started. We must use the RELATE command to link it by RecordNumber to these records.
8. To relate the new table to the other Views, click on the RELATE command, click the ALL choice in the middle of the dialog so that you can see data tables as well as Views, and highlight the table called FirstHIVPosDate, but do not click OK yet. We want to use the MotherRecordNumber key to guide the linking, so we have to tell RELATE what keys to use.
9. Click the BUILD KEY button. The dialog that appears is useful, but has some peculiar features that are not entirely intuitive. Note that Current Table is selected. In the Available Variables field, find MotherRecordNumber and select it. It appears in the field at the top, but belongs in the larger field below Current Table. Now click on Related Table, and MotherRecordNumber appears under Current Table. Use Available Variables again to see the variables in the Related

Table and choose MotherRecordNumber again. This is the copy of the Mother's RecordNumber that was written to the FirstHIVPosDate table. How do you get MotherRecordNumber to show up in the large field on the right under Related Table? Strangely enough, it is by clicking once again on Current Table, which will make the choice show up in Related Table!! Or you can simply type MotherRecordNumber in both of the large text fields in the BuildKey dialog. Click OK and you will see MotherRecordNumber::MotherRecordNumber at the bottom of the RELATE dialog. Again, after a little experience, you may decide just to type the key definition into this field rather than using Build Key. Now check the item Use Unmatched (ALL) and click OK in the RELATE dialog. Use LIST to see the results. What has been achieved?

10. Use the WRITE command to write the related table formed from Mother and FirstHIVPosDate tables to a new table called MotherPosDate1 in the file NewRelated.MDB. The achievement mentioned above is that each positive mother now has a date of her first HIV positive test attached to all of her records or virtual records.
11. Click the SAVE button in the program window and save the program as a TEXT FILE named GOALTwo.pgm

Goal 3:Finding Rate of HIV Transmission from HIV-Positive Mothers

The strategy for this section is to use the newly enhanced mother's record containing the date of first HIV positive test, and relate the child records. To find the number of children born to seropositive mothers, we select children whose BirthDate is on or after the Mother's FirstHIVPosDate. Since we have selected only exposed children, a simple frequency of those who are HIV positive (at 18 months, which rules out passive antibody transfer from the mother) and the total number at risk gives the HIV transmission rate.

1. Click the NEW button in the Analysis program editor and then YES to remove the previous program.
2. Use READ and Change Project to READ the data table, MotherPosDate1 in NewRelated.MDB, containing the dates of first positive for each mother.
3. RELATE MotherChildAge, containing a record for each child, by using the key combination MOTHERRECORD::MOTHERRECORD in the KEY field at the bottom of the RELATE dialog.

4. **SELECT records for which ChildBirthdate >= FirstHIVPosDate, in other words, the child was born on or after the mother's FirstHIVPosDate. There should be 4 records.**
5. **Do a Frequency of ChildHIVPositive to determine the rate of infection in children born on or after the date the mother was shown to have LabHIVAntibody. The result is a 25% rate of transmission, about what studies show occurs without antiretroviral preventive treatment.**
6. **Use the SAVE button in the program editor to save the program as GOALThree.pgm.**

This was a lot of keystrokes and dialogs, but there are two principles to remember:

1. The RELATE command gives you a lot of power to relate different types of tables, as long as they have key fields that can be used to make the link. Keys must be unique (no duplicates) within the tables in which they occur.
2. The SUMMARIZE command makes up for not being able to process records sequentially and count on a particular order of processing. Writing the summary information to a new table and relating that table back to the original table is a very useful trick for making the information (e.g., date of first positive HIV test) accessible from any record in the original table.

Self Assessment

In order to solidify your understanding of RELATE and SUMMARIZE, please give three hypothetical examples of related file systems below:

1.

2.

3.

If you had a clinical record system with tables called PATIENT and VISIT, how would you use SUMMARIZE to do the following:

1. Find the Hematocrit at the first visit for each patient, assuming that there is a field in VISIT called Hematocrit where the information is recorded and that each visit has a DateOfVisit.

1. Find the Hematocrit at the last visit for each patient.

2. Find the difference between first and last visit hematocrits for each patient.

3. Find the average number of minutes spent per visit for each patient (assuming that VISIT records have a Duration field)

4. Find the date of first hematocrit above 35 for each patient (in an anemia clinic).

Possible Answers

If you had a clinical record system with tables called PATIENT and VISIT related by PatientID, how would you use SUMMARIZE to do the following:

2. Find the Hematocrit at the first visit for each patient, assuming that there is a field in VISIT called Hematocrit where the information is recorded and that each visit has a DateOfVisit.
   a. READ VISIT
   b. SELECT Hematocrit >1 (logically "Hematocrit <>." should work, but does not always do so.)
   c. SUMMARIZE the Minimum DateofVisit for each PatientID to a new table called FirstVisit
   d. RELATE the FIRSTVISIT table to VISIT, using PatientID.
   e. SELECT those with DateOfVisit=FirstVisit and use LIST to show the PatientID and Hematocrit, or FREQuency to summarize the first Hematocrit values for all the patients.

3. Find the Hematocrit at the last visit for each patient.
   a. Use the same steps, but select Maximum rather than Minimum in step b.

