The EPIGRAM Computer Program for Analyzing Mortality and Population Data Sets

DANIEL A. GOLDMAN, MD, MPH

Dr. Goldman, a Chronic Disease Epidemiologist at the Texas Department of Health, Austin, wrote EPIGRAM, the computer program described in this article.

Ms. Sandy Crow, Systems Analyst, Texas Department of Health, assisted with the conversion to EPIGRAM format of the Texas death certificate data files.

Tearsheet requests to Daniel A. Goldman, MD, MPH; Texas Department of Health, 1100 West 49th St., Austin, TX 78756-3199; (512) 458-7534; fax (512) 458-7650.

Synopsis

EPIGRAM is a computer program designed to improve access to State-level underlying cause mortality data. The program produces results for population, deaths, death rate, age-adjusted death rate, years of potential life lost (YPLL), YPLL rate, and confidence intervals. Results can be compared variously among age groups, counties, causes of death, races, regions, and years.

The program's menu-driven interface facilitates the selection or modification of analysis parameters. Current selections are retained so the user can modify one parameter at a time. Based on the parameters that the user selects, the program produces a series of tables, one for each instance of a particular parameter. Each output table has columns for male, female, and both sexes combined, and an indefinite number of user-defined rows for age groups, causes of death, counties, races, regions, or years.

EPIGRAM has major advantages over other methods for analyzing mortality and population data. The program uses relatively small amounts of memory and disk space, executes rapidly, is flexible, can be used by inexperienced computer users, provides online help screens and tutorials, and runs under DOS or UNIX without modification. The program currently is used to analyze mortality and population data for Texas. Although it is not currently available for distribution, support is being sought for its evaluation and possible implementation in State health departments to analyze data for other States, or other data sets, such as hospital discharge data or cancer incidence data.

PUBLIC HEALTH WORKERS with responsibility for monitoring the progress of programs directed to mortality related goals often lack the capability for readily accessing available data. That difficulty was recognized in "Healthy People 2000" (*1*) in the section on surveillance and data systems.

> . . . Even when needed data exist, not all potential users of these data have ready and timely access to the information, and not all users are equally able to use the data to full advantage. Some national data files are too large and complex for easy use, while others are simply not available for public use. . .
>
> Many of the existing data systems suffer from the lack of comparability in how data are collected on common topics as well as in how data are presented to the public. . . .

Objective 22.6 addresses those problems in setting the following goal.

> Expand in all States systems for the transfer of health information related to the national health objectives among Federal, State, and local agencies.

Printed reports can be a convenient method for accessing health data. Except for simple tasks, however, accessing mortality and population statistics from printed reports is inefficient as they usually contain only a small subset of the results that can be generated from a large data set. Comparing data from different reports may be difficult because of incompatible definitions of age groups, race, or other variables. Population-based rates in reports become obsolete when population denominators are updated to reflect new census

figures. Copying data from reports and generating additional results from the report data is a time consuming process prone to error.

Since mortality and population data are widely used by health department employees and other public health workers, the need exists for computer software to efficiently analyze and disseminate these data sets. Essential criteria include flexibility for the user to specify the kinds of analysis to be performed, a variety of output formats, rapid execution, self-explanatory menu-driven interfaces, convenient modem access, and acceptability to users.

Commercially available statistical programs, such as SAS (*A*) and SPSS (*B*), commonly are used to analyze mortality data and to generate standard reports. However, such programs usually require the user to understand the structure of the data set as well as the program's specialized command syntax. Usually such programs are not useful in calculating age-adjusted rates. They are not specifically designed for public health data analysis.

Methods

EPIGRAM was written in the C programming language and designed to run under DOS or UNIX operating systems without modification. The program has about 11,000 lines of source code, organized into about 150 separate modules to facilitate adding or removing program functions. The development of program features and the user interface was guided by evaluation and feedback from Texas Department of Health users.

