Accessing NCI's SEER Cancer Data Base with SeerQuery and CD-ROM

SUDHIR SRIVASTAVA, PhD
CHARLES SMART, MD
THOMAS A. MARCINIAK, MD
LARRY DERRICK

Dr. Srivastava and Dr. Marciniak are with the National Cancer Institute's Division of Cancer Prevention and Control. Dr. Srivastava is Program Director of the Early Detection Branch. Dr. Marciniak is Chief of the Computer Systems Branch. Dr. Smart was Chief of the Early Detection Branch and is now retired. Mr. Derrick is Manager, Rocky Mountain Cancer Data System, University of Utah.

Tearsheet requests to Sudhir Srivastava, PhD; NCI DCPC EDB, Bethesda, MD 20892; tel. (301) 496-8544; fax (301) 496-8667.

Synopsis

The National Cancer Institute operates the Surveillance, Epidemiology, and End Results (SEER) cancer data base. SEER data are obtained from participating population-based registries that monitor cancer incidence and patient survival in a representative 10 percent sample of the general population. The data cover all cancers (except superficial skin cancers) in the defined regional populations. SeerQuery is a personal computer program for accessing that data on IBM-compatible personal computer compact diskettes in read-only memory (CD-ROM) form. SeerQuery facilitates rapid access to cancer data at minimal cost and effort to the user.

SeerQuery is menu-driven, enabling physicians and other health care professionals to query the data base directly. They can use the data to determine cancer frequency, perform cross-tabulation, determine incidence, and calculate survival using such variables as primary cancer site, histologic type, stage, sex, age, and race. The comprehensive data base lacks many selection biases that are inherent in data reported from other sources. SeerQuery has applications in professional education and in cancer control program planning and resource allocation.

A MAJOR STRATEGY of cancer prevention and control efforts is to disseminate information on cancer statistics to physicians, other health professionals, and health care administrators who are responsible for treatment and for resource allocation.

Major sources of cancer data are the American Cancer Society, which annually publishes national incidence and mortality estimates, such as those for 1992 (*1*), and the National Center for Health Statistics, which collects and publishes data on mortality from State vital statistics. The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) monitors the incidence of cancer in certain areas (*2*).

SEER collects information on all cases of cancer (excluding squamous and basal cell skin cancers) that have been diagnosed since 1973 in a representative 10 percent sample of the general population. SEER monitors five States, Connecticut, Hawaii, Iowa, New Mexico, and Utah; four metropolitan areas, San Francisco-Oakland, Detroit, Seattle-Puget Sound, and Atlanta (*2*); and 10 predominantly black, rural counties in Georgia, added in 1978, and American Indian residents of Arizona, added in 1980. Four counties in New Jersey added in 1983 participated until the 1989 reporting year. Puerto Rico participated through 1989, and New Orleans participated through 1977. The SEER Program publishes the annual NCI Cancer Statistics Review (*3*).

SeerQuery is an IBM-type personal computer (PC) program offering users access to the SEER data base through compact disk read-only memory (CD-ROM) technology. SeerQuery was developed in response to three concurrent factors: growing access to personal computers among health professionals, the emergence of efficient technology to store a complete data base on optically readable devices, and the capability for providing the data base to those who lack direct access and need to extract and analyze that data.

SeerQuery was developed in collaboration with Rocky Mountain Cancer Data Systems, University of Utah, Salt Lake City, UT.