4. Find the difference between first and last visit hematocrits for each patient.
   a. Relate the table from 1. above to the main table by PatientID
   b. Relate the table from 2. above to the main table by PatientID
   c. Define two new variables, FirstHCT and LastHCT
   d. Use IF statements to ASSIGN the variables
      i. IF VisitDate=FirstVisit THEN ASSIGN FirstHCT=Hematocrit
      ii. IF VisitDate=LastVisit THEN ASSIGN LastHCT=Hematocrit
   e. Define another variable called LastMinusFirstHCT and assign it the value of LastHCT-FirstHCT

5. Find the average number of minutes spent per visit for each patient (assuming that VISIT records have a Duration field)
   a. Use SUMMARIZE with AVERAGE of the DURATION field for each PatientID, writing to a variable and table called AverageVisitDuration.

6. Find the date of first hematocrit above 35 for each patient (in an anemia clinic).
   a. Select Hematocrit>35
   b. Use SUMMARIZE to write the Minimum of VisitDate for each PatientID

Adding Alternative Keys in a Relational System (Optional)

By now, you should have an idea of how Epi Info creates UniqueKeys for each parent record and an FKey (for foreign key) in each child record that “points” back to the parent record (by being a copy of the UniqueKey of the parent).

Since keys are so important to data integrity, it is reasonable to use two key systems in a large dataset—the built-in, automatic mechanism, and a second set of visible keys. The latter are entered by the user for the master (Mother) record, but can be copied to the children automatically using Check Code.

For those who prefer to use the extra key system, here’s how to install it in the database we have already constructed.

Installing an Alternate Key System

The objective is to insert new fields in the related Child and Lab Test views to match the MotherRecord number. In order to transmit the MotherRecord number to each record in the related Views, a global variable called MoGlobalID is created in check code and set to MotherRecord before each RECORD is accessed in the Mother View. MoGlobalID is then used to set the value of the Record Number in each related view before one of their RECORDs is accessed.

1. **To preserve the work already done, use MyComputer from the Windows Start menu or Desktop to copy your own MDB or the sample provided as Related.mdb. Click once on the filename to select it and then do Ctrl-C to copy and Ctrl-V to paste. This will make a copy of the file called ”Copy of…” Right click on its icon and use the RENAME feature to rename it to BetterRelated.MDB or another name that you prefer.**
6. **Return to Epi Info, run MakeView, and open BetterRelated.MDB and the MOTHER View.**
7. **To make sure that no record has a blank MotherRecord number, double click on the prompt words, Mother Record Number to bring up the field dialog and then click in the checkbox next to Required. Click the OK button. The user will now be required to enter a MotherRecord number before leaving each record.**
8. **With the MOTHER View visible, click the blue PROGRAM button in the left screen panel. This brings up the check code programming panel.**
9. **We want to transfer the MotherRecord value to a global variable so that it can be used to set the corresponding keys in the CHILD and LAB TEST Views. In the dropdown box under Choose field where action will occur, choose DEFINEDVARIABLES. Click the DEFINE command button, enter the name MoGlobalID, and check the variable type as GLOBAL. Click OK to create the global variable.**
10. **In order to cover both new records and existing records when a user browses back through the table, we must set the value of MoGlobalID in two places—**

immediately after a new MotherRecord number is entered, and also in the block called RECORD, which will set it before every record is displayed. A modifier called ALWAYS … END will make sure that it is set even when the user is paging through records and not entering a new record. In the space for FIELD WHERE ACTION WILL OCCUR, choose MotherRecord, then click on the Assign command and locate the MoGlobalID variable for the top blank and MotherRecord for the second one. Click on OK. This will set the global variable, MoGlobalID to MotherRecord immediately after the MotherRecord number is entered by the user.

11. Now choose RECORD as the location where action will occur, click BEFORE, and then make the same assignment, setting MoGlobalID equal to MotherRecordNumber. Note that the command appears before the marker ENDBEFORE, meaning that the assignment will occur before entering rather than after leaving the RECORD.
12. One more thing is necessary to make the assignment occur even when the user is scrolling back through records and making no entries. In the edit window, surround the assignment with the words ALWAYS and END. The code in the RECORD block should look like this:

```
ALWAYS
  Assign MoGlobalID=MotherRecord
END
ENDBEFORE
```

13. Click OK to return to Mother's main screen.
14. Hold the Ctrl key and click the CHILDREN button. In the child form, right click near the top to add a field. Name the field Mother's Number. Make it a NUMBER field with the pattern ###### (6 digits). This is important, as keys in one view must match by field type and pattern, or they cannot be used to relate the Views. Check the property called READ ONLY. Click the OK button. The user will not be able to enter data in this field, but we will use Check Code to do so automatically. The global variable MoGlobalID is visible from the CHILD form--that's why we made it GLOBAL—and we can now use it to set the value of MothersNumber in the CHILD View. Click the Program button and insert the following check code in the RECORD block:

