Epi Info Described.pdf

Epi Info: A General-Purpose Microcomputer Program for Public Health Information Systems

Andrew G. Dean, MD, MPH
Jeffrey A. Dean, BS
Anthony H. Burton, BS
Richard C. Dicker, MD

Epi Info is a general-purpose set of computer programs for word processing, database management, statistics, and graphics developed over the past five years at the Centers for Disease Control and the World Health Organization. The programs allow rapid questionnaire construction, data entry, and analysis during epidemic investigation. Both data entry and analysis can be programmed to provide customization and automatic operation for more permanent systems, such as those for disease or injury surveillance. Epi Info is in the public domain and copies may be freely distributed. It requires an IBM-compatible microcomputer with at least 512 kilobytes of memory. Translations into French and Spanish are in progress; a translation kit is available to facilitate translation into other languages. [Am J Prev Med 1991;7:178–82]

In the past few decades, and particularly since the widespread use of microcomputers in the 1980s, computer processing has become an epidemiologic tool as indispensable as laboratory work or statistics. Many of the computer needs of epidemiologists can be satisfied by microcomputers, if appropriate software is available. The software must be easy to use for those who are not computer experts or who, during an investigation, are too busy to attend fully to the requirements of a complex computer program.

Commercial database and statistics programs can be used for epidemiologic work, but most database packages, such as dBASE, lack statistical procedures, and many statistical packages lack epidemiologic procedures, such as Mantel-Haenszel and Fisher exact tests, and offer a large number of procedures for continuous data rarely used in epidemiologic work. Commercial programs usually cost hundreds of dollars per copy and may occupy up to 15 megabytes of hard disk space. Until recently most such programs had rather rudimentary data-entry programs or required extra time or skill to set up data entry.

An early program for epidemiologic work was SOCRATES, written by Richard Curtis for minicomputers at the Centers for Disease Control (CDC) around 1978. The Epidemiologic Analysis System (EAS), by Anthony Burton of CDC, and EPISTAT[1,2] and True Epistat, by Tracy Gustafson, were written in BASIC for IBM/PC-compatible computers. Both perform epidemiologic questionnaire processing, but have limitations in questionnaire length, ease of questionnaire construction, and degree of "user friendliness." A commercial program called EpiLog (Epicenter software, Pasadena, California) performs a number of advanced procedures as well as producing basic tables and frequencies, but it requires considerable knowledge to operate and is too expensive for the average public health user.

Epi Info is a noncommercial software system for general public health use that has been in development at CDC since 1984. The current Version 5 is a joint project of CDC and the Global Programme on AIDS, World Health Organization (WHO). It provides facilities for word processing, questionnaire construction, data entry, database management, epidemiologic statistics, and graphics. Version 5 can be used without programming for one-time functions such as epidemic investigations. It also offers extensive programming facilities for construction of permanent databases, such as those used in disease or injury surveillance. Extensive error checking on data entry, facilities for handling several linked files at the same time ("relational" file handling), and customized report generation are features of the latest version.

METHODS

Program Design and Development

Epi Info is designed to run on IBM/PC-compatible microcomputers, which are widely available in health department settings and offer portable models for field use. Turbo Pascal (Borland

From the Epidemiology Program Office, Centers for Disease Control, Atlanta, Georgia (A. Dean and Dicker), and the Global Programme on AIDS, World Health Organization, Geneva, Switzerland (J. Dean and Burton).

Address reprint requests to Dr. A. Dean, Epidemiology Program Office, Mailstop C08, Centers for Disease Control, Atlanta, GA 30333.

International, Scotts Valley, California) was used as the programming language because of its speed, widespread use, and freedom from restriction or royalty on distribution of the compiled programs. Other packages, Turbo Professional (TurboPower Software, Scotts Valley, California), the Borland Editor Toolbox, and GraphDrive Plus (Fleming Software, P.O. Box 528, Oakton, Virginia 22124) provided additional facilities for programming.

Epi Info began as a data entry program in 1985. The ANALYSIS program was added the following year, and other new features have been added gradually in Versions 2 through 5 since then.

The completed programs and manual have been tested during course exercises held for training of Epidemic Intelligence Service (EIS) officers during the summers of 1985–1990. Epi Info training has become a regular part of the EIS and other epidemiologic courses, and many groups have used the programs to process data from surveys done as part of teaching exercises.

A formal test of Version 5 was conducted by 23 volunteer "beta testers" in October 1989; approximately 150 "bugs" and usability problems were addressed as a result. Regular telephone support is provided, and hundreds of telephone queries and comments received from users over several years have been used to design new features and correct problems in the programs.