SeerQuery System

Hardware. Users need a PC with a CD-ROM drive, MS-DOS 3.0 or higher, a CD controller board, a minimum of 512 kilobytes of random access mem-

Figure 1. SeerQuery program on-screen dialogues for creating a subset, *Cervical*, showing all forms of cervical cancer among women between the ages of 35 and 40 years. This screen text has been edited for clarity

```
Screen Sequence 1:

Enter number of desired item # 1 (enter 0 if done): 24

Primary Site    ---- Field Length: 3

Enter minimum value for item  #1 -- 800
Enter maximum value for item # 1 -- 809

Enter number of desired item #2 (enter 0 if done): 19

Age             --- Field Length:  3

Enter minimum value for item #2 -- 035
Enter maximum value for item # 2 -- 040

Enter minimum value for item #2 (enter 0 if done): 16

Sex             --- Field Length: 1

Enter minimum value for item #3  --- 2
Enter Maximum value for item #3  --- 2

Enter number of desired item #4 (enter 0 if done): 0

Screen Sequence 2

To search a subset file enter its name.
Press <RETURN> to search the master file. :
Enter the name for this subset: Cervical

Screen Sequence 3

file to be searched is :  Master file
subset file is: Cervical
number of items selected :3
```

NOTE: Forms of cervical cancer are ICD codes 180.0–180.9. Reference (7). The first digit of the ICD code is not used and the decimal is dropped.
SOURCE: National Cancer Institute.

Figure 2. SeerQuery screen of dialogues for creating a summary report for the subset file *Cervical*. This screen text has been edited for clarity

```
Give name of Subset: Cervical

Give Title or Heading:
Summary Report for Cervical Cancer in Women Age 30-40

Is this report for individual SEER Registry(s)?  (answer y
or n) n

Follow-up Cutoff Date is 03/01/87
The years for the tables are 1973-1986
The year for the sex tables will be 1986
Table A and B will be printed for each site


Would you like to use the default values (y/n)? y

     Select type of break outs for the tables:

     (1)  All cases in one group of Tables by stages and all
stages
     (2)  New Tables for each site change (SEER Site Groups)
               Enter Breakout Type: 1
.
Now Sorting the File (This may take a while)
Generating the report
.
*Press 'e' to examine report on the screen
Press 'p' to print report on the printer
Press 'd' to delete the print file
Press 'r' to exit program


. : represents several intermediate dialogues
```

SOURCE: National Cancer Institute.

ory (RAM), and a minimum of 10 megabytes of free fixed disk space.

Software. Microsoft's CD-ROM extension program MSCDEX.EXE, v. 2.01 enables DOS to access the CD-ROM drive, which uses the High Sierra or ISO 9660 file format. The program enables the CD-ROM drive, with no file allocation table and with 2,048-byte sector sizes, to interface with the PC drive, which has 512 kilobyte sector sizes.

SeerQuery is written in ANSI C programming language and can be exported to other types of computers, such as Unix- or VMS-based minicomputers or mainframes, with few modifications. The program source code is available from the authors on request.

Statistical Analyses

SeerQuery uses standard techniques for statistical analyses. The age-adjusted incidence rate is a weighted average of the age-specific rate expressed as the number of deaths from cancer per 100,000 persons. The 1970 census is the standard population for calculation (*4*).

The survival programs use the actuarial method to calculate survival (*4, 5*). The relative survival calculation uses the actuarial observed survival rate adjusted using normal life expectancy tables (*4*). Explanations for statistical functions are provided at the beginning of the computer-generated report.

Report Functions

The main menu has 14 options selected by typing the letter preceding the narrative description of each option. Master file refers to a file of cancer incidence up to August 1988. Subset file refers to a file created using option 1 on the main menu. Subset files contain data on cancer cases based on variables specified by the user. The 14 main menu options follow.

(A) Create a subset enables the user to form a subset file of cases based on such variables as age, sex, and site. The variables appear on the screen as options when a subset is to be created. After the variables are selected and entered, the system searches the requested file, which may be either the master or a subset file.

(B) Sort a subset lists the variables or options available to sort a subset file. The user chooses major (primary) variables, or a sort file, and the

minor or secondary sort file. Two sort keys must be selected.

(C) Cross-tabulation and frequency report allows the user to produce a cross-tabulation, two-way or three-way tables, for any subset. This function takes the user through several self-explanatory steps or dialogues, providing 39 options in choosing the row or column, such as the site by stage or the county by site.