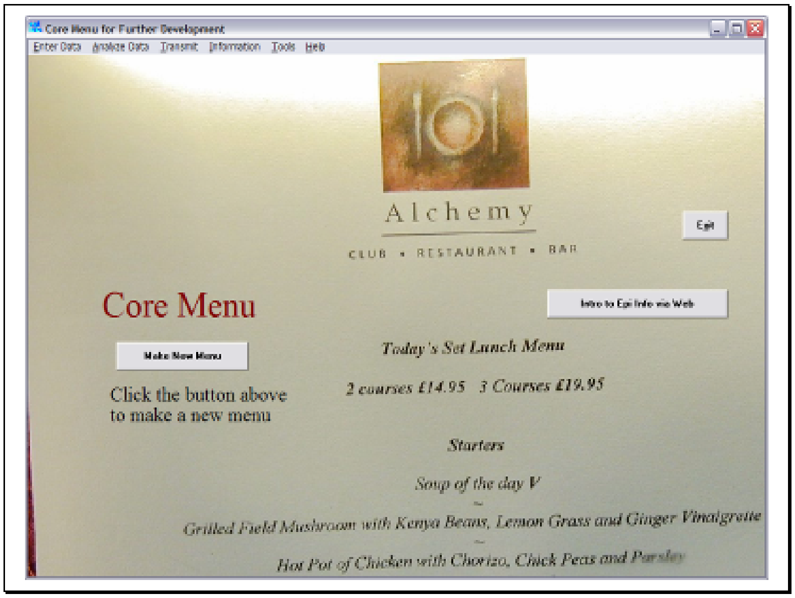
```
ALWAYS
   Assign MothersNumber=MoGlobalID
END
ENDBEFORE
```

15. Close the check code screen, and then click BACK to return to the MOTHER View.
16. Hold the Ctrl key and click on Lab Tests. Make the same changes in this view that you made in the child view, except that the 6-digit field should be called Patient Number and the ASSIGN should be adjusted accordingly.
17. Test your new code from the MOTHER View in Enter and make sure that the child forms always contain the Mother's number. If not, look for errors in the check code and consult with your colleagues or instructor. To insert

the new keys in related records you will have to visit each record.

18. **A working version of the code is found in EnhancedRelated.MDB.**

Menus and Permanent Applications



Click the button above to make a new menu

Today's Set Lunch Menu

2 courses £14.95 3 Courses £19.95

Starters

Soup of the day V

~

Grilled Field Mushroom with Kenya Beans, Lemon Grass and Ginger Vinaigrette

~

Hot Pot of Chicken with Chorizo, Chick Peas and Parsley

Concept

Levels of Epi Info Use

1. Interactive
2. Analysis and check code programs
3. Applications for convenient and repeated use

Having learned to make questionnaire views, enter data, import data, do analysis, and save programs, you probably have quite a number of pieces of data systems saved as files. In order to make these more accessible for easy use during a study or other regular activity, you can make a customized menu to tie the pieces together into a unified whole, called an APPLICATION.

The menu in an application provides point-and-click access to programs, links, and help files that make a defined series of tasks easy and even fun. The Epi Info menu can also be programmed or "scripted" to perform tasks like copying files, obtaining input from the user, and providing customized messages as a result of defined conditions.

In this exercise, we will use an application called MakeMenu to—guess what—make a menu and a new folder or directory in which we can develop our own application. The exercise has two parts:

- Making a Menu (and a folder for the application)
- Configuring the menu files from previous exercises into a sample application

Making a Menu

We will use an Epi Info program called MakeMenu to construct a new menu automatically. MakeMenu is not part of Epi Info itself, but is supplied with the class materials.

1. **From the Microsoft Windows desktop open My Computer and locate the EpiInfoCourseII materials and the folder called exMenu. Open this folder and then click on the icon labeled, "Click here to run MakeMenu". You should see a menu (with an image from a London restaurant) and instructions for making a new menu of your own.**
2. **Click the button or menu entry called MAKE NEW MENU. If there is no icon on the desktop called, "Mothers and Children," use this as the name of the new menu. Otherwise add a number or change the name to make a new menu. Follow the prompts until you have created the menu; then close the current menu. You should see an icon on the desktop that has the new name.**
3. **Click the new icon and examine the menu. Choose Edit This Menu and examine the text program (.MNU file) that configures the menu and specifies its activities.**
4. **Return to the menu and try out the sample data entry feature under ENTER CASE called ENTER CASE DATA. Exit from the entry program and again click the Edit This Menu button. In the .MNU file, find the corresponding text entry, MENUITEM "Enter Case Data", and the menu command block further down called "EnterCases", where the command to EXECUTE the Enter.exe program is given, together with the name of the MDB and View that are to be brought up in Enter.exe. This block provides a sample format from which you can develop your own call to enter data, once you have created an MDB and View that you want to access.**
5. **In the Analyze Data part of the menu, try the entry "Run Analysis Statistics". Be patient; Analysis may take a while to come up. After exploring the output of the program, exit from Analysis. In the menu program, locate the corresponding parts of the menu file (.MNU) and note the different syntax of the Analysis and Enter command lines.**

Configuring the Menu to Include New Features

In this part of the exercise, you will edit the .MNU text file that configures the menu to include the data entry and analysis programs we developed in the Related exercise and be able to run them from the menu.

1. **Find the icon for the menu on the desktop and click it to display the menu.**
2. **Click the button called "Edit New Menu." This opens the MNU or menu file in the Notepad or Wordpad editor. Lines beginning with an asterisk (*) are comments to explain how menu files work. You should read these and**

familiarize yourself with the structure of the file.

