7

USING A COMPUTER FOR FIELD INVESTIGATIONS

Andrew G. Dean and Consuelo M. Beck-Sagué

Computers are increasingly important tools for epidemiologic field investigations. Epidemiologists routinely use computers in field investigations along with questionnaires, statistics, laboratory tests, and other essential epidemiologic tools.

Computers, whether laptop, desktop, or palmtop, are machines, and, like most machines, require an investment of technical skill and setup time that must be balanced against the anticipated increase in quantity and quality of output.

Computers are most useful for:

- Tasks that are clearly defined and that will be done many times in the same way
- Rapid computation or counting involving large numbers of similar records
- Tasks matching the capabilities of existing software
- Numerically intensive calculations
- Accurate retention of details
- Investigators who have used the same system before

Manual processing is still indicated for:

- One-time or occasional tasks
- Small numbers of records

- Complex or changing tasks requiring human judgment, perhaps prior to computer entry
- Operators who are not familiar with computer use
- Situations where staffing for manual tasks is easier to obtain than computers or knowledgeable operators

Tasks that may be usefully performed on a computer during an outbreak investigation include searching for information, sample size calculation, questionnaire design, data entry, importing or exporting files in various formats, tabulation of results, statistical calculations, graphing, mapping, presentation graphics, and computer communication.

MICROCOMPUTERS

Progress in the miniaturization of computers has been nearly miraculous in the past three decades, and a description of microcomputer hardware is sure to be outdated as soon as it is printed. At present a portable computer and a printer can be carried to the field in a briefcase and operated either from batteries or standard electrical power. Palmtop computers that fit in a pocket and do not require a keyboard are becoming popular, although they are still limited compared with laptop models. A laptop or desktop computer may have a hard disk capable of storing millions of records. Portable modems make it technically possible to send files, access bibliographical databases, or search the Internet from any area with cable or telephone Internet service, although some countries place restrictions on modem use. Wireless connections are rapidly becoming available in some areas, and useful work can be performed from "Internet cafés" if private connections are not available.

The most common type of microcomputer is the Intel-compatible computer with a Microsoft Windows® operating system. Microsoft estimated in 2004 that there are 600 million copies of Windows, 36% of which are pirated.[1] Since microcomputers running some form of Microsoft Windows are ubiquitous and also permit fairly easy development of software, most epidemiologic software is available for these models. Macintosh and Linux computers require different software, but browsers in all three systems can access the Internet for searching, communication, and calculation. Windows programs can be run within or beside the other operating systems using Windows emulators or "dual boot" systems. Documents and spreadsheets can be created, edited, and shared on the Internet using only a browser, thus allowing those with different types of computer operating systems to participate.

Laptop computers are more expensive than desktop models of the same capacity but are fairly rugged and light enough to carry, and most models easily

adapt to international electrical variations. Because they have built-in batteries, they are easy to use during power outages.

The overall issues in choosing a computer include compatibility with other computers in the home office and field environment; availability of epidemiologic and statistical software; and the usual factors of cost, capacity, speed, durability, and repair service. As the age of computer communication progresses, the types of connections and provisions for security and virus checking assume greater importance.

SOFTWARE

The type of software available for epidemiologic investigation is more important than the brand of computer or operating system. During a field investigation, software may be needed for word processing, data entry, database management, data analysis and statistics, communications, bibliographical searching, and miscellaneous functions such as scheduling and note-taking. Commercial programs are available for word processing, scheduling, note-taking, graphing, and other functions that are common business applications. Data entry and database management can be done with commercial programs such as Microsoft Access, but these programs do not offer statistics for epidemiology, and setting up databases and manipulating records may require more attention than investigators are able to spare in a busy field situation. Commercial database software can also be quite expensive if multiple copies are required. Statistical software is available commercially, the most popular general-purpose programs being Statistical Analysis System (SAS, www.sas.com), and Statistical Programs for the Social Sciences (SPSS, www.spss.com). They perform a wide variety of statistical procedures for those familiar with the statistics and with programming in SAS or SPSS. Since their commands are different from those of the database programs, the use of both statistics and database programs requires learning two "languages." SAS and SPSS both offer facilities for data entry, and thus may be used without a database program, although data entry usually cannot be controlled to the extent that it can in a database program.

Epidemiologic fieldwork often requires statistics for categorical (coded or yes/no) data rather than continuous data. Mantel-Haenszel analysis of stratified data is important, and logistic regression may be desirable after preliminary Mantel-Haenszel analysis. It is important that entry, checking, coding, and editing of data be easy to perform. Setting up a new questionnaire is almost always required in a field investigation, and this should be easy to do in the software that is chosen.

The Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) have developed a program called Epi Info for use in

epidemiologic investigations that attempts to provide the best compromise between ease of use and flexibility. It is in the public domain, and versions for both DOS and Windows may be downloaded from the CDC website, copied for use by others, or translated. In this chapter, we use Epi Info to illustrate many of the tasks to be performed with computers in the field. Other free and inexpensive software for use in epidemiology can be found by searching the Internet. A recent search turned up links to free calculators for purposes as diverse as estimating caloric intake, civil engineering, producing random numbers, and doing specialized statistics, many of which could be useful in epidemiology.