(D) Listing report enables the user to see subset files on screen. The user can select up to six variables or use standard variables in two different formats. Standard variables include, for example, name, date of birth, age, cancer site, stage, history of disease, date of diagnosis, survival time, date last seen, and followup status.

(F) Summary report generates a summary report on the selected cancer or on all cancers. The user is given options for including or excluding cases.

(G) Survival report generates the observed survival rate using the actuarial method. Relative rates use a standard life table. Results represent the proportion of cancer patients surviving for a specified period after diagnosis.

(H) Incidence report provides an opportunity to generate an incidence report for a given cancer site. The types of incidence rates that can be generated are age-adjusted, age-specific, or expected rate for a given population. For the rate of a given population, the user provides a standard population. The program prompts the user to provide a range of years of interest. Cases outside this range are ignored. A range of up to 11 years is accepted. Two rate files (one can be used twice) are needed to calculate the expected value. The rate file contains age specific rates (per 100,000 persons) for each cancer site, patient sex, and age group. Usually the total SEER age-specific rates are used, but any age-specific rate file can be used. The report is generated by site using SEER recoded site groups. The report has three sections, male, female, and total, providing an average for given years.

(I) Generate population file provides a population file if needed to calculate incidence rates. Two population years are requested, and one may be used twice. The population files are stratified in 18 groups. The strata are 5-year groups starting with 0–4 years of age and ending with 85 years and older. The program will perform a linear interpolation or an extrapolation to approximate the populations for the requested range of years.

SeerQuery Program Subset File Variables Used in Cross-Tabulation and as Primary or Secondary Keys for Sorting

1. Identification number
2. Sequence number
3. Registry number
4. Reporting source
5. Residence
6. Census type
7. Census tract
8. Birthplace
9. Birthday
10. Age
11. Race
12. Ethnic type
13. Sex
14. Marital status
15. Date of diagnosis
16. Primary site
17. Laterality
18. Method of diagnosis confirmation
19. Histology
20. Behavior code
21. Grade code
22. Tumor size
23. Nodes examined
24. Nodes positive
25. Stage
26. American Joint Commission on Cancer stage
27. Anatomical extent of disease
28. Diagnostic procedure
29. First course prescription date
30. Treatment
31. Surgery
32. Radiation
33. Radiation to central nervous system
34. Chemotherapy
35. Hormonal
36. Biological Response Modifier
37. Other prescriptions
38. Radiation sequence
39. Date of Last contact
40. Patient status
41. Tumor status
42. Cause of death
43. Survival time in months
44. General, as specified

(J) Browse a subset permits viewing a subset by page scroll.

(K) Expand a subset to an ASCII file converts a subset to an ASCII file for transfer to another system.

(M) Delete subset files deletes subset files not in use.

(N) Combine subset files combines two or more subset files with the name of first, unless another name is specified. After the files are combined, they may be used in any of the report programs.

(P) Return to main menu brings the user to the first level of choices.

Reports can be generated for either a SEER

Figure 3. SeerQuery screen of part of the summary report generated for the cohort of all forms of cervical cancer among women between the ages of 35 and 40 years

SUMMARY REPORT Summary Report for Cervical Cancer in Women Age 35-40 Wed May 29 15:49:43 1991

SUMMARY REPORT
BY CALENDAR YEAR OF DIAGNOSIS

		1973<	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	Unk	Total
TOTAL		0	613	715	842	779	758	746	812	792	735	786	816	924	871	934	0	11123
% MICRO. CONFIRM.	%	0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	100
STAGE AT DIAGNOSIS(%)																		
IN SITU	%	0	79	82	85	82	83	82	82	84	83	83	82	82	82	79	0	82
LOCALIZED	%	0	15	12	11	13	10	10	10	8	8	10	11	11	12	14	0	11
REGIONAL	%	0	4	3	3	2	3	4	5	4	5	3	4	5	5	5	0	4
DISTANT	%	0	0	0	0	1	1	0	1	1	1	1	2	1	1	1	0	1
UNKNOWN	%	0	2	3	1	2	3	4	3	3	3	3	1	1	0	1	0	2
FOLLOW-UP STATUS																		
DEAD		0	68	53	52	44	44	46	60	44	39	41	37	43	31	26	0	628
ALIVE		0	545	662	790	735	714	700	752	748	696	745	779	881	840	908	0	10495
DEAD, CAUSE NOT CA		0	25	22	13	13	9	9	12	6	11	7	4	4	2	4	0	141
DEAD, CAUSE UNK		0	7	3	5	4	9	5	5	4	3	6	7	4	6	2	0	70
DEAD, CAUSE CANCER		0	36	28	34	27	26	32	43	34	25	28	26	35	23	20	0	417