3. **Find the menu item, "Enter &Case Data" and change it to, "Enter &Mother's Data". The "&" indicates that the following letter is a "hot key," so that this item can be activated when the menu is open by pressing the "M" key. Change the name of the menu block from "EnterCases" to "EnterMother".**
4. **Now scroll down until you find the program block called "EnterCases". Change its name to "EnterMother". Place an asterisk in front of the line that begins with "Execute Enter.exe " and copy its contents on the next line, changing "Sample.mdb:Surveillance" to the MDB and View where your relational file system, Mother, resides. If it is in a folder called exRelated that lies at the same level in the folder hierarchy as the MakeMenu folder, then the line should be:**

```
Execute Enter.exe @@MenuDir\..\..\exRelated\Related.mdb:Mother
```

You may have to experiment and do some searching in MyComputer to find the correct combination. The two dots followed by the backslash repeated twice mean, "Go up two levels, then back down to find the exRelated folder." The predefined variable MenuDir provides the location of the current menu as a starting place.

Save the MNU file with Ctrl-S or the SAVE command in the File menu, and test your new menu item to see if it brings up the Enter program and the correct database and View (The EpiInfo.exe menu program will load the edited file as soon as it is saved). If the edited version does not work, try giving the absolute position of the Related.MDB as in My Computer--for example, C:\Epi Info CourseII\exRelated\Related.mdb:Mother. Eventually, by trial and error or with the help of an instructor you should hit on the right combination, and clicking on the "Enter Data" menu and then "Enter Mother Data" will bring up the Mother form ready to enter data. This may seem like a lot of work, but you will never have to hunt around for the file on this computer again, since you can access it through the menu.

5. **Now let's provide a way to run your three analysis programs called Goal1, Goal2, and Goal3. Find the POPUP called Analyze Data, and add three items of your own to represent the three programs, as follows:**

```
MENUITEM "&Run Analysis Statistics", RunAnalysisPGM
MENUITEM "Run Goal&1", RunGoal1
MENUITEM "Run Goal&2", RunGoal2
MENUITEM "Run Goal&3", RunGoal3
```

6. **Save the MNU and test the menu. Note that the three items are now present in the Analyze Data menu, but that they do nothing when clicked, because there are no corresponding command blocks.**

7. **Now find the command block called RunAnalysisPGM and make a copy by the name, RunGoal1, with the same BEGIN and END statements and an Execute line between, as follows:**

```
RunGoal1
Begin
  Execute Analysis.exe
pgmname='@@menuDir\..\..\exRelated\Goal1.pgm'
End
```

If your Goal1.pgm is in a different location, you may have to make some changes in its path, as with the EnterMother block. Note that the file path must be in single quotes and that no MDB or View is specified, since they are already specified in the READ statement that comes first in the program. If this were not the case, you could add a viewname= statement as in the RunAnalysisPGM block and this specified view would become the default for the program.

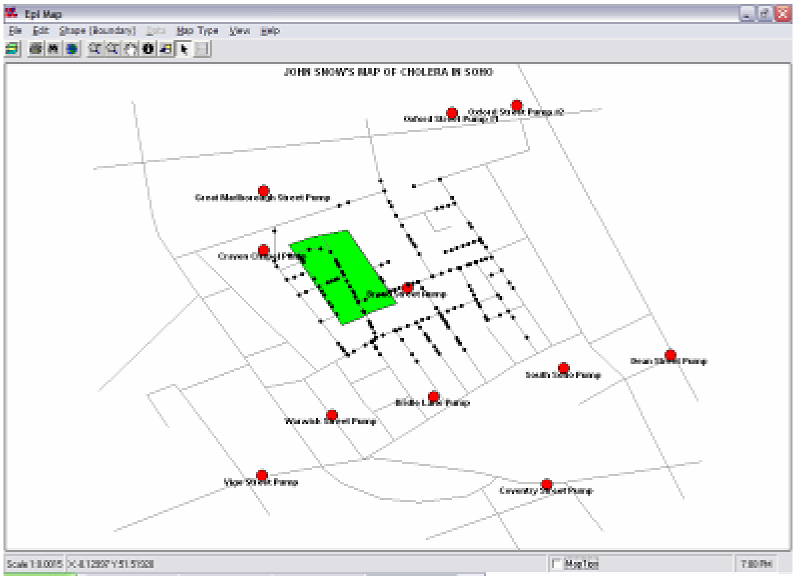
8. **Once you have things working for Goal1, you can copy and paste the Goal1 block twice and then change "1" to "2" or "3" to form the other two blocks needed to implement your menu items.**

```
RunGoal2
Begin
  Execute Analysis.exe                    (no line break here)
  pgmname='@@menuDir\..\..\exRelated\Goal2.pgm'
End

RunGoal3
Begin
  Execute Analysis.exe pgmname='@@menuDir\..\..\exRelated\Goal3.pgm'
End
```

9. **The menu has many commands other than EXECUTE, and any DOS batch file command can also be used, for example, for copying files automatically. Study the sample menu and read about the other commands in the Epi Info help file to begin making your own applications.**

Geographic Information Systems (GIS)



Introduction

Geographic Information Systems sometimes are regarded as special magic, to be managed by experts with expensive software and a vocabulary that differs from that of epidemiologic and statistical software. For public health purposes, however, geography means "place" in the classical triad of "time," "place," and "person" that is the basis of epidemiologic investigation. An epidemiologist needs not only GIS, but also a "Temporal Information System (TIS)," for time, and a "Personal Factor Information System(PFIS)" for other risk factors. These are both supplied by Epi Info (without the imaginary names) in the form of Graphs for Time and Tables and associated statistics for Personal Factors.