CDC programmers have written Epi Info programs for 35 state health departments to enter and tabulate disease reports and to forward these electronically to the CDC over a telecommunications network.[3] The experience has also led to design of new features and detection of bugs.

Using Epi Info

Epi Info Version 5, described here, was released in April 1990. The programs are designed for two levels of user sophistication and corresponding flexibility for special functions—"interactive" use and "programming."

Interactive use allows construction of questionnaires, immediate data entry, and analysis of data files to produce line listings, frequencies, cross tabulations, graphs, and accompanying statistics. A program called CHECK allows ranges, skip patterns, and automatic coding to be added to the data entry process. Simple commands such as SELECT, RECODE, and IF within the ANALYSIS program allow the interactive user to "clean up" the data by recoding, selecting, and performing other logical operations in preparation for tabulation.

Saving commands in text files constitutes "programming." Programming offers a convenient way to reuse the commands. In Version 5, the data entry process may be programmed to handle several files automatically as a single unit ("relational data entry"), calculate or check values based on several fields, or display help messages to the user. The ANALYSIS program permits sophisticated programming to set up permanent systems accessing many files and producing customized reports for disease surveillance or research.

Epi Info is installed on floppy disks or a hard disk computer by running a program called INSTALL. The user then types "EPI" to run the main menu, providing access to the rest of the system. The menu provides information about each program and allows several methods of choosing programs, depending on the preference and experience of the user.

Interactive use: rapid data entry and analysis. During an epidemic investigation or simple survey, speed of data entry and the need for immediate results may be more important than computerized error checking or sophisticated analysis. Epi Info facilitates rapid data entry and analysis without precluding later use of more advanced features.

The user can computerize a questionnaire or investigation form in a few minutes by completing these steps:

- Running the main menu.
- Typing the questionnaire or form in EPED, the word processor.
- Entering data in the questionnaire using the ENTER program.
- Analyzing the data using the ANALYSIS program to produce lists, frequencies, cross tabulations, and means, with accompanying statistics and graphs.

A general-purpose word processor called EPED is used to construct a questionnaire. Headings, text, and questions are typed on the screen as desired. Locations for data entry are indicated by special characters as shown below:

_____	Text entry blank. Accepts text, such as names or addresses.
### ##.#	Numeric entries, accepting only numbers.
<A >	Text entry converted to upper case characters.
<Y>	Yes/No field. Accepts "Y," "N," or a blank.
<mm/dd/yy> <mm/dd>, <dd/mm/yy> <dd/mm>	Dates in U.S. or European format.
<phonenum> <long distance>	Telephone numbers.
<IDNUM>	Automatically incremented identification number.

With these few exceptions, the completed questionnaire looks like one designed for handwritten entries. It may be up to 500 lines long and contain hundreds of data entry fields, although this is seldom epidemiologically advisable.

A miniature questionnaire might be created as shown below:

Anywhere Health Department
Someplace, U.S.A.

Name _____ Age ### Sex <A> (M,F,U)
Telephone number <phonenum>
Did you have any of the following symptoms during the week of August 20–26?

Nausea	<Y>	(Y, N, or blank)
Vomiting	<Y>	
Diarrhea	<Y>	

At the wedding reception for Jennifer and Jeremy Newlywed on August 20, which of the following foods did you consume?

Baked ham	<A>	(Y, N, or U for unsure)
Mashed potatoes	<A>	
Gravy	<A>	

Once a questionnaire has been typed on the screen and saved as a file, another program called ENTER reads the questionnaire, and, in a few seconds, presents it on the screen ready for data entry. All definition of fields, creation of field names, etc., is done automatically. Field names are constructed from the text immediately preceding the field—NAME, AGE, SEX, and TELEPHONEN in the above example.

During data entry, the cursor jumps from blank to blank on the screen, and the questionnaire pages up and down automatically or in response to appropriate keyboard commands. When one questionnaire has been completed, the question "Save data to disk?" appears; an affirmative answer clears the questionnaire, increments the record number on the screen, and allows entry of another record. Entries saved are placed in a disk file, which is used for later editing or analysis.

Function keys are provided to perform searches on the contents of any field, to browse through the data, to find a record by number, or to print the questionnaire and the data. Any field can be indexed for fast searching, and several related questionnaires can be moved on and off the screen automatically as entry proceeds.

Missing data values are recorded by pressing the ⟨Enter⟩ key. In ANALYSIS, these values can be ignored or selected specifically.