A set of utility programs was written to convert population and mortality data files to a format used by EPIGRAM, based on the specifications of the format of the files to be converted. The conversion of the mortality and population files for Texas required several hours. Files for Texas for underlying cause mortality for 1980–91 were obtained from the Bureau of Vital Statistics. The final format required 7 megabytes (MB) of disk storage space for 1.4 million deaths (5 bytes per death). Each data file was sorted and indexed to reduce access time. Files for the population of Texas from 1980 projected to 2000 were obtained from the Texas Department of Commerce.

Results

User interface. EPIGRAM has a menu-driven user interface for selecting analysis parameters. A main menu displays the current parameter selections (figure 1). A set of submenus and prompts is used to change the parameters. A sample submenu is shown (figure 1). The accompanying box shows the options available with each EPIGRAM menu.

The interface was designed for simplicity of use and to prevent user error and frustration. Menus and prompts have a consistent format. Optional parameters must be specified only when required. The program does not allow the user to select incompatible parameters, such as population statistics ranked by cause of death. EPIGRAM menus retain all current selections, allowing exploratory data analysis by changing one parameter at a time, without having to reselect all parameters. Online help is available from each menu. An online tutorial demonstrates each section of the program and gives examples of data analysis.

Output tables. EPIGRAM produces a variety of output tables, allowing the user to compare results variously among different age groups, causes of death, counties, races, regions, or years. Two sample output tables are shown (figure 2). The initial section of each output table describes how the table was specified. Each table includes a bar chart of the data. Tables can be saved to an ASCII file, may be edited and printed using almost any text editor, and imported into graphing, mapping, or spreadsheet programs. For large output tables, EPIGRAM allows the user to scroll through the output.

Analysis parameters. Each output table is specified by a set of analysis parameters: causes of death, counties and regions, races, years, a single statistic, rows, multiple tables, and other settings. The analysis parameters select a subset of the data to analyze and specify the format of the output table.

Counties and regions. Parameters may be specified or modified in any order. For discussion purposes, assume that a user starts by specifying which counties or regions (groups of counties) to analyze. From the main menu, the user chooses option D (figure 1). EPIGRAM shows the county-region menu (figure 1).

The county-region menu displays the current county selection, lists the available operations, and prompts the user for a response. Choosing menu options and responding to prompts, the user creates and modifies a list of selected counties. The user is not required to know county code numbers or exact spellings of county names. After selecting which counties and regions to analyze, the user

Summary of EPIGRAM Menus

Main menu (displays all parameters):

Select statistic (separate menu)
Select table rows (separate menu)
Select multiple tables (separate menu)
Select county-region menu
Select causes of death menu
Select race groups (separate menu)
Select range of years (separate menu)
Select age operations menu
Select other settings menu
Select online tutorial (separate menu)
Produce output table(s)
Exit EPIGRAM

County-region menu:

Add individual counties (separate menu)
Add all counties
Delete individual counties
Delete all counties
Add counties from a region (separate menu)
Delete counties in a region (separate menu)
List counties in a region (separate menu)
Change region type (separate menu)

Causes of death menu:

Add one or more ICD sets (separate menus)
Add all causes of death
Print some or all ICD codes
Delete an ICD set
Delete all ICD sets
Join several ICD sets into new set (separate menu)
Split an ICD set into components

Age operations menu:

Select 5 year age groups
Select 10 year age groups
Select 10 year groups, from age 5
Select 20 year age groups
Extend lower limit of age group
Extend upper limit of age group
Split an age group into components

Other settings menu:

Set age-adjustment standard (separate menu)
Change age for YPLL cutoff (separate menu)
Set confidence level (separate menu)
Enable display of confidence levels
Show confidence factors
Show standard U.S. population

NOTE: ICD, Reference *2*. YPLL = Years of potential life lost.

chooses option Z and is returned to the main menu.

Causes of death. Next, the user may specify which causes of death to analyze by choosing option E from the main menu. EPIGRAM allows access to more than 6,000 categories from the International Classification of Diseases (ICD–9) (*2*) (5,114 4-digit ICD–9 codes, 911 3-digit codes, and 127 higher level groupings).