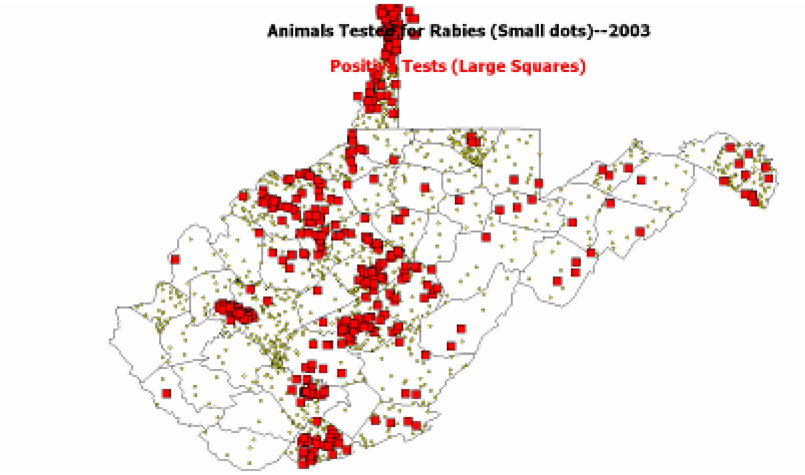
If we wish to combine place-related epidemiologic data with maps or other diagrams, it is necessary that the data be described so that they can be matched to place locations on a map. There are two main ways of doing this:

Type of Data	**Linked by**	**Type of Map**	**Example**
Counts or Rates	Polygon Name	Named Polygons	Case counts per county or state
Points or Cases	Latitude/Longitude or other coordinates	Any graphic that has a matching coordinate system	Individual cases on street maps

In practice, summary data is usually linked to polygons, such as states, counties or zip codes. Individual cases with exact locations can be plotted as symbols on maps of streets, or other backgrounds that do not necessarily have named polygons. In Course I in this series, there were two GIS exercises—one to associate cases of anthrax and controls with polygons overlying a map of New York and the other to plot disease reports by county. Both illustrated how to display counts or rates on suitable *shapefiles*, the standard for maps of polygons, lines, or points established by the ESRI, the largest GIS software provider (www.esri.com).

In this course, we will further explore plotting points to display cases or other individual events. Since plotting the exact residence address of an ill person may compromise confidentiality, these maps must be kept within the health department or the data can be summarized to display as counts or rates that do not reveal exact locations. For data on zoonoses or other non-human events, however, and for mortality data, point maps can be used whenever coordinates are available.

Mapping Points in Epi Map

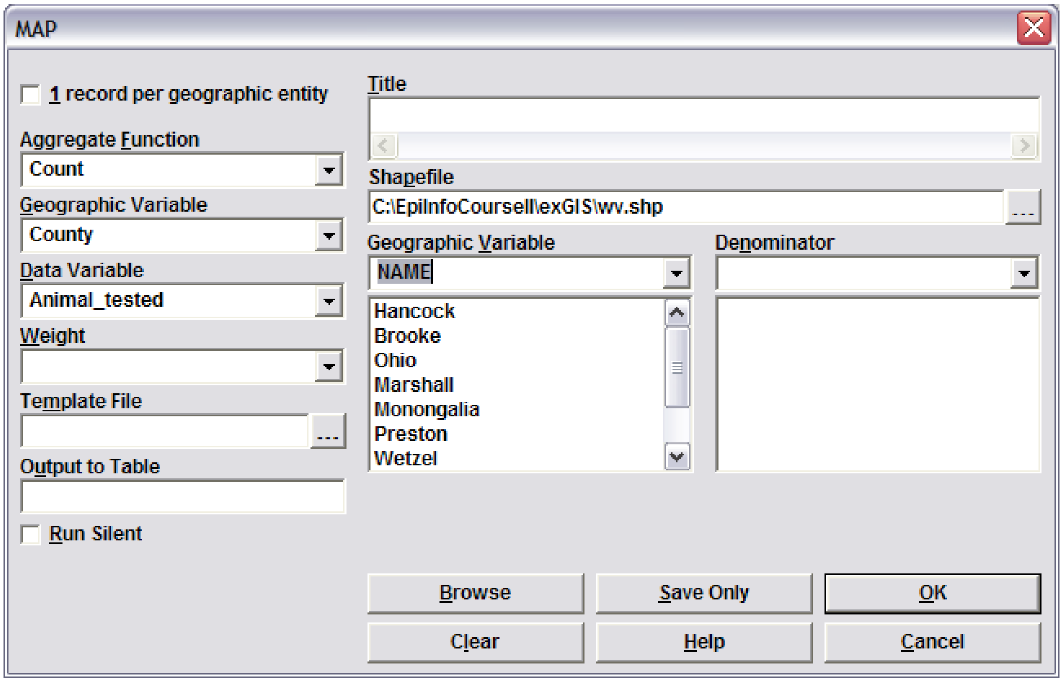

Animals Tested for Rabies (Small dots)--2003
Positive Tests (Large Squares)

Concept	Points are placed in a layer on a shape file using coordinates that match those of the shape file. The usual coordinates are longitude and latitude in decimal degrees. In this exercise, we will first use a dot density map to show animals tested for rabies in West Virginia. On top of this display, we will create another layer of points representing animals having positive tests plotted by their lat/long coordinates.

Displaying Animals Tested by County

As background for the display of animals with rabies, we will use a dot-density map of animals tested. To make this map, choose ANALYZE DATA from the main Epi Info menu. READ the GISDATA.MDB file and choose the table called RabiesTests2003. To see it, you may have to click ALL rather than VIEWS in the READ dialog.