The ANALYSIS program is used to produce lists, frequencies, tables, and graphics from Epi Info or dBASE data files. The commands are simple ("TABLES AGE SEX," "FREQ BAKEDHAM," etc.), and help screens giving syntax and examples are available by pressing a function key.

Frequencies of numeric variables are accompanied by the overall sum, standard deviation, and mean of the observations. Tables may be from two to 110 cells wide and two to hundreds of cells high. Two-by-two tables are accompanied by odds ratios, relative risks, confidence limits, chi-square with and without Yate's correction, the Mantel-Haenszel chi-square, and, if one or more cells has an expected value less than 5, the one-and-two-tailed Fisher exact tests. Larger tables produce a simple chi-square. Stratified tables may be produced, with automatic calculation of the Mantel-Haenszel weighted odds ratio, relative risk, chi-square, and confidence limits. A MEANS command provides a variety of statistics for continuous data sets, including one-way ANOVA and Kruskal-Wallis tests. A command called MATCH is provided for analyzing matched data sets.

In ANALYSIS, histograms; bar, line, and pie charts; and scattergrams with least squares regression lines are produced from data files using commands such as HISTOGRAM DATEONSET. They are displayed on computers with a variety of different graphics cards and may be sent to dot matrix or laser printers or to a plotter.

Other interactive functions produce more flexibility, but require more understanding of logic and of Epi Info's features on the part of the user. They allow data to be checked during entry and the cleaning of data sets prior to analysis and include these capabilities:

- Using a program called CHECK interactively, specifying range checks, legal codes, automatic coding, repeat fields, and skip patterns (conditional jumps) for any field during data entry.
- Selecting records, creating new variables, recoding data, manipulating dates, and carrying out conditional operations with IF statements during ANALYSIS, using simple commands such as SELECT, DEFINE, RECODE, and IF..THEN.
- Importing and exporting files from other systems such as SAS, SPSS, dBASE, and Lotus 1-2-3.
- Comparing two files entered by different operators to detect differences, using a program called VALIDATE.
- The ANALYSIS program allows reformatting of the data so that other files can be created with fewer fields, calculated fields, or fields in different order. Results of analyses can be sent to a printer or stored in disk files.

The CONVERT program creates files in formats suitable for a variety of other programs. This allows Epi Info to be used as a data entry program, with records subsequently converted to dBASE, Lotus 1-2-3, SAS, or SPSS files, for example.

A program called STATCALC accepts numeric entries in 2-by-2 up to 2-by-10 tables from the keyboard and produces the same statistics obtained from data files by ANALYSIS. Stratified tables can be entered. STATCALC does exact confidence limits[4] for single and stratified tables, a computationally intensive operation that would interrupt the flow of processing in ANALYSIS if done for every table. Sample size calculations for descriptive studies or surveys, and for case-control or cohort studies, are done in STATCALC, and the Mantel extension test for single or stratified dose-response studies is also offered.

The operations described so far can all be performed from the keyboard. They are frequently performed by epidemiologists during the investigation of an acute outbreak, perhaps using a portable computer in a hotel room. Programming consists of saving the commands for entry or analysis in text files that are then run, perhaps repeatedly, to perform the operations specified. Many users find it convenient to construct small programs for cleaning even small datasets.

Epi Info's advanced programming features are important to those setting up permanent database systems or large studies, or who want to customize Epi Info's operations to suit special needs. With a more advanced understanding of Epi Info programming, the user can perform these functions:

- Program the data entry process to include checking conditions in several related fields, calculating new values from one or more fields, changing colors, using pop-up boxes, or linking in custom routines written in other languages.
- Specify the format of reports from ANALYSIS to produce customized tables that contain text and values from any cell or marginal cell of a table that can be specified for ANALYSIS. The cell values can be placed anywhere on the report page.
- Enter data into more than one file during the same session, skipping automatically to the proper form within ENTER (relational data entry). Entering a particular value in a field, such as a disease code for "hepatitis," can cause a questionnaire from a related file to appear on the screen. When the user is through entering and saving the hepatitis questionnaire, the original disease report questionnaire reappears. Systems can be created for disease surveillance with a number of specialized reports for different diseases or for household interviews with linked records for each house, person, and followup visit, for example.
- Link several different types of files together in ANALYSIS so that questions can be answered that require data from more than one file (relational file analysis).