Whatever software you choose, it is important that you be familiar with its use and limitations before leaving for the field. A tense field situation with high stakes and an insistent press leaves little time for learning about software or devising programs to solve new problems. The analysis does not have to be sophisticated, but it should be correct with regard to the totals obtained and the elementary statistics. Logistic regression analysis can be refined later, but the basic data must return from the field intact, properly backed up, and well documented.

THE WORKING AND TRAVELING ENVIRONMENT

To minimize problems in the field, hardware, software, and operator skills should be practiced as much as possible before leaving the home office. A "dress rehearsal" should be conducted before leaving to be sure that all necessary elements are available.

Magnetic disks must be treated like fine phonograph records and protected from fingerprints, scratches, coffee, magnets, sharp bending, and denting by firm objects like ballpoint pens. They will not be harmed by a reasonable number of passes through a modern airport X-ray machine, but metal detectors and motor-driven moving belts do generate magnetic fields that could be harmful to diskettes. Diskettes should be protected from both heat and intense cold. They should never be left in a parked car in warm weather. If possible, a portable Compact Disk (CD-ROM) writer should be used to make permanent backups of data, as optical media, particularly the CD-R or write-once CD-ROMs, are not affected by magnetic fields and are more resistant to physical abuse than floppy disks. They can be damaged by fingerprints or scratches, however. Extra copies can be made and mailed home by more than one route in case of loss or theft of luggage. CD-ROM drives are delicate, however, and should be treated gently.

When traveling, it is important to be sure that the type of power at your destination (120 vs. 240 volts) and connecting plug are known and compatible with the equipment being used. Portable computers may be run from car batteries in remote locations with appropriate adapters. Whenever possible, the computer

should be protected from voltage surges with a voltage spike protector. Increasingly, uninterruptible power supplies (UPSs) are affordable, and will protect against power interruptions for long enough to allow you to save current work and shut down the computer. The device used must be designed for local voltage levels, as voltage spike protectors designed for 110 volts perish with a puff of smoke when plugged into 220 volts. Battery power is much less subject to voltage variations.

Some countries require prior clearance to bring a computer in or out. Others have restrictions on the use of modem communications. It is important to check on such regulations with appropriate embassies, scientific colleagues, or customs officials.

In the field, your work space should be shielded from direct sun and protected from dust. The power cord for a desktop computer can be fastened to the outlet with tape or other means so that power will not be accidentally interrupted.

Organization of a portable computer's hard disk can contribute greatly to ease of use. Some investigators recommend creating a new directory for each investigation, keeping all files pertaining to that investigation in the same directory. The 1.4 megabyte floppy diskette has become a universal standard, but for files larger than one diskette, it is important to have software such as PKZIP or WinZip, which compresses files and automatically spans more than one diskette. A number of higher-capacity removable-storage devices such as ZIP and flash drives are available, but with the more proprietary formats, it is important that more than one compatible drive be available, and that both generating and receiving machines use the same drive format. Files can be transferred via Local Area Network (LAN) connections or the Internet if these options are available.

AU: OK to add?

Sending backup files to the home office can provide protection against loss of data through theft or loss of luggage during a trip. A colleague should be asked to verify that the files arrived intact, however, as various e-mail systems may refuse to transmit an attachment, or worse yet, simply remove an attachment that they sense could carry a program with malignant intent (as Hotmail does with MDB—Microsoft Access files—used by Epi Info). There are Internet file storage facilities (search for "file storage") that are less stringent in their requirements, and that may be used for transmission, and sometimes "zipping" or compressing the file will solve the problem.

When transmitting files over the Internet or other networks, it is important to protect their confidential contents. This can be done by encrypting the files with a program such as Epi Info's EpiLock, which offers 128-bit or better encryption. One can also reduce the risk of disclosing personal material by omitting names and personal identifiers as much as possible in files to be transmitted or stored.

WORD PROCESSING

Word processing is used for producing questionnaires, plans, and reports, and for recording miscellaneous observations during the investigation. A word processing package previously used by the investigator is preferable, since it takes time to adjust to a new package.

If collaborators in the investigation use different software for word processing, a common format such as "Rich Text Format" (RTF) files can be used, but compatibility should be tested in both types of software, as even standard formats are sometimes version-dependent. Plain text or ASCII files can be used as the lowest common denominator if necessary. Sending files or text by e-mail can also bridge compatibility gaps.

DESIGNING A QUESTIONNAIRE FOR COMPUTER USE

A questionnaire is a tool or template for structuring data collection so that items to be tabulated by computer or by hand are all of the same type. An item called AGE, for example, will contain data expressing age in a uniform way, perhaps as a number representing years. A good questionnaire, like a computer program or written essay, begins with an outline of major topics to be addressed. Theoretically, it is even more desirable to begin with the type of output desired and work backward to define the necessary input elements. In practice, an iterative approach to consider both input and output is often used until a satisfactory "design" is achieved.