NOTE: A complete report includes tables for the stage at diagnosis, laterality, first course treatment, followup status, reporting source, and other factors by calendar year of diagnosis. Report tables are preceded by definitions and explanations of the terms used in the tables and the options selected for printing.
SOURCE: National Cancer Institute.

Figure 4. SeerQuery table in part of the summary report generated for the cohort of all forms of cervical cancer among women between the ages of 35 and 40 years

SURVIVAL SUMMARY

SURVIVAL RATES (%): 1,2,3,4,5,20 YEARS AFTER DIAGNOSIS (ANALYTIC CASES ONLY AND CASES INCLUDED BY OPTIONS (PAGE 1))
OBS. OBSERVED RATE IN % REL. = RELATIVE RATE IN % ERR. = STANDARD ERROR IN %

= HISTOLOGIES ALL =

ALL STAGES 1916 CASES				LOCALIZED 1188 CASES				REGIONAL 426 CASES				DISTANT 81 CASES			
	OBS.	REL.	(ERR.)		OBS.	REL.	(ERR.)		OBS.	REL.	(ERR.)		OBS.	REL.	(ERR.)
1 YR	93.0%	93.2%	(0.8)	1 YR	98.2%	98.4%	(0.7)	1 YR	85.7%	85.9%	(1.7)	1 YR	47.5%	47.6%	(5.6)
2 YR	84.5%	84.8%	(0.9)	2 YR	94.3%	94.6%	(0.8)	2 YR	67.9%	68.2%	(2.3)	2 YR	32.4%	32.5%	(7.9)
3 YR	79.6%	80.1%	(0.8)	3 YR	91.9%	92.4%	(0.8)	3 YR	57.2%	57.5%	(2.4)	3 YR	24.3%	24.4%	(9.7)
4 YR	77.6%	78.2%	(0.8)	4 YR	90.6%	91.3%	(0.7)	4 YR	54.7%	55.1%	(1.6)	4 YR	16.5%*	16.6%*	(13.2)
5 YR	75.4%	76.1%	(0.8)	5 YR	88.8%	89.7%	(0.8)	5 YR	51.4%	51.9%	(2.1)	5 YR	16.5%*	16.6%*	(0.0)
10 YR	70.1%	71.4%	(0.8)	10 YR	84.4%	86.0%	(0.9)	10 YR	43.5%	44.3%	(0.0)	10 YR	16.5%*	16.6%*	(0.0)

* = FEWER THAN 10 CASES USED FOR CALCULATIONS OR STANDARD ERROR WAS > 10%

Death Certificate or Autopsy only cases = 0 Benign cases = 0 Insitu cases = 9139 Unk age or P.S. 3 cases = 0
Unknown sex cases = 0 Alive & survival 0 = 656 Out of state = 0 non-analytic = 0 Multi Sequences = 143

********** TOTAL CASES 11123 REJECTED CASES 9207 ACCEPTED CASES 1916 **********

NOTE: A complete report lists tables on survival for all stages and histologic types, followed by survival reports based on the histological types. Tables are preceded by definitions and explanations of the terms used in the tables and options selected for printing.
SOURCE: National Cancer Institute.

location or all locations combined. For a specific location, the user selects *reporting source* (SEER participating site) as one of the variables for the subset file.