The interface helps the user create a list of selected causes of death. A set of submenus starts with broader disease categories and progresses to more narrowly defined groups. For example, to select *colon cancer,* the user first chooses *neoplasms* from the menu with the broadest categories. The program displays a menu listing types of neoplasms, from which the user chooses *cancers of the gastrointestinal tract.* In response, the program displays a menu that lists types of gastrointestinal cancers, from which the user chooses *colon cancer.*

The user is not required to know ICD code numbers or exact disease names, because the menus lead the user to the desired disease or condition. [To customize disease lists,] the user can combine any two selected ICD sets into a new ICD set, as for example, *myocardial infarction* (codes 410–414) and *cerebrovascular disease* (codes 430–438). Or, the user may select a standard list of causes of death, such as leading causes of death.

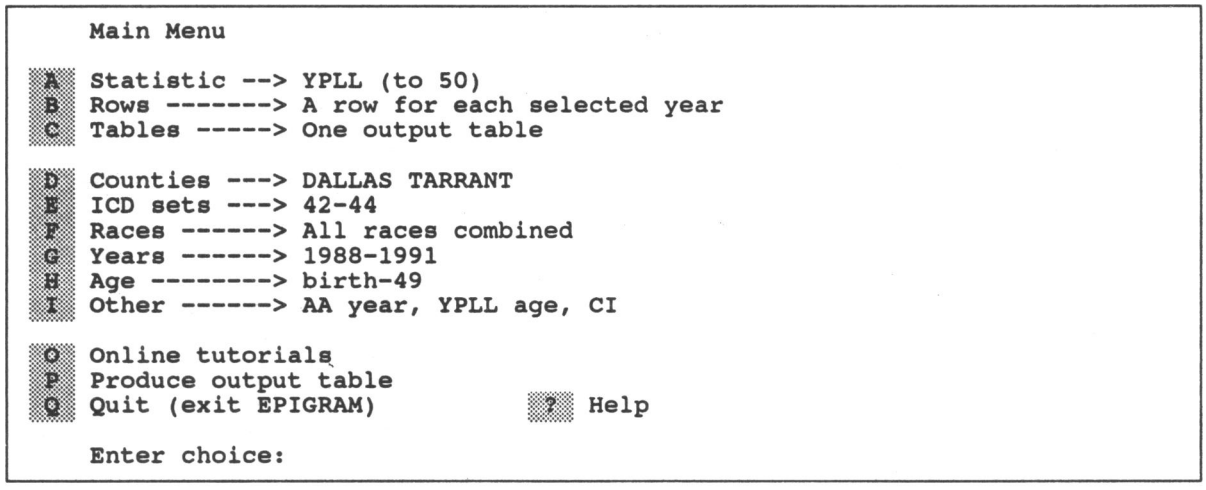
Race. The user may select from a submenu any combination of race-ethnicity. The groups currently used in Texas are white, black, and Hispanic.