Often the objective is to explore correlations between an illness or injury and one or more exposures or risk factors. The large topics in an outline could then be:

- A unique identifier for each record or questionnaire copy
- Identifiers and follow-up information
- Demographic information (age, sex, etc.)
- Outcome (disease or injury) as determined by the case definition
- Exposures
- Possible confounders

The desired outputs might include:

- Graph of case onset over time (Time)
- Map of cases by residence and/or workplace (Place)
- Tables of exposure by outcome (Person)

If the database design begins with a questionnaire, a series of questions is identified within each major section. These are usually given names that can also serve as field or variable names in the computer file-names, like First Name, Social Security Number, Diarrhea, and Potato Salad. Each of these can be developed into a question understandable to the subject or to the interviewer. Some, like Diarrhea, may require several questions (onset date and time, frequency, consistency, etc.) that may be summarized in a final yes/no conclusion for meeting the investigator's case definition of diarrhea.

In designing a questionnaire, it is useful to know what computer program will be used to enter and analyze the data. If Epi Info is used, the following computer terms will be useful in describing data entry and analysis.

A *field* or *variable* is one data item, such as first name or age. Usually Field is used to describe the blank in which data items are entered and Variable refers to the field name that may be manipulated later during analysis. A *record* is usually the information from one respondent to a questionnaire. Many records are stored together in a *file* or *table*. Epi Info for DOS data files end in **.rec** as in **data01.rec** and contain both data and a description of the questionnaire.

Epi Info for Windows records are stored in tables in Microsoft Access (.MDB) files. "Views" in Epi Info for Windows are separate tables containing a description of the questionnaire to be displayed on the screen. A file may be recalled for analysis or data entry, stored on floppy or hard disks, or copied from one disk to another. A file compression program such as WinZip can be used to compress files too big to fit on a single diskette, or a CD-ROM writer can be used to store larger files. The EpiLock program can be used to encrypt the file so that only those knowing the correct password can decrypt it.

In Epi Info, a field has a prompt or text question and a space for entering data. In Epi Info for DOS, the questions are typed into a text questionnaire in a word processor. In Epi Info for Windows, the MakeView program guides the design of a questionnaire through dialogues that appear after right-clicking a location on the screen. For each field, the dialogue requires a prompt and a field type, such as Text, Number, or Date. A variable name is created automatically after a question or prompt is supplied, but can be edited if desired.

Almost all data entry programs accept data of the specified type (e.g., numeric) and reject other entries (e.g., "Jones" in a numeric field). Many have sophisticated methods for evaluating entries and taking appropriate action to prevent erroneous entries. In Epi Info, for example, setting field properties or inserting commands in the Check Code scripting language for data entry allows specification of minima, maxima, legal codes, skip patterns, automatic coding, and copying of data from the preceding record. In Check Code command blocks, the user can set up more complex checks to issue an error message if a particular date precedes another date or a diagnostic code conflicts with the person's age

or gender. Check Code can also be written to do mathematics or to call another program to perform complex calculations and put the results in other parts of the data entry form.

Complex checking on data entry has a cost in terms of setup time and skill required. During an outbreak investigation with Epi Info, most epidemiologists would insert a few checks, such as ranges or legal codes, and would tell the program to skip questions shown to be irrelevant by previous answers (e.g., skip the section on symptoms if the person was not ill). If several different people will be entering the data, it may be worth spending extra time to set up checks for consistency and acceptability, but this may be less necessary if one person enters all the data and the number of records is small enough to allow manual checking after entry.

In some situations, it is preferable to enter data directly into the computer rather than using paper forms first. Direct entry has been used in door-to-door survey work and for abstracting records in medical record rooms. In most outbreak investigations, however, a paper form will be used for interviews and the results will be transferred to a computer later, perhaps in a health department office or in a motel room with a portable or laptop computer. In the future it is likely that palmtop hand-held computers will expand the possibilities for direct data entry in the field.

There are several styles of questionnaire images that may be used on the computer screen. The first is a telegraphic or "keypuncher's" form. It consists of field names and data entry blanks only, arranged on the screen to allow the fastest possible entry by a person thoroughly acquainted with both the paper and the screen forms. Such a questionnaire might begin as follows:

Idnum ____
Name _____
Age __
Sex (M/F) __
County _____
Disease _____
Chicken (Y/N) ___
Ham (Y/N) ___
Beef (Y/N) ___

In spreadsheet format, such as used in Microsoft Excel or OpenOffice.org's spreadsheet facility:

Idnum Name Age Sex (M/F) County Disease Chicken (Y/N) Ham (Y/N) Beef (Y/N)

The third style is an extended format offered by Epi Info that resembles the paper form as closely as possible, complete with headings, questions, instructions to the user, and blanks. With slight editing, the same form may be used in an actual interview. This format is most useful if there are relatively few questionnaires, if there are several people entering the data who do not have time to become "experts" on the data format (entering 100 questionnaires might produce an "expert"), or if those entering data will be frequently interrupted.