Click on the MAP command, and fill in values as follows:



☐ 1 record per geographic entity

Aggregate Function: Count

Geographic Variable: County

Data Variable: Animal_tested

Weight:

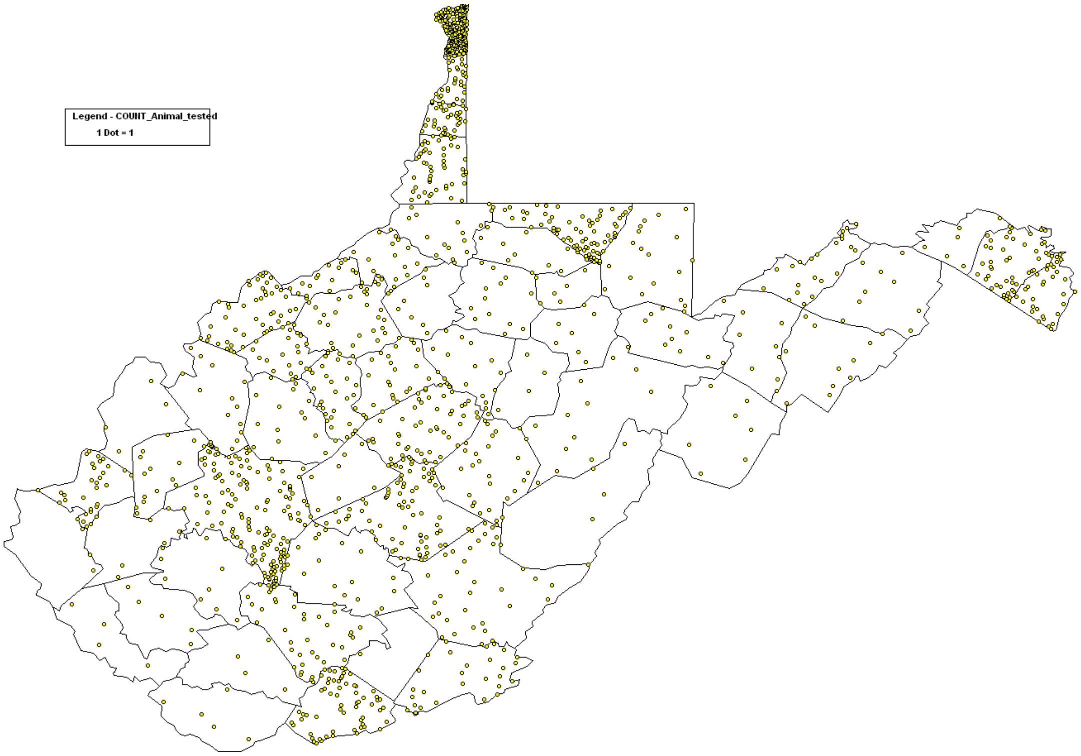
☐ Run Silent

Shapefile: C:\EpiInfoCourseII\exGIS\wv.shp

Geographic Variable: NAME

Hancock
Brooke
Ohio
Marshall
Monongalia
Preston
Wetzel

Choose OK, and then, in the warning dialog, CONTINUE, and you will see a choropleth map of rabies tests performed by county in 2003. In Epi Map, choose MAP TYPE from the menu and then DOT DENSITY. In the properties pages that appear, be sure you are on the one labeled DOT DENSITY, and then click OK. You should see yellow dots representing animals tested in each county, as follows:



Place the mouse cursor on the legend at the lower left, hold down the left mouse button, and move the legend up to a more aesthetic position. Note that, as you do so, the dots are redrawn many times, and that the number of dots in a county remains the same, but that their position is determined randomly. The tests performed, therefore, are accurately enumerated by county, but their position within the county is not displayed. This makes sense, since you instructed Epi Map to use COUNTY as the Geographic variable in the data table.

To see the name of the county with the large number of dots, click the toolbar button with the "I" (for "information") and then click on the county. County is given as NAME, but note the large number of other data items associated with the shape file (stored in its .DBF section).

Displaying Positive Rabies Tests by Point Coordinates

On this background, we will display the positive tests using latitude/longitude coordinates contained in the data table. The positive results are in a separate table called RabiesPositives2003.

Within Epi Map, with the tests still displayed, choose MAP MANAGER from the FILE menu, and click on the button ADD POINTS. Find the data file GISData.MDB and choose the table RabiesPositive2003. Another dialog pops up asking you to SELECT XY FIELDS AND POINT PROPERTIES… For the X FIELD, select DECIMAL_LONG, and leave the Y FIELD as DECIMAL_LAT. POINT LABEL allows you to display, for example, the species of ANIMAL_TESTED beside each point, but we found by experiment that this dataset produces too many labels to be good looking, so leave this item blank. For POINT TYPE choose SQUARE, click on the black rectangle and change the POINT COLOR to bright red, then click OK, select 10 for POINT SIZE, and click OK again.