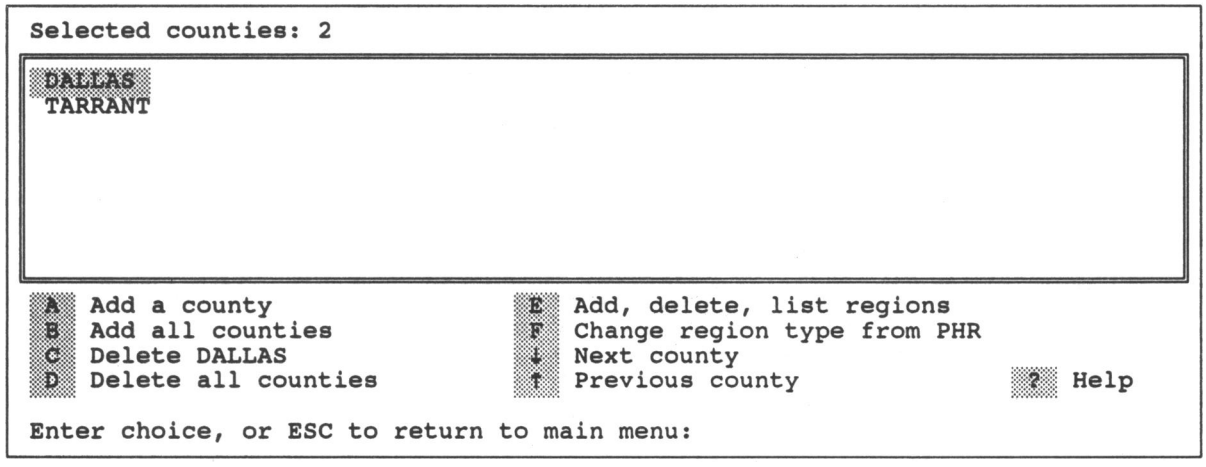
Statistic. Next, the user may choose option A from the main menu to specify which statistic will be calculated in the output table. The user selects from a submenu one of the following: population, deaths, crude mortality rate, age-adjusted death rate, years of potential life lost (YPLL), and YPLL rate.

Years. The user is prompted to select a single year or range of years for analysis. Results for a range of years represent the average during the period.

Age groups. The user selects age groups for analysis. A set of 33 age groups (0, 1, . . . , 20, 21, 22–24, 25–29, . . . , 70–74, and 75 or more) can be arranged into any set of contiguous groupings (for example, 0–9, 10–17, 18–44, 45–64, and 65 or more). A single age group (for example, 40–49) can be selected for data analysis.

Figure 1. Main menu and county-region menu for EPIGRAM mortality and population data analysis program





Rows and columns. Rows for the output table may be specified. For example, the upper screen of figure 2 shows a row for each selected year. Rows also can be specified for each selected age group, cause of death, county, race, or region. For counties and causes of death, rows can be sorted by male, female, or total data. As shown in figure 2, EPIGRAM output tables have total, male, and female columns. Figure 2 also shows how output tables display the number of deaths alongside a rate or YPLL statistic.

Confidence intervals. For each statistic except population, the program calculates and displays confidence limits based on the Poisson probability distribution (*3*). The user selects from a list of commonly used confidence levels.

Multiple tables. Ordinarily, only one output table is produced. However, option C from the main menu allows the user to select multiple tables. There can be one table for each selected age group, cause of death, county, race, region, or year.

Other parameters. For age-adjusted death rates, a menu option permits the user to select a standard U.S. census population. If YPLL or YPPL rate is being analyzed, a menu option is chosen for the age limit for calculating years of potential life lost.

Producing output. When all parameters are chosen, the user selects an option from the main menu to produce an output table.

Speed. On an IBM-type, 386/25 computer, with 4 MB of memory and a mathematics coprocessor, each output table in figure 2 was calculated and displayed in 1 second. On a 286/12, with 640 KB of memory and no coprocessor, each table took 2 seconds. The times were the same whether the data

Figure 2. Examples of output tables from EPIGRAM mortality and population data analysis program

EPIGRAM 1988-1991 TX YPLL to age 50
Deaths by underlying cause, by county of residence
Death data from TDH Bureau of Vital Statistics
Counties --> DALLAS TARRANT
ICD 042-044: Human Immunodeficiency Virus (HIV) Infection
Age group: birth-49
Race: All races combined

Year	M+F	Deaths	Male	Deaths	Female	Deaths
1988	4422.0	282	4261.0	272	161.0	10
1989	5164.0	352	5003.5	345	160.5	7
1990	6608.5	463	6476.0	452	132.5	11
1991	7511.0	520	7055.5	497	455.5	23
Total	23705.5	1617	22796.0	1566	909.5	51

Bar graphs of the data:

Year	M+F	Male	Female
1988	===========	===========	
1989	=============	=============	
1990	=================	=================	
1991	====================	==================	=