In Epi Info, either format may be used, according to the investigator's preference. With either form, the screen prompts can be more extensive than the brief name chosen for the variable to be manipulated during data analysis.

In using Epi Info and other programs, it is important to know how the program handles missing values before finalizing the questionnaire. Epi Info allows a missing value to be entered by pressing the <Enter> key to leave the field blank. Some programs record missing values as zero for numeric fields. In these programs the questions must be designed so that there is no confusion between a true code or value of zero and a missing value where this distinction is important. Zero glasses of water consumed and "unknown" glasses of water consumed, for example, are quite different, so a special code (often 9 or 99) should be assigned for the case of "unknown." Such codes are unnecessary in Epi Info and most current data entry programs, since missing data are stored as values distinct from zero.

In some investigations, particularly in research settings, it is useful to assign additional codes (e.g., 8's) to distinguish answers cited as "unknown" by the subject, those considered less accurate or unknown by the interviewer, and those somehow omitted during data entry. These extra codes can complicate the analysis considerably and should be assigned only after careful thought about the format of the table that will show the results. "Somebody might ask about it later" is not sufficient reason to burden the investigation with a series of extra codes unless they contribute meaningfully to the analysis. In a field investigation it is often sufficient to use only one kind of missing value, since the modest number of cases and rough-and-ready data-collection process may not permit analysis of bias that may have arisen due to more than one type of missing data.

To provide proper analysis of questions, codes should be assigned during data entry. Merely typing in the names of counties or diseases can result in a profusion of synonyms and misspellings that is impossible to analyze. Either numeric or text codes may be used. When producing tables during analysis, codes indicating the actual values are more useful than numeric codes, although numeric codes can be recoded to produce useful labels during analysis. Generally "Y" and "N" are less likely to produce errors in data entry than "0" and "1", and "URI" is more meaningful than "7002" for Upper Respiratory Infection.

A key issue in setting up data entry forms involves multiple-choice questions. The question:

How many glasses of water do you drink per day (choose one)?

0. None
1. 1–2
3. 3–4
5. 5 or more
9. Don't know
Water #

has five mutually exclusive answers including a blank entry, and the entire question, therefore, has a single answer. A one-digit numeric field called WATER is enough to record the answer.

Another type of question is:

What symptoms have you had in the past month?

1. Diarrhea
2. Fever
3. Chills

Note that all three symptoms might have been present. Each part of what looks like a single question requires a yes/no answer, and this question should be set up as follows:

What symptoms have you had in the past month?

Diarrhea <Y>
Fever <Y>
Chills <Y>

The same would be true of a list of foods possibly eaten at a meal. Each item is really a separate question, since the answers are not mutually exclusive. In the Analysis program in Epi Info, the first question is summarized with the command FREQ WATER, to display the codes for each level and the number of times each code is represented.

AU: Is this correct?

The symptom question is more complicated, however. By asking for a frequency distribution of the variable Diarrhea (FREQ DIARRHEA, in Epi Info), it is a simple matter to ascertain the number of persons with and without diarrhea. But discovering how many symptoms each person had takes more complex

programming—complex enough so that it may be easier to add another summary question below the list of symptoms, such as "Number of symptoms #," if this is important for the analysis. The person entering data can quickly scan the paper form, count symptoms, and enter this number rather than requiring the investigator to do extra programming during the analysis stage. The trade-off between intelligent data consolidation during data entry and having the computer do the work is evident at many points during the design of computer entry forms and paper questionnaires. If you will be using both, consider simplifying as much as possible the data transferred to the computer from the paper form. Names, addresses, and other follow-up information may be omitted, and complex case definitions may be summarized with a single yes/no question. Field investigation usually results in scores or hundreds of questionnaires, and the human mind and eye may be a simpler processing alternative for some kinds of questions than having a busy investigator with modest computer skills try to write a program to condense the data electronically.

In the end, the investigator must decide what to collect, how much of a completed questionnaire to process by hand, and in what form to code it for computer use. Although experience plays a major role, pilot testing can be a good substitute. A pilot test might consist of entering data from five or six instances of a questionnaire (preferably from people who will not be included in the final study). These are then processed to produce a model for the final analysis, saving the program that results. This procedure will often reveal gaps, inconsistencies, or ambiguities in the questionnaire and point out questions that do not contribute to the analysis, and is almost guaranteed to improve the final questionnaire design. Before finalizing the design, the investigator should examine each question with the additional questions, "What do I really want to know? " and "How am I going to process this variable?"

DATA ENTRY AND VALIDATION

Usually paper questionnaires from the field are far from ready for analysis after data entry. They contain misspellings, synonyms, abbreviations, upper/lower case mixtures, marginal notes, and missing data. Data entry is an opportunity for partial "cleaning" of the data set. It must be done with scrupulous dedication to preventing bias—the kind that could insert data favorable to a hypothesis or eliminate items detrimental to it. Since field investigations seldom have the luxury of "blind" coders and data entry personnel, only strict and literal attention to accuracy can prevent bias.