Extensive help files can be accessed from the word processor, ANALYSIS, or STATCALC by pressing the ⟨F1⟩ key. The

printed, 400-page *Epi Info Manual* contains tutorials on each major function. Sample programs and documentation include two epidemics, a matched case-control study, two disease-surveillance systems, and a number of customized report format files. There are interactive tutorials for learning to use the word processor, for designing a questionnaire, for designing a case-control or cohort study, and for elementary and intermediate functions of ANALYSIS. The entire *Epi Info Manual* is optionally available on IBM-compatible disks from the Epi Info distributor. A 76-page course manual and instructor's notes for teaching Epi Info in a computer laboratory are available from the senior author.

Applications of Epi Info

Epi Info can be used as a complete database and statistics system either for epidemic investigation or for disease or injury surveillance. It has been the primary computer package in hundreds of outbreak investigations conducted by CDC staff or by other agencies.

USD International, Inc., a small business in Stone Mountain, Georgia, has reproduced and distributed copies of Epi Info since 1987. In the first 12 months Version 5 was available, nearly 10,000 copies were distributed. Since copying of the disks and manual by others is encouraged, this is an underestimate of the number actually in circulation. Thirty-five state health departments in the United States have reportable disease surveillance systems built around Epi Info. A menu constructed from DOS batch files allows clerical personnel to enter data in a form customized to fit local needs. The Epi Info ANALYSIS program and programming language are used to implement other menu choices that reorganize data items and write new files in the uniform national reportable disease record format.[3] Other menu choices produce standard reports in a format tailored to local needs.

These 35 systems have been written, installed, and maintained by two programmers at the national level, with variable amounts of assistance from staff in state health departments. Installation, customization, staff training, and testing of a new state system usually requires a visit of less than a week by a CDC programmer. Over a period of five years, the National Electronic Telecommunications Surveillance System (NETSS) has grown from six states to 56 states and territories, with 35 of the new systems Epi Info installations. File sizes range from a few hundred to 18,000 records per year, and weekly data input and processing takes from a few minutes to several hours.

The Global Programme on AIDS of WHO in Geneva has used Epi Info to assist behavioral research studies in Chad, Sri Lanka, and Sudan and for analysis of databases of research projects and World AIDS Day activities. The government of Germany and sexually transmitted disease clinics in the Dominican Republic use Epi Info for human immunodeficiency virus (HIV) seroprevalence studies. AIDS surveillance records in Uganda, Portugal, and Ethiopia are analyzed with Epi Info.

Using the ability of the ENTER program to call a user-supplied Pascal or assembly-language program, a program that has been written for the Nutrition Division of the CDC performs anthropometric calculations. The user enters the age, sex, height, and weight of the subject, and the program almost instantly compares the values with the WHO international growth reference curves and returns the percentile, z-score, and percentage of median for height for age, weight for age, and weight for height. The nutritional calculation program is loaded into memory before using the ENTER program, and a series of standard procedures allows this program to be called from ENTER and the results returned to be placed in the data file for further processing by ANALYSIS.

Schools of public health at the University of Pittsburgh, the University of California at Berkeley, and Johns Hopkins and Emory Universities use Epi Info in courses to teach students microcomputer methods in epidemiology. The price of the manual and software is comparable to that of a textbook, and the student can leave the course with Epi Info ready for use in the next work environment.

French and Spanish versions of Epi Info, Version 5, are in preparation through collaboration with the Ecole Nationale de la Santé Publique, Rennes, France, the Pan American Health Organization, and the Consejeria de Salud, Junta de Andalusia, Seville, Spain. Arabic and Chinese versions and a Thai translation of the manual are planned.

A translator's kit is available for those who wish to convert the programs to another language, but it requires that the translator purchase copies of commercial programs, parts of which are included in the source code for Epi Info, at a cost of about $360–$600. Inquiries should be directed to the senior author.

DISCUSSION

Epi Info was originally designed to fill needs in a specialized area of epidemiology—investigation of epidemics in field settings. Commercial database and statistics packages at the time lacked the simplicity and rapid setup needed by new computer users in the field. The limited number of epidemiologists in the United States did not seem large enough to support development of commercial programs for this purpose.

However, versions of Epi Info since the first have appealed to wider circles of users. Version 5 contains many features for setting up permanent computer systems for disease surveillance, and potentially for databases of other types. CDC staff members have used Epi Info to develop systems for administrative planning, matching of personnel to positions, and processing of educational evaluation questionnaires.

A number of advantages have accrued in developing Epi Info within a government agency as a public domain program to be copied and distributed freely. One is the amount of user participation in testing new versions and suggesting improvements. There is an enormous difference between a computer program that works well in one computer with the author's hands on the keyboard and one that satisfies classes of beginning computer users or expert programmers under field and desktop use in many countries with hundreds of different combinations of monitors, graphics boards, disk drives, and computer brands. Our several conversations with users each day, and many letters received over the last five years, have been essential in approaching this goal.