EPIGRAM 1990 TX death rate (per 100,000)
Deaths by underlying cause, by county of residence
Death data from TDH Bureau of Vital Statistics
Population data from Texas Department of Commerce
Counties --> All (entire state)
1) 153: Colon Cancer
2) 154: Cancer Of Rectum And Anus
Race: All races combined

Age group	M+F	Deaths	Male	Deaths	Female	Deaths
birth-39	0.5	51	0.5	26	0.5	25
40-49	5.4	112	5.7	58	5.2	54
50-64	28.8	578	35.3	339	22.8	239
65-99+	121.4	2074	145.3	1005	105.2	1069
Total	16.6	2815	17.1	1428	16.1	1387

Bar graphs of the data:

Age group	M+F	Male	Female
birth-39			
40-49			
50-64	===	====	===
65-99+	================	====================	==============
Total	==	==	==

were being accessed from a 286/12 file server with Novell Netware, version 2.15, or from a local disk.

Storage space requirements. The total storage requirement for 12 years of Texas mortality data and 21 years of Texas population data was about 14 MB. Each death required 5 bytes, and the population for 1 county for 1 year required 1,088 bytes. The executable file was 200 KB. Accessory data files were about 700 KB.

Usage. Currently, EPIGRAM is being used extensively within the central offices of the Texas Department of Health. During the period January 27, 1992, to January 27, 1993, the program was used 1,143 times to generate 3,880 tables.

Discussion

EPIGRAM is a mortality and population data analysis program written for and used by the Texas Department of Health. EPIGRAM has made the department's analysis and dissemination of mortality and population data faster and easier than before, provides statistics that were not readily available previously, and permits rapid and convenient routine analyses, giving the user flexibility in specifying the kind of analysis wanted.

Access to mortality data, beyond that available in written reports, previously involved several steps that required days to weeks. Those steps included loading mortality data from a tape onto a mainframe computer, creating an SPSS routine, running the SPSS routine overnight, transferring the results from the mainframe computer to a microcomputer, creating a spreadsheet to manipulate the results, reading the mortality results into the spreadsheet, gaining access to a population data file, and reading the population data into the spreadsheet.

The same results now can be produced in seconds, without error. Users get immediate feedback and can carry out exploratory data analysis. Relatively sophisticated analyses that were rarely done, because of analytical difficulties, including confidence intervals and multiple age-adjusted rates, may now be carried out by any user with relative ease.

Alternative methods. A survey of directors of State vital statistics bureaus was conducted in January 1992 to determine whether other States were developing software similar to EPIGRAM. Directors were asked to route the survey form to the appropriate person with knowledge of how mortality data were analyzed and disseminated in that State. The questionnaire asked if the State had software for general analysis of mortality data, defined as being available to a wide range of users, producing output in a variety of user-defined formats, and not requiring special computer knowledge (for example, SPSS or SAS commands).

Responses were received from 50 States and the District of Columbia. California had two menu-driven programs (the Public Health Information System and the Microcomputer Injury Surveillance System), which produce tables with numbers of deaths according to categories of age, race, and sex. Massachusetts had a program (the Health and Environmental Assessment Database System—Massachusetts) that is intended for the general user and includes information on cancer mortality and population. Delaware had a menu-driven program that displays the number of deaths meeting the set of criteria of county, age group, race, sex, and cause of death. Illinois was developing a menu-driven system for disseminating aggregated mortality data, including rates. Some States reported having computer software to display preprepared fact sheets that include data on leading causes of death. Although survey comments indicated interest in accessing mortality and population data analysis software, no respondents reported having a program similar to EPIGRAM.

At the national level, the Centers for Disease Control and Prevention's Wide-ranging Online Data for Epidemiological Research (WONDER) (*4*) provides access to a range of data sets. The Texas Cancer Data Center at M.D. Anderson Hospital, Houston, TX, accesses cancer-related data bases (*5*). Although both systems analyze mortality and population data, neither functions as fully and efficiently as EPIGRAM. For example, WONDER does not include a Hispanic race-ethnicity group (essential for analyzing data in Texas), uses different population figures than Texas State agencies, and has a more difficult-to-use interface (for example, an incorrect keystroke can cause the program to malfunction). The Texas Cancer Data Center software does not provide flexibility in specifying what kind of analysis to do (for example, only a single county or region can be analyzed at a time), and it does not remember the parameters the user previously selected.