It is a good idea to alternate case and control forms during data entry to avoid bias from the small decisions and adaptations that occur during the course of

entering forms. If there is more than one data entry person, each should enter the same ratio of case to control forms.

In most data entry systems, including Epi Info, a cursor on the screen indicates where entry will occur. The cursor jumps automatically from field to field. When an entry is made, the item is checked for correct type (numeric, date, etc.) and additional checks programmed into the check file are performed. If a problem is encountered, the program indicates this and waits for correction before going on to the next field. At the end of each questionnaire, the record is saved automatically or by answering an explicit question such as "Save data to disk? (Y/N)". In Epi Info, a power failure (or someone tripping over the power cord) will not result in loss of records already saved, although the partial record being entered may have to be reentered. If other programs do not have this feature, save your work frequently. It is a good idea to mark each paper questionnaire as data entry is completed to avoid accidental reentry.

When all records have been entered, the entries should be carefully validated to be sure that they represent the source documents accurately. One person can read the data entered aloud while the other verifies that the entries represent the source document accurately.

Further checking may be done by performing frequencies on each field. FREQ * will accomplish this in Epi Info. Examining the results will often disclose outliers such as "*Gf!" that crept in during a moment of distraction. These may be edited in the data entry program before beginning the actual analysis.

Some investigators prefer to have the same set of questionnaires entered in duplicate by two different operators in separate files. The two files are then compared, and differences are reconciled by a person authorized to make data entry decisions. Programs for making the comparison are included with Epi Info.

At intervals during data entry and after it is completed, backup copies should be made and stored in a secure place away from the original computer. Placing encrypted copies in a file storage facility on the Internet will offer additional security.

ANALYSIS OF DATA IN FIELD EPIDEMIOLOGY

Analysis of a descriptive study or survey usually begins with a simple frequency for each variable (in Epi Info, FREQ *). Then, for a study with two or more groups, such as cases and controls, ill and well, exposed and unexposed, you would want to compare the two groups. For categorical (coded) data the TABLES command in Epi Info, for example TABLES * ILL, will produce cross-tabulations of each variable by illness status (Y/N), with appropriate statistics for each.

Often in a case-control or cross-sectional study, a histogram or epidemic curve is needed. In Epi Info, the case group would first be selected before doing the histogram, for example, SELECT CASE = "Y." The histogram might be performed with: HISTOGRAM ONSETDATE. Continuous variables such as age or diastolic blood pressure are analyzed with the MEANS command, for example, MEANS SBP ILL if SBP is systolic blood pressure and ILL is case status.

In most analytic programs it is necessary to use names of variables to do an analysis. Unlike algebraic notation, computer programs usually allow a descriptive name for each field. In some programs the length of these names is limited to, for example, 8 or 10 characters. If transfer of data from one program to another is contemplated, be sure that variable names truncated to the length allowed in the most restrictive program are unique. For example, "ADDRESSLINE1" and "ADDRESSLINE2" might both emerge as ADDRESSLIN in a program with variable names limited to 10 characters. "ADDRESS1" and "ADDRESS2" would survive truncation to eight characters, however. Many programs will object to names beginning with a number, however.

After doing frequencies for each field, you will have an idea of how many records are in each group, and how many missing values there are for each field. If missing values are displayed, many of the tables may be three-by-three rather than two-by-two tables, and the statistics that result are not as complete as those that accompany two-by-two tables. Some packages allow you to suppress missing values (in Epi Info, with the SET command). Repeating the analysis after giving this command will omit the missing values and focus the analysis solely on records that have data for the tables and frequencies being produced. Two-by-two tables in Epi Info are automatically accompanied by chi-square tests, odds ratios, risk ratios, confidence limits, and Fisher and mid-P exact tests.

Often one or more "significant" findings may be indicated by p values less than 0.05 or confidence limits that exclude 1.0 for odds ratios or risk ratios. Further analysis to consider confounding variables is indicated, at least for frequent confounders such as age and sex. This is done by stratifying the table of interest (say SALAD by ILL), producing a separate table for each value of the confounder.

In Epi Info, the crude table is produced by TABLES SALAD ILL and stratification by sex by TABLES SALAD ILL SEX to produce separate tables for males and females. The Mantel-Haenszel summary chi-square and p-value for the stratified result may be compared with the results of the crude analysis. If the odds ratios in the two or more strata are similar, interaction is not present, and a difference in the crude and Mantel-Haenszel odds ratios may be taken as an indication that sex was a confounder. Other potential confounders such as age, socioeconomic status, etc., can be evaluated similarly, either one by one or in combination (TABLES SALAD ILL SEX RACE).

Stratification does not work well for small data sets if there are many strata, and variables such as age may need to be recoded to produce fewer strata, such as

CHILD and ADULT, rather than a number of age groups. Examples of data manipulation, including automation of a complex case definition, are included in the *Epi Info for DOS* manual in a chapter on epidemic investigation.