Example

To demonstrate the use of SeerQuery to generate summary and survival reports, we select cancer of the uterine cervix, with cervical cancer as the primary site and age as the other variable. The dialogue for selection is shown in figure 1. After the queries are answered, the program searches the master file and creates a subset based on the requested variables. The user can generate summary and survival reports using this subset. Selected computer dialogues for the summary report are shown in figure 2. Similarly, the survival option

from the menu initiates computer dialogues (not shown) and computes the survival for the cohort selected. Partial printouts of survival and summary reports are shown in figures 3 and 4. A list of computer dialogue screens and printouts is available on request.

Advantages of PC-based System

The compact disk technology is attractive to users of large scale information storage and information systems because of growing familiarity with PCs and rapid strides in the technology of computer hardware, which have brought mainframe data base applications to the PC environment. Large data bases are available on CD-ROM for a fraction of the cost of storage on main-frame computers. CD-ROM technology is cost effective and capable of substituting for tape storage drives and even mainframe systems for some applications (*6*). For example, CD-ROM technology provides up to 720 million bytes of data retrieval capacity for personal computer use. CD-ROMs are read by laser and can store the content of 1,500 floppy diskettes, equivalent to 200,000 typewritten pages of material.

Low weight, low cost, portability, data security, and rapid access are features favoring the use of CD-ROM for storing data bases such as SEER. The prices of CD-ROM drives have dropped sharply, and they occupy about the same physical space as floppy diskette drives. ISO 9660 hardware confers international compatibility.

Disadvantages of a PC-based System

CD-ROM drives are slow in retrieving data, compared with magnetic disk drives, and have read-only capability. However, these characteristics are minor problems with static or infrequently changing data bases. Slow evolution of data bases can be handled by periodically replacing diskettes with updates. Although CD-ROM is in disk format, its throughput is modest (throughput being the time for the drive to go to the beginning of a record and to read that record).

Availability and Uses

The test version of the SeerQuery program and the SEER data set on CD-ROM are available from NCI through the authors at no cost. Users are encouraged to devise analyses in response to their particular data needs. For example, State or local health departments not in a SEER Program area may compare their population and demographic factors to those of SEER sites. SEER survival or incidence data could be used to describe the cancer problem in nonparticipating States or areas by comparing them to the SEER data.

Physicians and others, without having to learn mainframe commands, can access the SEER Program for survival or incidence data on a specific cancer. SeerQuery is a tool for cancer patient management. For example, a physician may want to know survival data for a specific cancer in specific age groups before making decisions on treatment, and survival data can influence that decision.

SeerQuery follows SEER coding systems. In current form, SeerQuery requires the user to have a hard copy of coding information for each variable. Information on coding is in the user's manual and later versions of the program are to show this information on-screen. Data updates are to be offered annually to registered users at no cost.

Collaboration with interested investigators on expansion and modification of the program is encouraged.

References

1. Boring, C. C., Squires, T. S., and Tong, T.: Cancer statistics, 1992. CA 42: 19–38, January–February 1992.
2. National Cancer Institute: Surveillance, Epidemiology, and End Results: incidence and mortality data, 1973–1977. NCI Monograph, Vol. 57, 1981.
3. Ries, L. A. G., et al.: Cancer Statistics Review, 1973–1987. NIH Publication No. 91–2789. National Cancer Institute, Bethesda, MD, 1991.
4. Elisa T. Lee: Statistical methods for survival data analysis. Wadsworth, Inc., Lifetime Learning Publications, Belmont, CA, 1986.
5. Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H.: Epidemiological research: principles and quantitative methods. Van Nostrand Rienhold Co., Inc., New York, NY, 1982.
6. Laub, L.: What is CD-ROM? *In* CD-ROM: the new papyrus, edited by S. Lambert and S. Ropiequet. Microsoft Press, Redmond, WA, 1986, pp. 47–72.
7. International Classification of Diseases: Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death. 9th Revision. World Health Organization, Geneva, 1977.