Field Epidemiology

Edited by
MICHAEL B. GREGG

Associate Editors
Richard C. Dicker
Richard A. Goodman

New York Oxford
OXFORD UNIVERSITY PRESS
1996

12

USING A MICROCOMPUTER FOR FIELD INVESTIGATIONS

Andrew G. Dean

In the past decade the microcomputer has become an important tool for epidemiologic field investigations. Epidemiologists routinely use portable computers in field investigations along with questionnaires, statistics, laboratory tests, and other more traditional epidemiologic tools.

A computer is a machine, and, like most machines, it requires an investment of technical skill and setup time that can be recovered through increased quantity and quality of output.

Computers are most useful for:

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

Manual processing is still indicated for:

1. One-time or occasional tasks
2. Small numbers of records

3. Complex or changing tasks
4. Operators who are not familiar with computer use
5. 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 word processing, processing of questionnaire data, analysis of existing electronic data, computer communication, and bibliographic searching.

MICROCOMPUTERS

The pace of progress in the miniaturization of computers in the past 15 years has been nearly miraculous; therefore a description of microcomputer hardware is sure to be outdated as soon as it is printed. At present, a portable computer and printer can be carried to the field in a briefcase and operated either from batteries or standard electrical power. The computer may have a hard disk capable of storing thousands of questionnaire records and generally would have most of the features of its desktop cousins at the office at a slightly higher cost. Portable modems make it technically possible to send files or access remote databases from any area with telephone service, although some countries place restrictions on modem use.

The most common type of microcomputer is the IBM-compatible computer with the DOS or Microsoft Windows operating system. Since IBM-compatible microcomputers are ubiquitous and also permit fairly easy development of software, most epidemiologic software is available for these models. The Macintosh computer, known for its ease of use and minimal barriers to learning, is now available in "notebook" size models that should increase its use in the field. Both types of computers are fairly rugged and light enough to carry, and some models are easily adapted to international electrical variations. 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.

SOFTWARE

The type of software available for epidemiologic investigation is of more importance 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, bibliographic 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 *dBASE* and *Paradox*, but these do not offer statistics for epidemiology, and setting up databases and manipulating records requires more attention than many investigators are able to spare in a busy field situation. This software is quite expensive if multiple copies will be required.

Statistical software is available commercially, the most popular general purpose programs being *Statistical Analysis System* (*SAS*)[1] and *Statistical Programs for the Social Sciences* (*SPPS*).[2] They both require considerable amounts of hard disk space but will 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." Both *SAS* and *SPSS* 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) rather than continuous data. Mantel-Haenszel analysis of stratified data is important, and for those who know how to use and interpret it, 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 and the World Health Organization have developed a program called *Epi Info*[3,4] for use in epidemiologic investigations; it attempts to provide the best compromise between ease of use and flexibility. It is in the public domain and may be copied for use by others. Versions are available in French, Spanish, Arabic, Russian, and Chinese, and translations of the manual in several other languages. In this chapter we will use *Epi Info* to illustrate many of the tasks to be performed with computers in the field. A companion program called *Epi Map*[5] provides features for producing and editing community maps in the field and for displaying epidemiologic data through shading, patterns, or dots on the map. Other free and inexpensive software for use in epidemiology is listed from time to time in the *Epidemiology Monitor*.[6]

Whatever software is chosen, it is important that the investigator 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 can wait until 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 have been used as much as possible before leaving the home office. At the very least, 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 the metal detectors through which passengers walk 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.

When traveling to other countries, it is important to be sure that the type of power (120 vs. 240 volts) and connecting plug are known and compatible with the equipment being used. With appropriate adapters, portable computers may be run from car batteries, or even by solar power in remote locations. Battery power is much less subject to effects from voltage variations found in many developing countries. Some countries require prior clearance for bringing 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, the computer work space should be shielded from direct sun and protected from dust. The power cord for the computer should 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 profusion of disk sizes and densities, even on IBM-compatible microcomputers, can lead to problems in transferring data from one computer to another. It is useful to carry appropriate cables and software to transfer files via serial-port connections. One such product is Laplink III (Traveling Software, Bothell, WA 98011).