At this point the analysis may be complete enough for field purposes, if confounding has been identified and eliminated through stratification, and interaction has been addressed (perhaps recording the results for more than one stratum rather than the overall results, as in "For people up to the age of 18, the effect was_____ . . .; those over 18 did not react the same way."). The significant findings must be evaluated from a biomedical point of view and distributed to interested parties.

Graphing important findings may be helpful in visualizing or explaining results, particularly those pertaining to temporal variables. Epi Info offers bar, histogram, pie, scatter, and line graphs. Two variables are required for scatter graphs, and one variable for other formats. A second variable can be represented by plotting a separate graph for each of its values, as in showing date of onset in separate histograms for males and females, for example.

In cases where there are several significant risk factors or several confounders, logistic regression may be helpful. Logistic regression is offered as part of Epi Info, and a number of other programs are available for this purpose after exporting data from Epi Info for DOS.

GEOGRAPHIC INFORMATION SYSTEMS AND THE ANALYSIS OF "PLACE"

In some outbreaks, the place of residence, work, visitation, food or water consumption, or aerosol inhalation is important in the analysis. Geographic Information Systems (GIS) are used to link database information with maps or other graphics to provide opportunities for spatial analysis. Locations can be recorded in variables such as city, state, mail code, or census tract. Exact locations can also be recorded as longitude and latitude, obtained through geocoding or from field measurements with a hand-held geographic position sensor (GPS). *Geocoding* means obtaining exact longitude and latitude coordinates from street addresses or other location information. It is done through the use of special geocoding databases or services available commercially or provided on the Internet.

One or more geographic variables must match the geographic information in the "map," called a *boundary file* in Epi Info 6 and Epi Map for DOS, and a *shape file* in Epi Info/EpiMap for Windows. City names must be spelled the same way in both data sets, for example, or point coordinates must be in the same units, such as latitude/longitude in decimal degrees. The GIS software combines the map image with the database to show locations as colors, patterns, dots, or other symbols that represent spatial information. An entire science has developed around

the analysis of geographic information, but the basic operations can be performed in Epi Info for Windows with the Epi Map program, and refined if necessary in dedicated GIS software. S.B. Eng describes the use of dot maps in investigating an outbreak.[2] Epi Info uses software produced by the makers of ArcView (Environmental Systems Research Institute, Inc., www.esri.com) so that maps can be displayed by either system.

OBTAINING AND USING EXISTING COMPUTERIZED DATA

Sometimes useful computerized information already exists at the site of an investigation. For example, hospital computer systems may have laboratory values, diagnostic information, or operative schedules; a water treatment plant may have results of water analysis. Such files may contain more information than is relevant and may be in a variety of file formats. Selection of relevant information can be done by the person managing the data system. If you specify a time period or category of record to be selected, it may be relatively easy for the data manager to create a file containing only the desired items, perhaps with only certain fields represented.

The file format is also important. Most computerized database and statistics programs, including Epi Info, will accept an ASCII file in fixed-field format. This means that only the first 128 standard characters are included, and each line represents a different record. A field is distinguished by its position on the line and either occupies a fixed number of characters or is terminated by a comma or other delimiter. It is important to obtain a list of the fields, their types and length, and the delimiter(s) used.

Epi Info for DOS will analyze files in the DBASE format directly and will import files in the Lotus 1-2-3, comma-delimited, DBASE, and fixed-field ASCII formats with a program called Import. The Analysis program in Epi Info will read and analyze files in more than 20 different formats and can write files in these same formats, allowing for extremely flexible data conversion.

Whenever external files of any kind are copied, the source disk should first be checked for computer viruses with a suitable program, no matter how reputable the supplier of the data. Reference data such as telephone lists or the Epi Info manual may be carried to the field as files on CD-ROMs or the computer's hard disk, so that heavier paper copies are not needed.

COMPUTER COMMUNICATIONS

A computer equipped with a wireless connection, an Ethernet connection to a local network, or a modem can be used to send files of any type to another

computer or the Internet. In many parts of the world, Internet access is available for modest charges in Internet cafés. If available, the Internet provides not only facilities for e-mail but also access to searches, guidelines, textbooks, calculators, reference data, and information (of variable quality) on almost any conceivable subject. J. Woodall provides a review of the use of computer networking in investigating disease outbreaks, with particular reference to biological and toxic weapon use.[3]

It is not always easy to connect a portable computer to a telephone line, as many businesses and hotels have digital telephones that do not work with standard types of modems. Hotels increasingly have made special provisions for wireless (Wi-Fi) connections or fast Internet by cable connection, and asking about these facilities in advance is a good idea before traveling to a field site.

OBTAINING INFORMATION FROM THE WORLD ELECTRONIC AND PRINT LITERATURE WHILE IN THE FIELD

Unless the investigator is a specialist in the type of problem being investigated, bibliographic searching may be of great importance. The MEDLARS database of the National Library of Medicine is the most comprehensive source of medical and public health information. It contains references and often abstracts describing millions of articles in thousands of biomedical journals. Searches can be performed free of charge at the National Library of Medicine portal http://gateway.nlm.nih.gov/gw/Cmd or at other sites that allow MEDLARS searches.