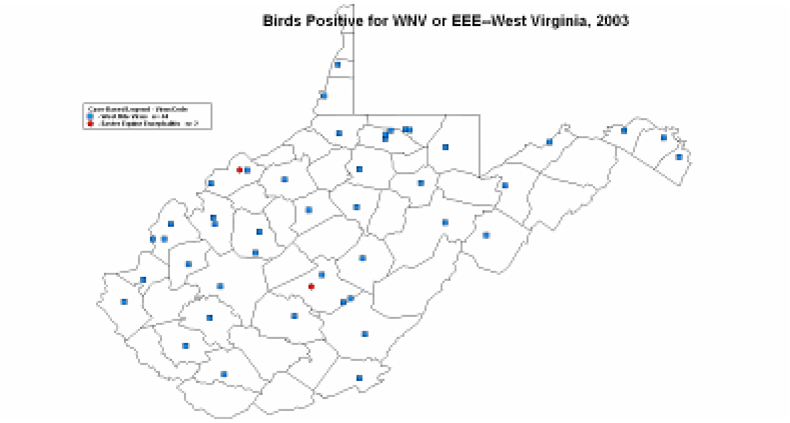
When you click OK, you should see the map represented at the beginning of this exercise. Note that it shows both numerator (Positives) and denominator (Animals Tested) data on the geographic background, with the positives having more precise locations than the animals tested. We have ignored the missing values, which should be considered in producing your final report, but the display is a good summary of the year's results. If you would like to display the animal species tested, repeat the ADD POINTS step, but request labels for ANIMAL_TESTED.

Try adding titles to your map by clicking on the graphics button in the toolbar—the one with the small yellow rectangle, circle, and triangle. Then click the button with the large "A" to insert a title ("Animals Tested for Rabies (Small Dots)—2003"). We found that a point size of 20 is about right. A subtitle can be inserted in red to indicate "Positive Tests (Large Squares)".

Move the legend around and note that the positive test points remain in the same locations, since they were specified by latitude and longitude, but that the randomly distributed dots representing all tests change location with each move. They are only linked to the counties by county name.

To save the image for your newsletter or annual report, choose SAVE AS BITMAP FILE from the FILE menu, and give the file a suitable name and location.

Mapping Categorical Data in Epi Map



Concept	Epi Map can create dot density maps with more than one category, based on numeric codes. In this exercise, we insert codes in a data table for West Nile Virus (1) or Eastern Equine Encephalitis (2), and show the two categories by West Virginia county.

Preparing Data by Inserting Numeric Category Codes

Using ANALYZE DATA from the Epi Info main menu, choose READ and then EXCEL 8 for DATA TYPE. Find the file POSITIVEBIRDS2003.XLS in the exGIS folder and select the range or worksheet provided. There should be 43 records.

LIST the records to see the variables and data values. Note that latitude and longitude are given in decimal degrees and that VIRUS has two values, WNV (for West Nile Virus) and EEE (for Eastern Equine Encephalitis). It would be easy to display the values as points as we did in the previous exercise with VIRUS as the POINT LABEL, so that the two EEE values would stand out. However, sometimes there are too many values to use the categorical label display, and we would like to show the different categories as different shapes, colors, and sizes of symbols. Let’s use POSITIVEBIRDS2003 to illustrate how to do this.

The categorical symbol display in Epi Map requires numeric codes for the categories. We must therefore convert “EEE” and “WNV” to 1 and 2 for example. Now that we have read the Excel file in Analysis, we can do this before writing a table in the MDB to be used for mapping.

First DEFINE a new variable with the DEFINE command, and call it VirusCode.

Choose the IF command and insert the condition VIRUS=”WNV” (with quotes). In the box labeled THEN, type the command ASSIGN VirusCode=1 (no quotes). Click OK. What will this command do if the value of VIRUS is “EEE” or missing? Use LIST to find out.

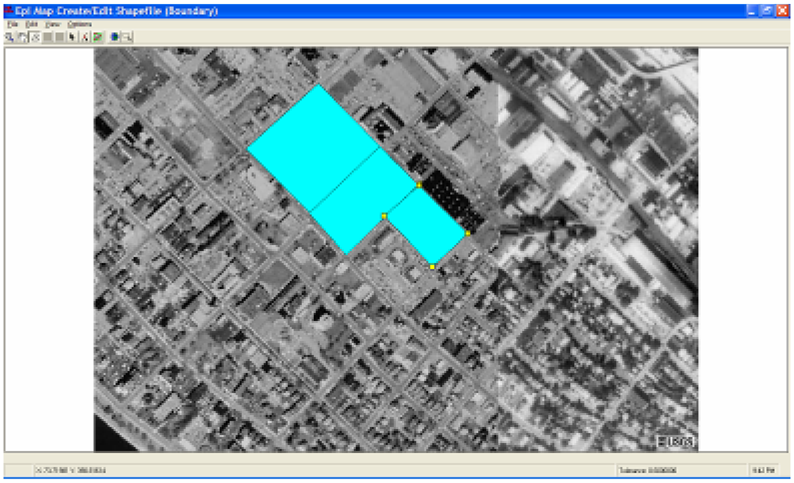
Now do another IF command for VIRUS=“EEE”, assigning VirusCode=2. What does this one do if VIRUS is “WNV” or missing?

Note that nothing visible happens until your do a LIST or FREQuency, but that the values will remain until you do another READ or leave the Analysis program. Use the WRITE command to write all variables to a table in GISData.mdb called POSITIVEBIRDS2003. READ this table and LIST it to confirm that the new variable called VirusCode is correctly set to 1 or 2 by virus type.