The disadvantages of developing public domain programs have also become apparent. The nondisclosure agreements and cloak of secrecy that surround the development of most commercial programs are not available, and preliminary test versions tend to be circulated widely from one user to another, causing confusion and sometimes discouragement when bugs

are encountered. Many users of programs in the public domain receive copies from friends, electronic bulletin boards, or software exchange sessions at meetings, and may not have access to the printed manual. When bugs are corrected, funding may not be available to send updated copies to users; nor are all users known, since many obtained the program through informal channels. We have partially compensated for these limitations by providing telephone consultation so that users with difficulties can request advice or information about updated programs.

We were fortunate in choosing a development system—the Turbo Pascal Compiler—that has undergone remarkable development in the past five years. Enhancements in graphics, program size, debugging, and other features have made the growth of Epi Info possible. The modular design encouraged by the Pascal language has also contributed to the ability to add new features. Version 5, with 139,000 lines of source code, is more than twice as large as Version 3.

Computer programs can be compared on the basis of cost, program size, features, ease of use, speed of processing, and many other characteristics. Systems balance program size against the number of features offered in different ways. Programs written specifically for epidemiology have the advantage of not offering statistical routines for other disciplines little used in epidemiology. The balance between data handling and statistical processing is another crucial decision. Epi Info is unusual in offering both relational file handling and epidemiologic statistical processing, and we are striving to achieve the best balance between these two areas.

After exploring a number of avenues for distribution of Epi Info, a private company offered to distribute the program and manual at low cost. Since the program is legally a "work of government" and may not be copyrighted in the United States, this is permissible and solves a number of problems that occur when a program is distributed through government channels. Neither the CDC nor CDC staff members receive money or other benefits from the sale of copies of Epi Info. This mechanism has proved to be quite efficient, flexible, and beneficial to users.

In the development of disease surveillance systems for 35 states, Epi Info has been useful because both data entry and analytic operations can be extensively tailored to local needs. Each state maintains a different data entry form and data file structure, and yet, during weekly processing, Epi Info programs write essential data into a standard national record format for forwarding to CDC via telecommunication facilities. In our experience, the degree of customization directly affects the acceptability of software to state epidemiologists. Epi Info allows complete programming of a new surveillance system and training of state personnel during a one-week visit by one or two people. Future plans for Epi Info include releasing a geographic mapping program, assisting programmers who wish to write programs that use Epi Info files, and providing additional features to support use of Epi Info on local area networks (LANs). The current programs function well on LANs, but a feature called "record locking" is needed to permit more than one person to edit records in a single file at the same time.

Version 5 offers a host of features to those wishing to develop permanent epidemiologic databases built around freely distributable software. Epi Info will be available as cities, counties, provinces, and public health clinics acquire microcomputers and those at state and national levels wish to supply standard but customizable software for public use. We will continue to maintain and enhance Epi Info to serve both the individual epidemiologist and agencies wishing to develop permanent specialized programs. During the next few years, we intend to use Epi Info as the basis for a series of computer software modules that will collectively form an "Epidemiologist's Workstation" for use in a variety of public health tasks.

Obtaining Epi Info

Copies of Epi Info may be obtained from U.S.D. Inc., 2075A West Park Court, Stone Mountain, GA 30087, telephone (404)469–4098, fax (404)469–0681, for $35 (+$10 outside North America) including manual, disks, and shipping. The material is not copyrighted, and users are free to make and distribute additional copies as desired. A floppy disk version of the manual is available.

We are grateful for the assistance of Dr. Kevin Sullivan, Division of Nutrition, Center for Chronic Disease Prevention and Health Promotion, CDC, and Dr. Jonathan Gorstein, Nutrition Unit, WHO, for developing the Nutritional Anthropometry component of Epi Info. The hundreds of users who called or sent suggestions, and particularly the beta testers of Version 5, contributed greatly to the quality of Epi Info.

REFERENCES

1. Goldstein R. Shareware and free statistical software. PC Magazine 1989;8(5):184.

2. Lee RS, Cohen J. Good with stats: SYSTAT and EPISTAT. PC Magazine 1985;4(5):179–86.

3. Graitcer PL, Burton AH. The epidemiologic surveillance project: a computer-based system for disease surveillance. Am J Prev Med 1987;3:123–7.

4. Mehta CR, Patel NR, Gray R. Computing an exact confidence interval for the common odds ratio in several 2 × 2 contingency tables. J Am Statistical Assoc 1985;80:969–73.