The website www.google.com provides free-text searching, returning first the references most heavily cited by others, thus filtering out much of the chaff from billions of possibilities. Other search engines, such as www.yahoo.com, provide classification hierarchies that may be better for reviewing a systematic field of knowledge. In either case, the Internet is becoming more and more a reflection of the state of the world and of both episodic and cumulative information that cannot be ignored. A quick search of the Internet is often a practical way to obtain a grasp on a new field, such as air handling, plumbing, laws, lay medical advice, organizations dealing with relevant problems, and even telephone numbers or methods of locating people.

Most computer products are supported by websites, including Epi Info, and it is possible to download a free copy of Epi Info, or an update, from www.cdc.gov/epiinfo/. Many hardware companies such as printer manufacturers provide free downloads of current drivers for their products. Many statistical calculators can be accessed from www.openepi.com and the links it provides to other sites. Maps in great detail are available for downloading, and Epi Info contains an on-line reference chapter with links to hundreds of sites that provide resources for mapping.

OBTAINING TECHNICAL ASSISTANCE DURING A FIELD INVESTIGATION

Occasionally a computer problem arises in the field that requires more expertise than the investigator possesses. Computer breakdowns, unfamiliar file formats, access to special printers or other equipment, and difficulties with telephone connections may all require assistance. Technical expertise is available in most communities from a variety of sources. If calling your home base support staff does not solve a problem, a search of local health departments, technical schools, computer stores, and computer clubs may lead to a person with the necessary knowledge or piece of equipment. The Internet and e-mail provide sources of information that can be accessed at any hour of the day or night because of the differences in time zones. Googling for information with words from error messages or your own description of the problem frequently turns up others who have met and conquered the problem.

COMPUTER VIRUSES AND DATA BACKUP

There is little satisfaction in having written a book whose only manuscript was lost in a fire. Similarly, proper backup of computer data is essential. Whatever can go wrong should be expected to do so—perhaps more than once. In the past few years computer viruses have been added to the list of things that can go wrong, but they are only an additional cause for careful backup procedures that already were necessary to protect against hard disk crashes, power outages, theft, and late-night human errors.

Computer viruses are becoming more and more prevalent. They cause a variety of problems, but the most serious destroy all data on disks used in a particular computer. They may be acquired from a source outside a previously uninfected computer, either by copying files or through communication with another system.

Commercial programs are available to detect and often remove these viruses, and one of these should be used to check all disks inserted into the computer before copying any files, processing data, or running programs. A suitable virus protection program should be active at all times in a computer, and special care should be taken to check diskettes that may have been in other computers and become infected. Portable computers are attractive to thieves, and their hard disks—like all hard disks—may "crash," making data difficult or impossible to recover. More than one floppy disk or CD-ROM copy of all data should be made on a regular basis, and the backups should be carefully stored in places separate from the computer itself, to rule out the possibility of complete loss from theft,

carelessness, or fire. Several well-verified disks, traveling by different routes, mailed home, and/or stored with different people are the best backup system. New backups should be made at intervals, perhaps every hour or two during data entry. It is also useful to have CD-ROM copies of important software in case a hard disk must be replaced in the field. As described in a previous section, Internet file storage facilities can be used for off-site backup if the files are first encrypted to protect confidentiality.

Generally in a field investigation it is practical to give new names to each new set of backup files so previous files are not written over. If anything goes wrong with a current file or disk, the previous set of files may provide a good copy of most of the data set. Although good commercial programs are available for backing up hard disks, they are usually not necessary in field investigations, since the data files are usually small. The files may simply be copied to floppy disks, maintaining several such carefully labeled disks to be used in sequence.

AU: Usually CDs or flash drives these days. Update?

When things go wrong, a frequent reaction is to make the problem worse through panic. If difficulties in recovering files are experienced, first obtain technical help in diagnosing the problem. If you decide to restore files from the backup disks, be sure that the WRITE-PROTECT function (see previous section) is set on these disks to avoid having the backups destroyed by a virus or faulty procedure. If files have been accidentally erased on the hard disk, it is important to avoid entering further records or copying files until an attempt has been made to recover the lost files. Programs such as Norton Utilities can restore erased files and repair many corrupted files if they have not been written over by further manipulations.

DATA CONFIDENTIALITY AND LEGAL ISSUES

Maintaining confidentiality of data on a portable computer is similar to protecting a stack of questionnaires. The best protection is through maintaining careful physical custody of any disks containing data, including the internal hard disk of the computer. With small data sets, files can be kept on floppy disks so the hard disk does not contain confidential data. In many investigations, names and addresses are not needed in data files, and such data should not be entered unless it is absolutely necessary. Arbitrary identification numbers are adequate for most computerized data sets. Frequently names and other identifiers may be left with the local health department and only code-identified data transported to a more central site. Encryption programs or compression programs with password protection should be used to protect data in case CD-ROMs or diskettes are stolen, lost in the mail, or the computer itself is stolen. A program for 128-bit encryption is provided with Epi Info as the utility EpiLock. If the password used for encryption is lost,

however, the encrypted file cannot be recovered, so adequate password management is essential.