Now exit from Analysis, and go directly to Epi Map by clicking the the main menu button, CREATE MAPS. Choose the Map Manager and ADD LAYER to add the WV.SHP file and bring up the map of West Virginia. Now choose the CASE-BASED button and locate GISData.mdb and the table POSITIVEBIRDS2003. A somewhat complicated dialog appears, in which you should choose NAME and COUNTY in the first two boxes, and VirusCode in the third. Now a choice of symbols and legend labels appears. Change the colors to something contrasting, the size to 10 or 20, and the shapes to two different choices. Type legend entries “West Nile Virus” and “Eastern Equine Encephalitis”. When you choose OK, the two types of virus are displayed as different symbols.

As a thought exercise, give some examples of datasets for which the technique of plotting more than one kind of symbol would be useful.

Editing or Creating Shape Files



Concept	The Partial Load feature in Epi Map can be used to customize shape files by including only selected polygons. Editing features are provided to create or divide polygons and insert lines or points. When shape files are not available, a scanned image or aerial photo can be loaded and used to guide creation of a new shape file. Use this feature for refugee camps, building floor plans, or geography that has undergone recent change.

Creating a New Shape File by Combining Polygons from Others

From the Epi Info main menu, choose CREATE MAPS to run EpiMap. In the FILE menu, choose MAP MANAGER… and then ADD LAYER. Find the shape file called WV.shp and open it to display a county map of West Virginia. Now do ADD LAYER again and add the shape file PA.shp, and again to add OH.shp. Now you have the adjacent states of West Virginia, Pennsylvania, and Ohio. If the map is not correctly framed in the window, put away the Map Manager and click on the world view icon in the toolbar.

Our goal is to add to the West Virginia shape file just the adjacent counties from the other two states. This requires the names of the counties to be added. In order to see the names, bring up the Map Manager again and select the PA layer. Click Properties and then choose the Std Labels tab. In TEXT Field, choose NAME, and click OK. You will see the names of the counties. Do the same for OH.shp and you can write down the names of the counties you want to add to the West Virginia map, or just print the labeled map to use in the selection process.

Now remove PA and OH by selecting each and clicking REMOVE LAYER. Choose ADD LAYER PARTIAL and PA.SHP. In the list of counties in PA.SHP, hold down the Ctrl key and click on the name of each of the border counties--Allegheny, Beaver, Fayette, Greene, Lawrence, and Washington. When asked for a shape file name, use PACounties. Do the same for OH.shp, select Athens, Belmont, Columbiana, Gallia, Jefferson, Lawrence, Meigs, Monroe, and Washington, and give the shape file the name OHCounties.

So far, so good, but the combined map is a jumble of counties, and it's hard to see where West Virginia ends and the other states begin. In the map manager, select the WV layer and click PROPERTIES. In the SINGLE tab, set the Fill color to light blue or gray. Leave the other properties as they are. Click OK, and you will have West Virginia in gray or blue and the other states with clear or white backgrounds. To save the map as it is, close the Map Manager and choose SAVE MAP FILE from the FILE menu. Give the .MAP file the name WVPlus.

Click CLEAR MAP(S) in the FILE menu. Choose OPEN MAP FILE and open WVPlus.MAP to confirm that its properties were saved.

Making an Entirely New Shape File Based on an Image

At times, no shape file is available, but there are images from which one could be constructed. This may be the case with building plans, refugee camps, and areas of recent change when an aerial photograph is available.

To illustrate how to make a shape file "from scratch", open Epi Map from the Epi Info menu and in the SHAPE(BOUNDARY) menu choose CREATE/EDIT. You should see a blank screen with a toolbar at the top. In the FILE menu, choose LOAD BACKGROUND IMAGE. In the resulting file dialog, change BMP files to ALL FILES, and then navigate to find CHARLESTON1996.JPG, an aerial image of part of Charleston, West Virginia, taken from an Internet site. It is in the exGIS folder. A dialog invites you to set up a coordinate system, but we do not know the exact latitude and longitude of the image, so click OK to the arbitrary coordinates provided. To illustrate how to draw polygons on top of the image, we will trace a few blocks to construct areas called Block 1, Block 2, and Block 3.

Click on the third toolbar button, the one with the irregular polygon, to draw a polygon on one of the blocks. Then place the mouse cursor at the corner of one of the blocks and click once. Not much happens. Then move the cursor to an adjacent corner. This time, you may be able to see a thin line from the site of the first click to the cursor. Click once again, and move the cursor to the third corner of the block. This time you should see a triangle forming. Click on the corner, and move to the fourth corner. This time click twice on the same spot, and a dialog should pop up asking for the name of the polygon. Enter Block 1.

It takes a little practice to carry out this process. If you run into problems while drawing a polygon, you can start it again by pressing ESCape. Previous polygons can be selected with the arrow button, and vertices can be moved by holding down the cursor until the vertex turns red and then dragging it to a new location. There is an ADD VERTEX button that inserts additional vertices if there are not enough to modify the shapes correctly. It is possible to divide a polygon, and therefore to draw the grand outline of the map and work by successively dividing it into the inner polygon areas. This helps to assure that boundaries between areas are single rather than double lines. There is also a snap function that merges two lines that are sufficiently close together, and it is possible to set the snap distance with another button.

When you have made shapes to your satisfaction, the shape file can be named and saved with the SAVE function on the FILE menu. This creates all the files that make up a shape file, with the polygon names in the .DBF, the shapes in the .SHP, etc. Data can be linked to the new shape file by polygon name or by coordinates that match those of the drawing space. As you saw when initiating the drawing process there is an opportunity to use latitude and longitude or other coordinates for the drawing.