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 preferred, since considerable time may be needed to adjust to a new package.

The software (and the investigator) must be capable of producing a plain text or "ASCII" file for transfer to another word-processor. Collaborators in the investigation may use a brand of word processor with an incompatible proprietary file format, but most word processors will accept a plain text file as input.

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. A good questionnaire, like a computer program or written essay, begins with an outline of major topics to be addressed. Often the objective is to explore correlations between an illness or injury and one or more exposures or risk factors. The large topics in the outline would then be:

Identifiers and follow-up information
Demographic information (age, sex, etc.)
Disease
Exposures
Possible confounders

Within each section, a series of questions is identified. 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 intelligible 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 on meeting the investigator's case definition of DIARRHEA.

In designing a questionnaire, it is useful to understand the computer program that will be used to enter and analyze the data. A few 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 questionnaire. Many records are stored together in a FILE. Files are given names of eight or fewer letters followed by a period and an optional three letters. *Epi Info* data files end in .REC, as in DATA01.REC. A file may be recalled for analysis or data entry, stored on floppy or hard disks, and copied from one disk to another.

A field usually has a textual question or prompt, a maximum number of characters (length), and a name (up to 10 characters in *Epi Info*). In *Epi Info* for ex-

ample, a questionnaire (and a database specification) might be a file, created on a word processor, that begins with the following lines:

```
        Division of Epidemiology
        Public Health Department

{Id}entification {Num}ber ###
Name _____ Age ## Sex <A> (M/F)
```

Epi Info would automatically create four fields as follows:

Field name:	IDNUM	NAME	AGE	SEX
Field type:	Numeric	Text	Numeric	Upper-Case Text
Field length:	3	22	2	1

In *Epi Info* a pound or number symbol (#) requires numerical input. An underline allows any kind of text. The <A> represents an upper-case text field that converts entries to upper case, to avoid having sexes ("m," "M," "f," and "F") during the analysis phase, for example.

Other data entry programs may ask the user to specify "field name," "field type," and "field length" for each individual field as the database or file is being set up. The end result is the same: the program displays a prompt on the screen and allows entry of data in a blank field. 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, a program called *CHECK* allows specification of minima, maxima, legal codes, skip patterns, automatic coding, and copying of data from the preceding record. By inserting statements in a special check file, 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 files can also be set up 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 set-up time and skill required. During an outbreak investigation with *Epi Info*, most epidemiologists would insert a few checks, such as maxima and minima 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 not be worth the trouble 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.