Occasionally outbreaks lead to legal proceedings for negligence or even homicide. Records of the investigation may be subpoenaed or otherwise required for legal purposes. This possibility and scientific documentation make it important to keep good records of the investigation and to store them in such a way that they can be accessed by appropriate parties even if the investigator moves on to another job. Analytic programs may be written with comments explaining important steps, which also facilitates reuse of the programs in another investigation.

Computer disks should be carefully labeled, and after the investigation stored in an organized way so others can access the files. Paper copies of the data may be made for permanent documentation and ease of filing, since computer disks lose their magnetic data after a few years and for archival purposes should be copied to new disks annually or stored on CD-ROMs.

THE FUTURE OF COMPUTERS IN EPIDEMIOLOGIC FIELD INVESTIGATION

Future computers for field investigation will be smaller, lighter, and more powerful. Soon both voice and handwritten input will be practical. Medical and other records will be computerized to a greater extent, offering opportunities for capturing relevant information in detail for the investigator with the skills and tools to convert data from diverse formats. Eventually, perhaps, better programs will alleviate some of the compatibility problems between various types of software, but the competitive marketplace will ensure that other types of incompatibility arise. Palmtop computers will extend direct data collection to environments such as earthquake sites, the bedside, or other field locations, and digital cameras will find more use in documentation.

The Internet has begun the process of providing access to the entire world from a portable computer as though all the world's resources resided on a single hard disk ("in the cloud"). Current Internet sites offer storage of files, document and spreadsheet sharing, statistical calculations, and many other functions that are independent of computer brand or operating system and accessible anywhere there is an Internet connection. Search capabilities are used, not only for bibliographical purposes, but through access to news articles, as a way of monitoring (for example) influenza activity. Actual investigations may be carried out via the Internet, as many have already been conducted or aided by e-mail communication.[4] There is a growing field of computer forensics dealing with the investigation of computer crime, computer viruses, and corporate malfeasance.

Like most aspects of field investigation, computer use will continue to require ingenuity and adaptation. Those who have acquired the skills for using a portable computer, however, will find that the rewards in quantity and quality of epidemiologic work accomplished make it an indispensable companion in field investigation, and that the communication and information access offered by the Internet are becoming more and more central to the epidemiologic process.

REFERENCES

AU: No longer accessible, and rerouted to Jupiter research (?) Please update.

Note: References to Internet addresses supplied in the text may change but can usually be recovered by searching for the topic of interest or a trade name with an Internet search engine.

1. Website: www.microsoftmonitor.com/archives/003453.html, accessed 20 Nov 2006.
2. Eng, S.B., Werker, D.H., King, A.S., et al. (1999). Computer-generated dot maps as an epidemiologic tool investigating an outbreak of toxoplasmosis. *Emerg Infect Dis* 5(6), 815–9.
3. Woodall, J. (1998). The role of computer networking in investigating unusual disease outbreaks and allegations of biological and toxin weapons use. *Crit Rev Microbiol* 24(3), 255–72.
4. Kuusi, M., Nuorti, J.P., Maunula, L., et al. (2004). Internet use and epidemiologic investigation of gastroenteritis outbreak. *Emerg Infect Dis* Mar [date cited]. Available from: http://www.cdc.gov/ncidod/EID/v0110n02/02–0607.htm

AU: Please supply date… but see below.

AU: Page not found: please update and provide date above.

FURTHER READING

Beck-Sague, C., Jarvis, W.R., Martone, W.J. (1997). Outbreak investigations. *Infect Control Hosp Epidemiol* 18, 138–45.

Dean, A.G., Arner, T.G., Sangam, S., et al. (2002). Epi Info 2002, a database and statistics program for public health professionals for use on Windows 95, 98, NT, 2000, ME, and XP computers. Centers for Disease Control and Prevention, Atlanta. (Can be downloaded from www.cdc.gov/epiinfo/).

Dean, A.G., Dean, J.A., Burton, A.H., et al. (1991). Epi Info: a general-purpose microcomputer program for public health information systems. *Am J Prev Med* 7(3), 178–82.

Epidemiologic Case Studies. Available from: http://www.cdc.gov/epiinfo/tutorials/ and http://www.epiinformatics.com/.

Gerstman, B.B. (2000). Data analysis with Epi Info. http://www.sjsu.edu/faculty/gerstman/EpiInfo/.

Zubieta, J.C., Skinner, R., Dean, A.G. (2003). Initiating informatics and GIS support for a field investigation of bioterrorism: the New Jersey anthrax experience. *Int J Health Geogr* 2, 8. Published online 2003 November 16. doi: 10.1186/1476–072X-2–8.

AU: Is this the URL for the Zubieta article? Please provide full information.