EPIGRAM strengths.

- *Effective analysis of data.* EPIGRAM was written and designed for a specific task. It is specially

structured to analyze population and underlying cause mortality data.

• *Efficient computer resource use.* Space requirements are comparatively small. The program produces simple tables rapidly, even on an IBM XT computer.

• *Flexibility.* The user can select which census year to use as an age adjustment standard. The user can create any desired list of causes of death to analyze and select any age groupings. A variety of types of output tables can be produced.

• *Accommodates inexperienced users.* The program was written in an environment of constant feedback from users at the Texas Department of Health. All operations are driven by menus and prompts written in plain language. The user need not enter special codes for counties or causes of death. If needed, a separate help screen is available at each menu. There are online tutorials for each aspect of the program. Each output table exactly describes the analysis parameters that were used to make the table.

• *Portability.* EPIGRAM currently is being run under DOS, but has been compiled to run under UNIX as well. EPIGRAM can run on computers ranging from personal computers to mainframes and can be accessed by modem, using almost any communications software.

EPIGRAM limitations.

• EPIGRAM cannot be used to produce maps or graphs. It produces only data tables and simple bar charts. While output tables are in ASCII format, and can be imported to mapping, spreadsheet, or graphics applications, users are required to know those programs.

• EPIGRAM is not currently available for distribution, and support for its use elsewhere is not available. The time requirements and costs of maintenance, upgrading, producing and revising documentation, distribution, and support are beyond the capabilities of the sponsoring organization. Efforts are being made to obtain funding for further development of the program and its distribution for use by other State health agencies. Use by agencies of other States would require converting population and mortality data files to the format used by EPIGRAM in a customized process that reflects the format specifications of the files being converted.

Major expansion of the use of EPIGRAM is possible because of its characteristic of extendability. The program is suitable for modification to work with mortality and population data for other States. EPIGRAM also has a framework for developing programs for analysis of related data sets, such as multiple cause mortality data, cancer registry data, and hospital discharge data. Racial categories (for example, using other categories) and county and regional specifications (for example, another State or nationally) could be modified by recompiling the program.

EPIGRAM advances the methodology for providing efficient access to large and complex health data sets. It has the potential for facilitating the analysis and dissemination of mortality and population data and their use in monitoring progress toward program goals by agencies at Federal, State, and local agency levels.

References

1. Public Health Service: Healthy people 2000: national health promotion and disease prevention objectives. DHHS Publication No. (PHS) 91-50212. Office of the Assistant Secretary for Health, Office of Disease Prevention and Health Promotion. U.S. Government Printing Office, Washington, DC, 1990.
2. International classification of diseases: manual of the international statistical classification of diseases, injuries, and causes of death. 9th revision. Clinical modification. DHHS Publication No. (PHS) 91-1260. Centers for Disease Control, National Center for Health Statistics, and Health Care Financing Administration, Hyattsville, MD, 1992.
3. Diem, K., and Lentner, C., editors: Scientific tables (equations 802a, 802b, and 804, p. 189). Geigy Pharmaceuticals, Ciba-Geigy Corp., Ardsley, NY, 1970.
4. Friede, A., Reid, J. A., and Ory, H. W.: CDC WONDER: a comprehensive on-line public health information system of the Centers for Disease Control and Prevention. Am J Public Health 83: 1289-1294 (1993).
5. Texas Cancer Data Center: Texas cancer data center fact sheet. M.D. Anderson Hospital, Houston, TX, 1992.

Equipment

A. SAS, SAS Institute, Inc., Cary, NC 27709-2194; tel. (919) 677-8000.

B. SPSS, Inc., 444 N. Michigan Ave., Chicago, IL 60611; tel. (312) 329-2400.