There are two 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       <A>
County    <A>
Disease   <Y>
Chicken   <Y>
Ham       <Y>
Beef      <Y>
```

The second style is an extended format 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, 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 those entering data will be frequently interrupted.

In *Epi Info*, either format may be used, according to the investigator's preference. With the extended form, field names may be explicitly chosen by placing curly brackets around the most significant 10 characters, as illustrated previously.

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 (and a previous version of *Epi Info*) 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 that a special code (often 9 or 99) should be assigned for the case of "unknown." Such codes are unnecessary in the current version of *Epi Info* (version 6), since missing data are stored as values distinct from zero.

In some investigations, particularly in research settings, it is useful to assign additional codes (for example, 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 only be assigned 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 cumbersome codes unless their analysis accomplishes a specific objective. 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 do not permit analysis of bias that may have arisen due to more than one type of missing data.

To provide proper analysis of many questions, codes must be assigned. Merely typing in the names of counties or diseases can result in a profusion of synonyms and misspellings that is impossible to analyze. In *Epi Info*, either numeric or text codes may be used. In 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; 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 *Epi Info*, analysis of the two sample questions would proceed as follows. To obtain information on the number of persons having various levels of water consumption in the first style of question, the command FREQ (frequency) WATER will display the codes for each level and the number of times each code is represented.

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. 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 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, rather than thousands, 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. With modern systems such as *Epi Info*, it is quite easy to enter data from five or six sample questionnaires (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; it is almost guaranteed to improve the final questionnaire design. Before finalizing the design, each question should be examined with the additional questions hovering in the background, such as "What do I really want to know? " and "How am I going to process this field?"

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 an 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 the 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*. An examination of 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 the actual analysis is begun.

Some investigators prefer to have the same set of questionnaires entered in duplicate by two different operators in separate files. The *Epi Info* program *VALIDATE* may be used to compare the two files and detect and correct any differences.

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* (e.g., 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 (e.g., 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 (e.g., 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 analysis. Unlike algebraic notation, computer notation usually allows a descriptive name (up to 10 characters in *Epi Info*) for each field.

At this point, you will have an idea 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, 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*, SET IGNORE MISSING = ON). 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 accompanied by chi-square tests, odds ratios, risk ratios, confidence limits, and, if indicated, Fisher 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 TABLES SALAD ILL GENDER.

In the stratified results, there are separate tables (strata) for males and females. The Mantel-Haenszel summary chi-square test and P value that summarize the combined tables 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 GENDER was a confounder. Other potential confounders such as AGE, socioeconomic status, and so on can be evaluated similarly, either one by one or in combination (TABLES SALAD ILL GENDER SES).

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

strata. A number of examples of data manipulation, including automation of a complex case definition, are included in the *Epi Info* manual in a chapter on epidemic investigation.[4]

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 of important findings may be helpful in visualizing or explaining results. *Epi Info* offers the commands BAR, HISTOGRAM, PIE, SCATTER, and LINE, each followed by the name of a variable (two variables for SCATTER and several for LINE), to produce graphs.

In cases where there are several significant risk factors or several confounders, logistic regression may be helpful. For logistic regression in *SAS* or *SPSS*, the *CONVERT* program in *Epi Info* may be used to produce files for importation into these programs.

MULTLR,[7] *LOGISTIC*,[8] and *CLOGISTIC*[9] are public domain or shareware logistic regression programs for microcomputers. Exportation of files for *MULTLR* is included in Version 6 of *Epi Info*. *LOGISTIC* and *CLOGISTIC* have recently been adapted by their author so that they read *Epi Info* files directly. A list of free and inexpensive programs available for microcomputers is available.[10]

OBTAINING AND USING EXISTING COMPUTERIZED DATA

Sometimes useful computerized information already exists at the site of an investigation. Hospital computer systems may have laboratory values, diagnostic information, or operative schedules; a water treatment plant may have results of water analysis, and so on. 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 128 standard characters are included and each line represents a different record. A field is distinguished by its position on the line and always occupies a fixed number of characters. It is important to obtain a list of the fields and their types and length.

Epi Info 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*.

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 transported as files on hard or floppy disks, so that heavier paper copies are not needed.

COMPUTER COMMUNICATIONS

A computer equipped with a modem can be used to send files of any type to another computer over the telephone system. Guidelines, library searches, memos, and even programs can be exchanged with the epidemiologist's home office. A modem can also provide access to electronic mail (E-mail) systems. Fax machines or fax boards installed in a computer may be used to communicate textual or graphic material.

Communication is sometimes difficult to set up due to the variety of protocols, telephone connections, and communications programs. If computer communication is important on an investigation, practice sessions should be conducted to work out the details before leaving for the field. A variety of telephone-jack adapters can be obtained at radio or electronic stores. Methods for obtaining technical help in a community are discussed in a later section.

OBTAINING INFORMATION FROM THE WORLD 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. In the United States (and many other countries where it is available), the *MEDLARS* database of the National Library of Medicine is the least expensive and most comprehensive source of information. It contains references and often abstracts describing millions of articles in thousands of biomedical journals. "*GRATEFUL MED*" is a computer program[11] that allows *MEDLARs* searches from a portable computer in the field. It is necessary to open an account with the National Library of Medicine and highly advisable to practice using *GRATEFUL MED* for searching before leaving for a field investigation. Most medical libraries can also perform *MEDLARs* searches; an alternative is to locate a medical library and willing librarian to perform searches near the site of the investigation.

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. With the number of microcomputers in the United States alone estimated at 35 million,[12] technical expertise is available in most communities from a variety of sources. If calling the epidemiologist's 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.

COMPUTER VIRUSES AND DATA BACKUP

Just as there is little satisfaction in having written a book whose only manuscript was lost in a fire, there is little satisfaction in having gathered a great deal of data that is then lost. Therefore, 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 reason for careful backup procedures, which were already 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. If you have brought disks of software that you will copy onto a local computer for use, be sure that the disk is write-protected (notch covered with tape for 5 1/4" disks, slider open for 3 1/2" disks) before inserting your disk in the local computer, so as to prevent your disk from becoming infected from the computer.

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 copy of all data should be made on a regular basis, and the backup disks 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 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 floppy disk copies of important software in case a hard disk must be replaced in the field.

Generally, in a field investigation, it is practical to give new names to each new set of backup files, so that 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.

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 them. Programs such as *Norton Utilities*[13] 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 microcomputer is similar to protecting a stack of questionnaires. The best protection is through maintaining careful physical custody of any disks containing data, including, if necessary, the internal hard disk of the computer. With small data sets, files can be kept on floppy disks so that 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 community health department and only code-identified data transported to a more central site.

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 and the interest of good 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. This 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 that 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. For archival purposes, the data should be copied to new disks annually.

THE FUTURE OF COMPUTERS IN EPIDEMIOLOGIC FIELD INVESTIGATION

Future computers for field investigation will be smaller, lighter, and more powerful. Eventually 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 to 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.

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, find that the rewards in quantity and quality of epidemiologic work accomplished make it an indispensable companion in field investigation.

NOTES

Use of trade names is for identification only and does not constitute an endorsement by the U.S. Public Health Service.

Epi Info and *Epi Map* are available without charge via the Internet computer network at address:

ftp.cdc.gov
/pub/epi/epiinfo
and
/pub/epi/epimap

Two suppliers of *Epi Info* and *Epi Map* and their printed manuals are:

USD, Inc.
2075A West Park Place
Stone Mountain, GA 30087 U.S.A.
(404) 469-4098
Fax (404) 469-0681

Brixton Books (North America)
740 Marigny Street
New Orleans, LA 70117
USA

Brixton Books
P.O. Box 4398
London SW9 9xJ
United Kingdom

PHN (504)944-1074
FAX (504)947-8899
gfegan@mailhost.tcs.tulane.edu

REFERENCES

1. SAS Institute, Inc. (1985). *Statistical analysis system*. SAS Institute Inc., Cary, North Carolina.
2. SPSS, Inc. (1975). *Statistical programs for the social sciences (SPSS)*. SPPS, Inc., Chicago, Illinois.
3. Dean, A. G., Dean, J. A., Burton, A. H., Dicker, R. C. (1991). Epi info: A general purpose microcomputer program for public health information systems. *American Journal of Preventive Medicine*, 7, 178–82.
4. Dean, A. G., Dean, J. A., Coulombier, D., et al. (1994). *Epi info, version 6: A word processing, database, and statistics program for epidemiology on microcomputers*. Centers for Disease Control and Prevention, Atlanta, Georgia.
5. Dean, J. A., Burton, A. H., Dean, A. G., Brendel, K. A. (1994). *Epi map: A mapping program for IBM-compatible microcomputers*. Centers for Disease Control and Prevention, Atlanta, Georgia.
6. *Epidemiology Monitor* (1994). Epidemiology Monitor, Roswell, Georgia.
7. Campos, N., Franco, E. (1989). MULTLR: A microcomputer program for multiple logistic regression by unconditional and conditional maximum likelihood methods. *American Journal of Epidemiology*, 129, 439–44.
8. Dallal, G. E. (1988). Logistic: A logistic regression program for the IBM PC. *The American Statistician*, 42, 272.
9. Dallal, G. E. (1989). cLOGISTIC: a conditional logistic program for the IBM PC. *The American Statistician*, 42, 125.
10. Sullivan, K., Foster, D.A. (1991). Epidemiologic software. *The Epidemiology Monitor*, 1–11.
11. U.S. Department of Health and Human Services. *Grateful med*. National Library of Medicine (distributors). Bethesda, Maryland
12. Hansel, S. (1994). Banks going interactive to fend off new rivals. *The New York Times*, Oct. 19, section D (column 3), 1.
13. Chambers, D. (1993). *Norton utilities*. Peter Norton Computing, Inc., Santa Monica, California.